Perception-and-Cognition-Inspired Quality Assessment for Sonar Image Super-Resolution

Weiling Chen, *Member, IEEE*, Boqin Cai, Sumei Zheng, Tiesong Zhao, *Senior Member, IEEE*, and Ke Gu, *Senior Member, IEEE*

***Abstract*—Due to the light-independent imaging characteristics, sonar images play a crucial role in fields such as underwater detection and rescue. However, the resolution of sonar images is negatively correlated with the imaging distance. To overcome this limitation, Super-Resolution (SR) techniques have been introduced into sonar image processing. Nevertheless, it is not always guaranteed that SR maintains the utility of the image. Therefore, quantifying the utility of SR reconstructed Sonar Images (SRSIs) can facilitate their optimization and usage. Existing Image Quality Assessment (IQA) methods are inadequate for evaluating SRSIs as they fail to consider both the unique characteristics of sonar images and reconstruction artifacts while meeting task requirements. In this paper, we propose a Perception-and-Cognition-inspired quality Assessment method for Sonar image Super-resolution (PCASS). Our approach incorporates a hierarchical feature fusion-based framework inspired by the cognitive process in the human brain to comprehensively evaluate SRSIs' quality under object recognition tasks. Additionally, we select features at each level considering visual perception characteristics introduced by SR reconstruction artifacts such as texture abundance, contour details, and semantic information to measure image quality accurately. Importantly, our method does not require training data and is suitable for scenarios with limited available images. Experimental results validate its superior performance.**

***Index Terms*—Sonar image, super-resolution (SR), image quality assessment (IQA), task-oriented, hierarchical feature fusion.**

Manuscript received 25 July 2023; revised 30 November 2023 and 29 December 2023; accepted 29 December 2023. Date of publication 4 January 2024; date of current version 10 April 2024. This work was supported in part by the National Science Foundation of China under Grants 62322302, 62273011, 62076013, and 62021003, in part by the Beijing Natural Science Foundation under Grant JQ21014, and in part by the Natural Science Foundation of Fujian Province, China under Grants 2022J05117 and 2022J02015. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Balu Adsumilli. *(Corresponding author: Tiesong Zhao.)*

Weiling Chen and Tiesong Zhao are with the Fujian Key Lab for Intelligent Processing and Wireless Transmission of Media Information, Fuzhou University, Fuzhou 350116, China, and also with the Fujian Science & Technology Innovation Laboratory for Optoelectronic Information of China, Fuzhou University, Fuzhou 350116, China (e-mail: weiling.chen@fzu.edu.cn; t.zhao@fzu.edu.cn).

Boqin Cai and Sumei Zheng are with the Fujian Key Lab for Intelligent Processing and Wireless Transmission of Media Information, Fuzhou University, Fuzhou 350116, China (e-mail: 221120099@fzu.edu.cn; 201127086@fzu.edu.cn).

Ke Gu is with the Faculty of Information Technology, the Engineering Research Center of Intelligent Perception and Autonomous Control of Ministry of Education, the Beijing Laboratory of Smart Environmental Protection, the Beijing Key Laboratory of Computational Intelligence and Intelligent System, and the Beijing Artificial Intelligence Institute, Beijing University of Technology, Beijing 100124, China (e-mail: guke.doctor@gmail.com).

Digital Object Identifier 10.1109/TMM.2024.3349929

I. Introduction

SONAR Image (SI) serves as a vital presentation of information in various tasks. The resolution of SIs has a significant impact on the execution of tasks like object detection and recognition [1]. Using highly resolved sonar equipment is a direct way to acquire high-resolution SIs. However, according to the basic principle of acoustic propagation, signal frequency is positively proportional to its attenuation rate [2]. Sonar equipment working in low-frequency bands can provide long working distance but limited imaging resolution. Super-Resolution (SR) algorithms aim to increase the pixel density of a Low-Resolution (LR) image to a detail-rich High-Resolution (HR) image. The introduction of SR can substantially improve the resolution of SIs without upgrading hardware.

Although SR is crucial for SIs, there are currently few SR algorithms specifically designed for sonar images, and none of the authors have made their algorithm code or executable models publicly accessible. As a result, commonly used SR algorithms in SIs are mainly based on those developed for Natural Scene Images (NSIs) [7]. However, applying these algorithms to SIs leads to unsatisfactory quality due to differences in image characteristics and imaging mechanisms between NSIs and SIs. Specifically, earlier studies based on local interpolation often result in blur due to their limited use of local information. Deep learning-based approaches using dictionary and Generative Adversarial Network (GAN) offer more details but struggle to ensure their authenticity [8]. Fig. 1 shows some examples of Super-Resolution reconstructed Sonar Images (SRSIs) generated by different SR algorithms. It can be found that the SRSIs present problems like checkboard noise, blurred structure, and pseudo-details. Moreover, task background dictates that SIs expect more target-relevant details through SR to increase the accuracy and reliability of task execution. Consequently, it is urgent to find an objective task-oriented Image Quality Assessment (IQA) method capable of assessing the reconstruction artifacts and estimating the utility of SRSIs.

Existing objective IQA methods can be categorized into three types based on their applicable scenarios: general-purpose, optimization-oriented and scene-specific IQA. General-purpose IQA methods have no limit to image type and application scenarios. Among them, reference-free approaches are more universal. Typical achievements include BLIINDS-II [9], BRISQUE [10], and NIQE [11], which utilize statistics-based features in the spatial or spectral domain. Optimization-oriented approaches

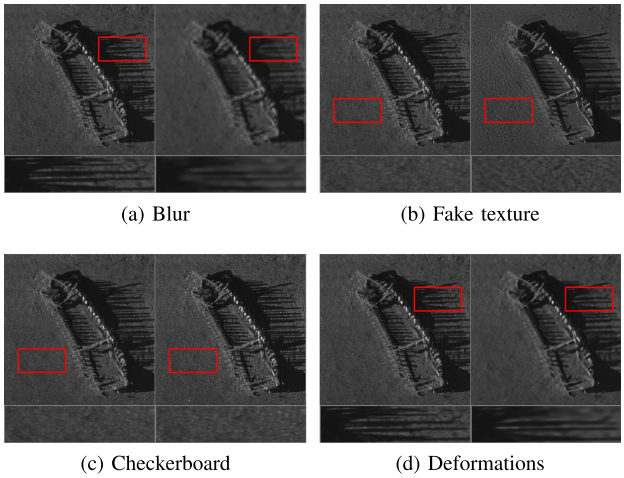


Fig. 1. Examples of SRSIs gennerated by different SR algorithms. In every image pair, the left is reference image and the right is SRSI. (a)–(d) are SRSIs via Bicubic [3], RealSR [4], VDSR [5] and RCAN [6] algorithms at scaling factor 4.

are tailored to specific optimization techniques, such as image SR, enhancement, and retargeting. They focus on the impact of the optimization process on image quality factors. In [12], a deep learning-based network has been proposed to effectively learn distinguishing feature representations from various LR and HR images and map the features to image quality prediction. Scene-specific IQA methods are designed for specific tasks such as marine detection and lesion segmentation. For instance, magnetic resonance IQA usually concentrates on measuring artifacts that relevant to false diagnosis, such as motion blur or ghosting artifacts [13].

As mentioned above, current objective IQA methods have limitations in evaluating the quality of SRSIs. On the one hand, general-purpose IQA methods do not effectively measure the mixed artifacts introduced by SR algorithms. On the other hand, evaluating the quality of SRSIs requires a combination of optimization-oriented and scene-specific IQA approaches since they are used for specific tasks after undergoing SR. However, there is currently no IQA method that successfully integrates image utility and optimization effects. To address these issues, we propose a Perception-and-Cognition-inspired quality Assessment method for Sonar image Superresolution (PCASS) based on our idea proposed in [14]. The PCASS considers both visual characteristics of sonar images and task-oriented framework. Furthermore, this work is training-free, i.e., it is independent of any training data. Compared to [14], this paper further dissects and justifies the proposed theory while reorganizing its core contribution. Additionally, new experimental results are included to validate the performance of our method. The main contributions of this paper are summarized as follows:

- Investigating the significance of task background and SR reconstruction artifacts for SRSI quality evaluation. A perception- and cognition-derived approach is proposed to combine task requirements and optimization processes.
- Designing a hierarchical feature fusion framework inspired by human visual object recognition. At each level, the representative features are extracted considering the visual perception characteristics introduced by SR.
- Making the proposed method requires no training. This is important for hard-to-obtain SIs. It is also particularly suitable for other data-constrained situations.

The rest of the paper is organized as follows. Section II summarizes previous works related to our research. Section III presents the framework design and feature representation of our proposed metric. Section IV shows and discusses experimental results, and final conclusions are drawn in Section V.

II. Related Works

A. SR IQA Methods

Image SR algorithms have flourished in the past several decades, but there has been little research on SR IQA. In fact, the most commonly used methods for SR IQA are still general-purpose methods, such as the Structural SIMilarity index (SSIM) [15], PSNR, and MSE, which have been shown to underperform in SR tasks [16]. Therefore, specific IQA methods for image SR are receiving increasing attention. According to the accessibility of the reference information, SR IQA methods can be divided into three categories: Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR). Wan et al. [17] considered orientation information when computing texture and gradient information due to its impact on the Human Visual System (HVS). Zhou et al. [18] independently calculated structural and textural similarity and incorporated high-frequency similarity to combine the final score. FR-IQA metrics require a practically inaccessible HR image as a reference.

Obviously, RR- and NR-IQA methods are more flexible with low demand for reference information. RR-IQA methods only employ practically accessible information as a reference, such as features of HR images [19], [20] or LR images [12]. NR-IQA methods can be derived from natural scene statistics models, where relevant statistical features in spatial and spectral domains are usually selected to quantify SR artifacts. Typically statistical features include DCT coefficients, steerable wavelet decomposition, and paired-product [21], [22]. NR methods can also be constructed with deep learning-based approaches. Bare et al. first applied Convolutional Neural Networks (CNNs) to the quality evaluation of single-image SR [23]. With a patch training strategy, the method achieved a significant performance improvement.

The aforementioned SR IQA methods mainly aim at human perception, which lack the consideration of the task background of SRSIs. There are also a few task-oriented SR IQA algorithms. For example, Zhou et al. [24] proposed an IQA method for face SR that considers several recognition-related factors, such as pose variation, lighting, and facial expressions. However, the tasks oriented by these algorithms differ from those of SI, leading to different emphases in corresponding IQA algorithms. Therefore, these IQA algorithms are also not applicable for evaluating the quality of SRSIs.

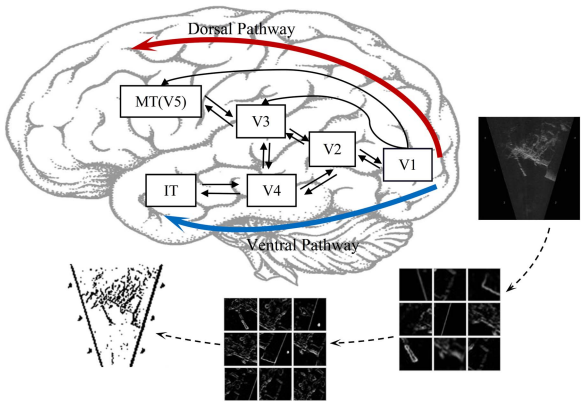


Fig. 2. Cognitive process of object perception.

B. Sonar IQA Methods

Although, to the best of our knowledge, no quality evaluation methods for SRSIs exist, there are some SI quality evaluation algorithms available. Chen et al. [25] proposed an FR method based on statistical information and structural information by comparing the local entropy and gradient spectrum of the most active block. However, obtaining reference images in underwater scenes is difficult. To address this problem, Chen et al. [26] proposed an RR method that utilizes contour information as reference data to obtain the similarity index and combines it with image comfort. In situations where channel conditions are poor, even small amounts of data cannot be reliably transmitted. Considering this issue, an NR method based on the degradation degree of image contour was proposed in [27]. Zhang et al. [28] devised a novel dual-path neural network to extract microscopic and macroscopic structures of SIs. Existing sonar IQA methods were all designed and trained based on the SIQD dataset [29]. The quality measurement criteria in this database are not strictly linked to utility. Moreover, these sonar IQA methods also lack the consideration for reconstruction artifacts introduced by SR.

C. Visual Object Recognition

The recognition mechanism of the human brain is a cognitive process in which external sensory information is received by different brain regions and an understanding of stimuli in the environment is generated [30]. Fig. 2 illustrates two main information processing streams during the cognitive process of object perception. The ventral pathway primarily processes information that is important for object recognition [31]. Information processing in the ventral pathway can be summarized in three steps. First, feature extraction is performed hierarchically, with the lower layer (V1) capturing elementary visual components and higher layers (V2 to IT) integrating these components to form even higher-level features. Second, the top-down feedback transmits high-level semantic information to early visual areas for a better understanding of objects. Finally, the brain makes the final decisions through directional competition resulting from this bidirectional flow of visual information [32]. In summary, the cognitive process involves the integration and analysis of features at different levels, as illustrated in the lower right of Fig. 2.

The method proposed in this paper draws inspiration from the cognitive process of feature analysis processing, whose framework is illustrated in Fig. 3. Its motivation is elaborated on in subsequent sections.

III. The Proposed Method

A. Framework Design

The quality evaluation model proposed in this paper is specifically designed for object recognition, which is crucial in human cognition. As mentioned in Section II-C, the visual cortex processes visual information for recognition tasks by integrating and analyzing visual features at different levels. Inspired by the mechanism, we construct a hierarchical feature fusion-based framework as shown in Fig. 3. It leverages the low-, mid-, and high-level features to thoroughly characterize SRSIs quality under object recognition. SR algorithms typically introduce diverse artifacts, including blur, noise amplification, aliasing artifacts, and fabricated details [33]. These artifacts are sophisticated and often appear in a combination way. As a result, it is important to consider the perceptual characteristics of reconstruction artifacts when selecting representative features at different levels. In the following part, experimental and theoretical analyses demonstrate that the selected features can reflect the perceptual characteristics of reconstruction artifacts and meet the requirements for object detection tasks.

B. Low-Level Features

The low-level representation of images refers to the local basic image information, such as color, edge, and texture [34], [35], [36]. One of the main challenges faced by SR algorithms is retrieving the aliased high-frequency information [37]. High-frequency information in an image encompasses fine details that exist within it. Failure to recover this high-frequency information can lead to loss of visual texture characteristics, amplification of noise, and a decrease in clarity [38]. SIs are primarily grayscale and heavily rely on texture information for recognition and analysis. Fig. 4 shows the SR results for NSI and SI with the same degree of blurriness. For NSI, even if the image is smooth, it does not fail our recognition. However, SI is difficult to recognize due to the absence of texture information and inherent color scarcity.

Sharpness measurement is capable of reflecting the abundance of image texture [39]. Thus we adopt image sharpness as the low-level feature. Currently, most sharpness evaluation algorithms are based on the gradient and high-frequency energy of the image. Successful methods like SSEQ [40] found that the Discrete Cosine Transform (DCT) entropy of NSI is skewed to the right. Similarly, we calculate the block DCT coefficient matrix D of SRSI on 8×8 blocks and then normalize the DCT coefficients to produce a spectral probability map:

$$P(i,j) = \frac{D(i,j)^2}{\sum_i \sum_j D(i,j)^2}, \tag{1}$$

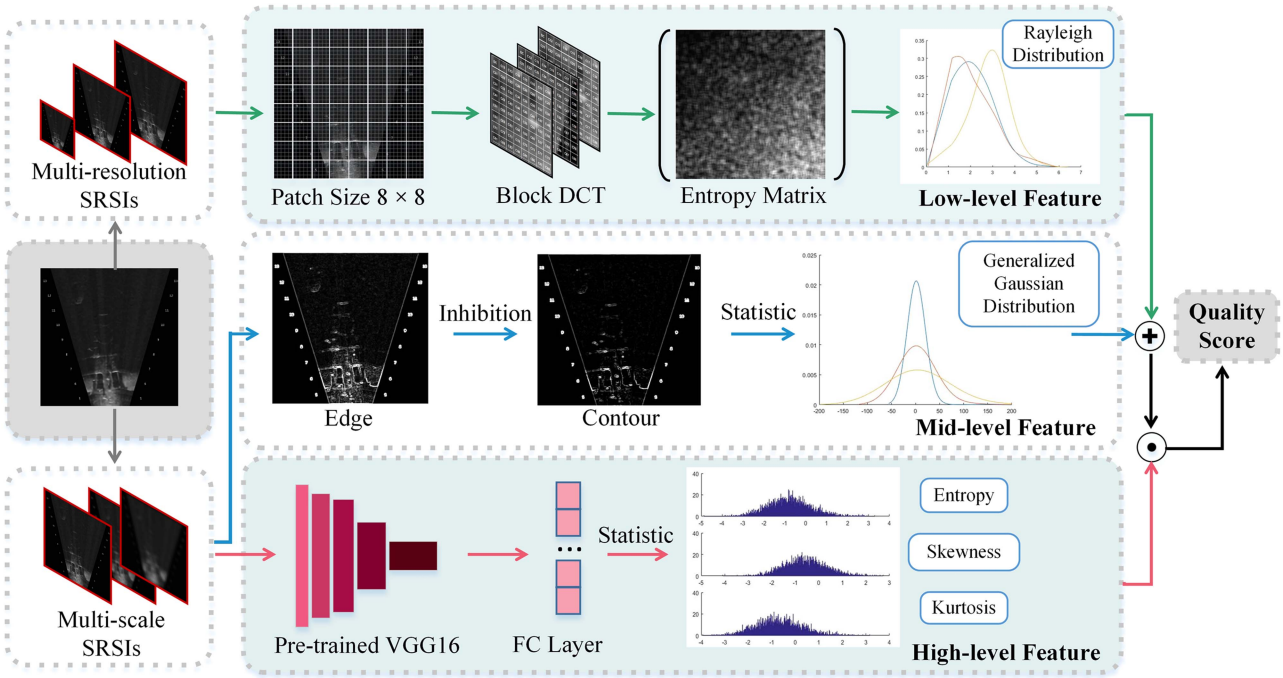


Fig. 3. Framework of PCASS method.

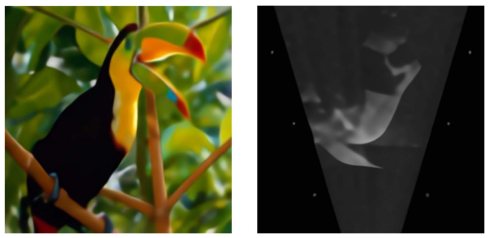


Fig. 4. Examples of NSI and SI reconstructed by the same SR algorithm at the same scaling factor.

where $1 \leq i \leq 8$, $1 \leq j \leq 8$, and $i, j \neq 1$. Then the entropy of each block is calculated as follows:

$$E = -\sum_{i}\sum_{j} P(i,j) log_2 P(i,j). \tag{2}$$

In this way, each image matches an entropy matrix. We drew the relevant histogram. It is found to present a Rayleigh distribution whose inclination increases with the decrease of utility. Under this discovery, we fit the DCT domain entropy with the probability density function of the Rayleigh distribution as shown in the second row of Fig. 5, which is formulated as follows:

$$f(x_1;\sigma) = \frac{x_1}{\sigma^2} e^{-\frac{x_1^2}{2\sigma^2}}, x_1 > 0, \tag{3}$$

where x_1 is DCT domain entropy, and σ^2 is the distribution parameter. We take σ^2 as the sharpness feature $f1$.

C. Mid-Level Features

Mid-level features serve as intermediate patterns that encode advanced concepts from basic visual elements [41]. Commonly employed mid-level representations include shape context, region-based descriptors, and bag-of-visual words [42], [43], [44], [45]. The key to successful image super-resolution lies in the up-sampling strategy. Due to the inherent loss of high-frequency components in LR images, it becomes impossible to fully reconstruct the expected HR image according to the Nyquist-Shannon sampling theorem. Improper representation of high-frequency components can result in artifacts such as aliasing, jagged edges, and ringing. These artifacts may cause duplicated target boundaries or create false shapes, leading to inaccuracies in target recognition.

Image contour offers a concise and informative representation of an object's overall geometric properties, making it a reliable indicator of utility [46], [47]. Consequently, we choose contour as the mid-level feature. Physiological studies have revealed that neuronal responses are substantially inhibited when there's a difference between the classical receptive field and its surroundings. Inspired by the biologically motivated surround inhibition model [48], [49], [50], we utilize distance to modulate the surround inhibition strength to obtain contour information. Sobel operators are used to extract gradient magnitude $M_\delta(x,y)$ first. Then inhibition weight is computed according to the Difference of Gaussian (DoG). Inhibition weight $W_\delta(x,y)$ is defined as follows:

$$W_\delta(x,y) = \frac{H(DoG_\delta^+(x,y))}{\left\| H(DoG_\delta^+(x,y)) \right\|_1}, \tag{4}$$

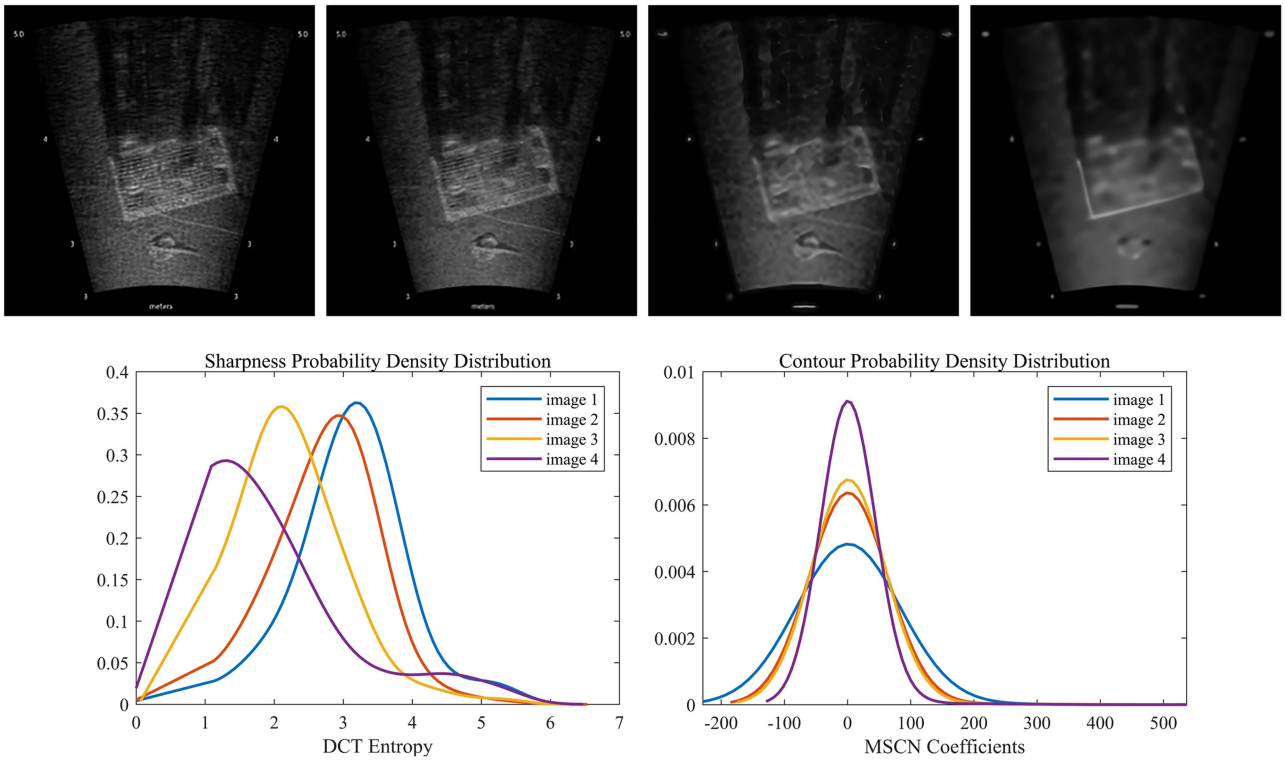


Fig. 5. Distribution curves of sharpness and contours correspond to SRSIs with different degrees of distortion. The utility of SRSIs in the first row gradually decreases from left to right (image1 to image4). It can be noticed that the hierarchical features of SRSIs vary with the degree of distortion.

where

$$DoG_{\delta}^{+}(x,y) = \frac{1}{2\pi(4\delta)^2} exp\left(-\frac{x^2+y^2}{2(4\delta)^2}\right) - \frac{1}{2\pi\delta^2} exp\left(-\frac{x^2+y^2}{2\delta^2}\right), \quad (5)$$

$$H(z) = \begin{cases} 0, & z < 0 \\ z, & z \geq 0 \end{cases}, \quad (6)$$

where δ is the standard deviation. $H(z)$ guarantees a negative response. Since the SI is not rich in edges, we exclusively consider distance-based modulation of surround inhibition strength. The isotropic suppression term $t_\delta(x,y)$ is defined as the convolution of gradient magnitude $M_\delta(x,y)$ and inhibition weight $W_\delta(x,y)$. The final contour operator $C_\delta(x,y)$ is formulated as follows:

$$t_\delta(x,y) = M_\delta(x,y) * W_\delta(x,y), \quad (7)$$

$$C_\delta(x,y) = H(M_\delta(x,y) - t_\delta(x,y)). \quad (8)$$

Similarly, we drew the probability density distribution of MSCN coefficients of suppressed contour images in the second row of Fig. 5. It presents a generalized Gaussian distribution, which is given by:

$$f(x_2;\alpha,\delta^2) = \frac{\alpha}{2\beta\Gamma(1/\alpha)} exp\left[-\left[\frac{|x|}{\beta}\right]^{\alpha}\right], \quad (9)$$

$$\beta = \delta\sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}, \quad (10)$$

where x_2 is the MSCN coefficient of the suppressed image. α is the shape parameter. β is scale, and $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$. ($\alpha$, δ^2) is taken as the second group of features, which is indicated as $f2$. Since the mean and variance are independent of each other, we combine them as follows:

$$f2 = \frac{\alpha * \delta^2}{\alpha^2 + (\delta^2)^2 + C}, \quad (11)$$

where C usually takes the value of 0.01.

D. High-Level Features

High-level features are object-based and related to the theme, content, or scene of an image [51]. Features extracted by deeper layers of CNNs are the most widely used high-level features [37]. As mentioned above, the unsuccessful or incomplete restoration of high-frequency components leads to the occurrence of distortions, either on a global or local scale. Global distortions can result in target occlusion, where important information or objects become obscured or indistinguishable. Local distortions can bring attribute-related information loss, where object-relevant characteristics may be altered or misrepresented. It poses a significant challenge to achieving accurate object recognition.

Semantic information facilitates finer-grained classification or provides guidance in the case where global information is incomplete [52]. High-level semantic information plays a pivotal role in shaping our subjective evaluation of image quality. Images with well-defined semantic information are perceived

as more coherent and pleasing. Furthermore, high-level features often serve as anchors for the viewer's visual attention and cognitive engagement [53]. This enhanced engagement positively influences IQA, as it leads to a more satisfying and informative viewing experience [54]. Therefore we employ semantic information as the high-level feature. CNN's convolutional units excel in object localization, making it easy to identify discriminative regions in an image [55]. At present, the VGG16 model [56] trained on ImageNet is proficient in achieving target recognition. Accordingly, it is utilized as an effective tool for extracting discriminative high-level features in this paper. In the deep CNN architecture, shallow layers tend to capture basic visual elements, while deep layers capture more contextual information and global patterns with a larger receptive field. Thus the Fully Connected (FC) layer after deep convolution layers is employed to incorporate task-related priori knowledge. The layer $fc7$ from the pre-trained VGG16 is finally selected. To reduce complexity and spatial dimension, three statistics including entropy S_{ent}, skewness S_{ske}, and kurtosis S_{kur} are calculated as the descriptors of high-level information, which are formulated as follows:

$$S_{\text{ent}} = -\sum_{1}^{n} P(i,j) log_2 P(i,j), \tag{12}$$

$$S_{\text{ske}} = \frac{1}{n}\sum_{1}^{n}\left[\left(\frac{(x_i-\mu)}{\delta}\right)^3\right], \tag{13}$$

$$S_{\text{kur}} = \frac{1}{n}\sum_{1}^{n}\left[\left(\frac{(x_i-\mu)}{\delta}\right)^4\right], \tag{14}$$

where n is the dimension of $fc7$. x_i is the parameter of $fc7$ in the i-th dimension. $P(i,j)$ is the probability of each value. μ and δ correspond to the mean and standard deviation.

The entropy can count useful information. The skewness and kurtosis, as high-order moments, play an important role in describing the statistical information of distribution. Since the three are of the same dimension, they are straightforwardly combined to obtain the third group of features $f3$:

$$f3 = S_{\text{ent}} + S_{\text{ske}} + S_{\text{kur}}. \tag{15}$$

E. Quality Evaluation

Most of the NR-IQA methods are opinion-aware, i.e., they require large numbers of subjective quality scores for model training. Since subjective quality scores are not easily available, it is more practical to design an opinion-unaware algorithm with better generalization ability. Features at different levels exert varying effects on image quality. Low- and mid-level features directly affect the image quality [57]. High-level features guide the aggregation of the lower-level features thereby influencing image quality [58]. Therefore, the quality expression of our metric adds up the low- and mid-level features $f1$, and $f2$, treating the high-level features $f3$ as an aggregated weighting factor. Additionally, considering the multi-scale property of HVS and the significance of image resolution in recognition, we perform a multi-scale fusion of images to obtain more complete image information. Thus, the predicted quality score Q_{PCASS} is finally defined as the following expression, through a simple linear fusion of features at each level:

$$Q_{\text{PCASS}} = \sum_{i=1}^{n} \frac{(f1_i + f2_i) * f3_i}{f1_i^2 + f2_i^2 + f3_i^2 + K}, \tag{16}$$

where n denotes the scale parameter. it is proved by the ablation experiment that $n = 3$ provides the best performance. Constant K is set to 0.01 to prevent the denominator from being 0.

IV. Experiments

A. Experimental Settings

Dataset: To evaluate the performance of various SR algorithms on SIs, we have proposed SRSID dataset, which comprises 1026 SRSIs and their corresponding Mean Opinion Scores (MOSs) [14]. In contrast to NSIs, the quality of SRSI is relative to a particular task. Therefore, we designed the SRSID dataset with a focus on its utility for object recognition tasks and visual perception characteristics related to SR. We defined a task- and perception-oriented evaluation criterion customized for SRSIs. We selected six representative and commonly used SR algorithms with multiple scale factors and different reconstruction artifacts to generate 1026 SRSIs. Among these algorithms, RealSR effectively reduces noise through a degradation framework based on kernel estimation and noise injection. TTSR [59] utilizes stacked modules to create a texture converter capable of capturing more accurate texture features. ESRGAN [60] introduces GANs to address blur issues in real scenes. Although some algorithms may generate pseudo-details, they do not significantly impact the overall recognition of the SR images. Furthermore, our dataset does not exhibit any significant semantic scene changes. It is worth noting that the selected SR algorithms are widely employed as benchmark algorithms in existing research on image SR (including sonar image SR). Importantly, all algorithms utilized in this study have been retrained using SR images.

Fifty-seven reference SIs are included in the SRSID dataset, which were selected from SIQD [29] and Marine Debris Dataset [61]. Despite the existence of numerous detection algorithms with remarkable performance, there is no widely acknowledged benchmark algorithm. Furthermore, human scoring is essentially more reliable in creating the benchmark. As a result, SRSID is established in a subjective form. Participants were told to give corresponding scores according to whether the target can be identified and its confidence degree. More details about the SRSID dataset can be seen in Table I.

Comparison Methods: In order to demonstrate the superiority of our method, we compared it with four classes of IQA metrics on the SRSID dataset. The first class contains six general-purpose NR-IQA metrics (BRISQUE, NIQE, LPSI [62], NFERM [63], SSEQ, BLIINDS-II). The second class is composed of two NR-IQA metrics designed for blurred images (JNB [64], CPBD [65]). The third class consists of two optimization-oriented IQA metrics (SFSN [8], SRmetric [21]). The fourth class is made up of two scene-specific IQA metrics (PSIQP [26], NRCDM [12]).

TABLE I
SOME DETAILS IN SRSID DATASET

Characteristics	Information
SRSIs amount	1026
Orignal resolution	320×320, 480×320
Image type	Side-scan sonar, acoustic lens sonar, forward-looking sonar
Scale factor	2, 3, 4, 6, 8
SR algorithms	Bicubic, RCAN, VDSR RealSR, TTSR, ESRGAN

Evaluation Criteria: To quantitatively measure the performance of the PCASS, we adopt four commonly used criteria, including two accuracy indices (Root Mean Square Error (RMSE) and Pearson Linear Correlation Coefficient (PLCC)), two monotonicity indices (Spearman Rank Order Correlation Coefficient (SROCC) and Kendall's Rank Order Correlation Coefficient (KROCC)). To measure PLCC and RMSE, we applied logistic regression [66] to establish a nonlinear mapping between the objective and subjective scores:

$$f(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (x - \beta_3)}} \right) + \beta_4 x + \beta_5, \tag{17}$$

where $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the fitting parameters.

B. Performance Evaluation

The performances of all IQA methods on SRSID are summarized in Table II, where the best and 2nd-best results are marked in bold and underlined. Generally, our method outperforms the other 12 methods by a large margin, surpassing the 2nd-best results by at least 4% in terms of PLCC, SROCC, and KROCC and obtaining a minimum RMSE of less than 9. The performances of general-purpose NR-IQA methods have a big span, where the worst and 2nd-best results are both occurring in this class.

Several conclusions can be drawn from the results. First, the properties of SRSIs are distinct from other types of images, making it difficult for both general-purpose and optimization-oriented, as well as scene-specific IQA methods to effectively assess. Notably, existing SR IQA and sonar IQA methods exhibit significant performance disparities when compared to ours. This may arise from their limited consideration of the task background of SRSIs and SR reconstruction artifacts. Second, a suitable representation of image structure is vital to the utility assessment of SRSIs. We noticed that gradient-based IQA methods such as LPSI and SSEQ yield somewhat satisfactory outcomes, whereas NSIs-based statistical methods exhibit inferior performance. Third, multi-level features contribute to a more comprehensive representation of SRSI quality. The other comparison methods ignore the important content of other levels, resulting in moderate performance. The proposed PCASS makes up for this defect and further improves the prediction ability.

We also conduct a statistical significance comparison to determine whether our method is significantly better than others. A bilateral F-test with a significance level of 0.01 is used to report the statistical significance of the performance gain. The last column of Table II lists the statistical significance results of comparison methods, where 1 indicates that our method is significantly better, and 0 indicates that the two methods are statistically comparable. It can be observed from the results that the PCASS is significantly better than others.

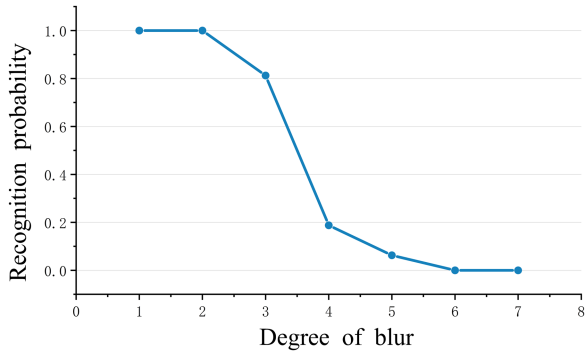


Fig. 6. Curve of recognition probability with the degree of blur.

C. The Effectiveness of Features

To verify the effectiveness of the selected features under task background, we added two subjective tests using 32 SRSIs. The first test aims to explore the correlation between sharpness-based low-level features and recognition accuracy. The test set consists of SRSIs with different levels of Gaussian blur. Observers were informed to vote according to whether they could identify the image content in a single stimulus way, with 1 being recognizable and 0 being not. Fig. 6 shows the negative correlation between the recognition probability and blur. It indicates a strong correlation between sharpness and recognition rate.

The second test is to verify the validity of the contour-based mid-level features. The test set contains the original SRSI and corresponding contour image through surround inhibition. The test was performed with a double stimulus way. Observers were required to classify the test image into $C_2^1 C_2^1 = 4$ cases shown in Table III according to whether the original SRSI and contour image could be identified. The hypothesis test was employed to ascertain the test image's attribute, i.e., identifiable or unidentifiable. Specifically, an initial hypothesis was made that the SRSI is identifiable. Following this, the Precision and Recall of subjective ratings were computed. Similarly, the opposite hypothesis was made and the Precision and Recall were also computed. Ultimately, the hypothesis yielding the highest F1-measure was selected as the ground truth of the SRSI attribute. The equation of Precision, Recall, and F1-measure are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{18}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{19}$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{20}$$

where TP, True Positives, are the instances that were correctly classified as positive. FP, False Positives, are the instances that were incorrectly classified as positive. FN, False Negatives, are

TABLE II
PERFORMANCE COMPARISON OF IQA METRICS

Class	IQA	PLCC↑	SROCC↑	KROCC↑	RMSE↓	F-test
IQA for SIs	PSIQP(RR) [26]	0.4744	0.4527	0.3236	11.7770	1
	NRCDM [27]	0.6979	0.6523	0.4582	9.5818	1
IQA for SR images	SFSN(FR) [8]	0.6419	0.5617	0.3838	10.2577	1
	SRmetric [21]	0.6979	0.6479	0.4677	9.5809	1
General-purpose NR-IQA	BRISQUE [10]	0.7052	0.6981	0.5118	9.4856	1
	NIQE [11]	0.7206	0.6881	0.5052	9.2751	1
	BLIINDS-II [9]	0.4321	0.3953	0.2692	12.0645	1
	SSEQ [40]	0.7231	0.7138	0.5265	9.2407	1
	NFERM [63]	0.6912	0.6901	0.5032	9.6676	1
	LPSI [62]	0.7258	0.7129	0.5278	9.2027	1
NR-IQA for blurred images	JNB [64]	0.6475	0.5642	0.3970	10.1944	1
	CPBD [65]	0.6778	0.6709	0.4871	9.8366	1
IQA for SRSIs	**PCASS**	**0.7804**	**0.7636**	**0.5703**	**8.3648**	–

The best result is marked in bold.

TABLE III
RECOGNITION RESULT CLASSIFICATION

Case	Original SRSI	Contour image
1	✓	✓
2	✓	✗
3	✗	✓
4	✗	✗

TABLE IV
ABLATION RESULTS OF FEATURE SELECTION

Features	PLCC↑	SROCC↑	KROCC↑	RMSE↓
$f1$	0.7733	0.7629	0.5676	8.4816
$f2$	0.7066	0.6435	0.4579	9.4662
$f3$	0.5184	0.4620	0.3169	11.4399
$f1$+$f2$	0.7756	0.7628	0.5686	8.4441
$f2$+$f3$	0.7788	**0.7643**	0.5702	8.3911
$f1$+$f3$	0.7018	0.6477	0.4627	9.5299
$f1$+$f2$+$f3$	**0.7804**	0.7636	**0.5703**	**8.3648**

The best result is marked in bold.

the instances that were incorrectly classified as negative when they were actually positive. The inconsistent judgment is considered a negative sample. We employ the following equation to calculate the total error probability:

$$P_{\text{error}} = \sum_{1}^{4} P(a_i)P(b_i|a_i), \tag{21}$$

where $P(a_i)$ is the probability of occurrence of each case. $P(b_i|a_i)$ is the error probability under the premise of event a_i. Our error probability is 0.0275, which is less than 0.05, indicating the effectiveness of the contour. With data saturation validation, the number of sonar experts invited in our subjective tests is seven and 25 respectively.

Additionally, in order to ascertain the effectiveness of the VGG model in extracting high-level features of SIs, we present the feature maps extracted by five convolution layers of the pre-trained VGG16 model in Fig. 7. The visualization illustrates that the deep layers effectively capture abstract information related to the object, which is reflected with a higher brightness in the feature map.

D. Ablation Study

Contributions of Components: The proposed PCASS comprises features from low- to high-level, namely sharpness, contour, and semantic information. We calculated the correlation between different feature groups and MOS at the same scale. Several findings are summarized from the results in Table IV. First, the overall performances of the combination of different features are better than the specific feature only. Multi-level features provide a more comprehensive description of visual content, enabling a more reliable quality prediction. Second, sharpness-based feature $f1$ yields the best results in single-level feature representation. Image sharpness captures crucial low-level information including structural texture, which is a solid foundation for the recognition task. Third, semantic features promote a performance improvement on the basis of low- and mid-level features, which further demonstrates the effectiveness of incorporating advanced priori knowledge. Meanwhile, it's observed that the introduction of $f2$ results in a marginal improvement in performance compared to $f1$. We deem that the level of abstraction in $f2$ may be a potential factor.

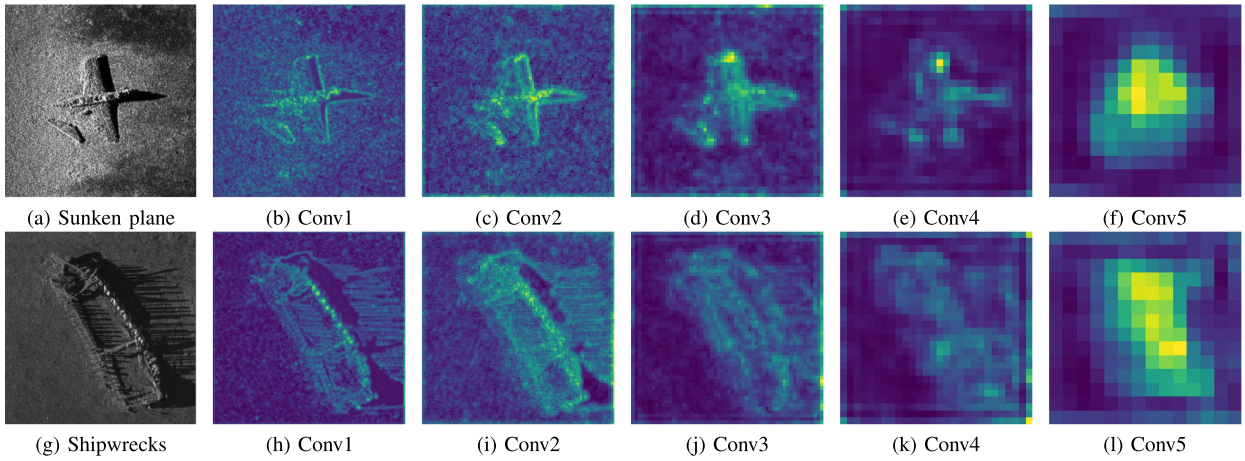


Fig. 7. Examples of feature map visualization of pretrained VGG model. The Conv1 to Conv5 correspond to the five convolution layers of the VGG16 model.

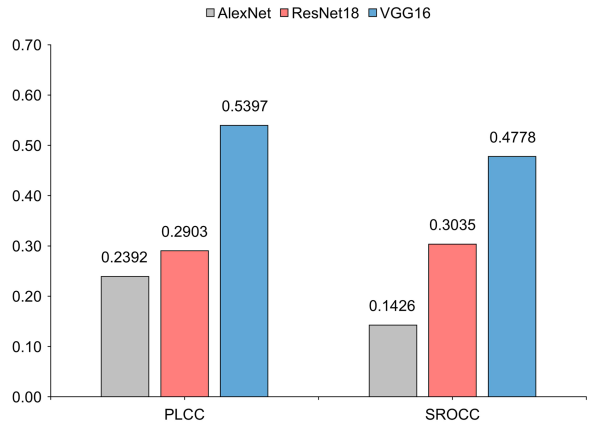


Fig. 8. Correlation between semantic features of each model and MOS.

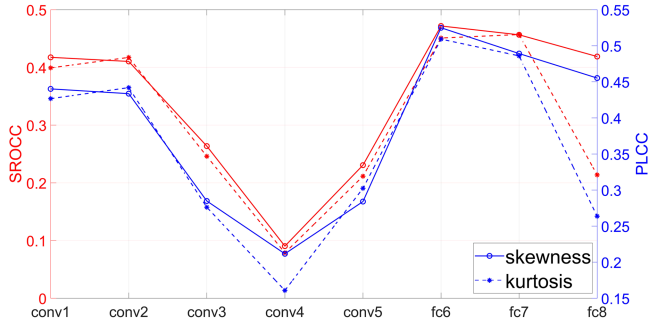


Fig. 9. Correlation between VGG16 layer features and MOS.

TABLE V
PERFORMANCE RESULTS IN DIFFERENT SCALES

Scale	PLCC↑	SROCC↑	KROCC↑	RMSE↓
1	0.7359	0.7312	0.5420	9.0584
2	0.7703	0.7527	0.5637	8.5315
3	**0.7804**	0.7636	**0.5703**	**8.3648**
4	0.7696	**0.7643**	0.5641	8.5420

The best result is marked in bold.

Discussion of Semantic Feature Construction: For high-level semantic feature construction, choosing the right model and layer is crucial. The basic principles of model and layer selection have been stated theoretically in Section III. To experimentally verify the effectiveness of our selection on prediction performance, two experiments were conducted. In the first stage, we select three classic CNN architectures for analysis, including AlexNet, ResNet18, and VGG16. Taking the first scale as an example, the semantic features are extracted according to the method mentioned in Section III. As shown in Fig. 8, VGG16 performs best on both PLCC and SROCC. In the second stage, we extract the output from $conv1$ to $fc8$ of VGG16 model. From Fig. 9, it can be seen that the features extracted from FC layers achieve better performance due to their ability to map the distributed semantic representation learned by convolution layers into sample label space. Among them, the $fc6$ and $fc7$ layers perform better, and $fc6$ performs best because it's close to the last convolutional layer. Through the combined performance index, it is proven that the $fc7$ layer can best reflect the utility-oriented quality of SRSIs.

Discussion of Parameter Setting: First, we discuss the impact of image scale on the performance, since our method is established under multi-scale. The scale denoted as 1 in Table V represents the original scale, and scale N represents N scale spaces including the original one. The Gaussian kernel function is applied to generate different scale spaces. The filter Gaussian kernel sizes corresponding to the other three scales are 3×3, 9×9, and 15×15, with standard deviations of 2, 4, and 6 respectively. The results in Table V show that there is a clear effect variation across different scale numbers. The best performance is achieved when using three scales, after which it starts to decline. It indicates that the multi-scale mechanism can discover image content from detail, region, and global, facilitating a comprehensive quality representation. However, excess scales can cause a high complexity and low generalization ability. Second, the impact of different K values on our quality expression is also explored. The range of K was set from 0.001 to 1000.

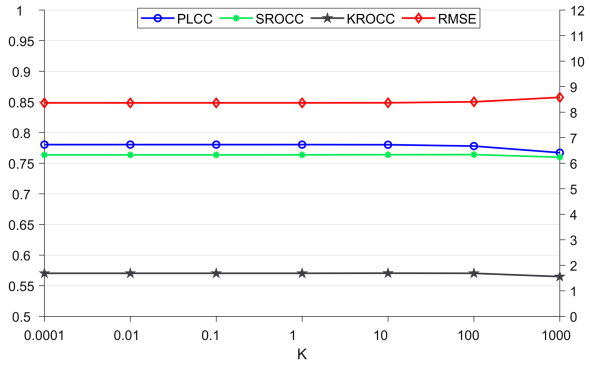


Fig. 10. Performance results with different K values.

TABLE VI
DETAILS OF SRSIS FOR OBJECT DETECTION

Characteristics	Information
Reference image amount	50
Image resolution	438×658, 500×605
Image type	Forward-looking sonar
Object category	10
SR algorithm	Bicubic, RealSR, TTSR, ESRGAN
Scale factor	2, 3, 4, 6, 8
Generated SRSI amount	300

The results of PLCC, SROCC, and KROCC correspond to the left axis of Fig. 10, and the results of RMSE correspond to the right axis. From Fig. 10, it can be seen that our method still has stable performance after scaling up the value of K by four orders of magnitude from the preset value of 0.01. It indicates that the proposed opinion-unaware method is insensitive to changes in K values.

E. Application

To explore the correlation between the proposed PCASS and the performance of detection algorithms, we made a comparison between this IQA method and the performance of the detection algorithm. This comparison was conducted on the Marine Debris Dataset (MDD) [67] using Underwater Forward-Looking Sonar Image Detection Network (UFIDNet) [68]. We randomly selected 50 sonar images from the MDD as a reference set, denoted as D_{Ref}. Then, the reference images were downsampled. Next, the SR algorithms were applied to generate multiple SRSIs.The details of the generated SRSIs and MDD dataset are shown in Table VI. Afterward, the PCASS was utilized to predict the quality of the generated SRSIs, and the distribution of prediction scores is shown in Fig. 11. The generated SRSIs were ranked based on their predicted quality scores, and divided into three subsets: D_{Q1}, D_{Q2}, and D_{Q3}. Each subset consists of 100 generated images. UFIDNet was then utilized to perform object detection on these subsets with varying quality. The detection performance for each SRSI subset is presented in Table VII. It is evident from the table that the mean Average Precision (mAP)

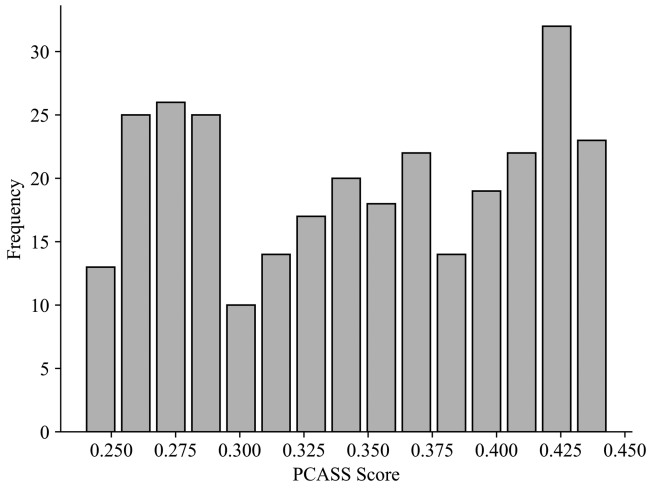
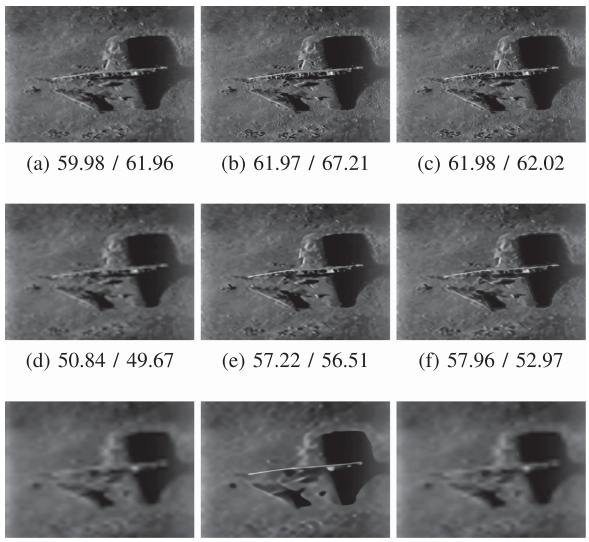


Fig. 11. Distribution of prediction scores for generated SRSIs.

TABLE VII
OBJECT DETECTION RESULTS ON SRSI SUBSETS WITH DIFFERENT QUALITY

	D_{Q1}	D_{Q2}	D_{Q3}	D_{Ref}
Score range	[0.239, 0.308]	[0.308, 0.385]	[0.385, 0.444]	-
mAP(%)	36.03	52.68	75.28	100



(a) 59.98 / 61.96 (b) 61.97 / 67.21 (c) 61.98 / 62.02

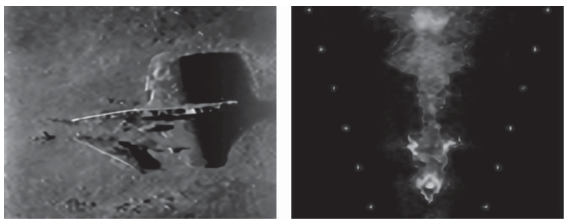
(d) 50.84 / 49.67 (e) 57.22 / 56.51 (f) 57.96 / 52.97

(g) 32.40 / 42.21 (h) 45.54 / 42.65 (i) 32.94 / 34.46

Fig. 12. Q_{PCASS} for several examples that illustrate the good performance of the proposed method. (a)–(i) are SRSIs via Bicubic, RCAN and VDSR algorithms at scale factors 2, 4, and 8 respectively. Quality scores below are uniformly noted in the form of the predicted score / MOS.

improves as the average quality of this subset increases. It indicates a higher likelihood of successful task completion for SRSIs with higher PCASS scores.

Furthermore, Fig. 12 showcases the correspondence between the SRSIs generated with different SR algorithms and the predicted scores. In general, there exists a strong correlation between the predicted scores and the MOS. When evaluating the reconstruction artifacts of the same SR algorithm, our predicted



(a) 50.99 / 46.16 / **57.22** / 56.51 (b) 48.62 / 50.27 / **46.22** / 45.10

Fig. 13. Examples that illustrate the performance gap between sonar IQA methods and the proposed method. Quality scores below are uniformly noted in the form of PSIQP score / NRCDM score / PCASS score / MOS. The best result is marked in bold.

scores exhibit a consistent trend to the MOS. It is revealed that the deep learning-based SR algorithms (RCAN, VDSR) perform better than traditional ones. In addition, the SR algorithms commonly do not perform well at a large scale factor. In order to highlight the performance gap between the sonar IQA methods and the proposed method, Fig. 13 illustrates the correspondence between MOSs and quality predicted by PCASS, along with two sonar IQA methods (PSIQP and NRCDM). The result indicates that PCASS demonstrates a higher correlation with MOSs compared to the sonar IQA methods.

V. Conclusion

The task requirements and complex distortions of SRSIs make most of the existing IQA methods insufficient to be directly applied. In this work, we exploit a hierarchical feature fusion-based IQA framework according to the cognitive process of HVS. Specifically, the hierarchical features include sharpness, contour, and semantic information, which are good reflectors of the visual perception characteristics associated with SR reconstruction artifacts. The efficacy of feature selection at each level is experimentally and theoretically validated. Extensive experiments on our established SRSID dataset show that our method performs well and is clearly distinguished from other methods. Moreover, investigating feature representations with a higher degree of abstraction could be a promising avenue for future research, given insights from ablation experiments. Additionally, transmission and compression significantly contribute to SI distortions, prompting us to explore their combined impact in future research.

References

[1] V. Myers and J. Fawcett, "A template matching procedure for automatic target recognition in synthetic aperture sonar imagery," *IEEE Signal Process. Lett.*, vol. 17, no. 7, pp. 683–686, Jul. 2010.

[2] H. R. Gordon, "Can the lambert-beer law be applied to the diffuse attenuation coefficient of ocean water?," *Limnol. Oceanogr.*, vol. 34, no. 8, pp. 1389–1409, 1989.

[3] H. Hou and H. Andrews, "Cubic splines for image interpolation and digital filtering," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 6, pp. 508–517, Dec. 1978.

[4] X. Ji et al., "Real-world super-resolution via kernel estimation and noise injection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 466–467.

[5] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.

[6] Y. Zhang et al., "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.

[7] T. T. C, A. M. Nambiar and A. Mittal, "A GAN-based super resolution model for efficient image enhancement in underwater sonar images," in *Proc. IEEE OCEANS*, 2022, pp. 1–8.

[8] W. Zhou, Z. Wang, and Z. Chen, "Image super-resolution quality assessment: Structural fidelity versus statistical naturalness," in *Proc. IEEE 13th Int. Conf. Qual. Multimedia Exp.*, 2021, pp. 61–64.

[9] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[10] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[11] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[12] T. Zhao, Y. Lin, Y. Xu, W. Chen, and Z. Wang, "Learning-based quality assessment for image super-resolution," *IEEE Trans. Multimedia*, vol. 24, pp. 3570–3581, 2022.

[13] Q. Chen et al., "MRIQA: Subjective method and objective model for magnetic resonance image quality assessment," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process.*, 2022, pp. 1–5.

[14] S. Zheng, W. Chen, T. Zhao, H. Wei, and L. Lin, "Utility-oriented quality assessment of sonar image super-resolution," in *Proc. OCEANS*, 2022, pp. 1–5.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[16] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM challenge on perceptual image super-resolution," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 334–355.

[17] W. Wan, J. Wu, G. Shi, Y. Li, and W. Dong, "Super-resolution quality assessment: Subjective evaluation database and quality index based on perceptual structure measurement," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2018, pp. 1–6.

[18] F. Zhou, R. Yao, B. Liu, and G. Qiu, "Visual quality assessment for super-resolved images: Database and method," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3528–3541, Jul. 2019.

[19] Y. Fang, J. Liu, Y. Zhang, W. Lin, and Z. Guo, "Quality assessment for image super-resolution based on energy change and texture variation," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 2057–2061.

[20] L. Yang, Y. Sheng, and L. Chai, "A machine learning based reduced-reference image quality assessment method for single-image super-resolution," in *Proc. IEEE Chin. Control Conf.*, 2019, pp. 3571–3576.

[21] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Comput. Vis. Image Understanding*, vol. 158, pp. 1–16, 2017.

[22] J. Beron, H. D. Benitez-Restrepo, and A. C. Bovik, "Blind image quality assessment for super resolution via optimal feature selection," *IEEE Access*, vol. 8, pp. 143201–143218, 2020.

[23] B. Bare, K. Li, B. Yan, B. Feng, and C. Yao, "A deep learning based no-reference image quality assessment model for single-image super-resolution," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 1223–1227.

[24] X. Zhou and B. Bhanu, "Evaluating the quality of super-resolved images for face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2008, pp. 1–8.

[25] W. Chen, K. Gu, W. Lin, F. Yuan, and E. Cheng, "Statistical and structural information backed full-reference quality measure of compressed sonar images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 334–348, Feb. 2020.

[26] W. Chen, K. Gu, T. Zhao, G. Jiang, and P. Le Callet, "Semi-reference sonar image quality assessment based on task and visual perception," *IEEE Trans. Multimedia*, vol. 23, pp. 1008–1020, 2021.

[27] W. Chen et al., "Reference-free quality assessment of sonar images via contour degradation measurement," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5336–5351, Nov. 2019.

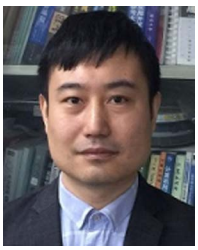
[28] H. Zhang et al., "Sonar image quality evaluation using deep neural network," *IET Image Process.*, vol. 16, no. 4, pp. 992–999, 2022.

[29] W. Chen, F. Yuan, E. Cheng, and W. Lin, "Subjective and objective quality evaluation of sonar images for underwater acoustic transmission," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 176–180.
[30] X. Yang et al., "Brain-inspired models for visual object recognition: An overview," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5263–5311, 2022.
[31] L. G. Ungerleider and J. V. Haxby, "What and where in the human brain," *Curr. Opin. Neurobiol.*, vol. 4, no. 2, pp. 157–165, 1994.
[32] N. Dijkstra, P. Zeidman, S. Ondobaka, M. A. van Gerven, and K. Friston, "Distinct top-down and bottom-up brain connectivity during visual perception and imagery," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 5677.
[33] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Oct. 2021.
[34] S. Ling, J. Li, P. Le Callet, and J. Wang, "Perceptual representations of structural information in images: Application to quality assessment of synthesized view in FTV scenario," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1735–1739.
[35] Y. M. Wang and H. Zhang, "Detecting image orientation based on low-level visual content," *Comput. Vis. Image Understanding*, vol. 93, no. 3, pp. 328–346, 2004.
[36] E. G. Danaci and N. Ikizler-Cinbis, "Low-level features for visual attribute recognition: An evaluation," *Pattern Recognit. Lett.*, vol. 84, pp. 185–191, 2016.
[37] Y. Li, L. Guo, and L. Jin, "A content-aware image retargeting quality assessment method using foreground and global measurement," *IEEE Access*, vol. 7, pp. 91912–91923, 2019.
[38] B. Yan, B. Bare, C. Ma, K. Li, and W. Tan, "Deep objective quality assessment driven single image super-resolution," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2957–2971, Nov. 2019.
[39] C. T. Vu and D. M. Chandler, "Main subject detection via adaptive feature selection," in *Proc. 16th IEEE Int. Conf. Image Process.*, 2009, pp. 3101–3104.
[40] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Process. Image Commun.*, vol. 29, no. 8, pp. 856–863, 2014.
[41] X. Liu, H. Yan, H. Huo, and T. Fang, "Mining mid-level visual elements for object detection in high-resolution remote sensing images," in *Proc. 10th IAPR Workshop Pattern Recognit. Remote Sens.*, 2018, pp. 1–6.
[42] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.
[43] S. Ling and P. Le Callet, "Image quality assessment for free viewpoint video based on mid-level contours feature," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2017, pp. 79–84.
[44] C. Yang, L. Fang, and H. Wei, "Learning contour-based mid-level representation for shape classification," *IEEE Access*, vol. 8, pp. 157587–157601, 2020.
[45] S. Xu, T. Fang, D. Li, and S. Wang, "Object classification of aerial images with bag-of-visual words," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 2, pp. 366–370, Apr. 2010.
[46] E. T. Scott and S. S. Hemami, "Image utility estimation using difference-of-Gaussian scale space," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 101–105.
[47] D. M. Rouse and S. S. Hemami, "Natural image utility assessment using image contours," in *Proc. IEEE 16th Int. Conf. Image Process.*, 2009, pp. 2217–2220.
[48] K.-F. Yang, C.-Y. Li, and Y.-J. Li, "Multifeature-based surround inhibition improves contour detection in natural images," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5020–5032, Dec. 2014.
[49] Z. Qu, P. Wang, Y. Gao, P. Wang, and Z. Shen, "Contour detection based on SUSAN principle and surround suppression," in *Proc. IEEE Int. Conf. Image Process.*, 2010, pp. 1937–1940.
[50] C. Grigorescu, N. Petkov, and M. A. Westenberg, "Contour and boundary detection improved by surround suppression of texture edges," *Image Vis. Comput.*, vol. 22, no. 8, pp. 609–622, 2004.
[51] Y. Cui, M. Yu, G. Jiang, Z. Peng, and F. Chen, "Blind tone-mapped HDR image quality measurement by analysis of low-level and high-level perceptual characteristics," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 5022015.
[52] H. Hong, D. L. Yamins, N. J. Majaj, and J. J. DiCarlo, "Explicit information for category-orthogonal object properties increases along the ventral stream," *Nature Neurosci.*, vol. 19, no. 4, pp. 613–622, 2016.
[53] C. Summerfield and T. Egner, "Expectation (and attention) in visual cognition," *Trends Cogn. Sci.*, vol. 13, no. 9, pp. 403–409, 2009.
[54] Y. Liu, K. Gu, S. Wang, D. Zhao, and W. Gao, "Blind quality assessment of camera images based on low-level and high-level statistical features," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 135–146, Jan. 2019.
[55] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.
[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Y. Bengio and Y. LeCun, Eds., in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, May 2015. [Online]. Available: http://arxiv.org/abs/1409.1556
[57] Y. Zhou et al., "No-reference quality assessment for view synthesis using DoG-based edge statistics and texture naturalness," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4566–4579, Sep. 2019.
[58] D. Li, T. Jiang, W. Lin, and M. Jiang, "Which has better visual quality: The clear blue sky or a blurry animal?," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1221–1234, May 2019.
[59] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5791–5800.
[60] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis.*, 2019, pp. 63–79.
[61] D. Singh and M. Valdenegro-Toro, "The marine debris dataset for forward-looking sonar semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3741–3749.
[62] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 339–343.
[63] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.
[64] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 717–728, Apr. 2009.
[65] N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2678–2683, Sep. 2011.
[66] H. Chen et al., "Perceptual quality assessment of cartoon images," *IEEE Trans. Multimedia*, vol. 25, pp. 140–153, 2023.
[67] M. Valdenegro-Toro, "Learning objectness from sonar images for class-independent object detection," in *Proc. Eur. Conf. Mobile Robots*, 2019, pp. 1–6.
[68] H. Long, L. Shen, Z. Wang, and J. Chen, "Underwater forward-looking sonar images target detection via speckle reduction and scene prior," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023.

Weiling Chen (Member, IEEE) received the B.S. and Ph.D. degrees in communication engineering from Xiamen University, Xiamen, China, in 2013 and 2018, respectively. She is currently an Associate Professor with the College of Physics and Information Engineering, Fuzhou University, Fuzhou, China. From September 2016 to December 2016, she was with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. Her research interests include image quality perception, computer vision, and underwater acoustic transmission.

Boqin Cai received the B.S. degree in electronic and information engineering in 2022 from Fuzhou University, Fuzhou, China, where she is currently working toward the M.S. degree with the College of Physics and Information Engineering. Her research interests include image processing and computer vision.

Sumei Zheng received the B.S. degree in communication engineering from the East China Institute of Technology, Jiangxi, China, in 2020 and the M.S degree in electronic and information engineering from Fuzhou University, Fuzhou, China, in 2023. Her research interests include image quality perception and computer vision.



Tiesong Zhao (Senior Member, IEEE) received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2006 and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2011. He was a Research Associate with the Department of Computer Science, City University of Hong Kong from 2011 to 2012, a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada from 2012 to 2013, and a Research Scientist with the Ubiquitous Multimedia Laboratory, The State University of New York at Buffalo, Amherst, NY, USA, from 2014 to 2015. He is currently a Minjiang Distinguished Professor with the College of Physics and Information Engineering, Fuzhou University, Fuzhou, China. His research interests include multimedia signal processing, coding, quality assessment, and transmission. Due to his contributions in video coding and transmission, he was the recipient of the Fujian Science and Technology Award for Young Scholars in 2017. Since 2019, he has also been an Associate Editor fro IET Electronics Letters.

Ke Gu (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2009 and 2015, respectively. He is currently a Professor with the Beijing University of Technology, Beijing, China. His research interests include industrial vision, environmental perception, image processing, and machine learning.