RFD-ECNet: Extreme Underwater Image Compression with Reference to Feature Dictionary

Mengyao Li[1], Liquan Shen[1*], Peng Ye[2], Guorui Feng[1], Zheyin Wang[1]
[1]Shanghai Universit, Shanghai, China [2]Fudan University, Shanghai, China
sdlmy@shu.edu.cn, jsslq@163.com, 20110720039@fudan.edu.cn
grfeng@shu.edu.cn, wangzy_0831@163.com

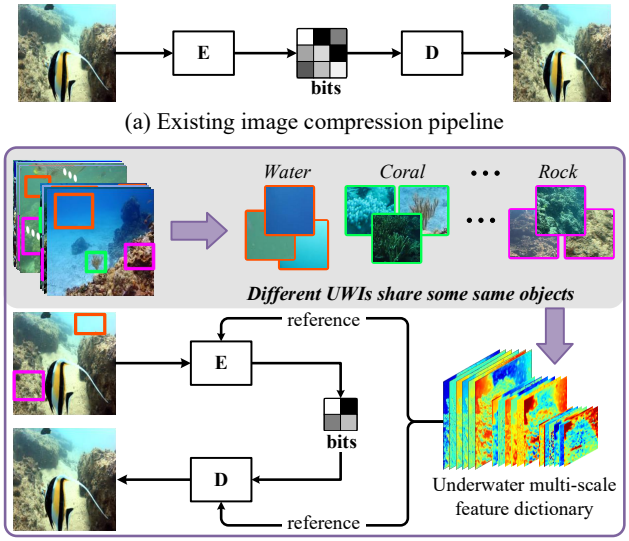
Abstract

Thriving underwater applications demand efficient extreme compression technology to realize the transmission of underwater images (UWIs) in very narrow underwater bandwidth. However, existing image compression methods achieve inferior performance on UWIs because they do not consider the characteristics of UWIs: (1) Multifarious underwater styles of color shift and distance-dependent clarity, caused by the unique underwater physical imaging; (2) Massive redundancy between different UWIs, caused by the fact that different UWIs contain several common ocean objects, which have plenty of similarities in structures and semantics. To remove redundancy among UWIs, we first construct an exhaustive underwater multi-scale feature dictionary to provide coarse-to-fine reference features for UWI compression. Subsequently, an extreme UWI compression network with reference to the feature dictionary (RFD-ECNet)[1] is creatively proposed, which utilizes feature match and reference feature variant to significantly remove redundancy among UWIs. To align the multifarious underwater styles and improve the accuracy of feature match, an underwater style normalized block (USNB) is proposed, which utilizes underwater physical priors extracted from the underwater physical imaging model to normalize the underwater styles of dictionary features toward the input. Moreover, a reference feature variant module (RFVM) is designed to adaptively morph the reference features, improving the similarity between the reference and input features. Experimental results on four UWI datasets show that our RFD-ECNet is the first work that achieves a significant BD-rate saving of 31% over the most advanced VVC.

1. Introduction

Recently, underwater exploration attracts great attention from governments, scientists, and the public due to the growing underwater applications. For example, mine search, nuclear-reactors detection, and underwater power inspection [1] are parts of homeland security operations. In addition, marine biology [2] and archaeology [3] are important scientific research for ocean resource development. Moreover, underwater entertainment [4] such as underwater live broadcasts is becoming greatly popular with the public.

In these underwater applications, images taken underwater play an essential role. To be transmitted from the deep sea to the board or ground, underwater images (UWIs) must be compressed at extremely low bitrates due to the very narrow underwater wireless acoustic bandwidth of about 20 kbps [5]. However, existing image compression approaches severely degrade the pixel fidelity (*e.g.*, MSE) of UWIs at extremely low bitrates (<0.1bpp), which greatly impairs the



(a) Existing image compression pipeline

(b) Our motivation and pipeline

Figure 1: Key ideas of our work. Different UWIs have similarities in texture, structure and semantic, which can be included in an underwater feature dictionary to provide reference for UWI compression. E/D indicates encoder/decoder.

* Corresponding Author.
[1]Code is available at https://github.com/lilala0/RFD-ECNet

broad application of UWIs. Hence, it is highly urgent to design a more powerful extreme UWI compression algorithm.

Different from general images, UWIs present two unique underwater characteristics: **(1) UWIs present multifarious underwater styles of color shift and distance-dependent clarity.** This is because the underwater imaging process is quite different from that in the open air [6]. When light travels through water, it suffers from absorption and scattering. Specifically, the blueish/greenish color shift is caused by the fact that the red light of the shortest wavelength is absorbed first, and then the green and blue light are followed [7]. In addition, light is scattered by the particles in water, which change the direction of light propagation [8], resulting in the spatially varying distance dependencies of clarity, $i.e.$, the image clarity decreases as the distance to the objects increases. **(2) Different UWIs contain some common underwater objects at diverse morphologies and sizes.** Due to the special underwater natural environment, there are some common underwater objects such as fish, corals, and rocks that widely exist in different UWIs in diverse morphologies and sizes. Hence, there are plenty of similar representations between UWIs. As shown in Fig. 1 (b), although the UWIs are captured at different times and in various underwater scenes, they contain some same underwater objects such as water, corals, and rocks, which share plenty of similarities in textures, structures, and semantics.

Given the characteristics of UWIs, there are two main drawbacks for existing image compression methods [9–21]: **(1) They only remove redundancy within an image and do not consider redundancy between UWIs.** Conventional image compression codecs such as JPEG2000 [9], BPG [10] and the latest VVC-intra [11] reduce the intra-redundancy through transform and intra-prediction. Recently, learnt image compression methods [12–18] have shown greater potential than conventional codecs. Based on variational autoencoder (VAE), [12] designs a hyperprior model to remove the statistic redundancy within an image. After that, [13] proposes an autoregressive model to further remove the local context redundancy. However, all of them are limited to removing redundancy in only an image. **(2) They utilize a unitary compression that cannot handle the multifarious underwater styles of color shift and distance-dependent clarity.** It is reasonable to use the unitary compression for the general terrestrial images because the colors (r, g, b) of terrestrial images are evenly distributed [22] and the internal clarity is steady broadly. However, it limits their performance on UWIs because the UWIs are full of multifarious color shift and distance-dependent clarity. To design an efficient extreme UWI compression method, it needs to both remove redundancy between UWIs and align the multifarious underwater styles.

Accordingly, we creatively propose to compress UWIs by referencing a dictionary to remove redundancy between UWIs. First, we construct an exhaustive underwater multi-scale feature dictionary to provide from coarse to fine reference information for the UWI compression. Specifically, the features in the dictionary are extracted from plenty of representative UWIs, which are strictly selected according to various underwater scenes, including different underwater objects, water types, and water depths. As such, our dictionary could cover a wide range of underwater common objects, assisting the UWI compression to fully remove redundancy between UWIs. Subsequently, we design a novel extreme UWI compression network with reference to the underwater feature dictionary (RFD-ECNet), shown in Fig 2, where the features of input UWI are matched with the dictionary to select the reference features, and then the residuals between the input and reference features are compressed, significantly reducing the coding bits of UWIs.

Additionally, an underwater styles normalization block-based feature match (USN-FMM) is proposed to ensure that the match is not interfered by the multifarious underwater styles and focuses on selecting the references that have the most similar texture to the input. To be specific, the USNB utilizes the underwater physical priors (UPPs) extracted from the underwater imaging physical model [6] to normalize the underwater style of reference features towards that of the input. Moreover, to improve the similarity between the reference and input features, a reference feature variant module (RFVM) is designed to adaptively morph the reference features based on their dependency relevance to the input feature. Extensive experimental results show that our RFD-ECNet outperforms state-of-the-art (SOTA) compression methods. Overall, our main contributions are summarized as follows:

- We make the first attempt to remove the massive redundancy between UWIs, and the proposed dictionary-based compression network RFD-ECNet is the first work that achieves significantly better BD-rate/PSNR than the latest VVC and other learnt methods.
- We construct an exhaustive underwater feature dictionary from representative UWIs manually selected based on various underwater scenes, water types, and water depths. Besides, detailed analyses are conducted to verify the comprehensiveness and compactness of the dictionary.
- To eliminate the effect of underwater styles on feature match, we utilize the underwater physical priors to design USNB to improve feature match accuracy. Moreover, we design RFVM to adaptively morph the reference features based on the dependency map, improving their similarity.

2. Related Works

Learnt Image Compression. Recently, learnt image compression approaches have developed rapidly and

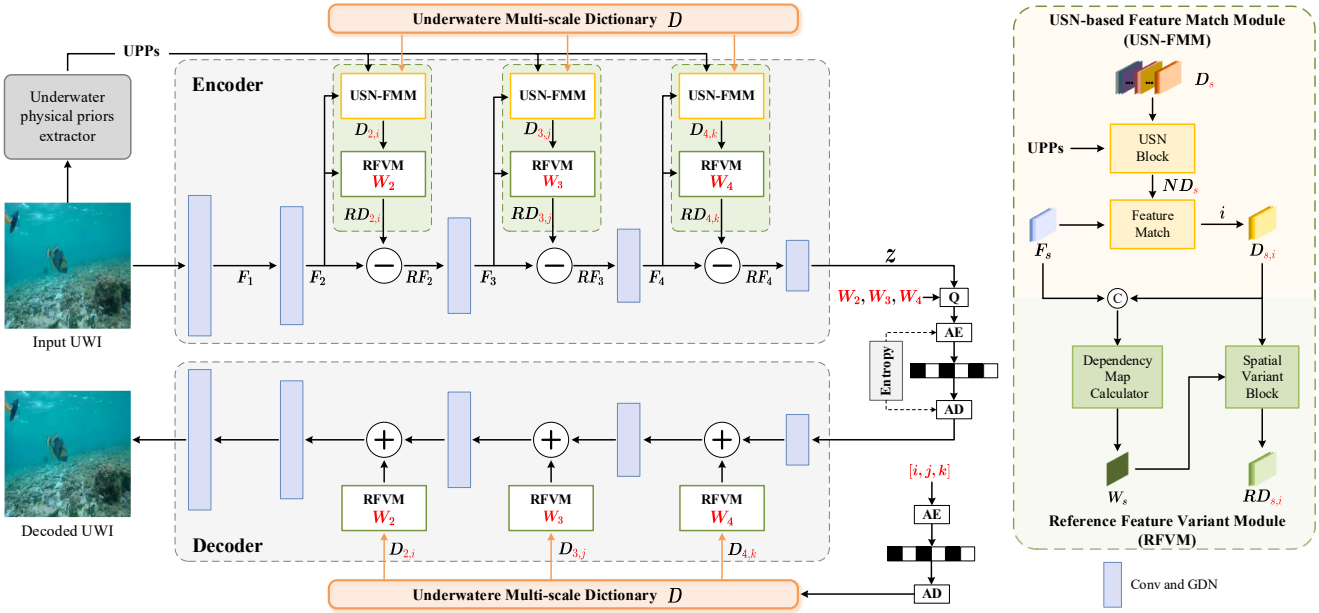


Figure 2: Overview of the proposed RFD-ECNet. Q indicates the quantizer. AE/AD represents the arithmetic en/decoding.

achieved great breakthroughs, which can be divided into VAE-based and GAN-based. Generally, VAE-based methods [12–18] utilize an encoder, which includes some parametric linear and nonlinear transforms, to compress image to compact latent features. After quantization and entropy coding, the latent features are compressed into bit-stream. Then a decoder, which usually has a symmetrical structure with the encoder, is used to reconstruct the image.

To improve the rate-distortion (R-D) performance, some works focus on designing more efficient learnt model to capture redundancy in an image. [13] proposes autoregressive model to combine the context model with the hyperprior model [12] for entropy estimation. [18] proposes window-based local attention block to capture local redundancy such as non-repetitive textures. [17] proposes Gaussian Mixture Model to estimate likelihoods of latent features more accurately. Although VAE-based methods achieve admirable R-D performance, they often present unpleasing artifacts at extremely low bitrates.

To this end, some works [19–21] combine VAE with the generative adversarial network (GAN) to generate visually pleasing textures to counteract compression artifacts. Nevertheless, the pixel fidelity (*e.g.*, MSE) of the image reconstructed by the GAN-based methods is quite low because the generated content is fake and deceptive. Hence, it is difficult for existing methods to achieve high pixel fidelity UWI reconstruction at extremely low bitrates by only removing redundancy within an image.

Image Set Compression. This direction aims to compress a batch of images captured at the same place from different view angles and focal distances [23], which can be divided into local-based and cloud-based. The local-based methods [24–26] construct the reference by first preprocessing all the images in a set. For example, [25] creates a new low frequency template as reference image by averaging the low frequency components of all images. The cloud-based methods such as [27] select reference images from the massive images in the cloud.

Obviously, the local-based methods must obtain all images before compression, which limits their usage flexibility. The cloud-based methods rely on the wide transmission channel to access the cloud disk, which cannot be achieved by the very narrow underwater wireless acoustic channel. Moreover, since these methods can only be applied to the small image subset captured at the same place, they cannot deal with UWIs captured at different underwater scenes and depths in the vast underwater world.

Reference-based Model. To the best of our knowledge, there exists no image compression method based on image/feature reference. The closest area to our work is reference-based super-resolution (RefSR), which utilizes the high-frequency information in high-resolution (HR) reference image/patch to rich the details of low resolution (LR) image. Given an arbitrary HR image, [28] uses the VGG network to simultaneously extract the referenced HR and LR features, and swaps the most similar features of reference and the LR. To overcome the inherent information deficit of a single reference, [29] uses multiple reference images to provide a more diverse pool of image features. Besides, [30] builds a face dictionary from HR face images

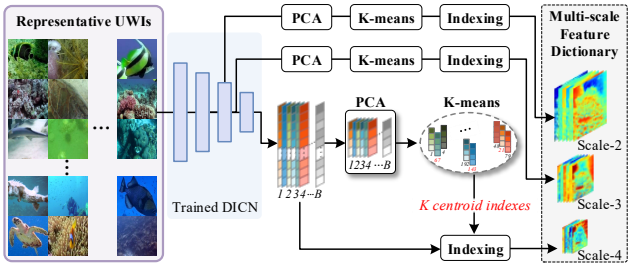


Figure 3: Construction of underwater multi-scale feature dictionary. B is the number of the representative UWIs.

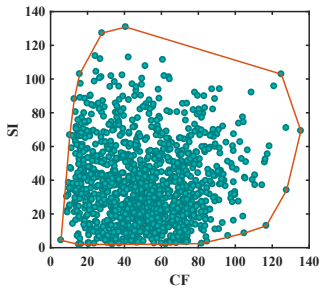


Figure 4: SI and CF of images used to build the dictionary.

to provide the high-frequency reference to construct more and richer features for the LR image.

The differences between our work and the RefSR works are as follows: (1) The RefSR task aims to learn more/richer features from the HR reference, which is a process of increasing information. In contrast, our compression task aims to learn fewer/more compact features, a process of removing redundant information. To our best knowledge, we are the first image compression work referring to the feature dictionary. (2) In [30], their dictionary only contains face images with eyes, nose and mouth, which are relatively more simple and similar than the vast UWI domain. Instead, UWIs contain multifarious underwater styles, where the common underwater objects are at diverse morphologies and sizes. To construct an exhaustive underwater dictionary, we manually select the representative UWIs based on various underwater scenes, water types, and water depths. (3) We design USNB to align the underwater styles of dictionary features and input, improving feature match accuracy. Additionally, we design RFVM to adaptively morph the reference features, further improving their similarity.

3. The Proposed Method

3.1. Exhaustive Underwater Feature Dictionary

To achieve extremely low bitrate UWI compression, we propose RFD-ECNet, a novel image compression network referring to underwater feature dictionary, to remove redundancy between UWIs. First of all, an exhaustive dictionary is constructed, which is offline and shared at the encoder and decoder. The pipeline of the construction of underwater multi-scale feature dictionary is shown in Fig. 3.

To build the feature dictionary, a series of representative UWIs are selected from the largest public underwater image dataset UGWI [2]. UGWI contains 2150 various UWIs, which are collected during oceanic explorations, human-robot collaborative experiments, or from websites such as Google and YouTube. To select representative UWIs that cover the richest and most comprehensive common contents of UWIs, we manually divided the two thousand images into 300 groups according to two rules. **1) Different Objects.** Based on [8], the common objects in UWIs can be divided into marine faunas of fish, turtles, sharks, snails and shellfish, *etc.*; marine plants of corals, seaweed, fringing reefs, *etc.*; non-biology of rocks, wreckage and sculpture, *etc.* **2) Different water types and depths.** The water type is distinguished from clear to turbid, and the water depth is distinguished from shoal water, deep sea and seabed. UWIs captured in clear and shoal water present lighter color shift and better clarity, and vice versa.

After grouping the UGWI manually, 1-2 images from each group are selected to build our dictionary, which is then cropped into non-overlapping 128×128 image patches P. For each image patch, we use the pre-trained popular deep image compression network DICN [12] to extract its compression features at multi-scales, shown in Fig. 3. The generation process of dictionary can be described as,

$$D_s = \mathcal{F}(P; \theta^s_{DICN}) \tag{1}$$

where $s \in \{2, 3, 4\}$ is the scale and θ^s_{DICN} is the parameters of the pre-trained DICN at scale s.

To verify the diversity of these representative UWIs, the Spatial Information (SI) and Colorfulness (CF) indices are computed respectively for the content variation along the spatial and color dimensions. Higher SI / CF indicates more complex image content / colorful. The CF versus SI distribution is plotted in Fig. 4. As shown, the distribution range is quite broad, which indicates that the content of these patches is diverse. Additionally, some points are distributed closely, which also verifies that there are many similar common content between UWIs.

To compact the features in dictionary, we adopt K-means to create K clusters. Considering the high computational complexity of directly performing K-means on massive features of multifarious UWIs, we first perform principal components analysis (PCA) on the features of per UWI to reduce their dimension. Finally, K typical features at each scale are preserved, constructing our multi-scale feature dictionary. The effect of cluster number K on the compression performance is shown in the supplementary materials.

[2]UGWI is a public underwater image dataset and is available online at https://github.com/Underwater-Lab-SHU

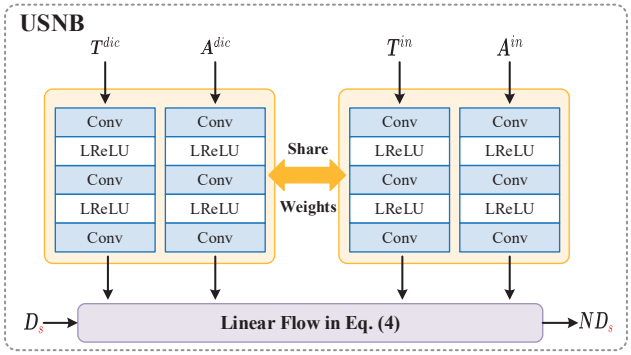


Figure 5: The structure of the proposed USNB.

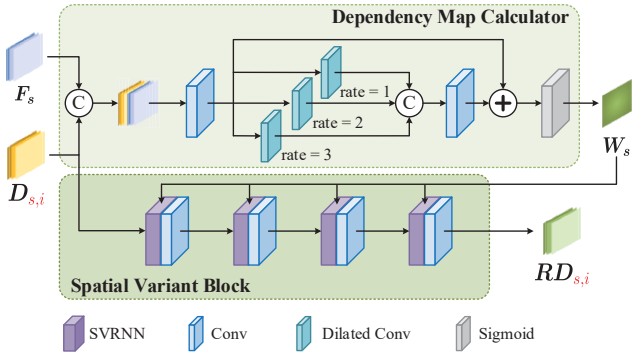


Figure 6: The structure of the proposed RFVM.

3.2. Overall of the proposed RFD-ECNet

The framework of our RFD-ECNet is shown in Fig. 2, which performs feature match at multiple scales ($s \in \{2, 3, 4\}$) in a progressive manner to remove the underwater common redundancy from coarse to fine. Let F_s be the input features at scale s, and $D_{s,i}$ denotes the i-th dictionary features at scale s. In the encoder, the input UWI is down-sampled by two convolutional sampling layers to obtain its features F_2, which are matched with the dictionary D to select its reference features $D_{2,i}$. To guarantee the input features F_2 and D in the same feature space, the first-two sampling layers of RFD-ECNet are filled with that of the pre-trained DICN, which are fixed during the training of our RFD-ECNet.

For more accurate feature match, USNB is proposed to normalize the dictionary features based on the underwater style of the input, eliminating the effect of the underwater styles diversity on feature match. Specifically, USNB utilizes the UPPs extracted from an UPPs extractor [7] to simulate the underwater imaging process [6], which dictate the underwater styles of UWIs. Moreover, the selected $D_{2,i}$ is further refined by the proposed RFVM to obtain $RD_{2,i}$ based on the similarity map W_2 between F_2 and $D_{2,i}$, greatly improving their similarity.

After removing redundancy on three scales, the latent features z and the dependency maps (W_2, W_3, W_4) are quantized by the uniform noise approximated method [12], and then entropy coded by the Gaussian mixture entropy model [17] to the bit-stream.

3.3. USNB-based Feature Match (USN-FMM)

Feature match aims to select the reference features that have the most similar textures to the input features. However, the UWIs sharing similarity may have multifarious underwater styles of color shift and distance-dependent clarity. Hence, it is necessary to normalize the underwater styles of dictionary features toward that of the input before the feature match.

Underwater Style Normalize. By studying the process of underwater imaging, it can be observed that the underwater styles are dictated by underwater physical priors such as ambient light and imaging transmission map, which can be described by the underwater physical scattering model:

$$\begin{aligned} J &= I \times T + A \times (1 - T) \\ T &= e^{-\alpha d} \end{aligned} \tag{2}$$

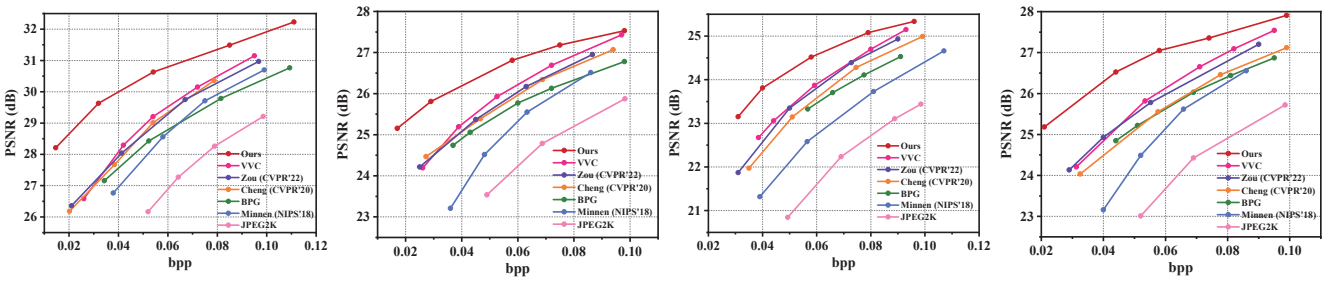
where I denotes the clear scene without underwater style and J is the captured UWI with multifarious underwater styles. Under the effect of underwater physical priors, including T and A, I is evolved to J. T and A respectively represent the transmission map and the global ambient light, and their visual examples are presented in the supplementary materials. T is calculated by α and d, which respectively refer to light attenuation coefficient and the distance between camera and object, which leads to the distance-dependent clarity. Derived from Eq. (2), the clear scene can be modeled as:

$$I = \frac{J - A \times (1 - T)}{T} \tag{3}$$

Based on Eq. (2) and Eq. (3), the dictionary features can be normalized towards the underwater style of input by

$$\begin{aligned} J^{dic\_input} &= I^{dic} \times T^{input} + A^{input} \times (1 - T^{input}) \\ &= \frac{J^{dic} - A^{dic} \times (1 - T^{dic})}{T^{dic}} \times T^{input} \\ &+ A^{input} \times (1 - T^{input}) \end{aligned} \tag{4}$$

Based on the above analysis, we propose the USNB to parametrically implement Eq. (4), as shown in Fig. 5, which is used before the feature match to normalize the underwater styles of dictionary features toward the input. In Fig. 5, D_s and ND_s respectively correspond to J_{dic} and J_{dic_input} in Eq. (4). Underwater physical priors of dictionary features (T^{dic} and A^{dic}) and that of the input (T^{in} and A^{in}) are fed to several convention layers to normalize D_s to ND_s by following the data flow in Eq. (4). Notably, the network parameters for processing T^{dic} and A^{dic} are shared



(a) UGWI, Rate-PSNR (b) EUVP, Rate-PSNR (c) UFO, Rate-PSNR (d) UIEB, Rate-PSNR

Figure 7: R-D curves of different image compression methods on UGWI, EUVP [31], UFO [32], and UIEB [8] datasets.

with that for T^{dic} and A^{dic} to guarantee that all underwater physical priors are in the same feature space. As such, the proposed USNB is able to normalize the underwater styles of the dictionary features toward the input for more accurate feature match. The effectiveness of USNB is verified in the ablation studies.

Feature Match. To select the reference features $D_{s,i}$ from the dictionary features D_s for input features F_s, we adopt the widely used inner product to measure the similarity score between F_s and D_s. The similarity scores between F_s and D_s is defined as:

$$S_s = \left\langle F_s, \frac{D_s}{\|D_s\|} \right\rangle \tag{5}$$

Following the procedure in [33], Eq. (5) is implemented by a convolution operation, where the kernel weights are set as D_s, to accelerate the similarity calculations. The i-th dictionary feature with the highest similarity score is selected as the reference feature $D_{s,i}$.

3.4. Reference Feature Variant Module (RFVM)

Although the reference features $D_{s,i}$ and the input features F_s share some similarities, the relative position and shape of the similar areas in $D_{s,i}$ and F_s may be dramatically different. To achieve the highest possible similarity, we develop the RFVM to adaptively morph the $D_{s,i}$ based on F_s, shown in Fig. 6. RFVM utilizes the spatial variant recursive convolution (SVRConv) [34] to implement the feature variant on a large receptive field, which is able to facilitate the dramatic spatial variant by the long-distance information propagation of recurrent convolution and the learned weights for different locations. The weights of SVRConv are learned by exploring the dependency between $D_{s,i}$ and F_s. To obtain the global dependency map W, dilated convolutions with different dilation rates are used to exploit the spatial information at different receptive fields. As such, our RFVM is able to morph the reference features in a large receptive field to improve the similarity with the input features.

4. Experiments

Extensive experiments are conducted to evaluate the extreme compression performance of our RFD-ECNet. First, we test RFD-ECNet on four different UWI datasets objectively and subjectively to verify our excellent generalization. After that, a series of ablation studies are presented to respectively verify the effectiveness of our multi-scale manner, USNB and RFVM.

4.1. Experimental Settings

Datasets. We perform experiments on four public UWI datasets, including the UGWI dataset, EUVP dataset [31], UFO dataset [32], and UIEB dataset [8]. These UWI datasets are collected during different oceanic explorations and hence include various underwater scenes and rich ocean species. As stated in Section 3.1, some representative UWIs selected from UGWI dataset are utilized to construct our underwater multi-scale feature dictionary. The rest of UGWI are divided into 1550 UWIs for training and 300 UWIs for testing. Additionally, we randomly select 100 images respectively from EUVP [31], UFO [32], and UIEB [8] to test the pre-trained RFD-ECNet to verify the generalization of our RFD-ECNet.

Training. As the most VAE-based methods, our RFD-ECNet is optimized by the rate-distortion trade-off loss function, which can be described as:

$$\mathcal{L} = \lambda \cdot D + R, \tag{6}$$

where λ is a trade-off parameter to balance bitrate R and distortion D. In our work, MSE is used as the D, and λ belongs to the set $\{512, 256, 128, 64, 32\}$ for 5 different compression ratios. The channel width of RFD-ECNet is set as 192, and the bottleneck layer is set as 64. Detailed network architecture is introduced in the the supplementary materials. All model are trained using the Adam optimizer [35] with standard parameters and learning rate of 1×10^{-4}.

Evaluation. The compression performance is measured by bitrate R and distortion D, which are respectively calculated by bits-per-pixel (bpp) and Peak Signal to Noise Ratio (PSNR). Lower bpp and higher PSNR indicate better com-

Table 1: Comparison of BD-rate and BD-PSNR on UGWI, EUVP [31], UIEB [8], and UFO [32] testsets. The BD-rate and BD-PSNR are calculated from PSNR-BPP curve over data points with bpp<0.1, when BPG [10] is set as the anchor. The performance of SOTA compression codec VVC is marked in red. The best performance is **bolded**.

Methods	BD-rate ↓ (%)					BD-PSNR ↑ (dB)					Computation complexity	
	UGWI	EUVP	UIEB	UFO	*Avg.*	UGWI	EUVP	UIEB	UFO	*Avg.*	Param. (M)	FLOPs (G)
VVC-intra [11]	-19.0	-16.2	-16.0	-16.1	-16.8	0.69	0.43	0.52	0.47	0.53	/	/
JPEG2000 [9]	61.7	68.4	67.2	66.9	66.1	-1.62	-1.71	-1.84	-1.56	-1.68	/	/
BPG [10]	0	0	0	0	0	0	0	0	0	0	/	/
Minnen (NIPS'18) [13]	-3.8	14.8	9.8	19.0	9.9	0.16	-0.36	-0.33	-0.59	-0.28	14.1	27.2
Cheng (CVPR'20) [17]	-12.8	-8.4	-1.3	-8.3	-7.7	0.43	0.20	0.04	0.24	0.23	20.5	61.3
Zou (CVPR'22) [18]	-13.8	-10.3	-12.4	-14.7	-12.8	0.48	0.25	0.36	0.43	0.38	75.0	33.3
Ours-slim	**-50.2**	**-44.6**	**-42.6**	**-34.7**	**-43.0**	**1.73**	**1.01**	**1.11**	**0.65**	**1.12**	15.4	40.7
Ours	**-53.7**	**-50.3**	**-52.9**	**-37.3**	**-48.5**	**1.97**	**1.18**	**1.56**	**0.99**	**1.43**	35.8	86.7

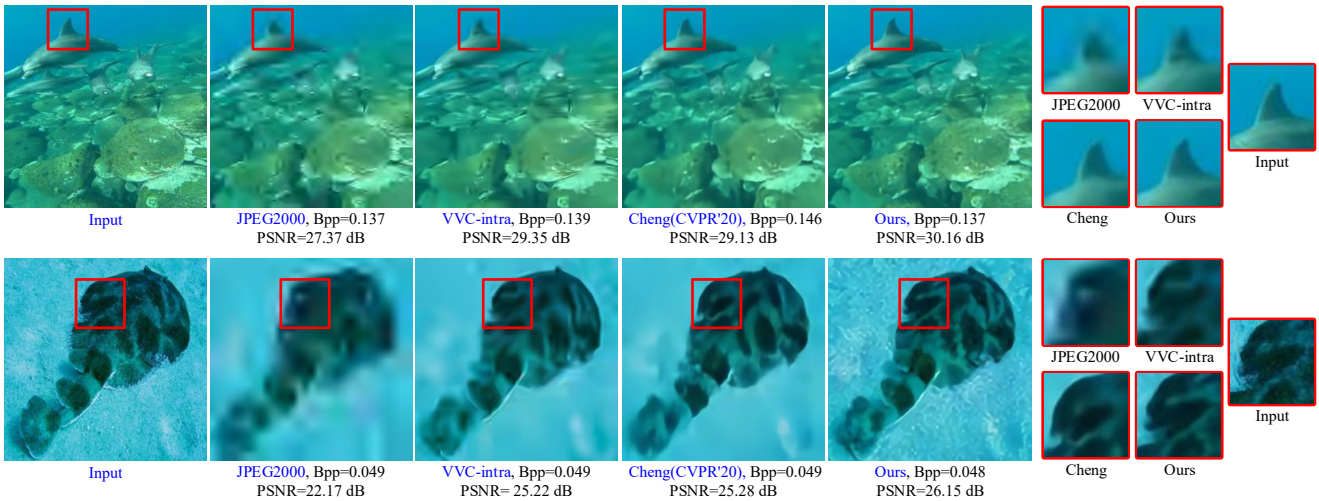


Figure 8: Visual examples of the reconstructed images. The examples in two rows are respectively from UGWI and EUVP [31] testset. The metrics are bpp (↓) and PSNR (↑). At extremely low bpp, our reconstructed images have more details and higher pixel fidelity (measured by PSNR) than others.

pression performance. Notably, our bpp is calculated by,

$$bpp = \frac{bits_z + bits_W + bits_{index}}{N_{pixels}} \tag{7}$$

where $bits_z$ and $bits_W$ respectively represent the bits of latent features z and dependency map W. $bits_{index}$ denotes the bits for coding the index numbers of reference features, which are directly coded by Huffman coding. In addition to R-D curves, the widely used Bjøntegaard delta rate (BD-rate) and BD-PSNR are also evaluated. Lower BD-rate and higher BD-PSNR indicate more efficient compression.

For the comparative methods, we compare our RFD-ECNet with influential and SOTA image compression networks, including Minnen (NIPS'18) [13], Cheng (CVPR'20) [17], and Zou (CVPR'22) [18]. For fair comparisons, all networks are trained on the UGWI dataset. Additionally, some conventional codecs are compared, including popular JPEG2000 [9], BPG (HEVC-intra) [10] and the latest VVC-intra (VTM 11.0) [11]. For BPG and VVC, the highest PSNR setting (4:4:4) is tested.

4.2. Comparison against SOTA methods

R-D Performance. Fig. 7 presents the R-D curves on four UWI testsets of UGWI, EUVP [31], UIEB [8], and UFO [32]. As shown, the R-D curves of our RFD-ECNet are above the curves of all comparative methods on four testsets, which demonstrates that our method achieves better compression performance than all comparative methods and our excellent generalization and robustness. Notably, as bpp decreases, our RFD-ECNet achieves higher PSNR improvements than comparative methods. This indicates that the reference features provided by our underwater feature dictionary are very efficient for the reconstruction, especially at extremely low bitrates.

Table 2: BD-rate (%) and BD-PSNR (dB) of VVC/HEVC at inter/intra modes, where Minnen [13] is set as the anchor.

Methods	VVC-inter	VVC-intra	HEVC-inter	HEVC-intra	Ours
BD-rate ↓	-22.1	-21.0	0.2	3.2	**-59.5**
BD-PSNR ↑	0.85	0.82	-0.10	-0.12	**2.25**

The BD-rate of different image compression methods with BPG [10] as the anchor is presented in Table 1. As shown, our method achieves the best BD-rate and BD-PSNR results on four testsets, which indicates that we use the smallest amount of bits to achieve the highest PSNR among all methods. Specially, our average BD-rate of -48.5 % and BD-PSNR of 1.43 dB are much better than the most advanced codec VVC-intra' -16.8 % and 0.53 dB. More intuitively, when VVC is set as the anchor, our RFD-ECNet achieves breakthrough BD-rate saving of 31.7 % and BD-PSNR of 0.90 dB, demonstrating the advanced performance of our RFD-ECNet on UWI extreme compression.

Visual Quality. Fig. 8 shows visual examples of reconstructed images by our RFD-ECNet, Cheng (CVPR'20), the most advanced VVC-intra, and the widely used JPEG2000. As shown, our reconstructed images retain more details, while the reconstructed images of VVC-intra and JPEG2000 present obvious blurring and blocking effects at extremely low bpp. As shown in the second row, when bpp=0.049, VVC-intra, Cheng (CVPR'20) and JPEG2000 completely loss the common object of sand, while our RFD-ECNet can roughly reconstruct the sand by referencing our dictionary. Overall, our pixel fidelity is higher than others at the similar extremely low bpp, which can be verified by our higher PSNR. More visual results tested on underwater videos are provided at https://github.com/lilala0/RFD-ECNet.

Comparison with video codecs. To verify that the redundancy between UWIs cannot be efficiently removed by video codecs, our RFD-ECNet is compared with the latest video codecs of VVC-inter and HEVC-inter. Their performance on UGWI testset measured by BD-rate and BD-PSNR with the Minnen [13] as the anchor is shown in Table 2. As shown, we achieve better results than both VVC-inter and HEVC-inter, verifying that our RFD-ECNet referencing to dictionary is more efficient than video codecs to remove the redundancy between UWIs. Moreover, both video codecs at inter-mode only achieve tiny improvement of BD-rate/PSNR than the intra-mode. This verifies that there is redundancy between UWIs indeed.

Comparison of complexity. The trainable parameters and FLOPs of all approaches are compared in Table 1. As shown, our parameters (35.8 M) is much less than that (99.8 M) of compression network [18]. Additionally, we also provide a slim version of Our RFD-ECNet by narrowing our network channels of 192 to 128 (denoted as Ours-slim), which only have 15.4M parameters and 40.7G FLOPs. Notably, ours-slim still achieves better performance than all SOTA methods with lower complexity, illustrating our excellent UWI extreme compression performance.

Table 3: BD-rate (%) and BD-PSNR (dB) of ablation studies about our USNB and RFVM, where BPG [10] is set as the anchor.

Setting	Ours	w/o USNB	w/o RFVM	w/o USNB&RFVM
BD-rate ↓	-53.7	-33.6	-43.1	-24.6
BD-PSNR ↑	1.97	1.14	1.49	0.83

Table 4: BD-rate (%) and BD-PSNR (dB) of ablation studies about the multi-scale manner, where BPG [10] is set as the anchor.

Scale setting	Scale-{2}	Scales-{2, 3}	Scales-{2, 3, 4}
BD-rate ↓	-46.1	-49.8	-53.7
BD-PSNR ↑	1.67	1.83	1.97

5. Ablation Study

Effectiveness of USNB and RFVM. To verify the effectiveness of the proposed USNB and RFVM, a series of ablation experiments are conducted, where the USNB and RFVM are removed from our RFD-ECNet in sequence. The performance of each ablation is measured by the BD-rate and BD-PSNR with the BPG as the anchor, shown in Table 3. *w/o USNB* indicates the network that directly performs feature match by inner product without using our USNB to normalize the dictionary features. *w/o RFVM* is the network where the matched reference features are not fed into our RFVM and therefore cannot be varied adaptively based on the input features. *w/o USNB&RFVM* is the network without both USNB and RFVM. As can be seen from Table 3, the removal of either USNB or RFVM degrades the performance of our RFD-ECNet. Moreover, the performance degrades most when USNB and RFVM are removed simultaneously, verifying the effectiveness of USNB and RFVM.

Effectiveness of multi-scale reference. In our RFD-ECNet, the input features are matched with the feature dictionary at three scales ($s = 2, 3, 4$), considering that the similarity between UWIs is reflected on the textures, structures, and semantics. Here we conduct ablations to verify the effectiveness of the multi-scale manner, and provide the results in Table 4. Since the features on scale-1 contain more noise less semantic information, we perform feature match starting from scale-2. As shown, our RFD-ECNet referring dictionary at scales-$\{2, 3, 4\}$ achieves the best performance, followed by scales-$\{2, 3\}$ and scale-$\{2\}$, which verifies that it is efficient to remove the underwater common redundancy at multiple scales from coarse to fine.

6. Discussion

This work creatively removes redundancy between UWIs and proposes a novel reference-based image compression framework. Actually, the redundancy between images exists not only in UWIs but also in other image domains where independent images contain some common objects specific to the image domain. For example, all Martian images [36] contain rocks and sand; all face images [37] contain eyes, nose and mouth; most remote sensing images [38] contain buildings, road and so on. However, existing image compression works neglect this nature and only remove redundancy in an image, which may limit the development of image compression.

To validate our hypothesis, we additionally conduct preliminary experiments on Martian images [36] by adjusting our network. Experimental results presented in the supplementary materials show that we also achieve better performance than VVC and Ding (VCIP'22) [36], illustrating the efficiency and feasibility of removing redundancy between images in other image domains.

7. Conclusion

In this paper, we explore the characteristics of UWIs and find that UWIs present multifarious underwater styles and different UWIs share some common ocean objects at diverse morphologies and sizes, resulting in plenty of redundancy between UWIs. Hence, we construct an exhaustive and compact underwater multi-scale feature dictionary from the manually selected representative UWIs, which provides coarse-to-fine reference for our RFD-ECNet to fully remove redundancy between UWIs. Meanwhile, to eliminate the effect of underwater styles on feature match, we utilize UPPs to design the USNB to align the multifarious underwater styles of dictionary towards that of input features. Moreover, to improve the similarity between the input and reference features, we propose RFVM to perform adaptive reference feature variant. Finally, extensive comparisons are conducted to verify our significant performance improvement than SOTA methods including the latest VVC, and comprehensive ablation studies verify the effectiveness of each module in our RFD-ECNet.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant 61931022 and Grant 62271301; in part by the Shanghai Science and Technology Program under Grant 22511105200; in part by the Shanghai Excellent Academic Leaders Program under Grant 23XD1401400; in part by the Natural Science Foundation of Shandong Province under Grant ZR2022ZD38.

References

[1] G.L. Foresti. Visual inspection of sea bottom structures by an autonomous underwater vehicle. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 31(5):691–705, 2001. 1

[2] Julia Åhlén and David Sundgren. Bottom reflectance influence on a color correction algorithm for underwater images. In *Image Analysis*, pages 922–926, 2003. 1

[3] Ya'acov Kahanov and Jeffrey G. Royal. Analysis of hull remains of the dor d vessel, tantura lagoon, israel. *The International Journal of Nautical Archaeology*, 30(2):257–265, 2001. 1

[4] Ching Yung Chen, Hsin An Hou, and Ching Min Shu. Hybrid-band relay based underwater video broadcasting systems with applications to entertainment and exploration. In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–5, 2010. 1

[5] Joseph Hall and Y. Rosa Zheng. Diversity schemes for a large mimo acoustic transmitter on a c2000 micro-controller. In *OCEANS*, pages 1–9, 2022. 1

[6] J. Y. Chiang and Y.-C. Chen. Underwater image enhancement by wavelength compensation and dehazing. *IEEE Trans. Image Process.*, 21(4):1756–1769, Apr.2012. 2, 5

[7] Y. Lin, L. Shen, Z. Wang, K. Wang, and X. Zhang. Attenuation coefficient guided two-stage network for underwater image restoration. *IEEE Signal Process. Lett.*, 28:199–203, Dec.2020. 2, 5

[8] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Trans. Image Process.*, 29:4376–4389, Nov.2019. 2, 4, 6, 7

[9] D. S. Taubman. Jpeg2000: Image compression fundamentals, standards and practice. *J. Electron. Imaging*, 11(2):286, Apr.2002. 2, 7

[10] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Trans. Circuits Syst. Video Technol.*, 22(12):1649–1668, Dec.2012. 2, 7, 8

[11] B. Bross, Y-K. Wang, Y. Ye, S. Liu, J. Chen, J. Sullivan, and J-R. Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Trans. Circuits Syst. Video Technol.*, 31(10):3736–3764, Aug.2021. 2, 7

[12] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston. Variational image compression with a scale hyperprior. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, pages 1–23, Jan.2018. 2, 3, 4, 5

[13] D. Minnen, J. Ballé, and George D. Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 31, Dec.2018. 2, 3, 7, 8

[14] Xiaosu Zhu, Jingkuan Song, Lianli Gao, Feng Zheng, and Heng Tao Shen. Unified multivariate gaussian mixture for efficient neural image compression. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 17591–17600, 2022. 2, 3

[15] M. Li, W. Zuo, S. Gu, J. You, and D. Zhang. Learning content-weighted deep image compression. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3446–3461, Oct.2021. 2, 3

[16] T. Chen, H. Liu, Z. Ma, Q. Shen, X. Cao, and Y. Wang. End-to-end learnt image compression via non-local attention optimization and improved context modeling. *IEEE Trans. Image Process.*, 30:3179–3191, Feb.2021. 2, 3

[17] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7936–7945, Jun.2020. 2, 3, 5, 7

[18] R. Zou, C. Song, and Z. Zhang. The devil is in the details: Window-based attention for image compression. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 17471–17480, Jun.2022. 2, 3, 7, 8

[19] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. Van Gool. Generative adversarial networks for extreme learned image compression. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 221–231, Oct, 2019. 2, 3

[20] F. Mentzer, G. Toderici, M. Tschannen, and E. Agustsson. High-fidelity generative image compression. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, pages 1–20, Dec.2020. 2, 3

[21] C. Huang, H. Liu, T. Chen, Q. Shen, and Z. Ma. Extreme image coding via multiscale autoencoders with generative adversarial optimization. In *IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, pages 1–4, Dec.2019. 2, 3

[22] Cosmin Ancuti, Codruta Orniana Ancuti, Tom Haber, and Philippe Bekaert. Enhancing underwater images and videos by fusion. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 81–88, 2012. 2

[23] L. Sha, W. Wu, and B. Li. Novel image set compression algorithm using rate-distortion optimized multiple reference image selection. *IEEE Access*, 6:66903–66913, Nov.2018. 3

[24] X. Zhang, W. Lin, Y. Zhang, S. Wang, S. Ma, L. Duan, and W. Gao. Rate-distortion optimized sparse coding with ordered dictionary for image set compression. *IEEE Trans. Circuits Syst. Video Technol.*, 28(12):3387–3397, Dec.2018. 3

[25] Chi-Ho Yeung, Oscar C. Au, Ketan Tang, Zhiding Yu, Enming Luo, Yannan Wu, and Shing Fat Tu. Compressing similar image sets using low frequency template. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6, 2011. 3

[26] Kosmas Karadimitriou and John M. Tyler. Min-max compression methods for medical image databases. *Association for Computing Machinery*, 26(1), 1997. 3

[27] Chen Zhao, Siwei Ma, and Wen Gao. Thousand to one: An image compression system via cloud search. In *2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2015. 3

[28] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7974–7983, 2019. 3

[29] Marco Pesavento, Marco Volino, and Adrian Hilton. Attention-based multi-reference learning for image super-resolution. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14677–14686, 2021. 3

[30] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *Computer Vision – ECCV 2020*, pages 399–415, 2020. 3, 4

[31] M. J. Islam, Y. Xia, and J. Sattar. Fast underwater image enhancement for improved visual perception. *IEEE Robot. and Autom. Lett.*, 5(2):3227–3234, Apr.2020. 6, 7

[32] M. J. Islam, P. Luo, and J. Sattar. Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception. *Robot. Sci. Syst. (RSS)*, Jul.2020. 6, 7

[33] Z. Zhang, Z. Wang, Z. Lin, and H. Qi. Image super-resolution by neural texture transfer. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7974–7983, Jun.2019. 6

[34] J. Zhang, J. Pan, J. Ren, Y. Song, L. Bao, R.W.H. Lau, and M-H. Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2521–2529, Jun.2018. 6

[35] J. Zhang, J. Pan, J. Ren, Y. Song, L. Bao, R.W.H. Lau, and M-H. Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2521–2529, Jun.2018. 6

[36] Qing Ding, Mai Xu, Shengxi Li, Xin Deng, Qiu Shen, and Xin Zou. A learning-based approach for martian image compression. In *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5, 2022. 9

[37] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, 2018. 9

[38] G. S. Xia, W. Yang, J. Delon, Y. Gousseau, and S. Hong. Structural high-resolution satellite image indexing. In *ISPRS TC VII Symposium*, 2010. 9