Codebook-Empowered Analysis-Friendly Extreme Underwater Image Compression

Jianhao Wu[1], **Yudong Mao**[2], **Qiuping Jiang**[1*]

[1]the School of Information Science and Engineering, Ningbo University
[2]the Department of Computer Science, City University of Hong Kong
3305604379@qq.com, jiangqiuping@nbu.edu.cn, yudongmao2-c@my.cityu.edu.hk

Abstract

While existing underwater image compression (UIC) methods optimize for human perception or basic redundancies, they neglect inter-image correlations and fail to prioritize machine-friendly features essential for automated analysis. This paper introduces a novel vector-quantized (VQ) codebook-driven framework for machine-centric UIC. We leverage VQ codebooks – pre-trained as external priors on diverse underwater data – to unify three critical stages: (1) Machine-friendly feature extraction via contrastive learning with high/low-quality codebooks, enhancing degradation robustness; (2) Compact compression using variable-size codebooks to map discriminative features to entropy-coded indices, enabling ultra-low bitrates (less than 0.04bpp); and (3) Feature refinement at the decoder, restoring semantic fidelity for downstream tasks. In addition, we contribute the first Underwater Visual Question Answering (UVQA) benchmark to holistically evaluate machine perception across object presence, counting, and localization. Extensive experiments demonstrate that our framework significantly outperforms state-of-the-art codecs in machine vision task performance at ultra-low bitrates. The VQ-codebook effectively harnesses inter-image redundancy, combats joint degradation, and delivers compact, analysis-friendly representations, establishing a new paradigm for machine-centric UIC.

Code and Datasets —
https://github.com/wjh-666666/VQ-UIC

1 Introduction

Underwater machine vision is rapidly emerging as a critical enabler for intelligent marine systems, yet accurate machine analysis remains challenging due to inherent degradations and the limited bandwidth of acoustic communication. Light absorption and scattering cause distortions such as color shifts and haze (2021; 2022), while aggressive compression – required for low-bandwidth transmission (2020) – introduces additional artifacts that further hinder perception. Given the central role of automated analysis in underwater applications, it is essential to develop underwater image compression (UIC) techniques tailored specifically for machine vision.

*Corresponding author

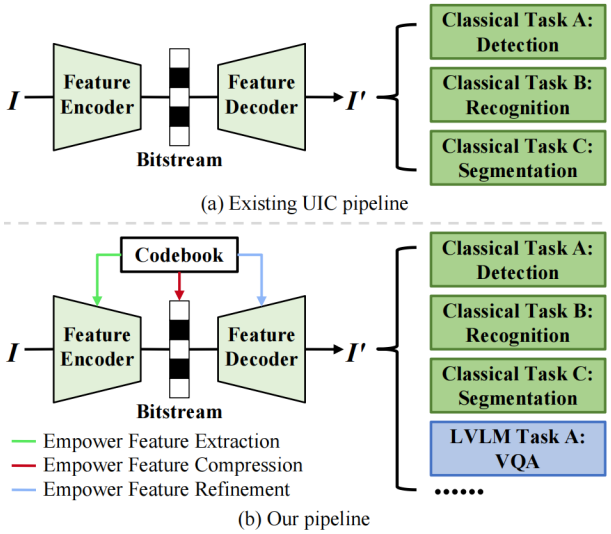

Figure 1: The key idea of our proposed method. We observe that a learned VQ codebook, with its strong representation capacity and high compression efficiency, can effectively support discriminative feature extraction, compact compression, and machine-friendly feature refinement. By leveraging this compact yet expressive representation, the reconstructed underwater images are expected to support robust performance across a variety of downstream machine vision tasks, such as semantic segmentation, object detection, salient object detection, and visual question answering (VQA) based on large vision-language models (LVLMs).

To address both bandwidth constraints and compounded degradations, machine vision-oriented UIC methods require dedicated strategies across different stages including machine-friendly feature extraction on the encoder side, compact feature compression, and machine-friendly feature refinement on the decoder side, forming a unified end-to-end machine-centric UIC pipeline. In the literature, some initial efforts have been made. To extract machine-friendly features from raw underwater images with unknown degradations, Fang et al. (2022) constructed external positive and negative datasets to support learning machine-friendly feature representations through contrastive learning. With a similar goal

of leveraging external priors, Fang et al. (2023) proposed a Feature Degradation Removal (FDR) module that mitigates degradation in a raw underwater image by incorporating features from its corresponding enhanced image – generated by an existing machine-friendly underwater image enhancement algorithm – as auxiliary information.

To achieve compact compression, existing UIC methods can be broadly categorized into three types: wavelet-based algorithms (1997; 2010; 2015; 2016; 2020), principal component-based algorithms (2014; 2016; 2021) and modern CNN-based algorithms (2019; 2020). While these methods have made progress in improving compression ratios, they predominantly optimize for pixel-level fidelity aligned with human visual perception, often overlooking the distinct physical and statistical properties of underwater environments. Critically, they focus almost exclusively on removing intra-image redundancy while neglecting the significant inter-image redundancy. This oversight is particularly impactful because underwater scenes exhibit strong inter-image correlations due to the repetitive appearance of common marine entities – such as fish, corals, and rocks – with varying morphologies and scales but share underlying visual and semantic features. Recognizing this limitation, recent advanced UIC methods explicitly leverage underwater-specific characteristics to achieve superior compression efficiency. For instance, Li et al. (2023) proposed RFD-ECNet, which uses a pre-built Underwater Feature Dictionary (UFD). It matches input image features to the UFD, finds similar reference features, and only compresses the differences (residuals) between them. This exploits cross-image similarity, slashing the coding bits. Similarly, Li et al. (2024) proposed the EUICN framework, which integrates universal correlation features derived from a comprehensive underwater dictionary with hyperprior and local features, enabling more accurate probability estimation during entropy coding, thereby minimizing bitrate overhead and achieving higher coding efficiency. Both methods exemplify a paradigm shift toward redundancy removal across images, directly addressing the unique compositional nature of underwater visual data to outperform conventional pixel-focused codecs. However, how to effectively incorporate the unique characteristics into machine vision-oriented UIC has rarely been explored.

In summary, the previous works demonstrate that both machine-friendly feature extraction and compact compression stages in the UIC pipeline can benefit from well-designed external priors. In this work, we exploit the powerful and robust representation capacity of vector-quantized (VQ) codebooks, pre-trained on diverse underwater datasets by the VQGAN model (2021), as effective external priors, to tackle the specific challenges in different stages, i.e., machine-friendly feature extraction, compact feature compression, and machine-friendly feature refinement for reconstruction. Compared with the priors used in previous works, the inherent properties of bandwidth efficiency, noise resilience, and semantic preservation make VQ codebooks uniquely suited for underwater applications. Specifically, we propose a novel VQ codebook-empowered UIC framework tailored for robust machine vision under low bitrate constraints and the compounded degradations (see Fig. 1). On the encoder side, we construct machine-centric positive (high-quality) and negative (low-quality) VQ codebooks, encouraging the extraction of analysis-friendly representations via contrastive learning. For compression, we introduce variable-size VQ codebooks trained to preserve generic, compact, and machine-relevant features, enabling the mapping of discriminative representations to VQ indices and entropy-encoded to enable flexible low-bitrate control. On the decoder side, the high-quality codebook facilitates feature refinement, mitigating degradations and restoring semantic fidelity critical for various downstream machine analysis tasks. Our contributions can be summarized as follows:

- We propose a novel machine-oriented UIC framework that leverages the VQ codebook to sequentially enable three stages within a unified end-to-end pipeline: discriminative feature extraction, compact compression, and machine-friendly feature refinement.
- We propose a dual-codebook contrastive learning strategy, where discriminative codebooks constructed from machine-oriented positive and negative samples guide the encoder to learn analysis-friendly features. This design discards the codebooks on the encoder side after training, enabling deployment on storage-constrained platforms such as underwater vehicles.
- We construct the first underwater scene VQA (UVQA) benchmark for comprehensive UIC evaluation and demonstrate that our framework consistently outperforms state-of-the-art codecs on classic vision and LVLM tasks under extremely low bitrates ($<$0.04 bpp).

2 RELATED WORK

2.1 Image Compression for Machine Vision

Unlike traditional image compression methods such as JPEG (1991), BPG (2012), and VVC (2021) – which are optimized for human perception – machine vision-oriented compression aims to preserve task-relevant semantics rather than pixel-level fidelity. It focuses on extracting and transmitting compact feature representations suitable for downstream tasks such as detection, segmentation, and recognition. Existing approaches can be broadly categorized into task-specific and task-agnostic schemes. Task-specific methods directly transmit features tailored for a particular task. For example, Chen et al. (2019b) leverage intermediate deep features for intelligent sensing, while Wang et al. (2019; 2021a) transmit compact facial representations to support recognition tasks. In contrast, task-agnostic methods aim to reconstruct a machine-friendly image that generalizes across tasks. For example, Yang et al. (2021) transmit sparse edges to reconstruct recognition-friendly face images, and Wang et al. (2021b) investigate network architecture and rate–accuracy trade-offs tailored for machine vision tasks.

In underwater scenarios, limited communication bandwidth and complex degradation patterns pose additional challenges to feature consistency and semantic preservation. Recently, Fang et al. (2022) explore a better machine-

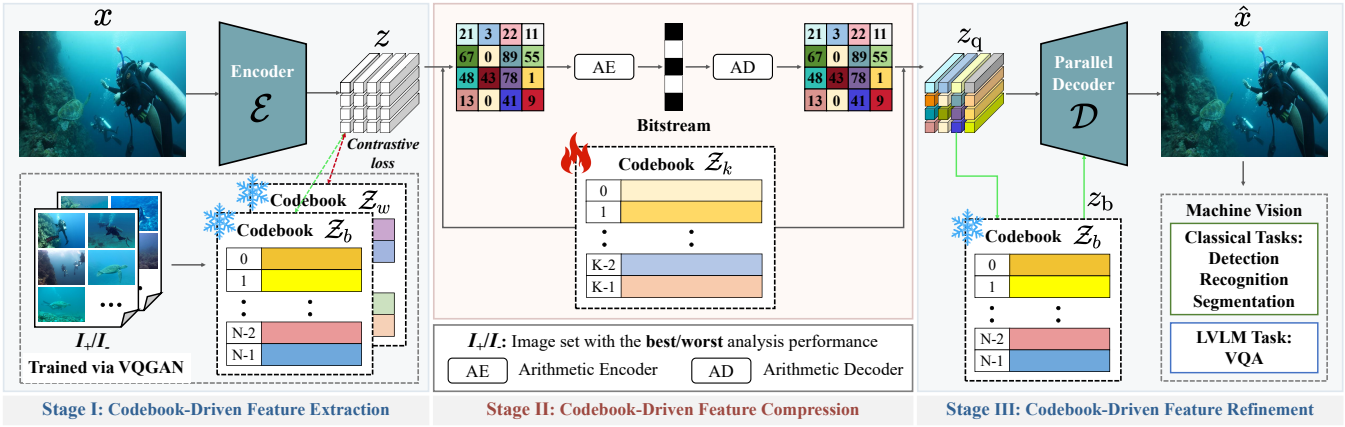

Figure 2: Framework of our proposed method. Functionally, it consists of three main components: Codebook-driven Feature Extraction, Codebook-driven Feature Compression, and Codebook-driven Feature Refinement.

oriented feature space through contrastive training, constructing a specialized dataset to learn feature discrimination based on machine-related metrics. Fang et al. (2023) further design a feature degradation removal module to alleviate the influence of underwater degradation on machine vision by taking analysis-friendly enhanced images as auxiliary information. These works motivate us to adopt more general external priors in order to learn more discriminative features that are critical for downstream visual tasks.

2.2 Verctor-Quantized Codebook

The concept of the vector-quantized (VQ) codebook is first introduced in the VQ-VAE (2017) for learning discrete image representations. Building upon this, VQGAN (2021), combining adversarial learning with refined codebook training, exhibits a powerful and robust representational capacity. In the next section, we briefly review the VQGAN architecture and its codebook training methodology, which serves as a foundation for our approach. Recently, pretrained codebooks have been leveraged as high-quality priors in various image refinement tasks, including blind face restoration (2022), adverse weather removal (2023), and low-light enhancement (2024). Additionally, Mao et al. (2024) further integrate VQ-indices compression into VQGAN, showing that the learned codebook enables effective extreme image compression and generalizes well across varying semantics and resolutions. Motivated by these findings, we explore applying the codebook to UIC.

3 Methodology

3.1 Overall Architecture

As illustrated in Fig. 2, our proposed framework comprises three key stages: Codebook-Driven Feature Extraction (CDFE), Codebook-Driven Feature Compression (CDFC), and Codebook-Driven Feature Refinement (CDFR). Our core insight is to leverage the VQ codebook to sequentially enable discriminative feature extraction, compact compression, and machine-friendly feature refinement within a unified end-to-end pipeline.

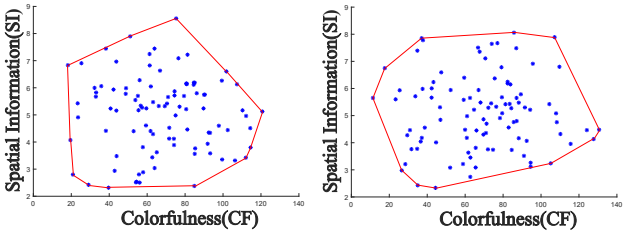

Figure 3: Spatial Information and Colorfulness distributions of the images used for codebook training. The left plot shows machine-friendly underwater images, while the right plot shows machine-unfriendly ones.

3.2 Stage I: CDFE

In this stage, given an input underwater image $\boldsymbol{x}$, the feature encoder $\mathcal{E}$ is trained to extract machine-friendly features $\boldsymbol{z}$ through contrastive learning guided by two VQ codebooks: $\mathcal{Z}_b$ and $\mathcal{Z}_w$, learned from positive sample set I_+ and negative sample set I_- with high and low machine analysis performance, respectively. Considering the high correlation among underwater images and viewing codebook training as a feature clustering process, the extracted features $\boldsymbol{z}$ are quantized to their nearest entries in the two codebooks, resulting in approximated high-quality features $\boldsymbol{z}_+$ and low-quality features $\boldsymbol{z}_-$. Contrastive feature learning then encourages the encoder to produce discriminative and analysis-friendly representations by optimizing a loss that pulls $\boldsymbol{z}$ closer to $\boldsymbol{z}_+$ and pushes it away from $\boldsymbol{z}_-$.

Image Set for Contrastive Learning. To train the VQGAN codebook, we construct an external underwater image set with both high and low machine analysis performance. Specifically, we sample $8,000$ machine-friendly and $8,000$ machine-unfriendly images from three publicly available datasets: SUIM (2020), RUOD (2023), and USOD (2023). From SUIM, we select 500 images with the highest and 500 with the lowest mIoU scores, based on its semantic segmen-

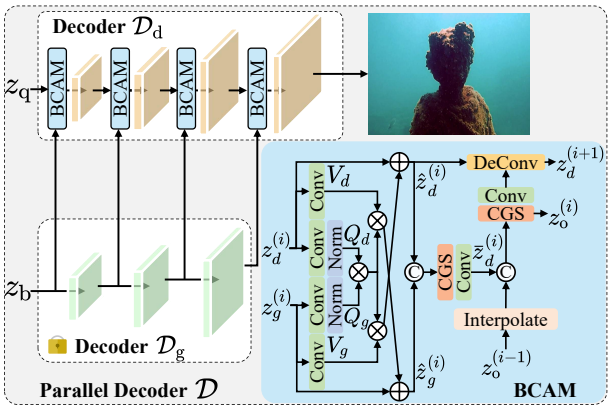

Figure 4: The architecture of the parallel decoder $\mathcal{D}$.

tation algorithm (2020). For RUOD, we select $3,500$ images with the highest and lowest mAP scores using MMDetection (2019a). For USOD, $4,000$ images with the highest and lowest F_β scores are chosen using the SVAM algorithm (2020). To verify the diversity of the selected images, we evaluate them using two metrics – Spatial Information (SI) and Colorfulness (CF) – which capture variations in spatial texture and color content, respectively (2012). Fig. 3 shows the distribution of 100 randomly sampled images in the SI-CF space. The broad distribution indicates significant diversity in both spatial texture and color.

Codebook Learning. Take $\mathcal{Z}_b$ as an example. We briefly describe the VQGAN's codebook learning mechanism, and more details can be found in (2021). Given an image $x_+ \in \mathbb{R}^{H\times W\times 3}$ from positive sample set I_+, the VQ codebook is defined as $\mathcal{Z}_b = \{z_k\}_{k=0}^{N-1} \subset \mathbb{R}^{N\times n_z}$, where N is the number of discrete codes and n_z the dimension of each. The encoder E encodes x_+ to a latent representation $f = E(x_+) \in \mathbb{R}^{h\times w\times n_z}$, where $h \times w$ denotes the spatial resolution. Each latent vector f_{ij} is quantized by the nearest codebook entry via a vector quantization module $\mathcal{Q}$:

$$f_\mathrm{q} = \mathcal{Q}(f) := (\arg\min_{z_k\in\mathcal{Z}_b} \|f_{ij} - z_k\|_2^2) \in \mathbb{R}^{h\times w\times n_z}, \quad (1)$$

where f_q is the quantized latent representation, which is then decoded by D to reconstruct the image $\hat{x}_+ = D(f_\mathrm{q})$. The model is trained end-to-end using the following objective:

$$\mathcal{L}_{VQ} = \|x_+ - \hat{x}_+\|^2 + \|\mathrm{sg}[E(x_+)] - f_\mathrm{q}\|_2^2 + \|\mathrm{sg}[f_\mathrm{q}] - E(x_+)\|_2^2, \quad (2)$$

where the first term is the reconstruction loss, and $\mathrm{sg}[\cdot]$ denotes the stop-gradient operation. The second term updates the codebook, while the third term is the so-called "commitment loss" (2017).

3.3 Stage II: CDFC

In this stage, the codebook $\mathcal{Z}_k$ maps the latent representation $\boldsymbol{z}$ into a sequence of VQ-indices via nearest neighbor search, which are then used to reconstruct the quantized latent $\boldsymbol{z}_\mathrm{q}$ accordingly. The VQ-indices provide a compact representation and are further compressed into a bitstream using lossless arithmetic coding (2004). After decoding the bitstream, the reconstructed latent vectors $\boldsymbol{z}_\mathrm{q}$ are generated by retrieving the corresponding codebook entries from the decoded indices.

3.4 Stage III: CDFR

In this stage, given the quantized features $\boldsymbol{z}_\mathrm{q}$, the learned feature decoder $\mathcal{D}$ is trained to combine the corresponding high-quality codebook features $\boldsymbol{z}_\mathrm{b}$ for image reconstruction, further mitigating the joint degradation caused by underwater imaging and compression. As shown in Fig. 4, our parallel feature decoder takes both $\boldsymbol{z}_\mathrm{q}$ and $\boldsymbol{z}_\mathrm{b}$ as input, consisting of a main branch $\mathcal{D}_\mathrm{d}$ and an auxiliary branch $\mathcal{D}_\mathrm{g}$. Their multi-level features are progressively fused by multiple bi-directional cross-attention modules (BCAMs).

Let us denote the multi-level features in the $\mathcal{D}_\mathrm{g}$ branch as $\boldsymbol{z}_\mathrm{g} = \{\boldsymbol{z}_\mathrm{g}^{(i)}\}$ (with $\boldsymbol{z}_\mathrm{g}^{(1)} = \boldsymbol{z}_\mathrm{b}$) which are fused with the quantized features $\boldsymbol{z}_\mathrm{q}$ by the BCAM, to get the multi-level features $\boldsymbol{z}_\mathrm{d} = \{\boldsymbol{z}_\mathrm{d}^{(i)}\}$ in the main decoder $\mathcal{D}_\mathrm{d}$ (with $\boldsymbol{z}_\mathrm{d}^{(1)} = \boldsymbol{z}_\mathrm{q}$). Mathematically, the process of multi-level features generation in $\mathcal{D}_\mathrm{d}$ and $\mathcal{D}_\mathrm{g}$ can be expressed as follows:

$$\boldsymbol{z}_\mathrm{g}^{(i)} = Res(Conv(\boldsymbol{z}_\mathrm{g}^{(i-1)})), \quad (3)$$

$$\boldsymbol{z}_\mathrm{d}^{(i)} = Res(Conv(BCAM(\boldsymbol{z}_\mathrm{d}^{(i-1)}, \boldsymbol{z}_\mathrm{g}^{(i-1)})), \quad (4)$$

where $Res(\cdot)$ denotes the ResBlock (2016), $Conv(\cdot)$ denotes the convolution, and the upsampled version of $\boldsymbol{z}_\mathrm{d}^{(4)}$ is the reconstructed underwater image $\hat{\boldsymbol{x}}$. Given the i-th level features $\boldsymbol{z}_\mathrm{d}^{(i)}$ and $\boldsymbol{z}_\mathrm{g}^{(i)}$, the $BCAM$ works as follows. Inspired by cross-attention for global context modeling (2017), we project these features through convolutional mappings with normalization to obtain query-value pairs $(Q_\mathrm{d}, V_\mathrm{d})$ and $(Q_\mathrm{g}, V_\mathrm{g})$, and then fuse them using the following bi-directional cross-attention formula:

$$\hat{\boldsymbol{z}}_\mathrm{d}^{(i)} = \lambda_\mathrm{d} \cdot Softmax\left(\frac{Q_\mathrm{d}Q_\mathrm{g}^\top}{\sqrt{D}}\right) V_\mathrm{g} \oplus \boldsymbol{z}_\mathrm{d}^{(i)}, \quad (5)$$

$$\hat{\boldsymbol{z}}_\mathrm{g}^{(i)} = \lambda_\mathrm{g} \cdot Softmax\left(\frac{Q_\mathrm{d}Q_\mathrm{g}^\top}{\sqrt{D}}\right) V_\mathrm{d} \oplus \boldsymbol{z}_\mathrm{g}^{(i)}, \quad (6)$$

where $Softmax(\cdot)$ denotes the softmax function. $\lambda_\mathrm{d}, \lambda_\mathrm{g}$ are trainable channel-wise scaling parameters initialized as zero vectors, $\sqrt{D}$ denotes the temperature scaling factor, and $\oplus$ represents element-wise addition. To better fuse the high-quality codebook prior features into the degraded features, we first generate an offset by concatenating the two output features, which can be formalized as follows:

$$\bar{\boldsymbol{z}}_{\boldsymbol{d}}^{(i)} = Conv\left(CGS\left(Cat(\hat{\boldsymbol{z}}_\mathrm{d}^{(i)}, \hat{\boldsymbol{z}}_\mathrm{g}^{(i)})\right)\right), \quad (7)$$

$$\boldsymbol{z}_\mathrm{o}^{(i)} = CGS\left(Cat\left(\bar{\boldsymbol{z}}_{\boldsymbol{d}}^{(i)}, Interpolate(\boldsymbol{z}_\mathrm{o}^{(i-1)})\right)\right), \quad (8)$$

where $CGS(\cdot)$ denotes a Conv-GroupNorm-SiLU sequence, $Cat(\cdot)$ denotes the concatenate operation, and

Q:How many flippers are on the robot? **A:**Six
Q:What is the color of the flippers on the robot? **A:**Yellow

Q:Is there a sea urchin in the image? **A:**Yes
Q:How many fish are visible in the image? **A:**One

Q:Are there any fish in this image? **A:**No
Q:What color are the flippers on the right diver? **A:**Blue

Q:Is there any coral in the image? **A:**Yes
Q:How many turtles are there in the image? **A:**One

Figure 5: Sample images from the UVQA dataset and the corresponding QA pairs.

$Interpolate(\cdot)$ denotes the Interpolation operation. Then, we use the generated offset in the deformable convolution to warp the codebook features to match the fidelity of the input, effectively removing the blurry degradation caused by compression. This process can be formalized as follows:

$$\boldsymbol{z}_{\mathrm{d}}^{(\boldsymbol{i+1})} = DeConv\left(\hat{\boldsymbol{z}}_{\mathrm{d}}^{(\boldsymbol{i})}, Conv(\boldsymbol{z}_{\mathrm{o}}^{(\boldsymbol{i})})\right), \tag{9}$$

where $DeConv(\cdot)$ denotes the deformable convolution (2017).

3.5 Loss Functions

The training objective of our framework comprises five components. Specifically, we adopt the triplet loss (2015) as a contrastive loss $\mathcal{L}_{con}$ to learn discriminative, machine-friendly features based on one positive and one negative sample, represented by:

$$\mathcal{L}_{con} = \max(0, d(\boldsymbol{z}, \boldsymbol{z}_{+}) - d(\boldsymbol{z}, \boldsymbol{z}_{-}) + margin), \tag{10}$$

where $d(\cdot)$ and $margin$ denote feature Euclidean distance and contrastive degree between the positives and the negatives. To efficiently enable end-to-end optimization of the model and the learnable discrete codebook $\mathcal{Z}_k$, we follow VQGAN (2021) and adopt its two loss terms, the reconstruction loss $\mathcal{L}_{rec}$ and the codebook loss $\mathcal{L}_{vq}$, which can be expressed as follows:

$$\mathcal{L}_{rec} = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2, \tag{11}$$

$$\mathcal{L}_{vq} = \|\mathrm{sg}[\mathcal{E}(\boldsymbol{x})] - \boldsymbol{z}_{\mathrm{q}}\|_2^2 + \|\mathrm{sg}[\boldsymbol{z}_{\mathrm{q}}] - \mathcal{E}(\boldsymbol{x})\|_2^2. \tag{12}$$

Considering that the purpose of the framework is to serve machine vision, we employ several feature-level and task-related loss functions instead of perceptual metrics to guarantee feature consistency and optimize network performance. The feature-preserving loss $\mathcal{L}_{fea}$ and task-related loss $\mathcal{L}_{task}$ can be expressed as follows:

$$\mathcal{L}_{fea} = d(VGG(\hat{\boldsymbol{x}}), VGG(\boldsymbol{x})), \tag{13}$$

$$\mathcal{L}_{task} = T(\hat{\boldsymbol{x}}, \boldsymbol{x}), \tag{14}$$

where VGG is the pretrained VGG-19 network (2014), and $T(\cdot)$ denotes difference of analysis performance measured by a machine analysis task. In this paper, we employ the cross entropy loss for semantic segmentation task (2020) as $\mathcal{L}_{task}$, given that semantic segmentation is a comprehensive analysis task. The total loss function is the combination of the above loss functions:

$$\mathcal{L} = \lambda_{con}\mathcal{L}_{con} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{vq}\mathcal{L}_{vq} + \lambda_{task}\mathcal{L}_{task} + \lambda_{fea}\mathcal{L}_{fea}, \tag{15}$$

where λ_{con}, λ_{rec}, λ_{vq}, λ_{task}, λ_{fea} are hyperparameters that balance the contributions of each loss component.

4 Experiments

4.1 Experimental Settings

Dataset. We train our method on the SUIM dataset (2020), which includes over $1,600$ underwater images with pixel-level semantic annotations across eight categories. We focus on five representative classes: reefs/invertebrates (RI), fish/vertebrates (FV), wrecks/ruins (WR), robots/instruments (RO), and human divers (HD). Evaluation is conducted on SUIM-100 for underwater semantic segmentation, RUOD-500 (2023) for object detection, and USOD-300 (2023) for salient object detection. Additionally, we introduce UVQA, a multiple-choice Visual Question Answering (VQA) task designed to evaluate the impact of UIC on large vision-language models. Unlike traditional detection or segmentation tasks, UVQA covers diverse aspects such as object presence, quantity estimation, and spatial localization in underwater scenes. Compared to open-ended VQA, the multiple-choice format constrains predictions to predefined candidate answers, enabling straightforward and quantitative evaluation via accuracy. Built upon SUIM and inspired by the design of (2015), the UVQA dataset includes over 8K questions and 40K candidate answers. We also use SUIM-100 to benchmark performance. Representative examples are shown in Fig. 5, with more details in the supplementary material.

Implementation Details. The model is trained end-to-end for $4,000$ epochs using the Adam optimizer (β_1 = 0.9, β_2 = 0.999). The learning rate is initialized at 10^{-4} and decayed by a factor of 10 after $1,000$ epochs. To ensure effective codebook training, we periodically replace inactive codes following (2023). The auxiliary decoder $\mathcal{D}_{\mathrm{g}}$ is initialized from the pre-trained VQGAN model (2021) and kept frozen. We empirically set λ_{con}, λ_{rec}, λ_{vq}, λ_{task}, and λ_{fea}

Method	Semantic Segmentation								Object Detection			Saliency Detection			UVQA	
	bpp	RI	FV	WR	RO	HD	mIOU(↑)	m$\mathcal{F}$(↑)	bpp	RA(↑)	mAP(↑)	bpp	S_α(↑)	F_β(↑)	bpp	Acc(↑)
Original	-	89.11	87.78	73.25	87.08	90.41	85.53	87.21	-	70.56	60.0	-	91.97	96.21	-	86.6
JP2K(2002)	0.025	21.43	19.57	1.58	5.74	19.45	13.55	13.39	0.025	44.59	23.1	0.026	62.47	65.79	0.025	81.4
BPG(2012)	0.026	29.71	41.58	11.13	8.29	45.58	27.26	25.84	0.025	37.92	17.2	0.025	77.06	84.34	0.026	**83.0**
VVC(2021)	0.026	36.70	**67.20**	12.92	27.51	39.16	36.70	33.97	0.024	47.57	24.3	0.025	79.57	85.16	0.026	82.2
Cheng(2020)	0.027	30.21	57.37	13.24	29.92	47.88	35.72	34.12	0.026	45.37	30.7	0.026	79.82	85.76	0.027	82.0
Zou(2022)	0.025	33.28	62.51	12.43	31.15	46.91	37.26	33.44	0.025	46.69	33.2	0.024	81.77	87.54	0.025	**83.0**
ELIC(2022)	0.025	29.96	53.75	2.80	**32.62**	54.97	34.82	35.62	0.026	48.32	**36.9**	0.027	**87.04**	89.88	0.025	81.2
Ours	0.022	**62.45**	52.37	**36.26**	29.92	**55.46**	**47.29**	**40.96**	0.022	**48.90**	34.4	0.024	82.21	**90.71**	0.022	82.4
JP2K(2002)	0.040	29.07	31.79	3.66	14.71	35.97	23.04	23.43	0.038	48.85	27.7	0.039	67.77	79.01	0.040	82.0
BPG(2012)	0.040	41.92	71.06	23.00	26.83	60.78	44.72	39.22	0.039	53.45	32.4	0.039	82.86	88.26	0.040	84.0
VVC(2021)	0.040	49.69	**74.52**	26.82	32.97	57.57	48.31	40.64	0.038	**53.98**	36.7	0.039	86.61	91.47	0.040	83.8
Cheng(2020)	0.042	37.84	64.19	24.59	44.83	56.50	45.59	40.28	0.040	48.75	34.5	0.043	84.31	89.33	0.042	83.0
Zou(2022)	0.041	39.54	68.13	24.88	47.63	57.10	47.46	41.12	0.039	49.91	36.8	0.040	86.45	90.42	0.041	**84.6**
ELIC(2022)	0.039	48.45	73.16	27.23	53.50	57.48	51.96	47.92	0.039	51.19	38.8	0.038	**88.08**	91.99	0.039	83.4
Ours	0.037	**61.75**	57.67	**33.72**	**56.48**	**64.53**	**54.83**	**48.01**	0.037	50.13	**39.5**	0.039	86.91	**92.41**	0.037	84.4

Table 1: Comparison results of different methods on machine vision tasks including underwater semantic segmentation (2020), underwater object detection (2023), underwater salient object detection (2023), and underwater visual question answering. The unit of all metric scores is $\%$. For all metrics, a higher score indicates a better performance. The top two scores are highlighted in **bold** and underline, respectively.

as $\{0.01,1,1,0.01,1\}$. To ensure an suitable bitrate range, we apply K-means clustering to the $4,096$ entries of the codebook $\mathcal{Z}_k$, each with 256 dimensions, from the pre-trained model. The resulting codebooks are reduced to sizes ranging from $1,024$ to 128, yielding average bitrates (bpp) of approximately 0.04 and 0.025. As a key design choice, the external codebooks $\mathcal{Z}_b$ and $\mathcal{Z}_w$ are set to 32 dimensions with $16,384$ entries, enabling finer-grained quantization and improved feature-codebook alignment. All experiments are performed on a PC equipped with a Nvidia RTX 4090 GPU.

Compared Methods. We compare our method with both traditional image compression methods and advanced deep learning-based image compression methods. The traditional methods include JP2K (2002), BPG (HEVC-intra) (2012), and VVC (VTM 11.0-intra) (2021). For BPG and VVC, we use 4:4:4 chroma format. The advanced deep learning-based image compression methods include Cheng (CVPR'20) (2020), Zou (CVPR'22) (2022), and ELIC (CVPR'22) (2022). The machine vision performance of different methods are compared under comparable bitrates.

Evaluation Metrics We evaluate all methods on SUIM-100, RUOD-500, and USOD-300 to assess machine vision performance across four tasks: underwater semantic segmentation (2020), underwater object detection (2019a), underwater salient object detection (2020; 2021c) and underwater visual question answering (UVQA). For semantic segmentation, we use mean Intersection-over-Union (mIoU) and mean Dice coefficient ($m\mathcal{F}$) to assess boundary localization. For object detection, recognition accuracy (RA) and mean Average Precision (mAP) measure detection quality. For salient object detection, mean F-measure (F_β) and S-measure (S_α) evaluate salient region localization. For UVQA, we use accuracy (Acc), defined as the ratio of correct predictions to the number of selected questions, evaluated using the LLaVA-7B model (2023). For all metrics, higher scores indicate better machine vision performance.

4.2 Machine Vision Performance Comparison

In this section, we assess the performance of our framework on four challenging underwater machine vision tasks: semantic segmentation, object detection, salient object detection, and visual question answering.

Quantitative Evaluations. Table 1 reveals that traditional codecs such as JP2K (2002), BPG (2012), and VVC (2021) suffer severe performance degradation under 0.025 bpp conditions, with semantic segmentation mIoU dropping to 13.55 (JP2K) and object detection mAP falling to 17.2 (BPG). Recent learned compression approaches (e.g., Cheng (2020), Zou (2022), ELIC (2022)) show improved robustness across downstream tasks, but often exhibit task-specific biases. For instance, ELIC achieves a high F_β of 89.88 in saliency detection, but struggles in semantic segmentation across complex regions such as WR. In contrast, our proposed method generally outperforms most baselines across tasks and bitrates. At 0.025 bpp, it achieves 47.29 mIoU, 48.90 RA, and 90.71 F_β, outperforming the next best method by a clear margin. Its performance remains strong at 0.04 bpp, demonstrating robustness and generalization. Meanwhile, accuracy (Acc) in UVQA varies only slightly across methods, suggesting that performance may be bottlenecked by the capacity of the underlying vision-language model (e.g., LLaVA-7B), rather than by the compression scheme itself.

Qualitative Evaluations. Fig. 6 shows qualitative examples of results of original images and reconstructed images

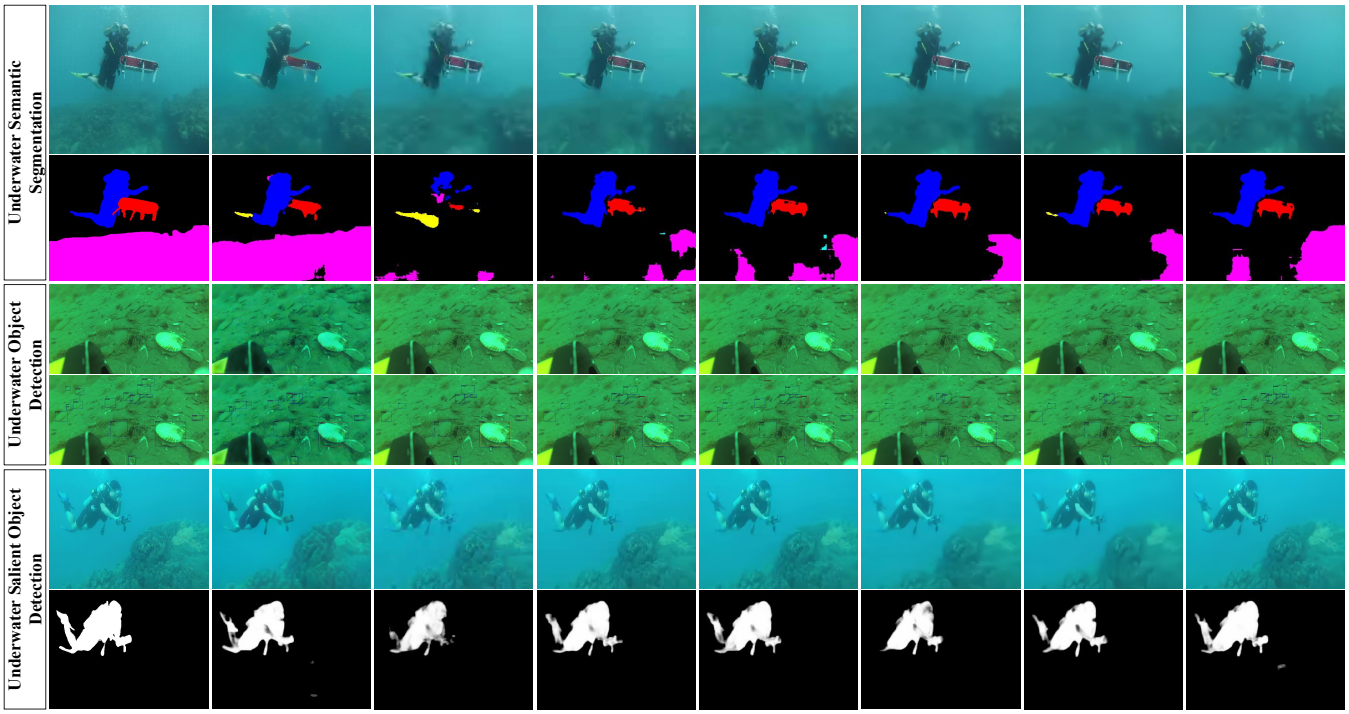

Figure 6: Qualitative comparison of different methods on machine vision tasks.

across three classical tasks at 0.04 bpp. For semantic segmentation, the second row in each group shows the predicted semantic maps, where different colors indicate different categories. Our method yields more accurate object boundaries compared to other approaches. For object detection, bounding boxes with confidence scores are overlaid on each image. Our method achieves higher detection accuracy and confidence, successfully identifying objects such as scallops that others miss. Competing methods often suffer from missed or false detections. For salient object detection, our reconstructed images produce saliency maps with sharper contours and more accurate localization.

Based on the above quantitative experiments and qualitative examples on four machine vision tasks, the superiority of our proposed method is demonstrated under extremely low bit-rate conditions.

4.3 Ablation Study

In this section, we conduct a series of ablation experiments to rigorously evaluate the effectiveness of the two proposed stages: Codebook-driven Feature Extraction (CDFE) and Refinement (CDFR). Table 2 presents the ablation study evaluating the individual and combined effects of the CDFE and CDFR. Comparing Rows (1) and (2), the introduction of CDFE brings moderate improvements across all metrics, indicating its positive contribution. However, CDFR alone (Row 3) yields significantly larger gains – most notably improving mIoU from 48.03 to 54.30 and F_β from 86.52 to 91.83 – demonstrating its dominant role in enhancing machine vision performance. Finally, combining both CDFE and CDFR in Row (4) achieves the best performance across all metrics. This clearly shows the complementary effects of the two stages, where CDFE provides a stronger representation basis and CDFR further refines it for optimal performance.

Models	mIOU(↑)	m$\mathcal{F}$(↑)	RA(↑)	mAP(↑)	S_α(↑)	F_β(↑)
(1) w/o E+R	48.03	41.96	41.23	30.8	80.23	86.52
(2) w/ E	49.06	42.75	42.86	32.4	81.57	87.43
(3) w/ R	54.30	47.10	48.72	38.9	86.08	91.83
(4) w/ E+R	**54.83**	**48.01**	**50.13**	**40.5**	**86.91**	**92.41**

Table 2: Contributions of CDFE(E) and CDFR(R) on machine vision. The bpp is set to 0.04.

5 Conclusion

In this work, we propose a novel codebook-driven framework for machine-oriented underwater image compression (UIC), addressing the dual challenges of inherent underwater degradation and extreme bitrate constraints. Our approach leverages the strong representational power and high compression efficiency of vector-quantized (VQ) codebooks to sequentially enable discriminative feature extraction, compact compression, and machine-friendly feature refinement within a unified end-to-end framework. We also introduce the first underwater visual question answering (UVQA) dataset and task, extending UIC evaluation beyond traditional benchmarks. Extensive experiments demonstrate the state-of-the-art performance of our method.

Acknowledgements

This work was supported in part by the Natural Science Foundation of Zhejiang (LR22F020002), in part by the Natural Science Foundation of China (62271277, 62471278), and in part by the Natural Science Foundation of Ningbo (2022J081).

References

Ahn, J.; Yasukawa, S.; Sonoda, T.; Nishida, Y.; Ishii, K.; and Ura, T. 2016. Image Enhancement and Compression of Deep-Sea Floor Imagefor Acoustic Transmission. In *OCEANS 2016-Shanghai*, 1–6. IEEE.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.

Bross, B.; Wang, Y.-K.; Ye, Y.; Liu, S.; Chen, J.; Sullivan, G. J.; and Ohm, J.-R. 2021. Overview of the Versatile Video Coding (VVC) Standard and its Applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10): 3736–3764.

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. 2019a. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155*.

Chen, Z.; Fan, K.; Wang, S.; Duan, L.; Lin, W.; and Kot, A. C. 2019b. Intermediate Deep Feature Compression:Toward Intelligent Sensing. *IEEE Transactions on Image Processing*, 29: 2230–2243.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7939–7948.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 764–773.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming Transformers for High-Resolution Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12873–12883.

Fang, Z.; Shen, L.; Li, M.; Wang, Z.; and Jin, Y. 2022. Prior-Guided Contrastive Image Compression for Underwater Machine Vision. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(6): 2950–2961.

Fang, Z.; Shen, L.; Li, M.; Wang, Z.; and Jin, Y. 2023. Priors Guided Extreme Underwater Image Compression for Machine Vision and Human Vision. *IEEE Journal of Oceanic Engineering*, 48(3): 888–902.

Fu, C.; Liu, R.; Fan, X.; Chen, P.; Fu, H.; Yuan, W.; Zhu, M.; and Luo, Z. 2023. Rethinking general underwater object detection: Datasets, challenges, and solutions. *Neurocomputing*, 517: 243–256.

Fu, Z.; Lin, H.; Yang, Y.; Chai, S.; Sun, L.; Huang, Y.; and Ding, X. 2022. Unsupervised Underwater Image Restoration: From a Homology Perspective. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 643–651.

Gailly, J.-l.; and Adler, M. 2004. Zlib compression library.

Gu, Y.; Wang, X.; Xie, L.; Dong, C.; Li, G.; Shan, Y.; and Cheng, M.-M. 2022. VQFR: Blind Face Restoration with Vector-Quantized Dictionary and Parallel Decoder. In *European Conference on Computer Vision*, 126–143. Springer.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5718–5727.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition*, 770–778.

He, Y.; Bu, X.; Jiang, M.; and Fan, M. 2020. Low Bit Rate Underwater Video Image Compressionand Coding Method Based on Wavelet Decomposition. *China Communications*, 17(9): 210–219.

Hoag, D. F.; Ingle, V. K.; and Gaudette, R. J. 1997. Low-Bit-Rate Coding of Underwater VideoUsing Wavelet-Based Compression Algorithms. *IEEE Journal of Oceanic Engineering*, 22(2): 393–400.

Hong, L.; Wang, X.; Zhang, G.; and Zhao, M. 2023. USOD10K: A New Benchmark Dataset for Underwater Salient Object Detection. *IEEE Transactions on Image Processing*.

Huh, M.; Cheung, B.; Agrawal, P.; and Isola, P. 2023. Straightening Out the Straight-Through Estimator: Overcoming Optimization Challenges in Vector Quantized Networks. In *International Conference on Machine Learning*, 14096–14113. PMLR.

Islam, M. J.; Edge, C.; Xiao, Y.; Luo, P.; Mehtaz, M.; Morse, C.; Enan, S. S.; and Sattar, J. 2020. Semantic Segmentation of Underwater Imagery: Dataset and Benchmark. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1769–1776. IEEE.

Islam, M. J.; Wang, R.; and Sattar, J. 2020. SVAM: Saliency-guided Visual Attention Modeling by Autonomous Underwater Robots. *arXiv preprint arXiv:2011.06252*.

Krishnaraj, N.; Elhoseny, M.; Thenmozhi, M.; Selim, M. M.; and Shankar, K. 2020. Deep learning model for real-time image compression in Internet of Underwater Things (IoUT). *Journal of Real-Time Image Processing*, 17(6): 2097–2111.

Li, M.; Shen, L.; Hua, X.; and Tian, Z. 2024. EUICN: An Efficient Underwater Image Compression Network. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, M.; Shen, L.; Ye, P.; Feng, G.; and Wang, Z. 2023. RFD-ECNet: Extreme Underwater Image Compression with Reference to Feature Dictionary. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12980–12989.

Li, Q.-Z.; and Wang, W.-J. 2010. Low-bit-rate coding of underwater color image using improved wavelet difference reduction. *Journal of Visual Communication and Image Representation*, 21(7): 762–769.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36: 34892–34916.

Mao, Q.; Yang, T.; Zhang, Y.; Wang, Z.; Wang, M.; Wang, S.; Jin, L.; and Ma, S. 2024. Extreme Image Compression Using Fine-tuned VQGANs. In *2024 Data Compression Conference (DCC)*, 203–212. IEEE.

Monika, R.; Samiappan, D.; and Kumar, R. 2021. Underwater image compression using energy based adaptive block compressive sensing for IoUT applications. *The Visual Computer*, 37(6): 1499–1515.

Naik, A.; Swarnakar, A.; and Mittal, K. 2021. Shallow-uwnet: Compressed model for underwater image enhancement (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15853–15854.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 815–823.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sullivan, G. J.; Ohm, J.-R.; Han, W.-J.; and Wiegand, T. 2012. Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12): 1649–1668.

Taubman, D. 2002. JPEG 2000: Image Compression Fundamentals, Standards and Practice.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural Discrete Representation Learning. *Advances in neural information processing systems*, 30.

Vaswani, A. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems*.

Wallace, G. K. 1991. The JPEG Still Picture Compression Standard. *Communications of the ACM*, 34(4): 30–44.

Wang, J.; He, B.; Zhang, S.; Nian, R.; Shen, Y.; and Yan, T. 2014. An Efficient Digital Image Compression Schemebased on ICA for Underwater Color Images. In *OCEANS 2014-TAIPEI*, 1–4. IEEE.

Wang, S.; Wang, S.; Yang, W.; Zhang, X.; Wang, S.; Ma, S.; and Gao, W. 2021a. Towards Analysis-friendly Face Representation withScalable Feature and Texture Compression. *IEEE Transactions on Multimedia*, 24: 3169–3181.

Wang, S.; Wang, S.; Zhang, X.; Wang, S.; Ma, S.; and Gao, W. 2019. Scalable facial image compression with deep feature reconstruction. In *2019 IEEE International Conference on Image Processing (ICIP)*, 2691–2695. IEEE.

Wang, S.; Wang, Z.; Wang, S.; and Ye, Y. 2021b. End-to-End Compression Towards Machine Vision:Network Architecture Design and Optimization. *IEEE Open Journal of Circuits and Systems*, 2: 675–685.

Wang, W.; Lai, Q.; Fu, H.; Shen, J.; Ling, H.; and Yang, R. 2021c. Salient Object Detection in the Deep Learning Era: An In-Depth Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3239–3259.

Winkler, S. 2012. Analysis of Public Image and Video Databasesfor Quality Assessment. *IEEE Journal of Selected Topics in Signal Processing*, 6(6): 616–625.

Yang, S.; Hu, Y.; Yang, W.; Duan, L.-Y.; and Liu, J. 2021. Towards Coding for Human and Machine Vision:Scalable Face Image Coding. *IEEE Transactions on Multimedia*, 23: 2957–2971.

Ye, T.; Chen, S.; Bai, J.; Shi, J.; Xue, C.; Jiang, J.; Yin, J.; Chen, E.; and Liu, Y. 2023. Adverse Weather Removal with Codebook Priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12653–12664.

Zhang, Y.; Li, Q.-Z.; and Negahdaripour, S. 2015. Seafloor Image Compression Using Hybrid Waveletsand Directional Filter Banks. In *OCEANS 2015-Genova*, 1–9. IEEE.

Zhang, Y.; Negahdaripour, S.; and Li, Q. 2016. Low bit-rate compression of underwater imagery based on adaptive hybrid wavelets and directional filter banks. *Signal Processing: Image Communication*, 47: 96–114.

Zhuang, M.; Luo, Y.; Ding, X.; Huang, Y.; and Liao, Y. 2019. A Robustness and Low Bit-Rate Image Compression Network for Underwater Acoustic Communication. In *International Conference on Neural Information Processing*, 106–116. Springer.

Zou, R.; Song, C.; and Zhang, Z. 2022. The Devil Is in the Details: Window-Based Attention for Image Compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17492–17501.

Zou, W.; Gao, H.; Ye, T.; Chen, L.; Yang, W.; Huang, S.; Chen, H.; and Chen, S. 2024. VQCNIR: Clearer Night Image Restoration with Vector-Quantized Codebook. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7873–7881.