frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 25 April 2024
DOI 10.3389/fmars.2024.1389553

OPEN ACCESS

EDITED BY
Xinyu Zhang,
Dalian Maritime University, China

REVIEWED BY
Feihu Zhang,
Northwestern Polytechnical University, China
Xuebo Zhang,
Northwest Normal University, China

*CORRESPONDENCE
Xin He
hexin6770@163.com

RECEIVED 21 February 2024
ACCEPTED 08 April 2024
PUBLISHED 25 April 2024

CITATION
He X, Li J and Jia T (2024) Learning hybrid dynamic transformers for underwater image super-resolution.
Front. Mar. Sci. 11:1389553.
doi: 10.3389/fmars.2024.1389553

Learning hybrid dynamic transformers for underwater image super-resolution

Xin He[1]*, Junjie Li[2] and Tong Jia[3]

[1]School of Basic Sciences for Aviation, Naval Aviation University, Yantai, China, [2]School of Electromechanical and Automotive Engineering, Yantai University, Yantai, China, [3]School of Art and Design, Yantai Institute of Science and Technology, Yantai, China

Underwater image super-resolution is vital for enhancing the clarity and detail of underwater imagery, enabling improved analysis, navigation, and exploration in underwater environments where visual quality is typically degraded due to factors like water turbidity and light attenuation. In this paper, we propose an effective hybrid dynamic Transformer (called HDT-Net) for underwater image super-resolution, leveraging a collaborative exploration of both local and global information aggregation to help image restoration. Firstly, we introduce a dynamic local self-attention to adaptively capture important spatial details in degraded underwater images by employing dynamic weighting. Secondly, considering that visual transformers tend to introduce irrelevant information when modeling the global context, thereby interfering with the reconstruction of high-resolution images, we design a sparse non-local self-attention to more accurately compute self-similarity by setting a top-k threshold. Finally, we integrate these two self-attention mechanisms into the hybrid dynamic transformer module, constituting the primary feature extraction unit of our proposed method. Quantitative and qualitative analyses on benchmark datasets demonstrate that our approach achieves superior performance compared to previous CNN and Transformer models.

KEYWORDS

underwater image, image super-resolution, local self-attention, sparse self-attention, deep learning, visual transformer

1 Introduction

Underwater imaging poses distinct challenges owing to the natural attenuation, scattering, and color distortion of light within aquatic environments. These factors contribute to degraded image quality, thereby constraining the effectiveness of underwater observation, exploration, and surveillance systems (refer to Figure 1). Consequently, underwater image enhancement techniques, notably super-resolution, have attracted considerable attention in recent years. Super-resolution aims to reconstruct high-resolution images from low-resolution counterparts, thereby improving

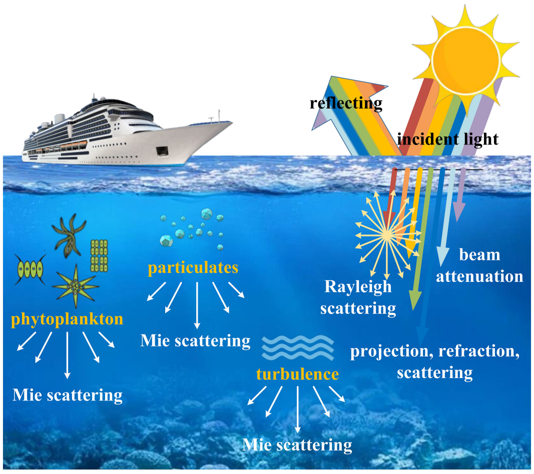

FIGURE 1
The physical imaging process of underwater conditions.

image clarity and detail. It holds immense potential for enhancing the visual quality of underwater scenes, facilitating better analysis, interpretation, and decision-making in various marine applications, such as ocean-observation and offshore engineering (Liu et al., 2024).

Despite recent advancements, underwater image super-resolution remains an active area of research, with ongoing efforts to develop robust and efficient algorithms capable of addressing the specific challenges.

posed by underwater environments (Islam et al., 2020a). Early efforts in underwater image super-resolution predominantly relied on traditional interpolation algorithms such as bicubic and bilinear. These approaches, while widely used in conventional image processing tasks, often yielded suboptimal results when applied to underwater imagery due to the unique characteristics of underwater environments.

In recent years, significant strides have been made in leveraging deep learning techniques, particularly convolutional neural networks (CNNs), for underwater image super-resolution. Unlike conventional interpolation methods, CNN-based approaches harness the capabilities of deep learning to discern intricate mappings between low-resolution and high-resolution image pairs directly from data. Various architectures, such as SRCNN (Super-Resolution Convolutional Neural Network) (Dong et al., 2015), VDSR (Very Deep Super-Resolution) (Kim et al., 2016), and EDSR (Enhanced Deep Super-Resolution) (Lim et al., 2017), have demonstrated remarkable performance improvements over those of conventional approaches. Subsequent research tends to focus on developing larger and deeper CNN models to enhance learning capabilities. However, due to the extensive number of network parameters, the computational cost of these methods is considerably high, limiting their applicability in real-world underwater scenarios (Jiang et al., 2021).

Later, Transformer-based architectures (Vaswani et al., 2017) have emerged as promising alternatives for underwater image super-resolution, offering unique advantages over CNN-based approaches. Unlike CNNs, Transformers leverage self-attention mechanisms to capture global dependencies and long-range dependencies within the input data (Han et al., 2022). For example, SwinIR (Liang et al., 2021) employs the window-based attention mechanism to better solve image super-resolution. Although the self-attention mechanism in the sliding window approach enables the extraction of local features, the discontinuity of the windows limits the ability to model local features within each window. In other words, these window-based image super-resolution methods are unable to aggregate information from outside the window, thus limiting the capability to model global information (Li et al., 2023a).

Indeed, the complexity and variability inherent in underwater environments elevate the challenges associated with underwater image super-resolution beyond those encountered in natural image superresolution tasks. The Transformer model, renowned for its adeptness in capturing global features, tends to introduce noticeable redundancy during the modeling process. Regrettably, this aspect has often been neglected in prior Transformer-based super-resolution approaches (Xiao et al., 2024). Therefore, developing a method to explore the characteristics of Transformers, aiming to better integrate both local and global features for joint modeling to achieve high-quality image reconstruction while reducing computational costs, holds significant promise.

To this end, we develop an effective hybrid dynamic Transformer (called HDT-Net) to solve underwater image super-resolution. The proposed method combines dynamic local self-attention with sparse non-local self-attention to synergistically enhance the representation capability of the Transformer model. The former dynamically explores local feature relationships based on a fully CNN model, mitigating errors induced by discontinuous windows. The latter aggregates features by selecting the most useful similarity values, alleviating redundancy caused by small self-attention weights. These strategies are carefully designed to address the challenges of complex underwater environments, thereby leveraging more effective feature information to improve the quality of image super-resolution. Finally, experimental validation on benchmark datasets confirms the effectiveness of the proposed approach.

In summary, the main contributions of this paper are as follows:

- We propose a lightweight deep model based on a hybrid Transformer for underwater image super-resolution tasks, aiming to enhance the quality of image reconstruction by jointly exploiting local and global features representation.
- We integrate a dynamic local self-attention and a sparse non-local self-attention to enable better capture of local and global feature information respectively, making the Transformer more effective and compact in long-range modeling.
- Experimental evaluation on commonly used benchmark datasets for underwater image super-resolution demonstrates that our method outperforms previous CNN and Transformer-based approaches both quantitatively and qualitatively.

2 Related work

In this section, we present a review of recent work related to underwater image super-resolution and vision transformer.

2.1 Underwater image super-resolution

Underwater image super-resolution is an uncertain task, and numerous studies have been conducted to explore suitable methods to address this challenge. Among the deep learning-based underwater image super-resolution models, CNN is one of the most common techniques. Shin et al. (Shin et al., 2016). proposed a CNN-based framework for estimating environmental light and transmission, featuring a versatile convolutional structure designed to mitigate haze in underwater images. Wang et al. (Wang et al., 2017). proposed a CNN-based underwater image enhancement framework called UIE-Net, comprising two sub-networks: CC-Net and HR-Net. CC-Net outputs color absorption coefficients for different channels to correct color distortion in underwater images. HR-Net outputs light attenuation transmission maps to enhance the contrast of underwater images. Li et al. (Li et al., 2017). proposed a novel generator network structure that combines the underwater image formation process to generate high-resolution output images. Subsequently, a dense pixel-level model learning pipeline is employed to perform color correction on monocular underwater images trained based on RGB-D and their corresponding generated images. The methods describe above address some aspects of underwater image super-resolution, yet they still exhibit a lack of robustness when handling highly complex underwater scenes.

Li et al. (Li et al., 2019). constructed an underwater image enhancement benchmark dataset, which provides a large-scale collection of real underwater images along with their corresponding reference images. This benchmark dataset facilitates comprehensive research on existing underwater image enhancement methods and enables easy training of CNNs for underwater image enhancement. But it lacks novelty in terms of algorithmic advancements compared to other methods. Guo et al. (Guo et al., 2019). proposed an underwater image enhancement method based on GAN. Additionally, the introduced MSDB combined with residual learning can improve network performance, while multiple loss functions can generate visually satisfactory enhancement results. Islam et al. (Islam et al., 2020b). proposed a simple yet effective underwater image enhancement model based on conditional genetic algorithms. This model evaluates image quality by incorporating global color, content, local texture, and style information to establish a perceptual loss function. Additionally, they provided a large-scale dataset consisting of paired and unpaired underwater image collections for supervised training. Chen et al. (Chen et al., 2020). proposed an improved deep reinforcement convolutional neural network based on deep learning principles. The main innovation involves incorporating wavelet bases into turbulence-based deep learning convolutional kernels, introducing an improved dense block structure. Further investigation is needed to assess the generalization of the methods utilized in the aforementioned studies to different underwater conditions.

Recently, Li et al. (Li et al., 2021). proposed a deep underwater image enhancement model. This model learns feature representations from different color spaces and highlights the most discriminative features through channel attention modules. Additionally, domain knowledge is integrated into the network by utilizing inverse media transmission maps as attention weights. Li et al. (Li et al., 2023b). proposed a novel method for realistic underwater image enhancement and super-resolution called RUIESR. Its purpose is to obtain paired data consistent with realistic degradation for training and to accurately estimate dual degradation to assist in reconstruction. In deep-sea or heavily polluted waters, the degradation characteristics may differ from those observed in the training data, potentially affecting the performance of the above methods. Dharejo et al. (Dharejo et al., 2024). investigated the integration of a typical Swin transformer with wave attention modules and reversible downsampling to achieve efficient multiscale self-attention learning with lossless downsampling. As a potential improvement over SwinIR, this model allows for faster training and convergence, as well as greater capacity and resolution. The computational complexity and resource requirements of this Transformer-based method may pose challenges.

2.2 Vision transformer

Vision Transformer (ViT) (Vaswani et al., 2017) is a model based on the Transformer architecture, initially proposed by Dosovitskiy et al. (Dosovitskiy et al., 2020). in 2020 to address image classification tasks in the field of computer vision. The introduction of ViT signifies the expansion of Transformer models from the domain of natural language processing to computer vision, ushering in a new paradigm for image processing tasks. Liang et al. (Liang et al., 2021). proposed an image restoration model called SwinIR. This model consists of three modules: shallow feature extraction, deep feature extraction, and HR reconstruction. It emphasizes the content-based interaction between image content and attention weights, achieved through a shifting window mechanism for long-range dependency modeling. The IPT (Chen et al., 2021) employs a multi-head, multi-tail, shared transformer body design, aiming to maximize the potential of the transformer architecture in serving various image processing tasks such as image super-resolution and denoising. The high computational complexity arising from this Transformer design may limit scalability to high-resolution images.

DRSAN (Park et al., 2021) proposes a dynamic residual network solution for lightweight super-resolution systems, leveraging different combinations of residual features considering input statistics. Additionally, it introduces residual self-attention, which, in collaboration with residual structures, enhances network performance without adding modules. Zamir et al. (Zamir et al., 2022). introduced Restormer, an image restoration transformer model known for its high computational efficiency in handling high-resolution images. They made critical design adjustments to the core components of the transformer block to enhance feature aggregation and transformation. To integrate the robustness of CNNs into the Transformer model, Restormer incorporates deep convolutions for encoding spatial local context. ELAN (Zhang et al., 2022) utilizes shift convolution (shift-conv) to effectively extract local structural information from the image. Subsequently, it introduces an intra-group multi-scale self-attention (GMSA) module to leverage the long-range dependency of the image. Further acceleration of the model's computation is achieved by employing a shared attention mechanism. In the task of image super-resolution, the effectiveness of integrating local and global feature representations in the aforementioned methods still requires further improvement.

Diverging from current approaches, we introduce a lightweight deep model rooted in a hybrid dynamic Transformer (HDT-Net). The goal is to bolster the quality of image reconstruction by synergizing local and global feature representations.

3 Proposed method

In this section, we first describe the overall pipeline of the model. Then, we provide details of the hybrid dynamic transformer module (HDTM), which serve as the fundamental building modules of the approach. HDTM is composed of four identical hybrid dynamic transformer blocks (HDTBs) connected end to end, as illustrated in the Figure 2. The HDTB mainly comprises three key elements: dynamic local self-attention (DLSA), sparse non-local self-attention (SNSA), and feed-forward network (FFN).

3.1 Overall pipeline

Figure 2 illustrates an overview of the proposed HDT-Net for underwater image super-resolution. Specifically, the low-resolution underwater image is first processed through a convolutional layer with a filter size of 3×3 pixels for shallow feature extraction. Subsequently, the feature information is sequentially processed through six identical HDTMs for deep feature extraction and fusion, both locally and globally. Within each HDTM, four internal HDTBs are connected end to end for processing, and the extracted features are finally passed to the next module through a 3 × 3 convolution. After the completion of HDTM processing, the features are further projected using a convolutional layer with a filter size of 3 × 3 pixels. Following that, high-resolution image reconstruction is performed through a 3 × 3 convolution and upsampling operation using PixelShuffle (Shi et al., 2016).

The process of the overall pipeline can be represented as Equations 1-4:

$$X^{'} = \mathrm{Conv}_{3\times 3}(X), \tag{1}$$

$$HDTM_s = HDTM_6(\dots(HDTM_1(\mathbf{X}'))), \tag{2}$$

$$\mathrm{X}_{\mathbf{low}} = \mathbf{X}' + HDTM_s(\mathbf{X}'), \tag{3}$$

$$X_{\mathrm{high}} = P(Conv_{3\times 3}(X_{low})), \tag{4}$$

where $X, Conv3 \times 3,\ P(\cdot), X_{low}, X_{high}$ represent the input features and 3 × 3 convolution, upsampling operation using PixelShuffle, low resolution image features and high resolution image features, respectively. The process of HDTM in the overall pipeline can be expressed as Equations 5, 6:

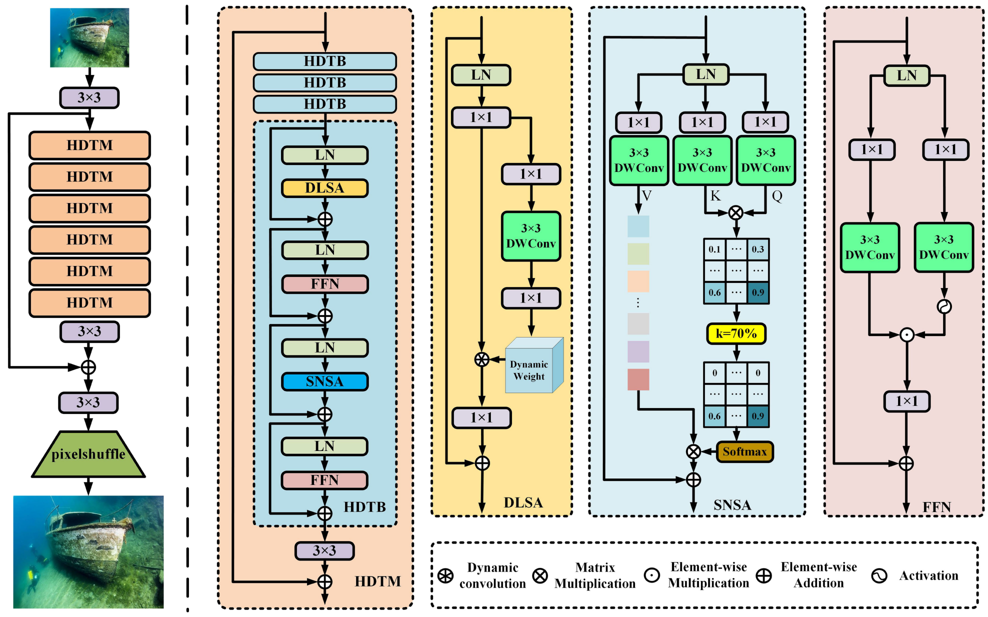

FIGURE 2
The overall architecture of the proposed network.

$$HDTB_s = HDTB_4(\dots(HDTB_1(X))), \tag{5}$$

$$HDTM = X + Conv_{3\times 3}(HDTB_s(X)) \tag{6}$$

3.2 Hybrid dynamic transformer block

We propose a hybrid dynamic transformer block consisting of DLSA, SNSA, and FFN. By combining DLSA and SNSA, the hybrid self-attention mechanism effectively weights each position against others in the input data, facilitating the integration of global information into each position's representation. Moreover, it enables the capturing of both global and local feature relationships at different positions in the image, allowing the model to capture long-range dependencies in the data. After each self-attention computation, the representation at each position undergoes non-linear transformation through FFN, mapping it to a new representation space to enhance the model's expressiveness. Formally, given the input features of the $(l-1)$-th block $\mathbf{X}_{l-1}$, the encoding of the HDTB process can be represented as Equations 7–10:

$$\mathbf{X}_l^d = \mathbf{X}_{l-1} + DLSA(LN(\mathbf{X}_{l-1})), \tag{7}$$

$$\mathbf{X}_l^f = \mathbf{X}_l^d + FFN(LN(\mathbf{X}_l^d)), \tag{8}$$

$$\mathbf{X}_l^s = \mathbf{X}_l^f + SNSA(LN(\mathbf{X}_l^f)), \tag{9}$$

$$\mathbf{X}_l = \mathbf{X}_l^s + FFN(LN(\mathbf{X}_l^s)), \tag{10}$$

where LN denotes the layer normalization, $\mathbf{X}_l^d$. d $\mathbf{X}_l^s$ represent the outputs of DLSA and SNSA, $\mathbf{X}_l^f$. d $\mathbf{X}_l$ represent the outputs of FFN, which are described below.

3.2.1 Dynamic local self-attention

To enhance the extraction and fusion of local features, we introduce a DLSA method aimed at capturing spatial relationships within an image, while also accommodating variable receptive fields. In contrast to conventional self-attention mechanisms, DLSA functions uniformly across the entire image. This dynamic approach empowers each spatial location to selectively attend to its nearby regions based on contextual cues. Specifically, given input features $X_{in} \in \mathbb{R}^{H\times W\times C}$ generated by layer normalization, 1 × 1 convolution is performed for feature aggregation. Similar to (Li et al., 2023a), we introduce a squeeze and excitation network (SENet) (Hu et al., 2018) as our dynamic weight generation network, which has no normalization layers and non-linear activations. Additionally, we employ a 3×3 depth-wise convolutional layer in SENet to encode features, ensuring better calculation of dynamic attention for local attention.

The proposed dynamic weight generation formula is as Equations 11-13:

$$\mathrm{X}_1 = \mathrm{DConv}_{3\times 3}(\mathrm{Conv}_{1\times 1}(\mathrm{X}_{in})), \tag{11}$$

$$\mathrm{X}_2 = \mathrm{Conv}_{1\times 1}(\mathrm{X}_1), \tag{12}$$

$$\mathrm{W}(\mathrm{x}) = \mathcal{R}(\mathrm{X}_2), \tag{13}$$

where $\mathcal{R}(\cdot)$ represents the reshaping function. In DLSA, we utilize learnable dynamic convolutions. Unlike traditional fixed kernels, learnable dynamic convolutional kernels offer greater flexibility and adaptability. Each pixel has a corresponding $K \times K$ dynamic kernel for dynamic convolution. We divide the number of feature channels into G heads, and learn separate dynamic weights in parallel. For the generated pixel-wise weights $\mathbf{W}$, we obtain the aggregated features using the following formula as Equation 14:

$$DLSA(\mathbf{X}) = \mathbf{W} \circledast \mathrm{X}_{\mathbf{in}}, \tag{14}$$

where $\circledast$ denotes the dynamic convolution operation using weight sharing across each channel.

3.2.2 Sparse non-local self-attention

Due to the fact that the dynamic estimated features generated by DLSA are based on fully convolutional operations, the efficiency of modeling global features is relatively low. To better perceive global features, we revisit the standard dot-product self-attention in Transformer (Zamir et al., 2022). However, this algorithm calculates attention maps based on fully connected operations for all query-key pairs. In our work, we develop SNSA to replace it, which leverages sparsity by selecting the top-k tokens (Chen et al., 2023) most relevant to the query, thus obtaining the most crucial information for computation. This approach avoids involving irrelevant information in the feature interaction process.

Specifically, we first perform feature aggregation by applying a 1 × 1 convolution, followed by a depthwise convolution with filter size of 3×3 pixels to encode per-channel contexts. This allows for self-attention computation across the three dimensions of query Q, key K, and value V, rather than spatial dimensions. Utilizing channel-wise similarity helps reduce memory consumption for efficient inference. Next, we compute the similarity between all pairs of queries and keys, and employ a selection strategy to mask out values with lower similarity, retaining those with higher similarity.

As shown in the Figure 2, k represents an adjustable parameter for dynamically setting the sparsity level. When k=70%, only the top 70% of elements with the highest scores are retained for activation, while the remaining 30% of elements are masked as 0. Finally, softmax is applied to normalize elements larger than the top-k, ensuring the output is a probability distribution. For elements with scores less than top-k, we use a scatter function to replace their probability at the given index with 0. This dynamic selection results in attention following a sparse distribution. Finally, matrix multiplication is used to multiply softmax with Value, which is then connected to the input residual through feature projection to obtain the final result.

The derivation formula for SNSA is as Equation 15:

$$SNSA(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathcal{S}\left(\mathcal{M}_k \odot \frac{\mathbf{Q}\mathbf{K}^{\top}}{\lambda}\right)\mathbf{V}, \tag{15}$$

where $S(\cdot)$ represents the softmax operation, λ is an optional temperature factor defined by $\lambda = \sqrt{d}$. Typically, multi-head attention is applied to each of the k new Q, K, and V, resulting in $d = C/k$ channel dimension outputs, which are then concatenated and projected linearly to obtain the final result of all heads.

$$[\mathcal{M}_k]_{ij} = \begin{cases} 1, \ \mathcal{M}_{ij} \in \quad \text{top} - \text{k} \quad (\text{row} \quad \text{j}) \\ 0, \quad \text{otherwise} \end{cases} , \tag{16}$$

where $\mathcal{M}_k$ denotes the top-k selection operator in Equation 16.

3.2.3 Feed-forward network

To extract sophisticated features from both the local and global self-attention data of the model and facilitate the learning of abstract representations, we introduce the FFN following the DLSA and SNSA modules. Specifically, we design two branches based on gating mechanisms. It first uses 1×1 convolutions for feature transformation and then employs 3 × 3 depth-wise convolutions to encode information from spatially adjacent pixel positions. One branch is used to expand feature channels, while in the other branch, it is activated along with the Gelu nonlinearity to reduce the channels back to the original input dimension and search for nonlinear contextual information in the hidden layers.

The FFN is formulated as Equations 17-19:

$$X_1 = GELU(\text{Conv}_{3\times3}(\text{Conv}_{1\times1}(X))), \tag{17}$$

$$X_2 = \text{Conv}_{3\times3}(\text{Conv}_{1\times1}(X)), \tag{18}$$

$$\hat{X} = \text{Conv}_{1\times1}(X_1 \odot X_2) + X\,. \tag{19}$$

In general, FFN plays a distinctly different role compared to self-attention. It controls the flow of information passing through various levels of our pipeline, allowing each level to focus on complementary contextual information to other levels.

3.3 Loss function

Building upon existing methods, we adopt the L1 loss function as the loss function for our model. The expression for the L1 loss function is defined as Equation 20 :

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \| y_i - \hat{y}_i \|_1, \tag{20}$$

where N is the number of samples in the dataset. y_i represents the ground truth value for the i-th sample. $\hat{y}_i$ represents the predicted value for the i-th sample.

The L1 loss function calculates the mean absolute error between the predicted values and the ground truth values, providing a measure of the average magnitude of the errors.

4 Experiments

In this section, we first introduce the implementation details, datasets and evaluation metrics. Then, we compare the proposed HDT-Net with 10 baseline methods, including bicubic, SRCNN (Dong et al., 2015), DSRCNN (Mao et al., 2016), SRGAN (Ledig et al., 2017), SRDM-GAN (Islam et al., 2020a), RFDN (Liu et al., 2020), LatticeNet+ (Luo et al., 2020), SMSR (Wang et al., 2021), IPT (Chen et al., 2021), and SwinIR (Liang et al., 2021). Finally, ablation experiments are conducted to validate the effectiveness of the proposed method. The experiments are trained on a server with two NVIDIA GeForce RTX 3090 GPUs.

4.1 Experimental settings

4.1.1 Implementation details

In the proposed SNSA, the threshold for top-k is set to 70%. We will analyze its impact in the ablation study. During the training, the batch size and patch size are configured as 16 and 64, respectively. The number of multi-head self-attention is set to be 6, and the number of feature is set to be 90. We utilize the Adam optimizer (Kingma and Ba, 2014) with default parameter configurations to train our model. The initial learning rate is established at 5 × 10^{-4}, employing a multi-step scheduler over 500K iterations.

4.1.2 Datasets and evaluation metrics

We validate the performance of various methods using the classic underwater image super-resolution benchmark datasets, USR-248 and UFO-120 (Liu et al., 2024). Each dataset showcases distinct underwater degradation characteristics, enabling comprehensive evaluation across diverse underwater imaging scenarios. Consistent with previous studies, we utilize PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) scores (Wang et al., 2004) to quantitatively compare the restoration results of different algorithms, enabling performance evaluation. In addition, we conduct evaluation calculations on the model parameter quantities of different deep networks.

4.2 Quantitative evaluation

Following (Dharejo et al., 2024), Table 1 presents the quantitative results of various methods on the USR-248 and UFO-120 datasets, including experimental setups with three different super-resolution scaling factors: ×2, ×4, and ×8. As shown, the experimental results demonstrate that our proposed HDT-Net consistently achieves the best quantitative performance. Compared to the state-of-the-art method SwinIR (Liang et al., 2021), our approach shows an average improvement of 0.5dB in PSNR, with a reduction in parameters by 58%. This indicates that our proposed hybrid transformer,

TABLE 1 Quantitative comparisons of different methods on the USR-248 and UFO-120 datasets.

Methods	Scale	USR-248		UFO-120		Average		Params(M)
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
Bicubic	x2	26.78	0.8263	27.01	0.8465	26.89	0.8364	–
SRCNN	x2	27.89	0.8467	27.12	0.8654	27.50	0.8560	0.067
DSRCNN	x2	28.12	0.8584	27.88	0.8731	28.00	0.8657	0.361
SRGAN	x2	28.41	0.8612	28.54	0.8815	28.47	0.8713	1.54
SRDM-GAN	x2	28.51	0.8592	28.58	0.8823	28.54	0.8707	0.586
RFDN	x2	28.72	0.8633	28.81	0.8841	28.76	0.8737	0.528
LatticeNet+	x2	28.74	0.8714	28.85	0.8854	28.79	0.8784	0.75
SMSR	x2	28.88	0.8712	28.91	0.8862	28.89	0.8787	0.985
IPT	x2	29.33	0.8831	29.05	0.8921	29.19	0.8876	11.3
SwinIR	x2	29.88	0.9018	30.01	0.9021	29.94	0.9019	11.45
Ours	x2	**31.23**	**0.9217**	**31.54**	**0.9168**	**31.38**	**0.9192**	4.71
Bicubic	x4	25.07	0.7823	25.12	0.8165	25.09	0.7994	–
SRCNN	x4	25.17	0.7978	25.21	0.8157	25.19	0.8067	0.067
DSRCNN	x4	25.78	0.8064	26.81	0.8177	26.29	0.8120	0.361
SRGAN	x4	26.09	0.8178	26.14	0.8188	26.11	0.8183	1.54
SRDM-GAN	x4	26.19	0.8211	26.51	0.8247	26.35	0.8229	0.586
RFDN	x4	26.66	0.8216	26.81	0.8350	26.73	0.8283	0.528
LatticeNet+	x4	26.78	0.8239	26.85	0.8245	26.81	0.8242	0.75
SMSR	x4	27.07	0.8296	27.15	0.8310	27.11	0.8303	0.985
IPT	x4	27.11	0.8626	27.16	0.8632	27.13	0.8629	11.3
SwinIR	x4	27.18	0.8634	27.27	0.8644	27.22	0.8639	11.45
Ours	x4	**27.69**	**0.8712**	**27.82**	**0.8745**	**27.75**	**0.8728**	4.71
Bicubic	x8	23.46	0.7684	23.84	0.7781	23.65	0.7732	–
SRCNN	x8	24.07	0.7877	24.12	0.7981	24.09	0.7929	0.067
DSRCNN	x8	24.12	0.7987	24.18	0.8031	24.15	0.8009	0.361
SRGAN	x8	24.22	0.8021	24.29	0.8024	24.25	0.8022	1.54
SRDM-GAN	x8	24.41	0.8162	24.47	0.8178	24.44	0.8170	0.586
RFDN	x8	24.55	0.8178	24.67	0.8218	24.61	0.8198	0.528
LatticeNet+	x8	25.08	0.8321	25.11	0.8324	25.09	0.8322	0.75
SMSR	x8	25.16	0.8344	25.23	0.8354	25.19	0.8349	0.985
IPT	x8	25.22	0.8353	25.34	0.8411	25.28	0.8382	11.3
SwinIR	x8	25.82	0.8555	26.04	0.8559	25.93	0.8557	11.45
Ours	x8	**26.37**	**0.8662**	**26.48**	**0.8655**	**26.42**	**0.8658**	4.71

Bold indicates the best results.

as opposed to window-based transformers, can better capture feature correlations. Particularly challenging is the task of image super-resolution at a scaling factor of ×8. In contract, our proposed solution, leveraging the efficient fusion of local and global information, exhibits robust performance advantages in complex underwater scenes.

4.3 Qualitative evaluation

Figures 3, 4 illustrate the visual comparison results of different methods on the USR-248 and UFO-120 datasets, respectively. Note that we do not compare RFDN (Liu et al., 2020) and LatticeNet+ (Luo et al., 2020) as their visual results are not available. It is evident

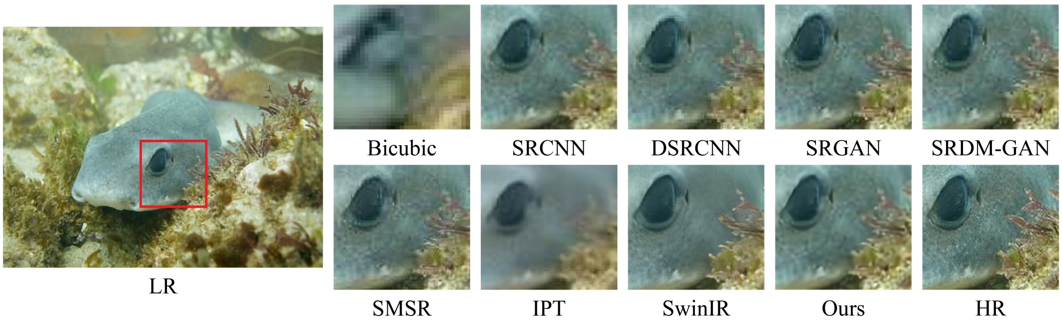

FIGURE 3
Image super-resolution comparisons for different methods on the USR-248 dataset.

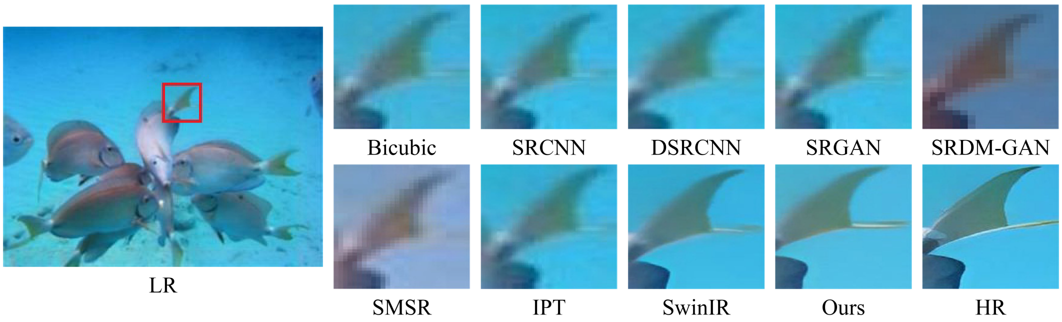

FIGURE 4
Image super-resolution comparisons for different methods on the UFO-120 dataset.

that effectively enhancing image resolution quality in complex underwater environments presents a formidable challenge compared to natural images. We find that the restoration results of most Transformer-based approaches tend to smooth out the details and textures of the images, which is attributed to the dense pattern of self-attention mechanisms. Furthermore, window-based self-attention global modeling methods fail to effectively aggregate information outside the window, thus affecting the quality of the restored images, as observed in SwinIR (Liang et al., 2021). In contrast, our proposed method achieves better image restoration by exploring the aggregation of local and global information. These quantitative and qualitative results indicate the effectiveness of the proposed hybrid dynamic Transformers, providing new insights into the challenging task of underwater image super-resolution.

4.4 Ablation study

In this section, we conduct a further analysis of the impact of the components proposed in our method and compare it against baseline models. To ensure a fair comparison, we employ the same settings used to train all baseline models as those of the proposed method. Here, we conduct ablation experiments with ×2 super-resolution on the USR-248 dataset. Specifically, the ablation study includes (1) effectiveness of the DLSA and SNSA, (2) effect of top-k values in the SNSA, and (3) effect of the number of HDTMs.

4.4.1 Effectiveness of the DLSA and SNSA

First, we analyze the effectiveness of the two key components proposed in the method, including DLSA and SNSA. To do this, we separately remove one of the components for comparative analysis. Table 2 presents the quantitative results of different variant models. It can be seen that our approach combining DLSA and SNSA achieves the best performance. Figure 5 illustrates the visual comparison results of different ablation models. It can be observed that, compared to using only a single self-attention mechanism for feature modeling, our proposed method can better restore the structure and detail regions of underwater images. The combination of local and non-local self-attention mechanisms enables the model to strike a balance between enhancing local details and preserving the overall scene context, resulting in more accurate and coherent super-resolved images.

4.4.2 Effect of top-k values in the SNSA

Next, we analyze the impact of the top-k value in SNSA. Regarding the choice of sparsity value, it also plays a crucial role in determining the performance of the model. A smaller sparsity value may result in a dense attention map, which could lead to

TABLE 2 Quantitative comparison of ablation results about the effectiveness of DLSA and SNSA.

Models	w/o DLSA	w/o SNSA	Ours
PSNR/SSIM	30.48/0.9060	29.63/0.8958	**31.23/0.9217**

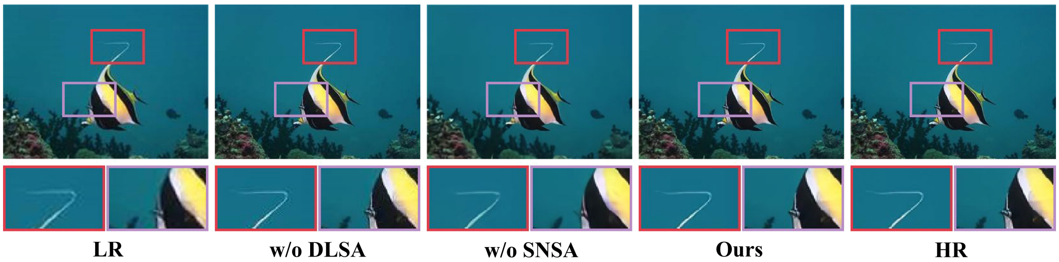

FIGURE 5
Visual comparison of ablation results about the effectiveness of DLSA and SNSA.

increased computational overhead and potential overfitting to noisy or irrelevant features. On the other hand, a larger sparsity value may cause the model to miss important global context or relevant features. Therefore, selecting an optimal sparsity value, such as k=70% in Figure 6, strikes a balance between capturing sufficient global information and maintaining computational efficiency, ultimately contributing to improved performance in underwater image super-resolution tasks.

4.4.3 Effect of the number of HDTMs

Finally, we analyze the impact of the number of HDTMs in the network backbone. Figure 7 presents the quantitative results using different numbers of HDTMs. It can be observed that when the number ranges from 6 to 8, the growth of PSNR value gradually converges. Therefore, to balance model efficiency and performance, we ultimately choose $N = 6$ as the configuration for the final network.

4.5 Limitations

While our proposed method demonstrates superior performance on classical underwater image super-resolution datasets (visible data) (Liu et al., 2024), its applicability is currently somewhat limited. The model's performance is significantly affected in scenarios with low light conditions, such as deep-sea environments or areas with poor visibility, where methods utilizing sonar (Yang, 2023; Zhang et al., 2024) for detection are more prevalent. To adapt our method to a wider range of underwater scenarios, we will explore the potential applications of the proposed method in sonar images.

5 Conclusions

In this paper, we have proposed an effective hybrid dynamic Transformer for underwater image super-resolution. We demonstrate the crucial importance of jointly exploring local features and global information in underwater image reconstruction for achieving high-quality results. At the technical level, we integrate dynamic local self-attention and sparse non-local self-attention to stack into the hybrid dynamic transformer module, forming the backbone of our proposed method. The former effectively captures details in underwater image regions, while the latter aids in the recovery of global image structure and color. Our proposed method achieves satisfactory reconstruction results on benchmark datasets. In future work, we will explore the extension of this hybrid transformer approach to other navigation-related visual tasks.

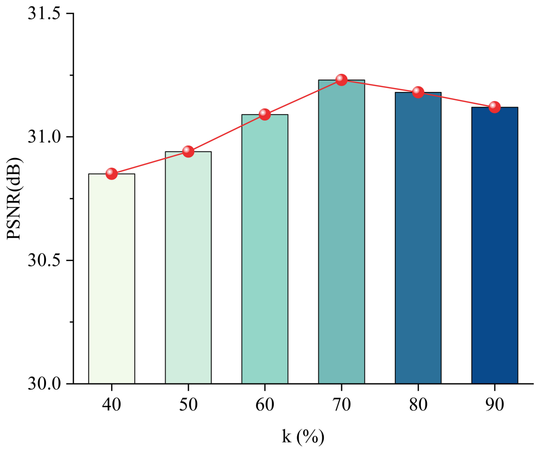

FIGURE 6
Comparison of ablation results about the effect of top-k values in the SNSA.

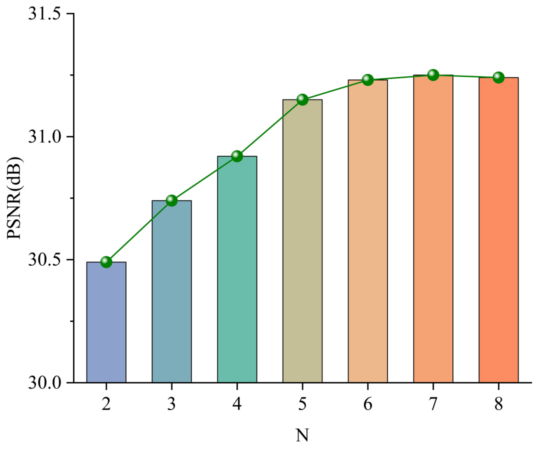

FIGURE 7
Comparison of ablation results about the effect of the number of HDTMs.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://drive.google.com/drive/folders/1dCe5rlw3UpzBs25UMXek1JL0wBBa697Q; https://www.v7labs.com/open-datasets/ufo-120.

Author contributions

XH: Data curation, Formal analysis, Methodology, Writing – original draft. JL: Investigation, Writing – review & editing. TJ: Data curation, Software, Visualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was partly supported by the Research Project of the Naval Staff Navigation Assurance Bureau (2023(1)).

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., et al. (2021). "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* (Virtual: IEEE), 12299–12310.

Chen, X., Li, H., Li, M., and Pan, J. (2023). "Learning a sparse transformer network for effective image deraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Canada: IEEE), 5896–5905. doi: 10.1109/CVPR52729.2023.00571

Chen, Y., Niu, K., Zeng, Z., and Pan, Y. (2020). "A wavelet based deep learning method for underwater image super resolution reconstruction," in *IEEE Access* (IEEE), 8, 117759–117769.

Dharejo, F. A., Ganapathi, I. I., Zawish, M., Alawode, B., Alathbah, M., Werghi, N., et al. (2024). Swinwave-sr: Multi-scale lightweight underwater image super-resolution. *Inf. Fusion* 103, 102127. doi: 10.1016/j.inffus.2023.102127

Dong, C., Loy, C. C., He, K., and Tang, X. (2015). "Image super-resolution using deep convolutional networks," in *IEEE transactions on pattern analysis and machine intelligence* (IEEE), 38, 295–307.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929.*

Guo, Y., Li, H., and Zhuang, P. (2019). "Underwater image enhancement using a multiscale dense generative adversarial network," in *IEEE Journal of Oceanic Engineering* (IEEE), 45, 862–870.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., et al. (2022). "A survey on vision transformer," in *IEEE transactions on pattern analysis and machine intelligence* (IEEE), 45, 87–110.

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition.* (USA: IEEE), 7132–7141.

Islam, M. J., Enan, S. S., Luo, P., and Sattar, J. (2020a). "Underwater image super-resolution using deep residual multipliers," in *2020 IEEE International Conference on Robotics and Automation (ICRA).* (Virtual: IEEE), 900–906. doi: 10.1109/ICRA40945.2020

Islam, M. J., Xia, Y., and Sattar, J. (2020b). "Fast underwater image enhancement for improved visual perception," in *IEEE Robotics and Automation Letters.* (Virtual: IEEE), 5, 3227–3234. doi: 10.1109/LSP.2016.

Jiang, N., Chen, W., Lin, Y., Zhao, T., and Lin, C.-W. (2021). "Underwater image enhancement with lightweight cascaded network," in *IEEE transactions on multimedia.* (IEEE) 24, 4301–4313.

Kim, J., Lee, J. K., and Lee, K. M. (2016). "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition.* (USA: IEEE) 1646–1654.

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., et al. (2017). "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition.* (USA: IEEE), 4681–4690.

Li, C., Anwar, S., Hou, J., Cong, R., Guo, C., and Ren, W. (2021). "Underwater image enhancement via medium transmission-guided multi-color space embedding," in *IEEE Transactions on Image Processing.* (IEEE) 30, 4985–5000.

Li, X., Dong, J., Tang, J., and Pan, J. (2023a). "Dlgsanet: lightweight dynamic local and global self-attention networks for image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision.* (France: IEEE), 12792–12801. doi: 10.1109/ICCV51070.2023.01175

Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., et al. (2019). "An underwater image enhancement benchmark dataset and beyond," in *IEEE Transactions on Image Processing.* (IEEE) 29, 4376–4389.

Li, Y., Shen, L., Li, M., Wang, Z., and Zhuang, L. (2023b). "Ruiesr: Realistic underwater image enhancement and super resolution," in *IEEE Transactions on Circuits and Systems for Video Technology.* (IEEE). doi: 10.1109/TCSVT.2023.3328785

Li, J., Skinner, K. A., Eustice, R. M., and Johnson-Roberson, M. (2017). "Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images," in *IEEE Robotics and Automation letters.* (IEEE) 3, 387–394.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. (2021). "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision.* (Virtual: IEEE), 1833–1844.

Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K. (2017). "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops.* (USA: IEEE), 136–144.

Liu, B., Ning, X., Ma, S., and Yang, Y. (2024). Multi-scale dense spatially-adaptive residual distillation network for lightweight underwater image super-resolution. *Front. Mar. Sci.* 10, 1328436. doi: 10.3389/fmars.2023.1328436

Liu, J., Tang, J., and Wu, G. (2020). "Residual feature distillation network for lightweight image superresolution," in *Computer Vision–ECCV 2020 Workshops*, UK, 2020. 41–55, Proceedings, Part III 16 (UK: Springer).

Luo, X., Xie, Y., Zhang, Y., Qu, Y., Li, C., and Fu, Y. (2020). "Latticenet: Towards lightweight image superresolution with lattice block," in *Computer Vision–ECCV 2020: 16th European Conference*, 2020. 272–289, Proceedings, Part XXII 16 (UK: Springer).

Mao, X.-J., Shen, C., and Yang, Y.-B. (2016). Image restoration using convolutional auto-encoders with symmetric skip connections. *arXiv preprint arXiv:1606.08921..*

Park, K., Soh, J. W., and Cho, N. I. (2021). "Dynamic residual self-attention network for lightweight single image super-resolution," in *IEEE Transactions on Multimedia.* 25, 907–918.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., et al. (2016). "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition.* (USA: IEEE), 1874–1883.

Shin, Y.-S., Cho, Y., Pandey, G., and Kim, A. (2016). "Estimation of ambient light and transmission map with common convolutional architecture," in *OCEANS 2016 MTS/IEEE Monterey (IEEE).* (IEEE) 2016, 1–7.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30.

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

Wang, L., Dong, X., Wang, Y., Ying, X., Lin, Z., An, W., et al. (2021). "Exploring sparsity in image super-resolution for efficient inference," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (USA: IEEE), 4917–4926.

Wang, Y., Zhang, J., Cao, Y., and Wang, Z. (2017). "A deep cnn method for underwater image enhancement," in *2017 IEEE international conference on image processing (ICIP)*. (IEEE) 2017, 1382–1386.

Xiao, Y., Yuan, Q., Jiang, K., He, J., Lin, C.-W., and Zhang, L. (2024). "Ttst: A top-k token selective transformer for remote sensing image super-resolution," in *IEEE Transactions on Image Processing*. (IEEE). doi: 10.1109/TIP.2023.3349004

Yang, P. (2023). An imaging algorithm for high-resolution imaging sonar system. *Multimedia Tools Appl.* 1–17. doi: 10.1007/s11042-023-16757-0

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M.-H. (2022). "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (USA: IEEE), 5728–5739.

Zhang, X., Yang, P., Wang, Y., Shen, W., Yang, J., Wang, J., et al. (2024). "A novel multireceiver sas rd processor," in *IEEE Transactions on Geoscience and Remote Sensing*. doi: 10.1109/TGRS.2024.3362886

Zhang, X., Zeng, H., Guo, S., and Zhang, L. (2022). "Efficient long-range attention network for image super-resolution," in *European Conference on Computer Vision*, (Israel: Springer), 649–667.