frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 15 July 2024
DOI 10.3389/fmars.2024.1411527

OPEN ACCESS

EDITED BY
Huiyu Zhou,
University of Leicester, United Kingdom

REVIEWED BY
Xuebo Zhang,
Northwest Normal University, China
Bangli Liu,
De Montfort University, United Kingdom

*CORRESPONDENCE
Zhibin Yu
yuzhibin@ouc.edu.cn
Bing Zheng
bingzh@ouc.edu.cn

RECEIVED 03 April 2024
ACCEPTED 24 June 2024
PUBLISHED 15 July 2024

CITATION
Zhang H, Fan S, Zou S, Yu Z and Zheng B (2024) Deep underwater image compression for enhanced machine vision applications. *Front. Mar. Sci.* 11:1411527. doi: 10.3389/fmars.2024.1411527

Deep underwater image compression for enhanced machine vision applications

Hanshu Zhang[1], Suzhen Fan[1], Shuo Zou[1], Zhibin Yu[1,2]* and Bing Zheng[1,2]*

[1]Sanya Oceanographic Institution, Ocean University of China, Sanya, China, [2]Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China

Underwater image compression is fundamental in underwater visual applications. The storage resources of autonomous underwater vehicles (AUVs) and underwater cameras are limited. By employing effective image compression methods, it is possible to optimize the resource utilization of these devices, thereby extending the operational time underwater. Current image compression methods neglect the unique characteristics of the underwater environment, thus failing to support downstream underwater visual tasks efficiently. We propose a novel underwater image compression framework that integrates frequency priors and feature decomposition fusion in response to these challenges. Our framework incorporates a task-driven feature decomposition fusion module (FDFM). This module enables the network to understand and preserve machine-friendly information during the compression process, prioritizing task relevance over human visual perception. Additionally, we propose a frequency-guided underwater image correction module (UICM) to address noise issues and accurately identify redundant information, enhancing the overall compression process. Our framework effectively preserves machine-friendly features at a low bit rate. Extensive experiments across various downstream visual tasks, including object detection, semantic segmentation, and saliency detection, consistently demonstrated significant improvements achieved by our approach.

KEYWORDS

underwater image compression, machine vision, frequency priors, feature fusion, deep learning

1 Introduction

The development of computer vision has greatly boost the advancement of underwater vision based marine research, including biological monitoring Gudimov (2020); Huo et al. (2021); Zhou et al. (2023), terrain mapping Rowley (2018); Nadai (2019); Jeyaraj et al. (2022), environmental surveillance Guo et al. (2020); Babić et al. (2023); Xue (2023),

fisheries management Hsu et al. (2019); Madia et al. (2023); Wang et al. (2023a), etc. In these research domains, underwater imagery is pivotal in acquiring marine visual information. Since underwater photography and image acquisition usually rely on potable devices, underwater image compression is always required.

Learning-free techniques like JPEG Wallace (1991), JPEG2000 Rabbani and Joshi (2002), BPG Sullivan et al. (2012), and VVC Bross et al. (2021) reduce intra-frame information redundancy through encoding, quantization, and intra-frame prediction. Recent advancements in image compression methods based on deep learning networks have revealed their superior potential compared to conventional approaches Ballé et al. (2016, 2018); Sullivan et al. (2012); Minnen et al. (2018); He et al. (2021, 2022); Bross et al. (2021). These deep learning-based image compression methodologies leverage deep neural networks to acquire image data's intrinsic features and compression strategies, aiming for higher compression rates and improved image quality. Unfortunately, current image compression methods are typically designed for terrestrial images. Applying these compression methods to underwater images makes it easy to trigger image information loss, which can be crucial to downstream visual tasks (e.g., image classification Deng et al. (2009); He et al. (2016); Sandler et al. (2018), object detection Redmon et al. (2016); He et al. (2017); Ren et al. (2017), and semantic segmentation models Long et al. (2015); Badrinarayanan et al. (2017); Chen et al. (2018a), as depicted in the Figure 1.

Due to the distinctive characteristics of the underwater environment, existing image compression methods suffer from two primary drawbacks during underwater image compression tasks. On the one hand, while these methods enhance the quality of reconstructed images to some extent, their primary focus is preserving pixel-level fidelity as perceived by the human visual system rather than facilitating feature recognition in machine visual applications Fang et al. (2023). Without considering the requirements of the underwater downstream visual tasks, the preserved information can be useless or even adverse to underwater downstream visual tasks.

On the other hand, current learning-based or learning-free compression methods are mainly designed to remove redundant information in terrestrial environments, in which typically exhibit uniform color distribution and high clarity Ancuti et al. (2012). However, underwater photos are highly susceptible to color bias, scattering, motion blur, and other distortions, which are quite different with the terrestrial environments Pei et al. (2018). The noise caused by the underwater environment can affect image compression and downstream visual tasks Jiang et al. (2020); Brummer and De Vleeschouwer (2023). Due to the enormous gap between the terrestrial and underwater domains, the experience for redundant information definition in terrestrial environments does not apply to underwater environments. In other words, the removed 'redundancy' information defined in these conventional compression methods may be useful in underwater downstream visual tasks.

Learning-based visual tasks are fundamental for underwater automation. For high-quality images, advanced visual tasks, such as image classification Deng et al. (2009); He et al. (2016); Sandler et al. (2018), object detection Redmon et al. (2016); He et al. (2017); Ren et al. (2017), and semantic segmentation models Long et al. (2015); Badrinarayanan et al. (2017); Chen et al. (2018a) can efficiently accomplish machine visual tasks by learning discriminative features. However, if we consider these tasks with underwater image compression, the accumulation loss of information due to underwater degradation and image compression can significantly impact the performance of reconstructed images in downstream machine visual tasks. Therefore, our primary concerns are effectively introducing underwater image transformation into the compression framework and obtain more machine-friendly feature representations. Following the learning-based compression framework, we introduce a task-driven feature decomposition fusion module (FDFM) to help the network understand and preserve machine-friendly information during the compression process. This allows the network to concentrate on information pertinent to the task, prioritizing task relevance over human visual perception. Furthermore, we propose a frequency-guided underwater image correction module (UICM) to reduce the impact of noise caused by the underwater environment and to accurately locate the redundant information that can be eliminated. To this end, we propose a novel underwater image compression framework that facilitates downstream visual tasks in underwater scenarios. The overall framework is illustrated in Figure 2. The primary contributions of this work are summarized as follows:

- We have proposed a novel machine-oriented underwater image compression framework, which has achieved high compression rates and ensured the performance of downstream underwater visual tasks. Extensive experiments on three different downstream visual tasks further demonstrate the consistent and significant improvements achieved by our method.
- To alleviate the impact of information loss caused by underwater degradation during the image compression process, we propose a frequency-guided underwater image correction module (UICM) that leverages frequency priors to remove the correct redundant information.
- We introduce a task-driven feature decomposition fusion module (FDFM). Under the guidance of downstream visual tasks, this module can effectively capture and keep machine-friendly information during the image compression process.

2 Related works

2.1 Image compression

Image compression uses reversible function mapping and encoding techniques to represent the original image data losslessly or lossily using fewer bits.

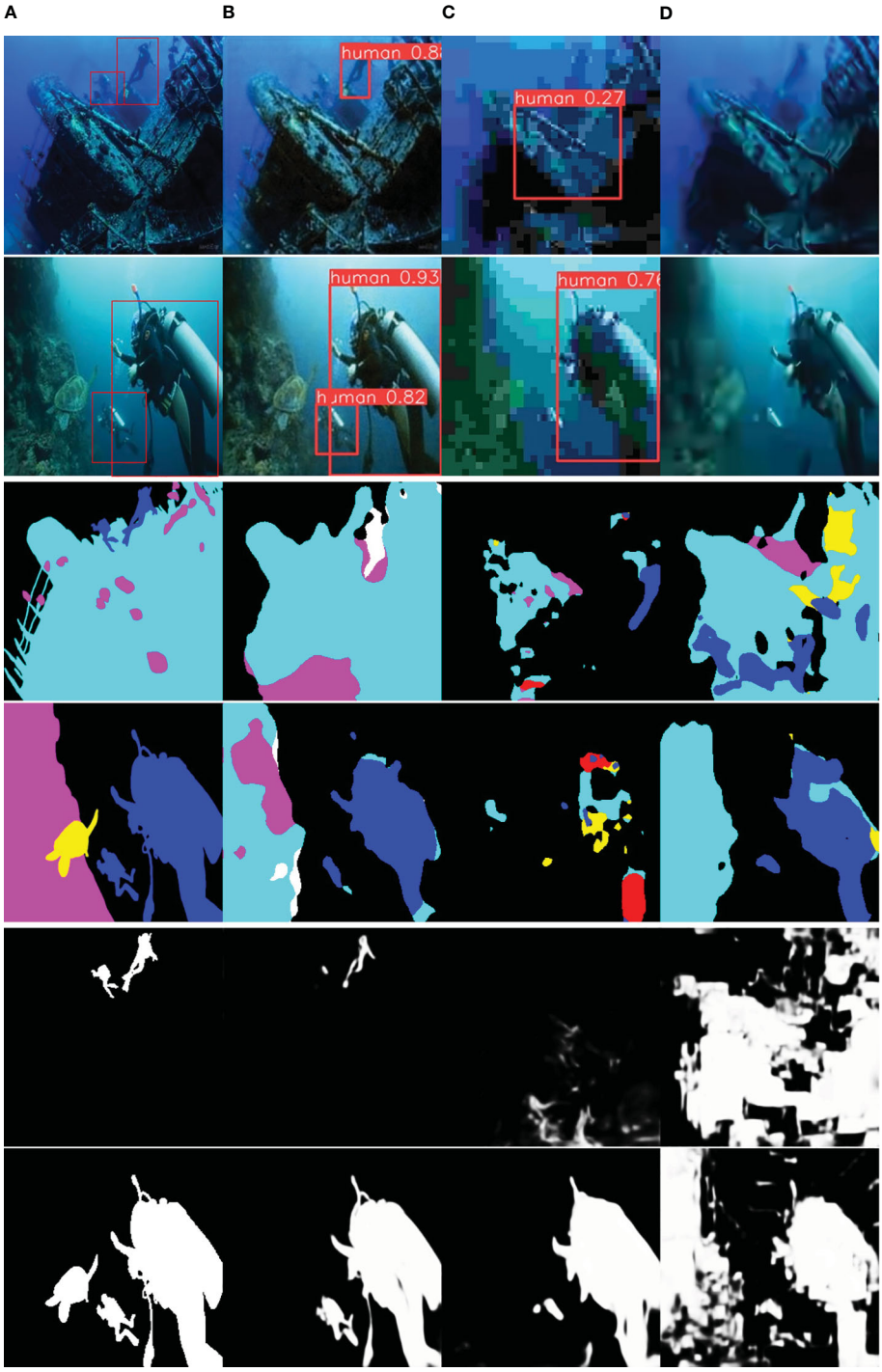

FIGURE 1
(A) source image, **(B)** the proposed method, **(C)** JPEG, and **(D)** BPG. We provide three groups of downstream visual tasks including object detection, semantic segmentation, and saliency detection. The initial subset of results pertains to the object detection task, where the first column exhibits the original image. Notably, our approach achieves superior accuracy and confidence in three tasks.

2.1.1 Learning-free image compression

In early years, learning-free image compression algorithms, including JPEG Wallace (1991), JPEG2000 Rabbani and Joshi (2002), BPG Sullivan et al. (2012), and VVC Bross et al. (2021), have gained widespread practical adoption due to their extensive development. These algorithms employ lossy compression techniques, such as transform Khayam (2003); Al-Haj (2007), quantization, entropy coding Di et al. (2003); Sze and Budagavi (2012), intra-frame prediction Brand et al. (2019), and deep hierarchical structure Motl and Schulte (2015), to process images. However, the individual components of these standards are manually designed in advance, with rate-distortion optimization applied to determine pixel signal fidelity. The rigid, hand-crafted nature of traditional codecs limits their adaptability and efficiency

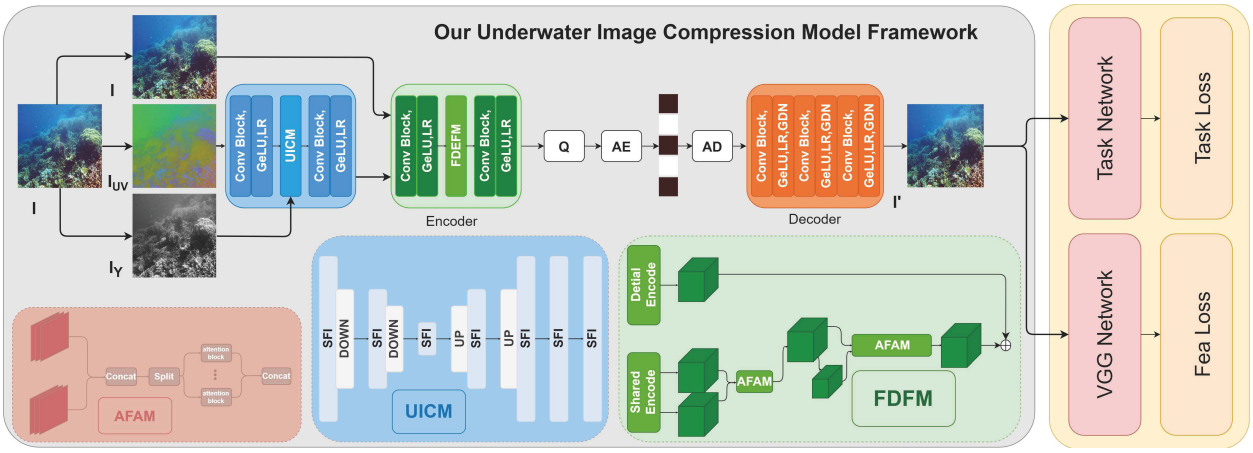

FIGURE 2
The details of the method we proposed. UICM, FDFM, AFAM respectively represent the feature recovery module, feature decomposition fusion module, and attention feature aggregation module, respectively. Q, AE and AD indicate quantization, arithmetic encoding and arithmetic decoding respectively. I_Y and I_{UV} represent the luminance and chrominance components of the image in the YCbCr color space.

in catering to diverse targets. Since they lack end-to-end optimization, they cannot dynamically adjust to image content characteristics. Consequently, compression requirements vary for different image types, scenarios and complexities, posing challenges to learning-free image compression methods.

2.1.2 Learning-based image compression

The rapid development of deep learning networks has significantly boost learning-based image compression methods. Notably, methods based on Variational Autoencoders (VAE) Ballé et al. (2016, 2018); Minnen et al. (2018); Cheng et al. (2020); Li et al. (2020); He et al. (2021; Chen et al. (2021); Zhu et al. (2022), 2022); Zou et al. (2022) employ encoders and decoders to compress images, focusing on compressing latent features. These approaches optimize the network in an end-to-end fashion, resulting in a high-performance compression framework. For instance, Ballé et al. (2016) introduced an image compression method incorporating a nonlinear analysis transform, a uniform quantizer, and a nonlinear synthesis transform. This method laid the foundation for the image compression model based on the VAE model. Similarly, Ballé et al. (2018) proposed an image compression model based on variational autoencoders, combining priors to capture spatial dependencies in latent representations and training the model in an end-to-end manner. When trained on appropriate losses, the model cannot fully achieve the performance of highly optimized traditional methods (such as BPG based on PSNR). This difference may indicate that the method has not yet reached the expressive power of traditional methods. In another approach, Minnen et al. (2018) enhanced an image compression method by refining the entropy model with an autoregressive model. The synergy between the autoregressive model and the prior model leads to improved image indicators, such as PSNR and MS-SSIM, outperforming the BPG Sullivan et al. (2012) method. However, the sequential computational approach of autoregressive models results in low operational efficiency. Moreover, Zhu et al. (2022) presented an image compression method using a multivariate Gaussian mixture, employing vector quantization to approximate the mean and solving it through cascaded estimation, avoiding the need for a context model and reducing complexity. However, this method is trained in an unsupervised manner, and the generated results may be biased. Li et al. (2020) introduced a content-weighted codec model, which generates an importance mask for local adaptive bit allocation through an importance mapping subnet, offering an alternative to entropy estimation. This method improves image compression efficiency while reducing the computational complexity of the context model. Chen et al. (2021) introduced an image compression method that combines non-local attention optimization with improved context modeling. This method utilizes local network operations as nonlinear transformations, estimating the corresponding latent features and priors by calculating local and global correlation information. This method leverages joint 3D convolution to enhance both the autoregressive model and the hyperprior model, improving the efficiency of the entropy model. Experimental results have demonstrated that this method outperforms JPEG, JPEG2000, and BPG in terms of image compression efficiency. Cheng et al. (2020) proposed an entropy model with enhanced flexibility in latent representation distribution estimation through a discretized Gaussian mixture model. Additionally, the performance was improved by incorporating an attention module to focus on complex regions. This method pays more attention to information-rich regions during the training process, thus improving the encoding performance. He et al. (2021, 2022) surpassed the compression efficiency of VVC by employing a checkerboard context model and unevenly grouped space channels. These two methods increase the decoding speed of the autoregressive model by more than 40 times, improving the parallelism and computational efficiency of the autoregressive model. Zou et al. (2022) presented a plug-and-play non-overlapping window local attention block, which calculates the attention map for each window using an embedded Gaussian function and normalization factors to focus on high-contrast

regions. Tolstonogov and Shiryaev (2021) present an underwater image compression method based on camera frames, involving semantic segmentation, semantic shape simplification, and binary data compression. Compared to the JPEG algorithm, this method achieves a threefold increase in frame rate. Anjum et al. (2022) introduces a data-driven underwater image compression method for transmitting images through water. This method effectively utilizes limited bandwidth to transmit images and exhibits robustness against disturbances caused by channel transmission. Burguera and Bonin-Font (2022) proposes a progressive underwater image compression method that divides images into small blocks that can be transmitted separately. Experimental results have shown that this method performs well in low bandwidth or unreliable communication channel environments. Liu et al. (2023) introduces an autoencoder-based underwater image compression technique. This method enhances the reliability of encoding through a multi-step training strategy and multi-description encoding policy. Despite the remarkable performance of VAE-based methods, they are primarily designed to preserve pixel-wise signal fidelity rather than high-level semantic features, which are required in downstream visual tasks.

In parallel to the approaches above, certain studies Agustsson et al. (2019); Wu et al. (2020); Liu et al. (2021a) have explored generative adversarial networks (GANs) to generate visually pleasing textures at low bit rates. GAN-based image compression offers several notable advantages. Firstly, GANs can compress full-resolution images, showcasing the versatility of this approach. Secondly, GANs are capable of achieving extreme bit-rate image compression. However, it is essential to note that the generated images may exhibit significant differences from the original ones, resulting in a potentially deceptive perception of clarity and high resolution.

2.2 Underwater downstream visual tasks

2.2.1 Object detection

The authors of Ellen et al. (2023) utilized underwater drones with YOLOv5 to detect submerged objects, achieving considerable accuracy. In Zhang and Zhu (2023), the authors improved YOLOv5 by implementing coordinate attention mechanisms and bidirectional feature pyramids, resulting in enhanced precision in ship detection. The work in Ranolo et al. (2023) compared the detection results of seaweed using YOLOv5 and YOLOv3, with YOLOv3 exhibiting higher accuracy. The method proposed in Gao et al. (2023b) significantly increased the detection accuracy in sonar imagery by denoising sonar images and enhancing YOLOv5. The approach in Ercan et al. (2022) involves detecting targets in swimming pools through cloud-based computing. In Fu et al. (2022), the authors utilized K-means to recluster target anchor frames, improving YOLOv5's accuracy in detecting small objects in side-scan sonar images. The authors of Hu and Xu (2022) reduced the backbone size of YOLOv5 and restructured the feature pyramid, introducing a novel method for underwater debris detection. The method presented in Xu and Matzner (2018) conducts a comparative analysis of fish detection across multiple datasets and suggests using different datasets during the detector training process. The sonar is an essential tool for the underwater image target detection. Zhang et al. (2024) developed a chirp scaling algorithm based on the reformulated Loffeld's bistatic formula. Compared with the traditional method, the proposed method is much more efficient and can be directly applied to multichannel and tandem synthetic aperture radar. Yang (2024) proposes an imaging algorithm based on Loffeld's bistatic formula for a multireceiver synthetic aperture sonar system. The presented method can produce high-resolution images.

2.2.2 Semantic segmentation

The authors of Nezla et al. (2021) used a deep convolutional encoder-decoder model based on the UNet architecture to segment the Fish4Knowledge image dataset, achieving commendable scores. Using a self-supervised approach, the method proposed in Singh et al. (2023) addresses the lack of large labeled datasets in underwater scenarios. This approach allows pretraining on extensive terrestrial datasets and fine-tuning on smaller underwater datasets. Kabir et al. (2023) introduced a novel underwater dataset centered around animals, with pixel-level annotations for various fine-grained animal categories. In Pergeorelis et al. (2022), the authors tackled the issue of class instance imbalance in underwater datasets by employing a scheme that involves cutting and pasting objects from one image to another. Chicchon et al. (2023) presented a combination loss function based on active contour theory and level-set methods to enhance underwater object segmentation accuracy. Wang et al. (2023b) employed a semi-supervised K-means clustering algorithm to train and validate objects like coral, sea urchins, starfish, and seagrass. Islam et al. (2020) proposed the first underwater semantic segmentation dataset, containing pixel annotations for eight object categories, and suggested that deep residual models can accurately segment underwater objects. Thampi et al. (2021) analyzed the impact of different thresholds on predicted masks for the underwater semantic segmentation of five different fish species in the Fish4Knowledge image dataset.

Despite the widespread application of advanced visual tasks in underwater environments, most require clear input images. Information loss caused by underwater degradation and image compression can affect the performance of these methods.

3 The proposed method

3.1 The overall architecture

The details of the proposed methodology are illustrated in Figure 2. To the impact of information loss caused by underwater degradation during the image compression process, we introduce the frequency-guided underwater image correction module (UICM). This module aims to reduce the impact of noise caused by the underwater environment and remove redundant information accurately. The subsequent advancement toward enhancing encoding efficiency at low bit rates involves the utilization of the task-driven feature decomposition fusion module (FDFM) for decomposing features according to their relevance to downstream

underwater visual tasks. This procedure preserves machine-friendly data while eliminating redundancy, yielding a concise, machine-friendly feature representation and reduced bit rate while retaining key features. Finally, a machine-friendly image is reconstructed in the decoder stage to facilitate diverse downstream visual tasks.

3.2 Frequency-guided underwater image correction module

Due to the complexity of optical imaging in underwater environments compared to terrestrial environments, underwater images are often subject to noise interference. Since image noise is non-compressible and irrelevant to downstream visual tasks, the compressed image bit rate will be lower than the standard Brummer and De Vleeschouwer (2023). The work on Xu et al. (2020) suggests the varying significance of different frequency channels in visual tasks. We have designed a frequency-guided underwater image correction module (UICM) to address this issue to eliminate noise and pinpoint removable redundant information. The structure of UICM is illustrated in Figure 2.

Firstly, we revisit the operations and properties of the discrete cosine transform(DCT). DCT is an orthogonal transformation method. Compared with the fast Fourier transform (FFT) and the discrete wavelet transform (DWT), DCT can save computation and maintain good performance Wen et al. (2022). Given a single-channel image f of size $N \times N$, the discrete cosine transform D transforms it into the discrete cosine space as X, which is expressed as Equation 1:

$$\begin{cases} X(u,v) = D_{f(i,j)} = c(u)c(v)\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \\ f(i,j)\cos\left[\frac{(i+0.5)\pi}{N}u\right]\cos\left[\frac{(j+0.5)\pi}{N}v\right] \\ c(u) = \begin{cases} \sqrt{\frac{1}{N}}, u = 0 \\ \sqrt{\frac{2}{N}}, u \neq 0 \end{cases} \\ c(v) = \begin{cases} \sqrt{\frac{1}{N}}, v = 0 \\ \sqrt{\frac{2}{N}}, v \neq 0 \end{cases} \end{cases} \tag{1}$$

where i and j are the coordinate bases in the spatial space; u and v are the coordinate bases in the discrete cosine transform space and D^{-1} denotes the inverse discrete cosine transform.

The image features affected by the underwater environment $I_{degradation}$ can be expressed as:

$$I_{degradation} = \sum_i I^{s_i}_{degradation} \tag{2}$$

where $I^{s_i}_{degradation}$ represents image features affected by the underwater environment at different scales and s_i represents different scale ranges. DCT can effectively model noise signals and redundant signals. Let $s^{s_i}_{uv}$ represent component of $I^{s_i}_{degradation}$ in the DCT space. Equation 2 can be reexpressed as Equation 3:

$$I_{degradation} = D^{-1}\left(\sum_{s_i}\sum_u\sum_v s^{s_i}_{uv}\right) \tag{3}$$

where s_i represents different scale ranges; u and v are the coordinate bases in the discrete cosine transform space.

Let Q represent the expected image features with low noise and low redundancy, and its component in the DCT space is denoted as $q^{s_i}_{uv}$. Q can be formulated as follows:

$$Q = D^{-1}\left(\sum_{s_i}\sum_u\sum_v q^{s_i}_{uv}\right) \tag{4}$$

where s_i represents different scale ranges; u and v are the coordinate bases in the discrete cosine transform space.

The difference between $I_{degradation}$ and Q in the DCT space, namely the spectral loss $z^{s_i}_{uv}$, can be represented as Equation 5:

$$z^{s_i}_{uv} = \sum_{s_i}\sum_u\sum_u(q^{s_i}_{uv} - s^{s_i}_{uv}) \tag{5}$$

where s_i represents different scale ranges; u and v are the coordinate bases in the discrete cosine transform space.

Equation 4 can be reexpressed as Equation 6:

$$Q = D^{-1}\left(\sum_{s_i}\sum_u\sum_u(s^{s_i}_{uv} + z^{s_i}_{uv})\right) \tag{6}$$

where s_i represents different scale ranges; u and v are the coordinate bases in the discrete cosine transform space.

Conventional approaches reliant on DCT space aim to directly adjust DCT coefficients, posing significant challenges for practical implementation. Drawing inspiration from Zheng et al. (2019), we leverage a convolutional neural network (CNN) to estimate $z^{s_i}_{uv}$. Acknowledging the influence of diverse-scale features and frequencies on images, our approach entails image adjustment across multiple scales.

UICM employs frequency-space interaction blocks (FSI) as illustrated in Figure 3 as fundamental units. The FSI block consists of a frequency branch and a spatial branch to learn global and local information, respectively. The frequency domain representation emphasizes global attributes, while the local attributes are learned in the spatial branch. These two branches interact to obtain complementary information. The frequency branch estimates the spectrum loss $z^{s_i}_{uv}$ in the DCT space via the CNN block and then converts it to the color space through block-IDCT. Block-IDCT uses a predefined convolutional layer with weights fixed as the D^{-1} coefficient. The spatial branch processes information in the spatial domain through convolutional blocks. We then interweave features from the spatial and frequency branches, facilitating the acquisition of more information by different branches. The FSI will then repeat the same calculation once more. Finally, we merge the outputs of the two branches using 1×1 convolution to obtain the output of the FSI block.

3.3 Task-driven feature decomposition fusion module

To ensure the image compression network prioritizes machine-friendly features over preserving pixel-level fidelity as perceived by the human visual system, we employ the task-driven feature decomposition fusion module (FDFM). This module facilitates the preservation of machine-friendly information while eliminating redundancies. Guided by downstream visual tasks,

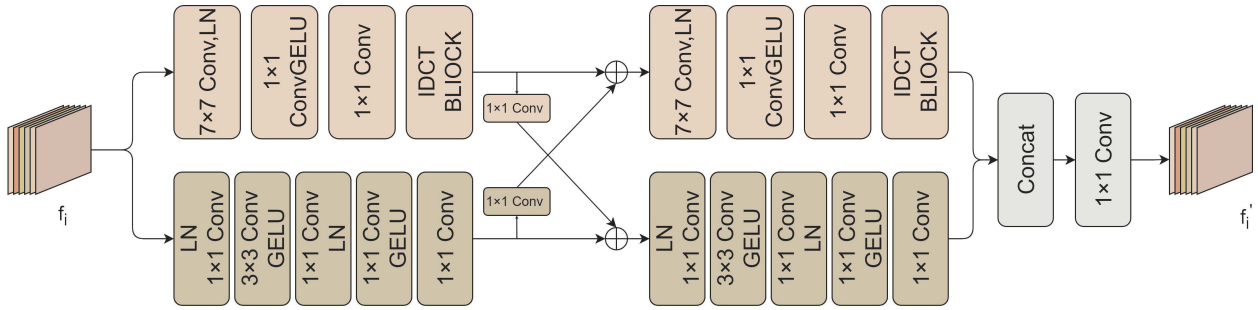

FIGURE 3
The illustration of the amplitude format of the FSI block. The FSI block consists of frequency and spatial branches to learn global and local information. The frequency domain representation emphasizes global attributes, while the local attributes are learned in the spatial branch. These two branches interact to obtain complementary information.

FDFM extracts machine-friendly details from both the original image and the image processed by UICM, effectively removing redundant information. Additionally, attention mechanisms are applied to discern the significance of pixels at various spatial positions. In alignment with downstream underwater visual tasks, distinct weights are assigned to individual pixels to mitigate information redundancy.

The detailed workflow is illustrated in Figure 2. The FDFM comprises three essential modules: a shared encoder Φ_l dedicated to extracting low-frequency features, a detailed encoder Φ_h specialized in capturing high-frequency features, and a decoder Ψ employed for the reconstruction of features with enhanced semantic information.

In a detailed approach, the FDFM model initiates the process by utilizing the shared encoder Φ_l and the detailed encoder Φ_h to dissect the low-frequency and high-frequency components of both the original source image I_{origin} and the image $I_{restored}$ processed through UICM. This results in the extraction of low-frequency information F_{lo} F_{le} and high-frequency information F_h, which is formulated as Equation 7. Drawing inspiration from recent advancements in backbone networks Ding et al. (2023, 2022, 2021); Liu et al. (2021b), we adopt the ConvNeXt Woo et al. (2023) structure for the detailed encoder.

$$F_{lo} = \Phi_l(I_{origin}), F_{le} = \Phi_l(I_{restored}), F_h = \Phi_h(I_{origin}) \tag{7}$$

In the current context, we possess low-frequency information denoted as F_{lo} and F_{le} extracted from both the source image I_{origin} and the restored image $I_{restored}$. The imperative task is to devise an efficient approach for integrating these information sets. Motivated by the positional attention mechanism discussed in Hou et al. (2021), which simultaneously empowers the neural network to assimilate information from diverse channels, we have formulated an attention feature aggregation module(AFAM). This module is specifically designed to handle features originating from various channels collaboratively. Moreover, this module analyzes pixel significance across various positions, utilizing coordinates to mitigate information redundancy. Initially, we engage in channel concatenation for the low-frequency information F_{lo} and F_{le}. Subsequently, we conduct computations employing operation *Coo*, culminating in the utilization of a 1×1 convolution operation to produce the final output, which is formulated as Equation 8:

$$F_{lc} = Pw(Coo(Cat(F_{lo}, F_{le}))) \tag{8}$$

where the F_{lc} means the integrated low-frequency information; *Pw* is indicative of the 1×1 convolution operation; *Coo* represents positional attention, and *Cat* stands for channel concatenation.

Multiscale learning enables the network to autonomously acquire global and local information from features at higher and lower resolutions. Consequently, we conduct scale decomposition on the acquired F_{lc}. We streamline Chen et al. (2022) and incorporate it as the feature extraction network. Subsequently, through AFAM fusion, we derive representations imbued with more profound semantic information. To encapsulate, the process above can be summarized as Equation 9:

$$\begin{aligned} F_{lc}^{Si} &= \phi(F_{lc}) \\ F_{lce} &= \Delta_{AFAM}\left(\textstyle\sum_i (F_{lc}^{Si})\right) \end{aligned} \tag{9}$$

where the ϕ signifies scale decomposition; F_{lc}^{Si} denotes the features subsequent to scale decomposition; and F_{lce} represents the augmented representation with enriched information post scale fusion.

In conclusion, we integrate F_{lce} with F_h. Drawing inspiration from He et al. (2016), we utilize skip connections to seamlessly amalgamate F_{lce} and F_h. Subsequently, the acquired features undergo reconstruction into features I_r endowed with more profound semantic information and less redundant information through the decoder Ψ. The process above can be summarized as Equation 10.

$$I_r = \Psi(F_{lce} + F_h) \tag{10}$$

3.4 Training

3.4.1 Loss function

In light of our approach to designing for downstream visual tasks, we employ four distinct loss functions to facilitate the training of our network.

L_{mse} is the reconstruction loss between the input image L and the reconstructed image L', used to constrain the pixel-level fidelity

of the reconstructed image L' to the input image L, which is formulated as Equation 11:

$$L_{mse} = MSE(I, I') \tag{11}$$

Inspired by the work of Johnson et al. (2016), we integrate the perceptual loss, denoted as L_{fea}, to accentuate the perceptual quality of the reconstructed image. Employing the initial three layers of a pre-trained VGG-19 Simonyan and Zisserman (2014) as feature extractors, we input both the original images I and reconstructed images I' to derive the corresponding output features. The loss is formulated by leveraging these features, expressed mathematically as Equation 12:

$$L_{fea} = \sum_i^N (F_i(I) - F_i(I')) \tag{12}$$

where the $F_i(I)$ and $F_i(I')$ denote the feature representations at the i layer within their pre-trained neural network; and the N represents the total number of layers.

In order to enhance the performance of the reconstructed image in sophisticated visual tasks, we incorporate diverse downstream task losses under the designation of the task loss L_{task}. The application of multiple loss constraints ensures that the reconstructed image aligns with the specific demands of a variety of downstream visual tasks. Throughout the training phase, the cumulative loss is denoted as Equation 13:

$$L_{total} = \lambda_1 L_{mse} + \lambda_2 L_{fea} + \lambda_3 L_{task} + L_{bit} \tag{13}$$

where L_{bit} represents the bit-rate of latent code; λ_1, λ_2 and λ_3 are hyperparameters that mediate the compression ratio of the network. The hyperparameters λ_1, λ_2 and λ_3 will all affect the results of the method. Typically, we set hyperparameters λ_1, λ_2 and λ_3 based on experience. Please refer to section 4.1 for detail.

3.4.2 Adaptive training strategy

The single-stage training strategy encounters challenges in achieving a harmonious equilibrium between low-level and high-level visual tasks. Current approaches for low-level visual tasks, propelled by their high-level counterparts, frequently employ pre-trained high-level visual models to direct the training of models dedicated to low-level visual tasks. Alternatively, some methodologies opt for concurrently training low-level and high-level visual tasks within a unified stage. Our strategy upholds the performance synergy between image fusion and semantic segmentation by subjecting the compression network and semantic segmentation network to alternating training. This method mitigates potential issues, such as mode collapse, commonly observed during Generative Adversarial Network (GAN) training Tang et al. (2022).

4 Experiments

4.1 Experimental setup

4.1.1 Datasets

SUIM Islam et al. (2020) is a dataset for semantic segmentation of underwater. It comprises over 1500 images, each pixel annotated for eight distinct object categories: vertebrate fish, invertebrate coral reefs, aquatic plants, sunken ships/ruins, human divers, robots, and the seabed. Following a predefined partitioning scheme, the dataset is divided into 1525 images for training and 110 for testing. The hyperparameters λ_1, λ_2 and λ_3 will all affect the results of the method. Typically, we set hyperparameters lambda based on experience.$\lambda_1/\lambda_2/\lambda_3$ are empirically set to 0.051/0.15/1, 0.051/0.5/1 and 0.051/2/1 under 0.1, 0.3 and 0.5 bpp respectively.

URPC2018 is a dataset for object detection of underwater. It compasses four distinct categories: sea cucumber, sea urchin, starfish, and scallop, comprising 2901 training images and 800 testing images. Our approach adheres to a pre-established partitioning scheme. The hyperparameters λ_1, λ_2 and λ_3 will all affect the results of the method. Typically, we set hyperparameters lambda based on experience. $\lambda_1/\lambda_2/\lambda_3$ are empirically set to 0.051/0.17/1, 0.051/0.5/1 and 0.051/2/1 under 0.028, 0.86 and 0.237 bpp respectively.

4.1.2 Compared methods

We assessed the efficacy of our proposed method through a comparative analysis with traditional and CNN-based compression methods. The entropy model is based on Zou et al. (2022). The traditional methods encompass JPEG Wallace (1991), JPEG2000 Rabbani and Joshi (2002), BPG (intra-frame, 4:4:4 chroma format) Sullivan et al. (2012), and VVC intra-frame (4:4:4 chroma format) Bross et al. (2021). Additionally, CNN-based methods such as Hyperprior (ICLR2018) Ballé et al. (2018), Devil (CVPR2022) Zou et al. (2022) and Gao Gao et al. (2023a) were included in the comparison.

We conducted an extensive series of experiments to assess the performance of the proposed underwater image compression model in downstream visual tasks downstream, encompassing object detection, semantic segmentation, and saliency detection.

4.2 Downstream visual tasks performance comparison

4.2.1 Object detection

We employed the Yolov8s framework for downstream object detection to present our findings. We fine-tune the detector using a pre-trained model on the COCO dataset Lin et al. (2014) for identifying targets such as humans, robots, invertebrates, vertebrates, and fish. The image dimensions were standardized to 640×640, and the detector underwent training using the Adam Kingma and Ba (2014) optimizer for 100 epochs, initialized with a learning rate of 0.00001. Notably, consistent settings were applied across various image compression methods. Evaluation of detection performance was based on the recall rate (RA) and the mean average precision (mAP_{50}).

Table 1 illustrates that our proposed method achieves high accuracy in target detection even under low bit rates. Specifically, under 0.1bpp, on SUIM dataset, our method surpasses JPEG, JPEG2000, BPG, and VVC in RA/mAP_{50} by 0.237/0.03267 points, 0.162/0.105 points, 0.102/0.065 points, 0.226/0.165 points, respectively. In comparison to Hyperprior Ballé et al. (2018), devil

Zou et al. (2022) and Gao Gao et al. (2023a) under 0.01bpp, our method demonstrates notable improvements of 0.091/0.059 points, 0.189/0.198 points and 0.028/0.053 points in RA/mAP_{50}. Noteworthy, under 0.3 and 0.5bpp, our method has a comfortable lead over the alternatives.

As shown in Table 2, our method demonstrates outstanding performance in the URPC2018 dataset. Specifically, under 0.028 bpp, our approach yields RA/mAP_{50} scores 0.066/0.09 points higher than those of VVC. In contrast, the reconstructed images produced by the JPEG2000 method suffer severe degradation, leading to a loss in analytical efficacy. Across different bitrates, our proposed method keeps ahead of various compression methods in underwater object detection tasks. Due to the influence of the underwater environment, the images in the URPC2018 dataset are blurry.

TABLE 1 Comparison on object detection tasks on SUIM dataset.

Method	bpp	*RA*	mAP_{50}
original	–	0.495	0.55
JPEG	0.230	0.279	0.235
JPEG2000	0.103	0.354	0.397
BPG	0.106	0.414	0.437
VVC	0.103	0.290	0.337
Hyperprior	0.100	0.425	0.443
Devil	0.114	0.327	0.304
Gao	0.105	0.488	0.449
Ours	0.103	0.516	0.502
JPEG	0.306	0.402	0.369
JPEG2000	0.304	0.438	0.473
BPG	0.321	0.487	0.489
VVC	0.289	0.432	0.492
Hyperprior	0.299	0.519	0.522
Devil	0.301	0.432	0.492
Gao	0.298	0.541	0.529
Ours	0.303	0.554	0.533
JPEG	0.502	0.443	0.514
JPEG2000	0.495	0.527	0.544
BPG	0.507	0.492	0.497
VVC	0.512	0.550	0.549
Hyperprior	0.504	0.510	0.541
Devil	0.512	0.428	0.532
Gao	0.509	0.570	0.551
Ours	0.501	0.578	0.558

*The top two scores are highlighted in red (the best) and blue (the second best).
*The minimum bpp of JPEG is 0.23.
We conducted experiments under different bits per pixel (bpp) settings.

TABLE 2 Comparison on object detection tasks on URPC2018 dataset.

method	bpp	*RA*	mAP_{50}
origin	–	0.622	0.688
JPEG	–	–	–
JEPG2000	0.030	0.015	0.01
BPG	0.027	0.222	0.196
VVC	0.028	0.229	0.202
Hyperprior	0.024	0.130	0.114
Devil	0.026	0.220	0.194
Gao	0.027	0.260	0.256
Ours	0.028	295	0.292
JPEG	0.209	0.135	0.136
JEPG2000	0.087	0.288	0.285
BPG	0.085	0.358	0.357
VVC	0.086	0.391	0.393
Hyperprior	0.087	0.321	0.314
Devil	0.088	0.349	0.341
Gao	0.089	0.398	0.410
Ours	0.086	0.416	0.436
JPEG	0.232	0.277	0.264
JEPG2000	0.237	0.439	0.466
BPG	0.229	0.47	0.487
VVC	0.234	0.482	0.519
Hyperprior	0.237	0.429	0.44
Devil	0.233	0.398	0.413
Gao	0.233	0.499	0.535
Ours	0.237	0.503	0.541

*The top two scores are highlighted in red (the best) and blue (the second best).
*The minimum bpp of JPEG is 0.209.
We conducted experiments under different (bpp) settings.

Qualitative analysis is shown in the Figure 4. Experimental results demonstrate that our method performs well even on blurry images.

4.2.2 Semantic segmentation

We employed DeepLabV3+ Chen et al. (2018b) as the semantic segmentation framework to present our findings. The segmentation framework underwent fine-tuning, utilizing a pre-trained model from Imagenet dataset Deng et al. (2009), for the precise segmentation of targets including vertebrate fish, invertebrate coral reefs, aquatic plants, sunken ships/ruins, human divers, robots, and the seabed. Standardizing the image size to 256×256, the segmentation framework underwent training with the Adam Kingma and Ba (2014) optimizer for 100 epochs, commencing with an initial learning rate of 0.0001. Notably, consistent settings were applied across diverse image compression methods. Evaluation of

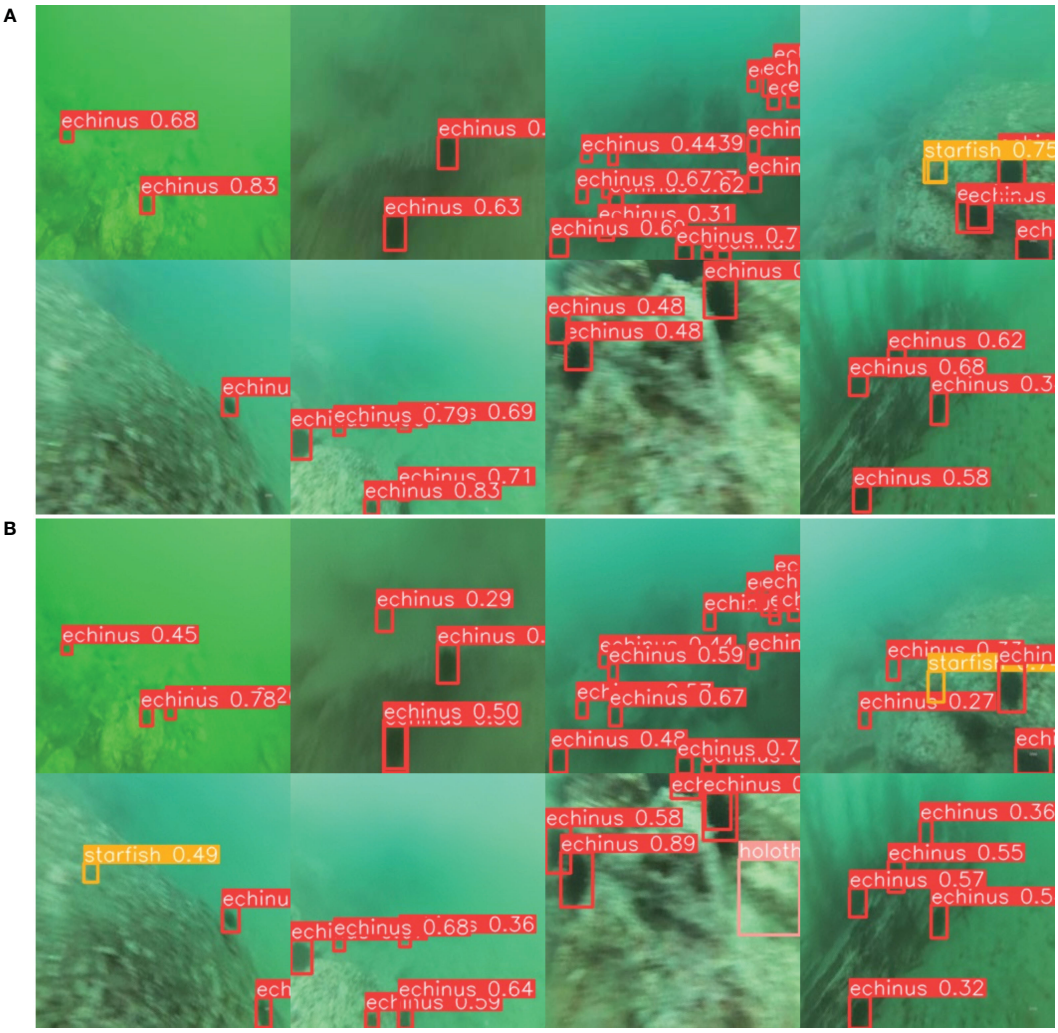

FIGURE 4
Qualitative analysis of object detection conducted on the URPC2018 dataset. Where **(A)** are original images, and **(B)** are reconstructed images using the proposed method under 0.237 bpp.

segmentation performance was conducted using mean Intersection over Union (*mIOU*), mean Pixel Accuracy (*mPA*), and Pixel Accuracy (*PA*).

Table 3 shows the comparisons among different methods in semantic segmentation tasks. It is evident that our approach outperforms the other methods. As we discussed in section 1, the high-level features in underwater images can be easily affected, posing challenges for downstream visual tasks. Unlike our methods, current CNN-based compression methods still focus on mitigating pixel distortion without considering the key features required by semantic segmentation and other downstream visual tasks. For example, under approximately 0.1bpp, our proposed method attains higher *mIOU*/*mPA*/*PA* scores than JPEG, JPEG2000, BPG, and VVC by 12.84/13.22/6.8 points, 5.09/3.98/1.78 points, 3.96/4.63/1.34 points, 0.44/1.21/1.08 points, respectively. In comparison to Hyperprior Ballé et al. (2018), devil Zou et al. (2022) and Gao Gao et al. (2023a) under 0.01bpp, our method remains ahead of 0.42/1.63/1.38 points, 10.56/10.36/6.3 points and 0.46/1.83/1.26 points in *mIOU*/*mPA*/*PA*.

4.2.3 Saliency detection

We employed the U^2net Qin et al. (2020) framework for underwater saliency detection with different compression frameworks. The saliency detection framework underwent fine-tuning utilizing a pre-trained model on the DUTS dataset Piao et al. (2020), specifically targeting human divers, robots, fish, and vertebrates. The dimensions of the images in the detection framework were standardized to 320×320, and the training process utilized the AdamW Loshchilov and Hutter (2017) optimizer for 360 epochs, initializing with a learning rate of 0.001. Consistency was maintained across various image compression methods as we adhered to the same settings. Our evaluation of the detection performance relies on mean absolute error (*MAE*) and maximal F-measure ($maxF_\beta$) Achanta† et al. (2009).

Table 4 reveals that our proposed method achieved the best performance in saliency detection even under low bit rates. Specifically, under 0.1bpp, our method's $MAE/maxF_\beta$ outperforms JPEG, JPEG2000, BPG, and VVC by 0.025/0.115

TABLE 3 Comparison on semantic segmentation tasks on SUIM dataset.

Method	bpp	*mIOU*	*mPA*	*PA*
original	–	62.1	72.67	84.13
JPEG	0.230	41.71	53.19	73.90
JPEG2000	0.103	49.46	62.43	78.92
BPG	0.106	50.59	61.78	79.36
VVC	0.103	54.11	65.20	79.62
Hyperprior	0.100	54.13	64.78	79.32
Devil	0.114	43.99	56.05	74.40
Gao	0.109	54.09	64.58	79.44
Ours	0.103	54.55	66.41	80.70
JPEG	0.306	52.80	64.64	79.20
JPEG2000	0.304	57.59	68.73	81.71
BPG	0.321	54.15	65.16	80.66
VVC	0.289	54.98	65.51	79.96
Hyperprior	0.299	57.85	68.68	82.59
Devil	0.301	57.01	67.56	81.42
Gao	0.305	57.69	68.69	82.40
Ours	0.303	58.73	69.48	83.04
JPEG	0.502	57.98	68.44	81.61
JPEG2000	0.495	59.33	70.31	82.49
BPG	0.507	57.88	69.02	81.68
VVC	0.512	58.72	69.49	82.71
Hyperprior	0.504	60.42	70.84	82.26
Devil	0.512	60.02	70.52	83.30
Gao	0.498	59.98	70.44	83.02
Ours	0.501	61.56	71.43	83.65

*The top two scores are highlighted in red (the best) and blue (the second best).
*The minimum bpp of JPEG is 0.23.
We conducted experiments under different (bpp) settings.

TABLE 4 Comparison on saliency detection tasks on SUIM dataset.

Method	bpp	*MAE*	*maxF*$_\beta$
original	–	0.031	0.785
JPEG	0.23	0.070	0.614
JPEG2000	0.103	0.057	0.671
BPG	0.106	0.059	0.671
VVC	0.103	0.049	0.706
Hyperprior	0.100	0.047	0.698
Devil	0.114	0.067	0.622
Gao	0.108	0.047	0.711
Ours	0.103	0.045	0.729
JPEG	0.306	0.054	0.686
JPEG2000	0.304	0.046	0.718
BPG	0.321	0.042	0.728
VVC	0.289	0.042	0.732
Hyperprior	0.299	0.042	0.739
Devil	0.301	0.043	0.733
Gao	0.308	0.038	0.760
Ours	0.303	0.036	0.769
JPEG	0.502	0.039	0.729
JPEG2000	0.495	0.036	0.759
BPG	0.507	0.039	0.745
VVC	0.512	0.038	0.758
Hyperprior	0.504	0.036	0.75
Devil	0.512	0.044	0.738
Gao	0.509	0.034	0.775
Ours	0.501	0.034	0.780

* The top two scores are highlighted in red (the best) and blue (the second best).
* The minimum bpp of JPEG is 0.23.
We conducted experiments under different (bpp) settings.

points, 0.012/0.058 points, 0.014/0.058 points, 0.004/0.023 points, respectively. In comparison to Hyperprior Ballé et al. (2018), devil Zou et al. (2022) and Gao Gao et al. (2023a) under 0.01bpp, our method showcases improvements of 0.002/0.031 points, 0.022/0.107 points and 0.002/0.018 points in *MAE*/*maxF*$_\beta$. Under 0.3 and 0.5bpp, our method consistently maintains superior performance. It is evident that our method can efficiently support underwater saliency detection tasks.

Figure 5 illustrates a qualitative analysis of the outcomes obtained from various methods across three tasks: object detection, semantic segmentation, and saliency detection. In the initial row of each set, we display the bounding boxes and confidence levels associated with the object detection results, with the initial image serving as a representation of the original image. Evidently, underwater images compressed with our approach remain high detection accuracy and confidence scores. The effectiveness of the object detector can be easily constrained by the impact of underwater degradation with compression. Nevertheless, our UICM systematically removes noise and redundant information, resulting in specific detection outcomes surpassing those of the original images. In the subsequent row of each set, the initial image serves as the ground truth for the semantic segmentation task, with varied colors denoting distinct categories. For semantic segmentation task, our approach obtained segmentation accuracy compared to alternative methods, producing contours that align more closely with the ground truth. Additionally, our approach demonstrates comparable efficacy in salient object detection, as depicted in the third-row results, where the initial image serves as the ground truth.

Through a comprehensive examination of both qualitative and quantitative outcomes in the three tasks above, our proposed method has exhibited superior performance on various downstream visual tasks.

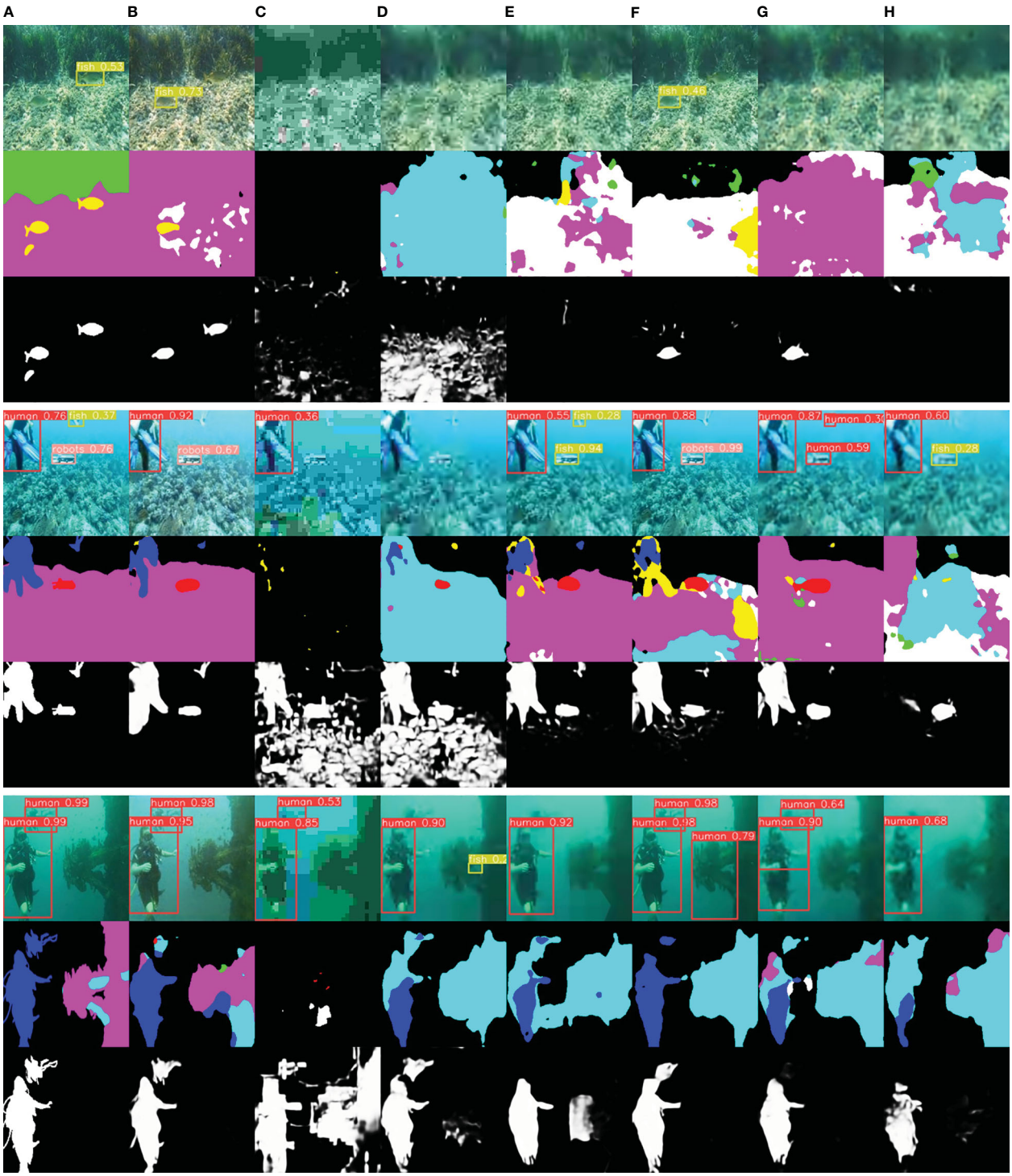

FIGURE 5
Examples results of **(A)** original image, **(B)** our method, **(C)** JPEG, **(D)** JPEG2000, **(E)** BPG, **(F)** VVC, **(G)** Hyperprior, **(H)** Devil. For each group, the results of object detection, segmentation and saliency detection are shown respectively. Our method has better performance compared to other methods.

4.3 Ablation study

We conducted some ablation experiments to validate the contributions of the proposed UICM and FDFM. We compared the results of object detection, semantic segmentation, and saliency detection for three network structures: (a) without UICM, (b) without FDFM, and (c) without both UICM and FDFM.

The ablation experiments' outcomes for the target detection task are presented in Table 5. Among the experimental setups, (b) demonstrates superior performance in RA and mAP_{50}. In

comparison to (c), (b) exhibits a notable enhancement of 0.081/0.037 points in both *RA* and mAP_{50}. Similarly, when contrasted with (a), (b) manifests an improvement of 0.009/0.005 points in *RA* and mAP_{50}. Furthermore, in contrast to (c), (a) displays an increase of 0.072/0.032 points *RA* and mAP_{50}.

We obtained similar performance on semantic segmentation tasks, as depicted in Table 6. Compared to (c), (b) demonstrates a notable improvement of 0.93/0.63/1.28 points in *mIOU*, *mPA*, and *PA*, respectively. Compared with (a), (b) displays a modest enhancement of 0.06/0.03/0.13 points in *mIOU*, *mPA*, and *PA*. Similarly, compared to (c), (a) exhibits an increase of 0.87/0.6/1.15 points in *mIOU*, *mPA*, and *PA*.

The outcomes of the ablation experiments conducted for the saliency detection task are presented in Table 7, unveiling consistent patterns in the results of the saliency detection task. In comparison to (c), (b) demonstrates a noteworthy improvement of 0.004/0.015 points in $MAE/maxF_\beta$. Similarly, contrasted with (a), (b) exhibits a modest enhancement of 0.001/0.012 points in $MAE/maxF_\beta$. Furthermore, when compared to (c), (a) shows an increase of 0.003/0.003 points in $MAE/maxF_\beta$.

The aforementioned experimental results validate the effectiveness of UICM and FDFM. UICM incorporates underwater prior knowledge into the image compression framework by leveraging frequency information, which is beneficial for noise and redundant information removal. Meanwhile, FDFM employs a task-driven approach to decompose image features, effectively assisting the network in understanding and preserving machine-friendly information during the compression process.

4.4 Human perception performance

In order to evaluate the proposed methodology can also apply to the human visual system, we prepared comprehensive evaluations to measure human perceptual performance. To refine the evaluation of our proposed approach within the context of the human visual system, a deliberate shift from pixel fidelity was made.

TABLE 5 Ablation study on object detection tasks on SUIM dataset under 0.3 bpp.

Method	UICM	FDFM	*RA*	mAP_{50}
(a)	×	✓	0.540	0.527
(b)	✓	×	0.549	0.532
(c)	×	×	0.468	0.495

TABLE 6 Ablation study on semantic segmentation tasks on SUIM dataset under 0.3 bpp.

Method	UICM	FDFM	*mIOU*	*mPA*	*PA*
(a)	×	✓	58.58	69.37	82.84
(b)	✓	×	58.64	69.40	82.97
(c)	×	×	57.71	68.77	81.69

TABLE 7 Ablation study on saliency detection tasks on SUIM dataset under 0.3 bpp.

Method	UICM	FDFM	*MAE*	$maxF_\beta$
(a)	×	✓	0.039	0.743
(b)	✓	×	0.038	0.755
(c)	×	×	0.042	0.740

This involved the utilization of metrics such as *PSNR* and *MS − SSIM* for natural images, along with the *UIQM* Panetta et al. (2015) metric tailored for underwater imagery. The ensuing outcomes have been meticulously compiled and are delineated in Tables 8–10. In the *PSNR* metric, our proposed method demonstrates comparable performance to the JPEG2000 approach. Within the *MS − SSIM* metric, the effectiveness of our proposed method aligns with that of the BPG method. Moreover, in *UIQM*, our proposed method outperforms alternative approaches.

TABLE 8 Comparison on human perception performance tasks on SUIM dataset in terms of PSNR metric.

Method	*PSNR*		
	0.1bpp	0.3bpp	0.5bpp
JEPG	21.435	28.746	30.157
JPEG2000	25.423	29.255	31.266
BPG	26.865	30.409	32.721
VVC	27.391	30.604	33.434
Hyperprior	26.173	29.340	30.501
Devil	24.248	27.870	29.166
Gao	25.984	30.101	31.899
Ours	25.493	27.741	31.745

*The higher the value, the better the reconstructed image quality.
We conducted experiments under different (bpp) settings.

TABLE 9 Comparison on human perception performance tasks on SUIM dataset in terms of MS-SSIM metric.

Method	*MS − SSIM*		
	0.1bpp	0.3bpp	0.5bpp
JEPG	0.704	0.929	0.952
JPEG2000	0.832	0.929	0.955
BPG	0.886	0.948	0.965
VVC	0.901	0.954	0.974
Hyperprior	0.884	0.952	0.968
Devil	0.798	0.927	0.951
Gao	0.870	0.944	0.961
Ours	0.897	0.951	0.983

*The higher the value, the better the reconstructed image quality.
We conducted experiments under different (bpp) settings.

TABLE 10 Comparison human perception performance tasks on SUIM dataset in terms of UIQM metric.

Method	*UIQM*		
	0.1bpp	0.3bpp	0.5bpp
JEPG	1.070	1.647	2.104
JEPG2000	1.952	2.223	2.299
BPG	2.058	2.197	2.239
VVC	2.114	2.230	2.273
Hyperprior	2.235	2.395	2.475
Devil	2.154	2.417	2.479
Gao	2.344	2.589	2.799
Ours	2.776	2.784	2.821

*The higher the value, the better the reconstructed image quality.
We conducted experiments under different (bpp) settings.

From the qualitative analysis examples presented in Figure 6, it is evident that, at low bit rates, the reconstructed images generated by our method exhibit enhanced clarity in fulfilling the task objectives. Specifically, in the first row, the human targets in our approach are markedly more distinct than alternative methods. In the second row, the small fish in the lower left corner of the reconstructed images from other methodologies appear more indistinct, whereas, in our proposed method, the small fish in the corresponding position is relatively well-defined. Progressing to the third row, our proposed method's reconstructed image of the sea urchin object displays more defined boundaries compared with alternative methods. In summary, despite its primary design for machine analysis tasks, our method preserves fundamental functionality for human recognition.

5 Conclusion

This paper proposes a new machine-oriented underwater image compression framework, introducing a frequency-guided underwater image correction module (UICM) and a task-driven feature decomposition fusion module (FDFM). The UICM progressively removes noise and redundant information. A Frequency-Spatial Interaction block (FSI) is used to learn complementary global and local attributes in the frequency domain. Additionally, the FDFM can effectively locate and keep useful features for downstream visual tasks through task-driven decomposition of image features. Extensive experiments on downstream visual tasks demonstrate that the proposed framework can effectively reduce the performance loss of the downstream visual tasks caused by compression at low bit rates.

In our future endeavors, we are committed to advancing the study of image compression techniques within more visual tasks. Moreover, we aim to investigate strategies for harnessing the potential advantages derived from large-scale models.

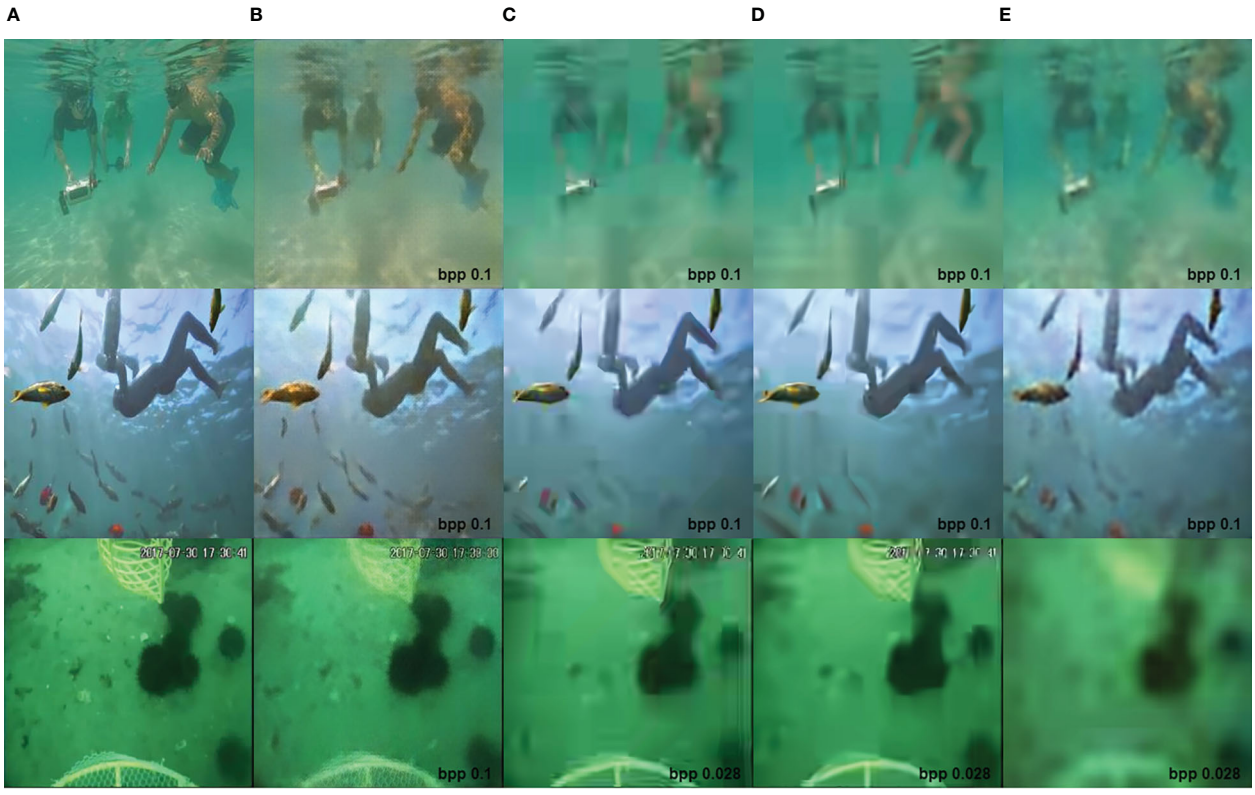

FIGURE 6
The visual comparison of **(A)** original image, **(B)** the proposed method, **(C)** BPG, **(D)** VVC, and **(E)** Hyperprior. **(A)** is designated as the original image, while the remaining columns depict reconstructed images using various methods at different bpp. The reconstructed images generated by the method proposed in this paper exhibit relatively high clarity.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Author contributions

HZ: Writing – original draft, Conceptualization, Methodology, Software. SF: Writing – original draft, Data curation, Validation. SZ: Writing – original draft, Data curation, Visualization. ZY: Writing – review & editing, Conceptualization, Funding acquisition, Resources, Supervision. BZ: Writing – review & editing, Funding acquisition, Resources, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by Hainan Province Science and Technology Special Fund, China (ZDYF2022SHFZ318); the National Natural Science Foundation of China under grant number 62171419; and Natural Science Foundation of Shandong Province of China under grant number ZR2021LZH005.

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

Achanta, R., Hemami, S., Estrada, F., and Süsstrunk, S. (2009). "Frequency-tuned salient region detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. (Miami, FL, USA: IEEE). doi: 10.1109/CVPR.2009.5206596

Agustsson, E., Tschannen, M., Mentzer, F., Timofte, R., and Gool, L. V. (2019). "Generative adversarial networks for extreme learned image compression," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*. (Seoul, Korea (South): IEEE), 221–231.

Al-Haj, A. (2007). Combined dwt-dct digital image watermarking. *J. Comput. Sci.* 3, 740–746. doi: 10.3844/jcssp.2007.740.746

Ancuti, C., Ancuti, C. O., Haber, T., and Bekaert, P. (2012). "Enhancing underwater images and videos by fusion," in *2012 IEEE conference on computer vision and pattern recognition*. (Providence, RI, United States: IEEE), 81–88.

Anjum, K., Li, Z., and Pompili, D. (2022). "Acoustic channel-aware autoencoder-based compression for underwater image transmission," in *2022 Sixth Underwater Communications and Networking Conference (UComms)*. (Lerici, Italy: IEEE), 1–5. doi: 10.1109/UComms56954.2022.9905691

Babić, A., Ferreira, F., Kapetanović, N., Mišković, N., Bibuli, M., Bruzzone, G., et al. (2023). "Cooperative marine litter detection and environmental monitoring using heterogeneous robotic agents," in *OCEANS 2023-Limerick*. (Limerick, Ireland: IEEE), 1–6.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495. doi: 10.1109/TPAMI.34

Ballé, J., Laparra, V., and Simoncelli, E. P. (2016). End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*. doi: 10.48550/arXiv.1611.01704

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. (2018). Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*. doi: 10.48550/arXiv.1802.01436

Brand, F., Seiler, J., and Kaup, A. (2019). "Intra frame prediction for video coding using a conditional autoencoder approach," in *2019 Picture Coding Symposium (PCS)*. (Ningbo, China: IEEE), 1–5.

Bross, B., Wang, Y.-K., Ye, Y., Liu, S., Chen, J., Sullivan, G. J., et al. (2021). Overview of the versatile video coding (vvc) standard and its applications. *IEEE Trans. Circuits Syst. Video Technol.* 31, 3736–3764. doi: 10.1109/TCSVT.2021.3101953

Brummer, B., and De Vleeschouwer, C. (2023). "On the importance of denoising when learning to compress images," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. (Waikoloa, HI, United States: IEEE), 2439–2447. doi: 10.1109/WACV56688.2023.00247

Burguera, A., and Bonin-Font, F. (2022). "Progressive hierarchical encoding for image transmission in underwater environments," in *OCEANS 2022, Hampton Roads*. (Hampton Roads, VA, United States: IEEE), 1–6. doi: 10.1109/OCEANS47191.2022.9976987

Chen, L., Chu, X., Zhang, X., and Sun, J. (2022). Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*. doi: 10.1007/978-3-031-20071-7_2

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018a). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 834–848. doi: 10.1109/TPAMI.2017.2699184

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018b). "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*. (Munich, Germany: Springer), 801–818.

Chen, T., Liu, H., Ma, Z., Shen, Q., Cao, X., and Wang, Y. (2021). End-to-end learnt image compression via non-local attention optimization and improved context modeling. *IEEE Trans. Image Process.* 30, 3179–3191. doi: 10.1109/TIP.83

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. (2020). "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the 2020 IEEE/CVF conference on computer vision and pattern recognition*. (Seattle, WA, United States: IEEE), 7939–7948.

Chicchon, M., Bedon, H., Del-Blanco, C. R., and Sipiran, I. (2023). Semantic segmentation of fish and underwater environments using deep convolutional neural networks and learned active contours. *IEEE Access* 11, 33652–33665. doi: 10.1109/ACCESS.2023.3262649

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. (Miami, FL, United States: IEEE), 248–255. doi: 10.1109/CVPR.2009.5206848

Di, W. D. W., Wen, G. W. G., Mingzeng, H. M. H., and Zhenzhou, J. Z. J. (2003). "A vlsi architecture design of cavlc decoder," in *ASIC, 2003. Proceedings. 5th International Conference on (IEEE)*. (Beijing, China: IEEE), Vol. 2. 962–965.

Ding, X., Zhang, Y., Ge, Y., Zhao, S., Song, L., Yue, X., et al. (2023). Unireplknet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition. *arXiv preprint arXiv:2311.15599*. doi: 10.48550/arXiv.2311.15599

Ding, X., Zhang, X., Han, J., and Ding, G. (2022). "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proceedings of the 2022 IEEE/CVF conference on computer vision and pattern recognition*. (New Orleans, United States: IEEE), 11963–11975.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., and Sun, J. (2021). "Repvgg: Making vgg-style convnets great again," in *Proceedings of the 2021 IEEE/CVF conference on computer vision and pattern recognition*. (Nashville, TN, United States: IEEE), 13733–13742.

Ellen, D. A. R., Kristalina, P., Hadi, M. Z. S., and Patriarso, A. (2023). "Effective searching of drowning victims in the river using deep learning method and underwater drone," in *2023 International Electronics Symposium (IES)*. (Denpasar, Indonesia: IEEE), 569–574. doi: 10.1109/IES59143.2023.10242589

Ercan, M. F., Muhammad, N. I., and Bin Sirhan, M. R. N. (2022). "Underwater target detection using deep learning," in *TENCON 2022 - 2022 IEEE Region 10 Conference (TENCON)*. (Hong Kong, Hong Kong: IEEE), 1–5. doi: 10.1109/TENCON55691.2022.9977994

Fang, Z., Shen, L., Li, M., Wang, Z., and Jin, Y. (2023). Prior-guided contrastive image compression for underwater machine vision. *IEEE Trans. Circuits Syst. Video Technol.* 33, 2950–2961. doi: 10.1109/TCSVT.2022.3229296

Fu, S., Xu, F., Liu, J., Pang, Y., and Yang, J. (2022). "Underwater small object detection in side-scan sonar images based on improved yolov5," in *2022 3rd International Conference on Geology, Mapping and Remote Sensing (ICGMRS)*. (Zhoushan, Chinapublisher: IEEE), 446–453. doi: 10.1109/ICGMRS55602.2022.9849382

Gao, C., Liu, D., Li, L., and Wu, F. (2023a). Towards task-generic image compression: A study of semantics-oriented metrics. *IEEE Trans. Multimedia* 25, 721–735. doi: 10.1109/TMM.2021.3130754

Gao, R., Yan, Y., and Liu, X. (2023b). "Target recognition method of sonar image based on deep learning," in *2023 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*. (Zhengzhou, China: IEEE), 1–6. doi: 10.1109/ICSPCC59353.2023.10400377

Gudimov, A. (2020). "The first it systems for ecological online monitoring in water environment," in *2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon)*. (Vladivostok, Russia: IEEE) 1–5.

Guo, Y., Li, F., and Du, Q. (2020). Research on key technologies of spatio-temporal analysis and prediction of marine ecological environment based on association rule mining analysis. *J. Coast. Res.* 115, 302–307. doi: 10.2112/JCR-SI115-095.1

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*. (Venice, Italy: IEEE), 2980–2988. doi: 10.1109/ICCV.2017.322

He, D., Yang, Z., Peng, W., Ma, R., Qin, H., and Wang, Y. (2022). "Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (New Orleans, LA, United States: IEEE) 5718–5727.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (Las Vegas, NV, United States: IEEE) 770–778. doi: 10.1109/CVPR.2016.90

He, D., Zheng, Y., Sun, B., Wang, Y., and Qin, H. (2021). "Checkerboard context model for efficient learned image compression," in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (Nashville, TN, United States: IEEE), 14771–14780.

Hou, Q., Zhou, D., and Feng, J. (2021). "Coordinate attention for efficient mobile network design," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (Nashville, TN, United States: IEEE), doi: 10.1109/CVPR46437.2021.01350

Hsu, W. W., Wang, S.-Y., Hong, W.-S., Hu, R.-H., Yu, C.-J., and Tasi, H.-Y. (2019). "Portable fisheries assistant systems for small scale fisheries management," in *2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*. (Yunlin, Taiwan: IEEE) 10–13.

Hu, Z., and Xu, C. (2022). "Detection of underwater plastic waste based on improved yolov5n," in *2022 4th International Conference on Frontiers Technology of Information and Computer (ICFTIC)*. (Qingdao, China: IEEE), 404–408. doi: 10.1109/ICFTIC57696.2022.10075134

Huo, J., Liu, S., Sun, L., Yang, L., Song, Y., and Li, C. (2021). "Research on biological disaster early warning and decision support system of nuclear power plant," in *2021 China Automation Congress (CAC)*. (Beijing, China: IEEE), 8120–8124.

Islam, M. J., Edge, C., Xiao, Y., Luo, P., Mehtaz, M., Morse, C., et al. (2020). "Semantic segmentation of underwater imagery: Dataset and benchmark," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. (Las Vegas, NV, United States: IEEE) 1769–1776. doi: 10.1109/IROS45743.2020.9340821

Jeyaraj, S., Ramakrishnan, B., and Ramsankaran, R. (2022). "Application of unmanned aerial vehicle (uav) in the assessment of beach volume change–a case study of malgund beach," in *OCEANS 2022-Chennai*. (Chennai, India: IEEE), 1–4.

Jiang, Q., Chen, Y., Wang, G., and Ji, T. (2020). A novel deep neural network for noise removal from underwater image. *Signal Processing: Image Communication* 87, 115921. doi: 10.1016/j.image.2020.115921

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. (Amsterdam, Netherlands: Springer) 694–711.

Kabir, I., Shaurya, S., Maigur, V., Thakurdesai, N., Latnekar, M., Raunak, M., et al. (2023). "Few-shot segmentation and semantic segmentation for underwater imagery," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. (Detroit, MI, United States: IEEE), 11451–11457. doi: 10.1109/IROS55552.2023.10342227

Khayam, S. A. (2003). The discrete cosine transform (dct): theory and application. *Michigan State Univ.* 114, 31.

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. doi: 10.48550/arXiv.1412.6980

Li, M., Zuo, W., Gu, S., You, J., and Zhang, D. (2020). Learning content-weighted deep image compression. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 3446–3461. doi: 10.1109/TPAMI.2020.2983926

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. (Zurich, Switzerland: Springer), 740–755.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021b). "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. (Montreal, QC, Canada: IEEE), 9992–10002. doi: 10.1109/ICCV48922.2021.00986

Liu, Y., Shu, Z., Li, Y., Lin, Z., Perazzi, F., and Kung, S.-Y. (2021a). "Content-aware gan compression," in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (Nashville, TN, United States: IEEE), 12156–12166. doi: 10.1109/CVPR46437.2021.01198

Liu, J., Yuan, F., Xue, C., Jia, Z., and Cheng, E. (2023). An efficient and robust underwater image compression scheme based on autoencoder. *IEEE J. Oceanic Eng.* 48, 925–945. doi: 10.1109/JOE.2023.3249243

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (Boston, MA, United States: IEEE), 3431–3440. doi: 10.1109/CVPR.2015.7298965

Loshchilov, I., and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv [Preprint] arXiv:1711.05101*. doi: 10.48550/arXiv.1711.05101

Madia, M., Bottaro, M., Vorsi, A. L., Amico, M. R., Gristina, M., Bizzarri, S., et al. (2023). "Reducing fishery impact on benthic community: new data by the use of guarding nets from the marine protected area of egadi islands," in *2023 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea)*. (La Valletta, Malta: IEEE), 94–98.

Minnen, D., Ballé, J., and Toderici, G. D. (2018). Joint autoregressive and hierarchical priors for learned image compression. *Adv. Neural Inf. Process. Syst.* 31. doi: 10.48550/arXiv.1809.02736

Motl, J., and Schulte, O. (2015). The ctu prague relational learning repository. *arXiv preprint arXiv:1511.03086*. doi: 10.48550/arXiv.1511.03086

Nadai, A. (2019). "Ocean wave measurement using sar cross-track interferometry," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. (Yokohama, Japan: IEEE), 7965–7967.

Nezla, N. A., Mithun Haridas, T., and Supriya, M. (2021). "Semantic segmentation of underwater images using unet architecture based deep convolutional encoder decoder model," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*. (Coimbatore, India: IEEE), Vol. 1. 28–33. doi: 10.1109/ICACCS51430.2021.9441804

Panetta, K., Gao, C., and Agaian, S. (2015). Human-visual-system-inspired underwater image quality measures. *IEEE J. Oceanic Eng.* 41, 541–551. doi: 10.1109/JOE.2015.2469915

Pei, Y., Huang, Y., Zou, Q., Zang, H., Zhang, X., and Wang, S. (2018). Effects of image degradations to cnn-based image classification. *arXiv preprint arXiv:1810.05552*. doi: 10.48550/arXiv.1810.05552

Pergeorelis, M., Bazik, M., Saponaro, P., Kim, J., and Kambhamettu, C. (2022). "Synthetic data for semantic segmentation in underwater imagery," in *OCEANS 2022, Hampton Roads*. (Hampton Roads, VA, United States: IEEE), 1–6. doi: 10.1109/OCEANS47191.2022.9976962

Piao, Y., Rong, Z., Xu, S., Zhang, M., and Lu, H. (2020). Dut-lfsaliency: Versatile dataset and light field-to-rgb saliency detection. *arXiv preprint arXiv:2012.15124*. doi: 10.48550/arXiv.2012.15124

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., and Jagersand, M. (2020). U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognit.* 106, 107404. doi: 10.1016/j.patcog.2020.107404

Rabbani, M., and Joshi, R. (2002). An overview of the jpeg 2000 still image compression standard. *Signal processing: Image communication* 17, 3–48. doi: 10.1016/S0923-5965(01)00024-8

Ranolo, E., Gorro, K., Ilano, A., Pineda, H., Sintos, C., and Gorro, A. J. (2023). "Underwater and coastal seaweeds detection for fluorescence seaweed photos and videos using yolov3 and yolov5," in *2023 2nd International Conference for Innovation in Technology (INOCON)*. (Bangalore, India: IEEE), 1–5. doi: 10.1109/INOCON57975.2023.10101342

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (Las Vegas, NV, United States: IEEE), 779–788. doi: 10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Rowley, J. (2018). "Autonomous unmanned surface vehicles (usv): A paradigm shift for harbor security and underwater bathymetric imaging," in *OCEANS 2018 MTS/IEEE Charleston*. (Charleston, SC, United States: IEEE), 1–6.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (Salt Lake City, UT, United States: IEEE), 4510–4520. doi: 10.1109/CVPR.2018.00474

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. doi: 10.48550/arXiv.1409.1556

Singh, K., Rypkema, N., and Leonard, J. (2023). "Attention-based self-supervised hierarchical semantic segmentation for underwater imagery," in *OCEANS 2023 - Limerick*. (Limerick, Ireland: IEEE), 1–6. doi: 10.1109/OCEANSLimerick52467.2023.10244736

Sullivan, G. J., Ohm, J.-R., Han, W.-J., and Wiegand, T. (2012). Overview of the high efficiency video coding (hevc) standard. *IEEE Trans. circuits Syst. video Technol.* 22, 1649–1668. doi: 10.1109/TCSVT.2012.2221191

Sze, V., and Budagavi, M. (2012). High throughput cabac entropy coding in hevc. *IEEE Trans. Circuits Syst. Video Technol.* 22, 1778–1791. doi: 10.1109/TCSVT.2012.2221526

Tang, L., Yuan, J., and Ma, J. (2022). Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Inf. Fusion* 82, 28–42. doi: 10.1016/j.inffus.2021.12.004

Thampi, L., Thomas, R., Kamal, S., Balakrishnan, A. A., Mithun Haridas, T. P., and Supriya, M. H. (2021). "Analysis of u-net based image segmentation model on underwater images of different species of fishes," in *2021 International Symposium on Ocean Technology (SYMPOL)*. (Kochi, India: IEEE), 1–5. doi: 10.1109/SYMPOL53555.2021.9689415

Tolstonogov, A. Y., and Shiryaev, A. D. (2021). "The image semantic compression method for underwater robotic applications," in *OCEANS 2021: San Diego – Porto*. (San Diego, CA, United States: IEEE), 1–9. doi: 10.23919/OCEANS44145.2021

Wallace, G. K. (1991). The jpeg still picture compression standard. *Commun. ACM* 34, 30–44. doi: 10.1145/103085.103089

Wang, N., Hou, X., Ma, L., Zhang, H., and Zhang, Z. (2023a). "Research on design of advanced marine scientific survey vessel based on computer data engineering and intelligent information system," in *2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA)*. (Tokyo, Japan: IEEE), 1604–1608. doi: 10.1109/ICPECA56706.2023.10075824

Wang, S., Mizuno, K., Tabeta, S., and Kei, T. (2023b). "Semantic segmentation of seafloor images in Philippines based on semi-supervised learning," in *2023 IEEE Underwater Technology (UT)*. (Tokyo, Japan: IEEE), 1–4. doi: 10.1109/UT49729.2023.10103432

Wen, H., Ma, L., Liu, L., Huang, Y., Chen, Z., Li, R., et al. (2022). High-quality restoration image encryption using dct frequency-domain compression coding and chaos. *Sci. Rep.* 12, 16523. doi: 10.1038/s41598-022-20145-3

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., et al. (2023). "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (Vancouver, BC, Canada: IEEE), 16133–16142. doi: 10.1109/CVPR52729.2023.01548

Wu, L., Huang, K., and Shen, H. (2020). "A gan-based tunable image compression system," in *Proceedings of the 2020 IEEE/CVF winter conference on applications of computer vision*. (Snowmass, CO, United States: IEEE) 2334–2342.

Xu, W., and Matzner, S. (2018). "Underwater fish detection using deep learning for water power applications," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*. (Las Vegas, NV, United States: IEEE), 313–318. doi: 10.1109/CSCI46756.2018.00067

Xu, K., Qin, M., Sun, F., Wang, Y., Chen, Y.-K., and Ren, F. (2020). "Learning in the frequency domain," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (Seattle, WA, United States: IEEE), doi: 10.1109/CVPR42600.2020

Xue, H. (2023). "Dynamic integration and analysis of marine environmental monitoring data based on support vector machine," in *2023 Asia-Europe Conference on Electronics, Data Processing and Informatics (ACEDPI)*. (Prague, Czechia: IEEE), 54–57.

Yang, P. (2024). An imaging algorithm for high-resolution imaging sonar system. *Multimedia Tools Appl.* 83, 31957–31973. doi: 10.1007/s11042-023-16757-0

Zhang, X., Yang, P., Wang, Y., Shen, W., Yang, J., Ye, K., et al. (2024). Lbf-based cs algorithm for multireceiver sas. *IEEE Geosci. Remote Sens. Lett.* 21, 1–5. doi: 10.1109/LGRS.2024.3379423

Zhang, A., and Zhu, X. (2023). "Research on ship target detection based on improved yolov5 algorithm," in *2023 5th International Conference on Communications, Information System and Computer Engineering (CISCE)*. (Guangzhou, China: IEEE), 459–463. doi: 10.1109/CISCE58541.2023.10142528

Zheng, B., Chen, Y., Tian, X., Zhou, F., and Liu, X. (2019). Implicit dual-domain convolutional network for robust color image compression artifact reduction. *IEEE Trans. Circuits Syst. Video Technol.* 30, 3982–3994. doi: 10.1109/TCSVT.76

Zhou, H., Men, Y., Yang, L., and Wang, J. (2023). "Design of control module for marine biogenic monitoring system in nuclear power plants," in *2023 IEEE 16th International Conference on Electronic Measurement & Instruments (ICEMI)*. (Harbin, China: IEEE), 304–308.

Zhu, X., Song, J., Gao, L., Zheng, F., and Shen, H. T. (2022). "Unified multivariate gaussian mixture for efficient neural image compression," in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (New Orleans, LA, United States: IEEE), 17612–17621.

Zou, R., Song, C., and Zhang, Z. (2022). "The devil is in the details: Window-based attention for image compression," in *Proceedings of the 2022 IEEE/CVF conference on computer vision and pattern recognition*. (New Orleans, LA, United States: IEEE), 17492–17501.