

Deformable Neighborhood Feature Fusion Network for Underwater Image Super-Resolution

Han Wang *, Yuzhang Chen, Tingting Jia and Yuqi Ge

School of Artificial Intelligence Hubei University, Wuhan 430415, China

Abstract

Due to the combined effects of light absorption and scattering in water, as well as complex environmental noise, underwater images commonly suffer from severe degradations, including low contrast, blurred edges, and loss of structural information. These characteristics pose significant challenges for underwater image super-resolution. To address these issues, we propose a Deformable Neighborhood Feature Fusion Network (DNFFNet) for underwater image super-resolution. The proposed network incorporates a Deformable Neighborhood Enhancement Module (DNEM) to strengthen the perception of local structural details by adaptively modeling content-aware neighborhoods. In addition, a Spatial-Channel Feature Fusion Module (SCFFM) is designed to effectively integrate fine-grained local details with global semantic information. The experimental results show that DNFFNet achieves consistent improvements in PSNR and SSIM over representative baseline methods on the USR-248 and UFO-120 datasets, indicating its robustness and generalization under complex underwater degradation conditions.

Keywords

Underwater images, super-resolution reconstruction, feature fusion, attention mechanism.

1. Introduction

With the growing demand for marine exploration and scientific research, underwater image processing has become increasingly important in applications such as ocean observation, ecological monitoring, underwater robotics, and marine archaeology. However, complex optical properties of seawater often lead to severe image degradation. Image super-resolution (SR) provides an effective and promising solution to this challenge. Recent advances in deep learning have led to significant progress in image super-resolution. Early work such as SRCNN[1] introduced convolutional neural networks into this task, establishing an end-to-end learning framework for super-resolution reconstruction. Transformer-based models have recently gained prominence in image restoration due to their ability to capture long-range dependencies, with methods such as SwinIR[2], Restormer[3], and CAT[4] demonstrating strong performance. However, despite rapid progress in general super-resolution, these approaches remain limited when applied to underwater images with complex degradation. CNN-based methods are constrained by fixed receptive fields, while Transformers, although effective at global modeling, often rely on fixed window mechanisms that struggle with spatially non-uniform underwater textures and incur high computational costs.

To address these challenges, we propose DNFFNet, a deformable neighborhood feature fusion network that jointly models local and global information to improve image reconstruction. The network is composed of a Deformable Neighborhood Enhancement Module (DNEM) and a Spatial-Channel Feature Fusion Module (SCFFM). Specifically, DNEM adaptively identifies informative spatial locations, thereby strengthening the representation of complex textures

and edge structures. SCFFM integrates Spatial Feature Refinement Attention (SFRA) and Dynamic Sparse Channel Attention (DSCA) to jointly refine spatial details and correct color distortions, leading to super-resolved images with improved visual consistency. In addition, both modules incorporate a depthwise gated feed-forward network (DGFN), which applies nonlinear transformations and adaptive modulation to pixel-level features, further enhancing representational capacity and modeling stability.

2. Method

2.1. Architecture

The overall architecture of DNFFNet is organized into three principal stages: shallow feature extraction, deep feature fusion, and high-resolution image reconstruction, as illustrated in Fig. 1.

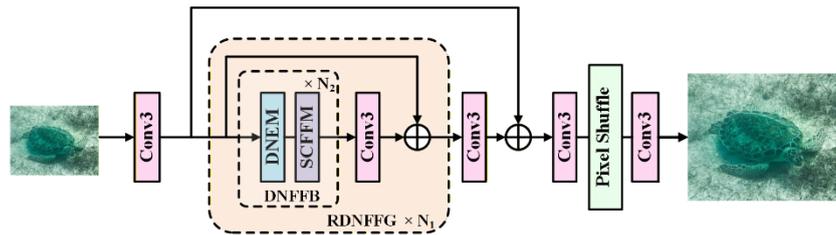


Fig. 1. The overall architecture of DNFFNet.

The network adopts an end-to-end learning framework to map the input low-resolution image I_{LR} to the reconstructed high-resolution output I_{HR} . In the shallow feature extraction stage, a 3×3 convolution is employed to extract the shallow feature $F_S \in \mathbb{R}^{H \times W \times C}$ from the low-resolution (LR) input image $I_{LR} \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the height and width of the input image, respectively, while C indicates the number of feature channels. The shallow feature F_S is then passed to the deep feature extraction stage, in which both low-frequency structural information and high-frequency details are further explored to obtain the deep feature representation $F_D \in \mathbb{R}^{H \times W \times C}$. A residual connection is introduced to effectively integrate the shallow and deep feature. Finally, in the image reconstruction stage, the extracted features are upsampled through convolution followed by sub-pixel rearrangement to produce the final high-resolution (HR) output $I_{HR} \in \mathbb{R}^{H_{out} \times W_{out} \times 3}$, where H_{out} and W_{out} denote the height and width of the output image, respectively.

2.2. Deformable Neighborhood Enhancement Module

Underwater images are often affected by complex environmental disturbances, leading to degraded details and blurred textures during acquisition. Such degradations not only impair visual quality but also substantially reduce the reliability and usability of the imagery. Window-based self-attention mechanisms rely on fixed partitioning schemes, which restrict the receptive field and limit information exchange across windows, making them less effective in modeling long-range dependencies and spatially distributed structural patterns in complex underwater scenes. To address these limitations, we design the DNEM module, which combines deformable neighborhood window attention (DNWA) with a depthwise gated feed-forward network (DGFN). By enabling adaptive neighborhood modeling and enhanced feature representation, DNEM improves the recovery of complex textures and edge structures in underwater images. The architecture of DNEM is illustrated in Fig. 2.

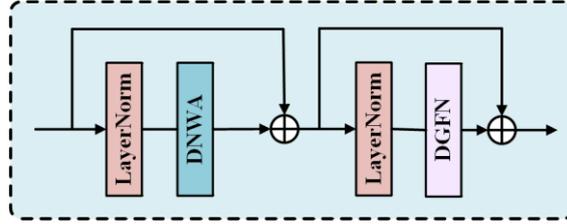


Fig. 2 Illustration of the proposed Deformable Neighborhood Enhancement Module

Specifically, the DNWA module draws inspiration from deformable convolution by introducing learnable spatial offsets to dynamically adjust sampling locations within local attention windows. This design enables the attention mechanism to adaptively focus on content-relevant regions of the image. Moreover, by incorporating a sliding-window strategy, DNWA facilitates smooth information exchange across neighboring windows while preserving the efficiency of local modeling, thereby improving the continuity and consistency of cross-window feature representation.

The shallow features F_s are first normalized by Layer Normalization (LN) and then fed into the DNWA module. A 1×1 convolution is applied to linearly project the input features into query representations, which serve two purposes: constructing the query vectors for attention computation and conditioning the offset prediction branch. The offset branch models local spatial relationships through lightweight convolutional transformations and nonlinear activations, adaptively predicting a two-dimensional spatial offset for each token. Subsequently, the predicted offsets are applied to resample the input feature map via bilinear interpolation, producing content-aware representations dynamically guided by the query features. Based on these representations, the corresponding keys and values are derived through linear projection. An unfolding operation is then employed to construct token-centered local neighborhoods. Scaled dot-product attention is used to estimate the correlations between the query and local key features, and the resulting attention weights are applied to aggregate the neighborhood values, yielding enriched local contextual representations. Finally, a linear projection followed by a residual connection restores the channel dimensionality, enabling adaptive feature aggregation and effective cross-window information interaction while preserving the efficiency of local modeling. The process is formulated as follows:

$$\begin{aligned}
 F_1 &= LN(F_s) \\
 Q &= Conv_{1 \times 1}(F_1) \\
 \Delta P &= Conv_{1 \times 1}(\varphi(LN(DWConv_{5 \times 5}(Q))) \quad (1) \\
 P_{ref}(i, j) &= (x_{i,j}, y_{i,j}) \\
 \Delta P(i, j) &= (\Delta y_{i,j}, \Delta x_{i,j}) \\
 P(i, j) &= P_{ref}(i, j) + \Delta P(i, j)
 \end{aligned}$$

where $Conv_{1 \times 1}(\square)$ denotes a two-dimensional convolution with a kernel size of 1×1 , while $DWConv_{5 \times 5}(\square)$ represents a depthwise separable convolution with a 5×5 kernel, employed to model local spatial relationships. $\varphi(\square)$ denotes the GELU nonlinear activation function, and ΔP corresponds to a learnable offset vector associated with each spatial location. $(x_{i,j}, y_{i,j})$ denotes the original spatial coordinates of the token at position (i, j) in the feature map, which define the reference sampling location $P_{ref}(i, j)$ on a regular grid. $\Delta y_{i,j}$ and $\Delta x_{i,j}$ represent the continuous displacements along the horizontal and vertical directions, respectively, and $\Delta P(i, j)$ denotes the predicted two-dimensional offset. The final sampling position $P(i, j)$ for

deformable sampling is obtained by adding the predicted offset to the reference position. Based on the continuous sampling coordinates described above, the input feature map is resampled using a differentiable bilinear interpolation operator (`grid_sample`). The operation is formally expressed as follows:

$$\begin{aligned}
 F_1' &= \text{GridSample}(F_1, P) \\
 K, V &= \text{Conv}_{1 \times 1}(F_1') \\
 \alpha_{i,j,p} &= \text{Softmax}(Q_{i,j}^T K_{i,j,p}) \\
 Y_{i,j} &= \sum_{p=1}^P \alpha_{i,j,p} V_{i,j,p}
 \end{aligned} \tag{2}$$

where $\text{GridSample}(\square)$ denotes the bilinear interpolation function based on the continuous sampling coordinates P , and F_1' represents the resulting feature map. For each query feature $Q_{i,j}$ at position (i, j) , its correlation with the p -th key feature $K_{i,j,p}$ within the local neighborhood is computed. The correlations are then normalized via a Softmax function to yield the corresponding attention weights $\alpha_{i,j,p}$. $V_{i,j,p}$ denotes the value feature vector of the p -th sampled point in the deformable local neighborhood centered at (i, j) . The output feature $Y_{i,j}$ is obtained by aggregating the value features weighted by the attention coefficients, providing a structurally consistent and stable representation that facilitates the reconstruction of high-frequency details in subsequent stages.

2.3. Spatial-Channel Feature Fusion Module

The SCFFM achieves collaborative modeling by integrating Spatial Feature Refinement Attention (SFRA) with Dynamic Sparse Channel Attention (DSCA), thereby enabling dynamic alignment between local feature representations and global contextual semantics. Through this coordinated spatial-channel interaction, the SCFFM improves the reconstruction quality and visual consistency of underwater image super-resolution in the following aspects: (1) The SFRA employs a lightweight convolutional self-attention mechanism to adaptively recalibrate local pixel-level features, thereby strengthening the representation of high-frequency information, including texture details, edge responses, and structural continuity. (2) The DSCA captures long-range dependencies across feature channels by computing self-attention along the channel dimension, enhancing global semantic coherence. A dynamic gating strategy is further introduced to adaptively regulate channel interactions, selectively emphasizing cross-channel dependencies that are most relevant to the reconstruction task while suppressing redundant or noisy responses commonly present in complex underwater environments. (3) The AFM adaptively aligns and fuses the dual-branch features across both spatial and channel dimensions, alleviating representation inconsistencies arising from the differing modeling paradigms of pixel-level local modeling and channel-wise global modeling. This coordinated fusion further improves the overall image reconstruction quality. The architecture of the SCFFM is illustrated in Fig. 3.

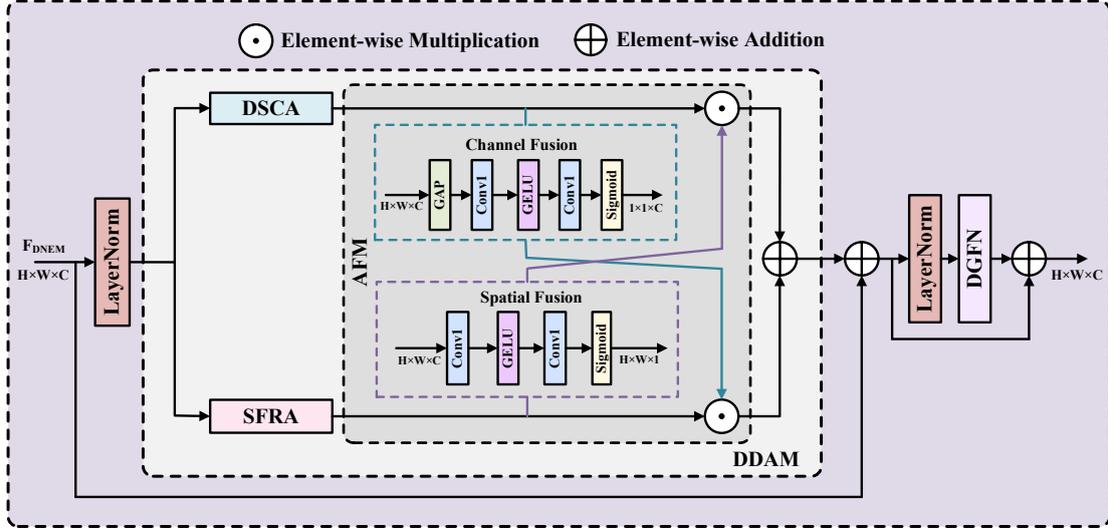


Fig. 3 Structure of the Spatial-Channel Feature Fusion Module.

Specifically, the feature F_{DNEM} is first processed by and then separately fed into the DSCA and SFRA modules, which are responsible for modeling local pixel-level spatial structures and cross-channel global semantic dependencies, respectively. After dual-branch feature modeling, the outputs are jointly input into the AFM module to enable cross-paradigm interaction. Within AFM, the SFRA output is fed into the Spatial Fusion branch, where a 1×1 convolution followed by GeLU activation is applied to learn spatial position weights, generating a spatial attention map of size $H \times W \times 1$ for recalibrating the DSCA features and highlighting critical regions. Meanwhile, the DSCA output enters the Channel Fusion branch, where global average pooling is first applied to compress spatial dimensions, followed by a 1×1 convolution and GeLU activation to generate a channel attention map of size $1 \times 1 \times C$ for reweighting the SFRA features along the channel dimension, emphasizing more informative semantic channels. Sigmoid activation is then applied to normalize the attention maps, ensuring stable and interpretable feature recalibration. Through this cross-attention mechanism, the network achieves coordinated restoration of high-frequency details and semantic consistency, thereby significantly improving the accuracy and robustness of super-resolution reconstruction. The corresponding formulation is as follows:

$$\begin{aligned}
 F_2 &= LN(F_{DNEM}) \\
 F_{SFRA} &= SFRA(F_2) \\
 F_{DSCA} &= DSCA(F_2) \\
 Map_S &= f(\text{Conv}_{1 \times 1}(\varphi(\text{Conv}_{1 \times 1}(F_{SFRA})))) \in \square^{H \times W \times 1} \\
 Map_C &= f(\text{Conv}_{1 \times 1}(\varphi(\text{Conv}_{1 \times 1}(GAP(F_{DSCA})))))) \in \square^{1 \times 1 \times C} \\
 F_{DDAM} &= F_{SFRA} \square Map_C + F_{DSCA} \square Map_S \\
 F_{SCFFM} &= F_{DNEM} + F_{DDAM} + DGFN(LN(F_{DNEM} + F_{DDAM}))
 \end{aligned} \tag{3}$$

where F_2 represents the output of LN after layer normalization, and F_{SFRA} and F_{DSCA} are the spatial enhancement features from the SFRA branch and the channel enhancement features from the DSCA branch, respectively. $GAP(\square)$ denotes global average pooling, $f(\square)$ represents the activation function *Sigmoid*, and Map_S and Map_C denote the spatial attention map and the channel attention map, respectively. \square represents element-wise multiplication. Map_S and Map_C are used to perform weighted summation on F_{SFRA} and F_{DSCA} to obtain the output F_{DDAM} . After

cross-attention modulation, the features F_{DDAM} are fused with the input features F_{DNEM} under the effect of residual connection, and after layer normalization LN, are input into the DGFN module to further enhance nonlinear representation capability. F_{SCFFM} represents the final output feature information of the entire SCFFM module.

3. Experiments

To evaluate the effectiveness of the proposed DNFFNet, experiments are conducted using two publicly available underwater image datasets, USR-248 and UFO-120, released by the Interactive Robotics and Vision Laboratory at the University of Minnesota. These datasets are widely adopted benchmarks in underwater image super-resolution research. Network performance is quantitatively assessed using three complementary image quality metrics: peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and underwater image quality measure (UIQM).

3.1. Comparison with State-of-the-Art Methods

To rigorously assess the efficacy of the proposed algorithm and its generalization capability across different datasets, DNFFNet is compared with several representative and state-of-the-art super-resolution methods, including SRCNN[1], DSRCNN[5], EDSR[6], SwinIR[2], IPT[7], DRCT[8], AGDN[9], CATANet[10], and DAT[11]. The reconstruction performance is analyzed from both quantitative and qualitative perspectives. Table 1 presents the quantitative results of different methods on the USR-248 and UFO-120 datasets under scaling factors of x2 and x4, evaluated using PSNR, SSIM, and UIQM. The best performance for each metric is highlighted in bold.

Table 1 Quantitative comparison with state-of-the-art methods

Method	Scale	Params(M)	USR-248			UFO-120		
			PSNR (dB)	SSIM	UIQM	PSNR (dB)	SSIM	UIQM
SRCNN	×2	0.067	29.82	0.8120	2.5220	25.75	0.7031	2.3837
DSRCNN	×2	1.11	30.35	0.8337	2.5913	26.34	0.7438	2.4189
EDSR	×2	1.37	31.53	0.8469	2.6531	27.26	0.7689	2.4556
SwinIR	×2	11.45	31.56	0.8589	2.6759	27.40	0.7731	2.4801
IPT	×2	11.3	31.58	0.8591	2.6532	27.19	0.7716	2.4778
DRCT	×2	13.991	31.61	0.8629	2.7175	27.58	0.7811	2.5296
AGDN	×2	0.289	31.57	0.8617	2.7142	27.51	0.7791	2.5241
CATANet	×2	0.477	31.55	0.8620	2.7089	27.54	0.7806	2.5234
DAT	×2	14.8	31.63	0.8630	2.7211	27.57	0.7815	2.5312
DNFFNet	×2	10.41	31.61	0.8634	2.7242	27.62	0.7826	2.5289
SRCNN	×4	0.067	26.07	0.6758	2.3471	24.49	0.6189	2.2754
DSRCNN	×4	1.11	26.58	0.6923	2.3936	25.87	0.6354	2.3176
EDSR	×4	1.479	27.51	0.7137	2.4152	25.98	0.6932	2.3864
SwinIR	×4	11.45	27.66	0.7137	2.4801	26.32	0.6890	2.3928
IPT	×4	11.3	27.69	0.7168	2.4978	26.30	0.6911	2.3899
DRCT	×4	14.139	27.70	0.7121	2.5143	26.29	0.6943	2.3945
AGDN	×4	0.303	27.63	0.7182	2.4978	26.28	0.6952	2.3905
CATANet	×4	0.535	27.69	0.7201	2.4957	26.32	0.7008	2.3910
DAT	×4	14.91	27.73	0.7215	2.5098	26.35	0.7089	2.3989
DNFFNet	×4	10.54	27.72	0.7221	2.5140	26.38	0.7108	2.4012

As shown in Table 1, Under the $\times 2$ super-resolution setting, traditional convolutional networks such as SRCNN and DSRCNN exhibit relatively limited reconstruction performance and fail to adequately recover fine structural details in underwater images. In contrast, the proposed DNFFNet achieves consistently strong performance across all evaluation metrics. On the USR-248 dataset, DNFFNet attains a PSNR value approximately 0.02 dB lower than that of DAT, while achieving higher SSIM and UIQM scores of 0.8634 and 2.7242, respectively, indicating superior structural preservation and underwater visual quality. On the UFO-120 dataset, DNFFNet improves PSNR by approximately 0.05 dB compared to DAT, while the difference in UIQM is negligible, at around 0.0023. These quantitative results demonstrate that DNFFNet possesses robust structural modeling capability and stable reconstruction performance under complex underwater imaging conditions, effectively balancing objective reconstruction accuracy with overall perceptual image quality.

For the $\times 4$ super-resolution task, the evaluation metrics of all compared methods on both USR-248 and UFO-120 datasets are generally lower than those obtained under the $\times 2$ setting, which is consistent with the well-known performance degradation observed in higher upscaling factors. On the USR-248 dataset, DNFFNet attains the best performance in terms of SSIM, while its PSNR is only 0.01 dB lower than that of DAT, and its UIQM remains comparable to DRCT. On the UFO-120 dataset, DNFFNet achieves the best results across all three metrics, further validating its robustness and generalization capability under complex underwater degradation conditions and high upscaling factors. Overall, DNFFNet consistently outperforms the existing comparison methods.

In addition to the quantitative evaluation based on objective metrics, a qualitative comparison is conducted to further assess the reconstruction performance of DNFFNet from the perspective of visual perception. The proposed method is compared with several representative super-resolution networks to highlight differences in detail recovery and texture restoration. For intuitive visualization, representative samples with scaling factors of $\times 2$ and $\times 4$ are selected from the USR-248 test set, and the regions marked by red boxes in the low-resolution images are enlarged for local comparison. Under the $\times 2$ super-resolution setting, as shown in Fig. 4, DNFFNet effectively restores fine-grained texture details on the fish tail, demonstrating superior detail preservation capability. When the scaling factor is increased to $\times 4$, as illustrated in Fig. 5, the images reconstructed by SRCNN exhibit noticeable edge blurring and artifact contamination, indicating limited capability in high-magnification reconstruction. In contrast, DNFFNet successfully preserves the structural continuity and hierarchical characteristics of the sea anemone, while producing more natural color reproduction and richer texture details. From a visual perception standpoint, DNFFNet achieves higher reconstruction quality compared with the competing methods.

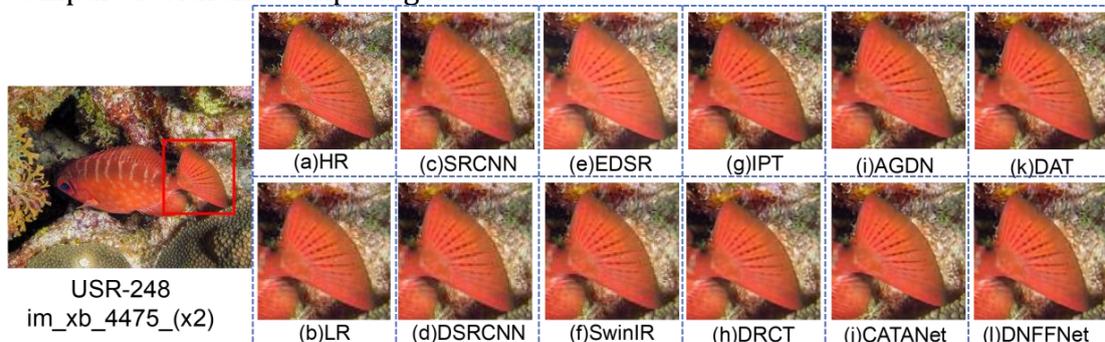


Fig. 4 Qualitative comparison of $\times 2$ super-resolution reconstruction for im_xb_4475 from the USR-248 dataset.

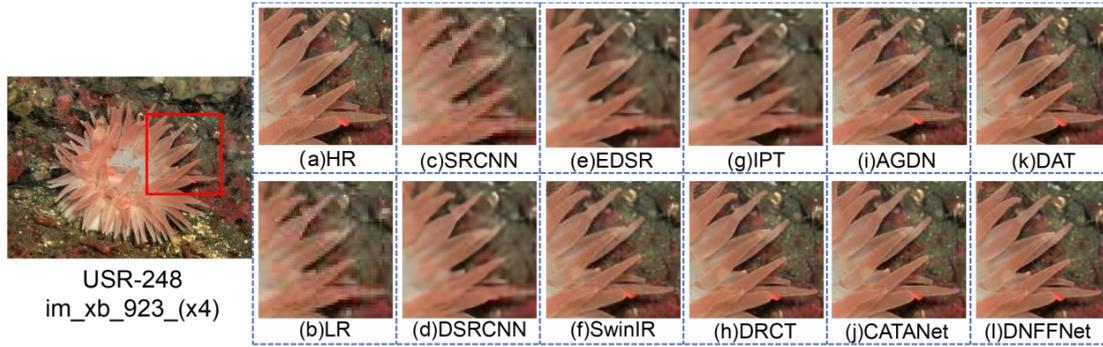


Fig. 5 Qualitative comparison of x4 super-resolution reconstruction for im_xb_923 from the USR-248 dataset.

3.2. Ablation Study

To further evaluate the effectiveness of DNFFNet, ablation studies are conducted on the USR-248 and UFO-120 datasets with a reconstruction scale of x2. The analysis focuses on the contribution of individual components within the deep feature extraction stage. By progressively incorporating the DNEM, DSCA, and SFRA modules, we evaluate their respective impacts on reconstruction performance. Quantitative results in terms of PSNR, SSIM, UIQM for different module configurations are reported in Table 2.

Table 2. Quantitative ablation study of different modules on DNFFNet

Datasets	DNEM	SCFFM		PSNR	SSIM	UIQM
		DSCA	SFRA			
USR-248	×	×	×	31.45	0.8578	2.7089
	√	×	×	31.56	0.8621	2.7178
	×	√	×	31.53	0.8615	2.7154
	×	×	√	31.50	0.8601	2.7132
	×	√	√	31.57	0.8624	2.7186
	√	√	√	31.61	0.8634	2.7242
UFO-120	×	×	×	27.45	0.7715	2.5034
	√	×	×	27.57	0.7805	2.5251
	×	√	×	27.54	0.7794	2.5208
	×	×	√	27.52	0.7768	2.5166
	×	√	√	27.58	0.7812	2.5267
	√	√	√	27.62	0.7826	2.5289

The experimental results indicate that removing either the DNEM or the SCFFM leads to noticeable reductions in PSNR and SSIM across both datasets, highlighting their substantial contributions to the overall performance of DNFFNet. Specifically, DNEM adaptively adjusts neighborhood sampling positions and receptive fields according to feature content, thereby effectively enhancing the recovery of edge and structural details. In contrast, SCFFM facilitates comprehensive fusion of local features and global semantic information, resulting in improved perceptual image quality. When DSCA and SFRA operate jointly and are further integrated with DNEM, the network achieves the optimal PSNR values of 31.61 dB and 27.62 dB, respectively. These results further confirm the effectiveness and necessity of each module within the DNFFNet architecture.

4. Conclusion

This work addresses the severe degradation of local structures and the challenges of cross-scale feature modeling in underwater image super-resolution by proposing a deformable neighborhood feature fusion network, termed DNFFNet. By incorporating DNEM, the network adaptively adjusts neighborhood sampling ranges, thereby strengthening the representation of underwater object boundaries and fine-grained texture details. Meanwhile, SCFFM facilitates coordinated spatial-channel interactions to achieve effective integration of local pixel-level information and global semantic features, resulting in improved structural stability and visual consistency in the reconstructed images. Experimental results on public benchmark datasets, including USR-248 and UFO-120, demonstrate that DNFFNet consistently achieves strong and stable performance across multiple scaling factors, with particularly notable advantages in suppressing structural distortions and reconstruction artifacts under high upscaling settings. Furthermore, ablation studies and qualitative visualizations confirm the critical contributions of the proposed components to the overall performance gains. Overall, DNFFNet provides an effective solution for super-resolution reconstruction in complex underwater scenarios and demonstrates promising potential for practical deployment.

References

- [1] Dong C, Loy C C, He K, et al. Learning a deep convolutional network for image super-resolution[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13. Springer International Publishing, 2014: 184-199.
- [2] Liang J, Cao J, Sun G, et al. Swinir: Image restoration using swin transformer[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 1833-1844.
- [3] Zamir S W, Arora A, Khan S, et al. Restormer: Efficient transformer for high-resolution image restoration[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5728-5739.
- [4] Chen Z, Zhang Y, Gu J, et al. Cross aggregation transformer for image restoration[J]. Advances in Neural Information Processing Systems, 2022, 35: 25478-25490.
- [5] Mao X J, Shen C, Yang Y B. Image restoration using convolutional auto-encoders with symmetric skip connections[J]. arxiv preprint arxiv:1606.08921, 2016.
- [6] Lim B, Son S, Kim H, et al. Enhanced deep residual networks for single image super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017: 136-144.
- [7] Chen H, Wang Y, Guo T, et al. Pre-trained image processing transformer[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 12299-12310.
- [8] Hsu C C, Lee C M, Chou Y S. Drct: Saving image super-resolution away from information bottleneck[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 6133-6142.
- [9] Wang H, Wei Z, Tang Q, et al. Attention guidance distillation network for efficient image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 6287-6296.
- [10] Liu X, Liu J, Tang J, et al. CATANet: Efficient Content-Aware Token Aggregation for Lightweight Image Super-Resolution[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 17902-17912.
- [11] Chen Z, Zhang Y, Gu J, et al. Dual aggregation transformer for image super-resolution[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 12312-12321.