This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Self-Supervised Compression and Artifact Correction for Streaming Underwater Imaging Sonar

Rongsheng Qian[1], Chi Xu[1], Xiaoqiang Ma[2], Hao Fang[1], Yili Jin[3,1], William I. Atlas[4], Jiangchuan Liu[1]
[1]Simon Fraser University, BC, Canada, [2]Douglas College, BC, Canada,
[3]McGill University, QC, Canada, [4]Wild Salmon Center, OR, USA
{rqa4, chix}@sfu.ca, mxqcs@ieee.org, fanghaof@sfu.ca, yili.jin@mail.mcgill.ca,
watlas@wildsalmoncenter.org, jcliu@sfu.ca

Abstract

Real-time imaging sonar is crucial for underwater monitoring where optical sensing fails, but its use is limited by low uplink bandwidth and severe sonar-specific artifacts (speckle, motion blur, reverberation, acoustic shadows) affecting up to 98% of frames. We present ***SCOPE****, a self-supervised framework that* jointly *performs compression and artifact correction without clean–noise pairs or synthetic assumptions. SCOPE combines (i) Adaptive Codebook Compression (ACC), which learns frequency-encoded latent representations tailored to imaging sonar, with (ii) Frequency-Aware Multiscale Segmentation (FAMS), which decomposes frames into low-frequency structure and sparse high-frequency dynamics while suppressing rapidly fluctuating artifacts. A hedging training strategy further guides frequency-aware learning using low-pass proxy pairs generated without labels. Evaluated on months of in-situ ARIS sonar data, SCOPE achieves a structural similarity index (SSIM) of 0.77, representing a 40% improvement over prior self-supervised denoising baselines, at bitrates down to* ≤ 0.0118 *bpp. It reduces uplink bandwidth by more than 80% while improving downstream detection. The system runs in real time, with 3.1 ms encoding on an embedded GPU and 97 ms full multi-layer decoding on the server end. SCOPE has been deployed for months in three Pacific Northwest rivers to support real-time salmon enumeration and environmental monitoring in the wild. Results demonstrate that learning frequency-structured latents enables practical, low-bitrate sonar streaming with preserved signal details under real-world deployment conditions.*

1. Introduction

Traditional sonar operated at high power for longer range and wider coverage, but had limited resolution, making it difficult to identify small targets and resolve fine details of objects. In contrast, modern imaging sonars provide high-resolution imagery at lower power, enabling detailed observation of underwater scenes and real-time monitoring in conditions where optical and infrared sensors are unreliable [1, 2]. By transmitting acoustic pulses and reconstructing echoes into spatial imagery, sonar allows perception in turbid or low-light environments where cameras fail [3, 4]. Advances in hardware affordability and access to 5G and satellite connectivity have expanded sonar-based analytics in remote settings, including offshore rescue [3], subsea infrastructure inspection [4], and seasonal fishery management [5, 6], highlighting its societal value in safety, industrial support, and ecological monitoring [7].

Despite its growing importance, it remains difficult to reliably perform streaming imaging sonar in wild, uncontrolled environments. Limited infrastructure and unstable uplinks restrict transmission capacity [8], while the imagery itself is degraded by complex artifacts. For instance, in our field studies (see Suppl.), a single satellite uplink (e.g., Starlink) averaged 16.1 Mbps with a standard deviation of 5.7 and fluctuated between 1.1 and 39.9 Mbps. When shared across multiple observation points, the per-site bandwidth dropped to an average of 4.7 Mbps with a standard deviation of 1.9 and a range of 0 to 14 Mbps. This is well below the more than 24 Mbps required to stream compressed sonar video at full resolution (1146x2138@15 FPS), leaving a substantial portion of data untransmitted (see Suppl.). H.264/H.265 at the same CRF reach only $\sim$8.03 Mbps on standard video, showing imaging sonar data is harder to compress (see Suppl.), due to pixel-wise noise fluctuations [9]. Meanwhile, imaging sonar suffers from persistent speckle, motion blur, reverberations, and acoustic shadows [10, 11]. Motion blur affects 98% of frames [1, 6], while reverberations and shadows can obscure objects by up to 6× their actual size [5]. These conditions reduce perceptual quality, increase entropy, and hinder both compression efficiency and downstream analysis.

The combined effects of bandwidth limitations and image artifacts show that compression and correction cannot be treated separately. Artifacts increase entropy and reduce coding efficiency, while compression that is not

artifact-aware can obscure small or camouflaged targets. Existing research provides only partial answers. Deep learning–based image and video compression has advanced rapidly, with VAE-based models learning compact latent spaces [12–18], and diffusion or transformer-based approaches achieving strong rate–distortion trade-offs [19–22]. Yet these methods are tuned for natural imagery and often fail to preserve the small, sparse motions characteristic of sonar. In parallel, self-supervised denoising approaches, including blind-spot networks [23–26] and diffusion models [27, 28], reduce dependence on clean labels, but typically assume noise is independent or Gaussian. This prevents them from handling environment-dependent distortions such as motion blur, reverberation, and acoustic shadows. These limitations highlight the need for a sonar-specific approach that unifies compression and artifact correction within a single framework.

To address these challenges, we present **SCOPE**, a real-time framework for **S**elf-supervised **CO**mpression and artifact **CO**rrection in imaging sonar **P**rocessing and str**E**aming. SCOPE is built on a Variational Autoencoder (VAE) backbone that performs latent-space compression, but extends it with mechanisms tailored for imaging sonar. Standard VAEs can be inefficient because of large latent spaces, while generic codebook methods often collapse or fail to capture subtle patterns. We introduce *Adaptive Codebook Compression (ACC)* to learn frequency-encoded latent representations that stabilize training and preserve small, concealed objects. We further propose *Frequency-Aware Multiscale Segmentation (FAMS)* to separate low-frequency background structures from sparse high-frequency dynamics, suppressing artifacts in the process. Finally, a *hedging training strategy* leverages low-pass proxy pairs to guide frequency-aware learning without requiring clean–noise supervision. Together, these components enable SCOPE to jointly compress and correct sonar video, providing efficient low-bitrate transmission while retaining the details needed for reliable underwater monitoring.

SCOPE was evaluated on six months of ARIS sonar data collected from three Pacific Northwest rivers, providing a diverse and unbiased basis for testing. It was then deployed for an additional three months at the same sites (Fig. 1), supporting real-time salmon tracking, counting, and environmental monitoring. The framework achieved a structural similarity index (SSIM) of 0.77, representing a 40% improvement over prior self-supervised denoising methods. It compressed sonar video to ≤ 0.0118 bits per pixel, reducing uplink bandwidth by more than 80% while preserving image fidelity. These improvements also enhanced downstream detection and analysis. With 3.1 ms encoding on an embedded GPU and 97 ms full multi-layer decoding on the server, SCOPE operates in real time and demonstrates practical value under real-world deployment conditions.

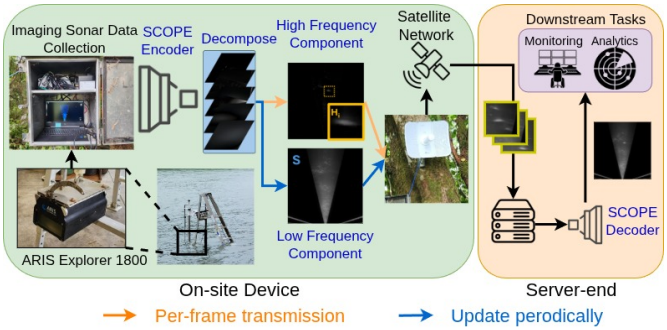


Figure 1. SCOPE is a self-supervised framework for joint compression and artifact correction of imaging sonar streaming.

The main contributions are summarized as follows:

- We develop and field-deploy SCOPE, a self-supervised framework that integrates compression and artifact correction to support real-time imaging sonar streaming under practical constraints.
- We propose Adaptive Codebook Compression (ACC) (Sec. 4.2) and Frequency-Aware Multiscale Segmentation (FAMS) (Sec. 4.3) to jointly compress sonar data and suppress artifacts while preserving small, concealed signals.
- We design a hedging training strategy (Sec. 4.4) that uses low-pass proxy pairs to guide frequency-aware learning without requiring clean–noise supervision.
- Extensive experiments and in-situ deployments (Sec. 5) demonstrate that SCOPE achieves efficient low-bitrate transmission, improves image fidelity, and enhances downstream performance.

2. Related Work

2.1. Speckle-Affected Imaging Framework

Change detection on speckle-contaminated images has been widely explored [5, 29, 30]. Xu et al. [5] used Mixture of Gaussians (MOG) [31] with guided filtering [32] to detect changes in sequential sonar frames, but pixel-wise modeling is computationally expensive, disrupts background consistency, and yields coarse, noisy outputs. Recent deep methods [29, 30] rely on pseudo labels from log-ratio images for CNN training, yet label noise introduces artifacts and hampers signal-preserving compression.

Despeckling remains challenging [33–37]. Li et al. [34] combined change detection and attention but required dual-polarization inputs and labeled maps. Zhao et al. [33] used contrastive and adversarial learning on unpaired data but struggled to generalize. Self-supervised methods show promise: MERLIN-Seg [35] still depends on labels, while Speckle2Void [36], based on Blind-Spot CNNs [26], oversimplifies speckle as uncorrelated noise.

2.2. Learned Image Compression (LIC)

Learned image compression (LIC) uses neural networks for efficient lossy compression, optimizing quality at low

bitrates. Ballé et al. [38–40] established the foundation for neural LIC. Recent methods include VAE-based models [12–18], diffusion models [19, 21], and transformers [20, 22, 41]. Guo et al. [21] introduced a diffusion model supporting flexible bitrate–quality trade-offs. VCT [41] tokenizes frames to capture spatio-temporal cues. ST-XCT [22] employs cross-covariance attention for improved coding. These designs improve performance but require heavy computation.

VAE-based CNNs remain dominant for latent modeling [42]. GLC [13] enhances perceptual quality via three-stage training in the VQ-VAE [43] latent space. STCM [17] captures temporal dynamics in remote sensing, while HybridFlow [14] reduces index map size without sacrificing quality. Bit Plane Slicing [15] combats posterior collapse [44] by emphasizing global features in hierarchical latents. Despite improved compression, these models involve multi-stage pipelines and high decoding cost, and often miss small or occluded signals—critical in sonar video analysis.

2.3. Self-Supervised Image Denoising

Self-supervised denoising has gained popularity due to scarce clean-noise pairs [23–27, 45–47]. Noise2Noise [24] trains on noisy pairs, while Noise2Void [26] and its extensions [45] use blind-spot networks (BSNs) to learn from single noisy inputs. However, these rely on ideal noise assumptions and degrade under real-world conditions. Neighbor2Neighbor [47] relaxes these assumptions via subsampling but still assumes IID noise, which fails for spatially correlated sonar noise [10, 11] (Fig. 2).

To handle non-IID noise, newer approaches have emerged. Diff-Unmix [27] uses spectral unmixing with diffusion models, but is computationally intensive and assumes Gaussian noise. LAN [46] aligns real-world noise with pretrained priors, while MASH [23] shuffles residuals, potentially introducing artifacts. AT-BSN [25] adjusts blind-spot size under weaker correlation assumptions—still ineffective for sonar data, where spatially dependent noise often obscures fine signals. In unpublished preprint work, SAVeD [48] exploits temporal cues from multiple frames to enhance low-SNR videos without clean labels. SCOPE instead integrates compression and artifact correction for real-time, low-bitrate analytics under deployment constraints.

3. Background and Motivations

Sonar devices have become increasingly affordable, from under $2,000 for handheld units [49] to $25,000 for advanced imaging sonar systems like ARIS [50], which are cheaper and higher-resolution than traditional sonar. Improved 5G and satellite connectivity enables broader real-time applications in remote areas, such as offshore rescue, disaster warnings, and fishery management. In turbid, low-light underwater environments where optical sensors fail,

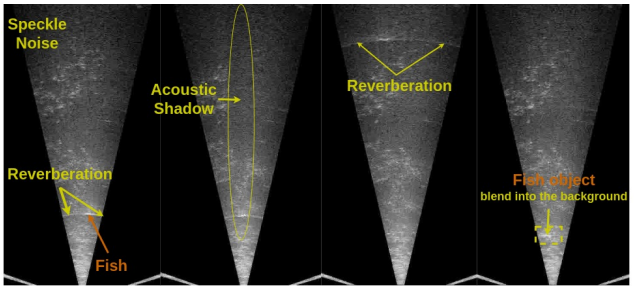


Figure 2. Artifacts in data from underwater imaging sonar.

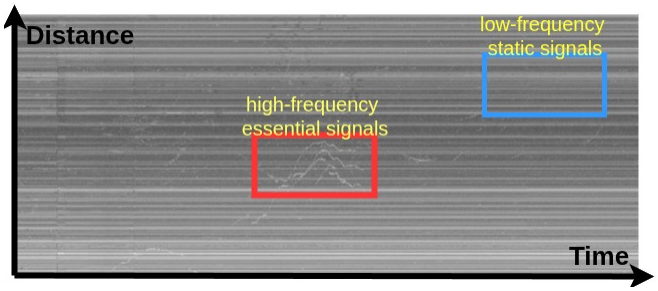


Figure 3. Echograms show high-frequency signals as dynamic patterns and low-frequency signals as consistent static regions.

sonar provides a reliable alternative for rescues and fishery conservation [5, 51], highlighting its social impact in emergency response and environmental monitoring.

Challenges from wild environments. Despite wider availability, deploying sonar in remote areas faces hurdles from limited infrastructure. Terrestrial networks cover only 15% of the Earth [52], while satellite links suffer from high latency (up to 600 ms) and unstable uplink bandwidth, especially in dense vegetation, mountains, or harsh weather [53]. These constraints hinder real-time imaging sonar transmission. E.g., an ARIS Explorer 3000 [54] at 1.8MHz outputs 1280×800@15 FPS, a downsample from higher native resolution used in previous deployments that degrades performance, still requiring $\geq$10.7 Mbps after codecs, whereas typical satellite uplinks provide only 4.7 Mbps (47% of the required bandwidth), making compression essential.

Artifacts of sonar data. Imaging sonar data from wild environments often contains artifacts such as speckle noise (high-frequency granular patterns), motion blur, acoustic shadows, and multipath reverberations (Fig. 2) that obscure object boundaries, reduce contrast, lower perceptual quality, increase entropy, hinder compression, and lower analytic accuracy in tasks such as detection, tracking, and segmentation. Effective correction is essential for maintaining signal fidelity and reliable sonar imagery.

4. Method

We develop SCOPE, a self-supervised real-time imaging sonar streaming framework that jointly compresses data and corrects artifacts without requiring clean-noise pairs or synthetic data, and is deployed across three rivers (Fig. 1).

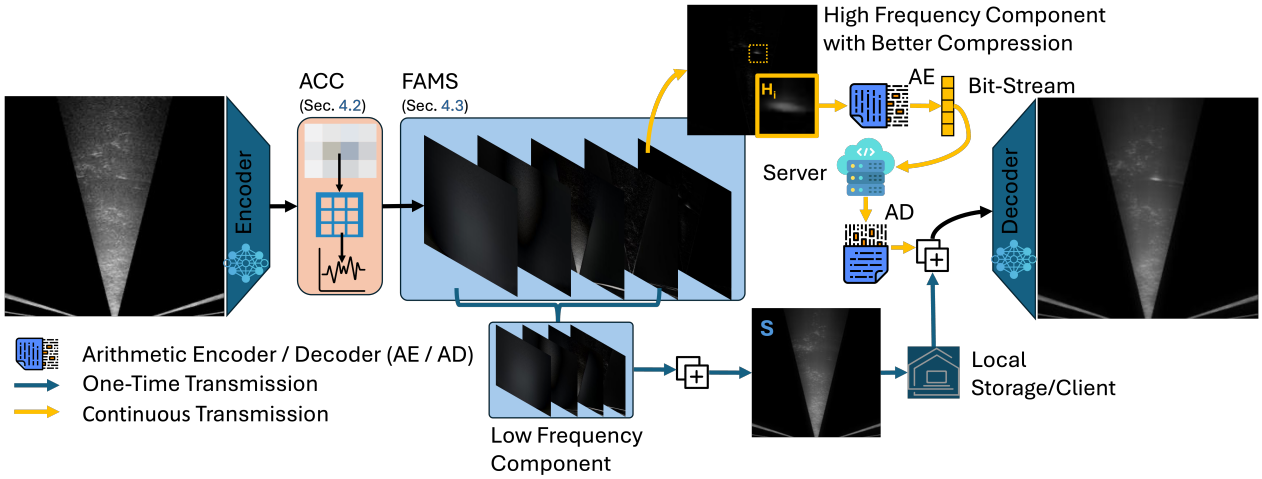


Figure 4. The overall underwater imaging sonar image compression and artifact correction architecture of SCOPE.

4.1. SCOPE Overview

SCOPE is motivated by temporal patterns observed in sonar echograms, where each vertical line represents a frame over time. As shown in Fig. 3, low-frequency static signals (blue box) dominate the scene, representing background and structured artifacts, while sparse high-frequency components (red box) capture dynamic targets such as marine life. The noise patterns in the echogram correspond to those in Fig. 2, indicating strong spatial but weak temporal correlation. This observation suggests that frequency-domain segmentation is well-suited for denoising. Each frame I_i is modeled as:

$$I_i = S + H_i + artifacts \tag{1}$$

As shown in Fig. 4, a noisy input I_i is encoded into a latent representation $\hat{I}_i$, which is processed by *Adaptive Codebook Compression (ACC)* and *Frequency-Aware Multiscale Segmentation (FAMS)*. ACC employs a VQ-VAE-based lossy LIC scheme to enhance representation learning of characteristic frequency components tailored for underwater imaging sonar data, producing k index maps for hierarchical sonar compression. FAMS performs soft frequency segmentation to isolate high-frequency components H_i (e.g., fish), generating k layers per frame by decoding the index maps k times (Fig. 4, upper blue region). Compression is achieved by transmitting only H_i for each frame, while the static low-frequency background S is transmitted weekly. At the receiver, S and H_i are recombined to reconstruct a clean, artifact-free frame. Training is guided by a self-supervised hedging strategy (Sec. 4.4).

4.2. Adaptive Codebook Compression

Learnable codebooks in VQVAE-based models [43, 55, 56] face challenges such as large codebook sizes and posterior collapse [44], where an overly powerful decoder bypasses

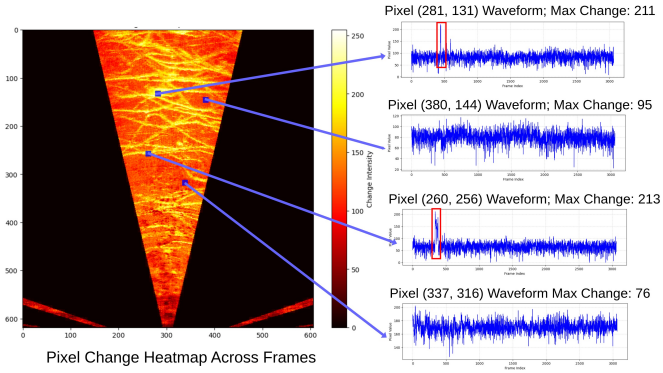


Figure 5. Pixel-wise temporal analysis of intensity fluctuations across imaging sonar frames.

the latent space, limiting the model to a few mapped features. A rich codebook is essential for high-quality reconstruction, as its entries B_k are derived from latent features $\hat{I}_i$. This work addresses two key questions: (1) How to preserve small, concealed objects blending into the background while avoiding collapse? (2) How to shrink the codebook to reduce bitrate?

Many signal representations rely on handcrafted transformations to extract frequency components, but such fixed bases often fail to capture diverse temporal dynamics. As shown in Fig. 5, sonar signals include distinct high-frequency essential components H_i (max change $\geq$ 200), low-frequency static components, and artifacts such as speckle noise, motion blur, and acoustic shadows (max change $\leq$ 100).

Our Adaptive Codebook Compression (ACC) departs from such fixed designs by introducing learnable codebook entries $\mathbf{b}_k$, jointly optimized via the hedging training strategy (Sec. 4.4) to capture diverse frequency patterns. The codebook $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_K]$ is trained to represent characteristic frequency components, enabling selective transmission of high-frequency H_i and low-frequency S, yielding a compact and efficient representation.

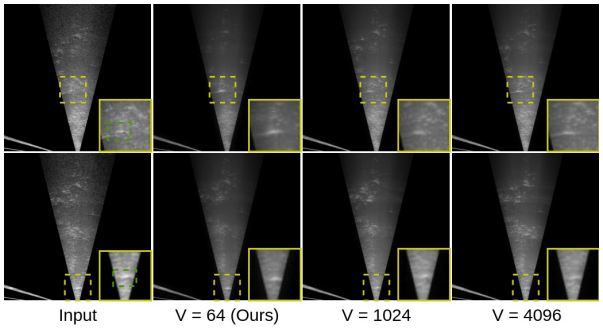


Figure 6. Quantitative comparisions on different codebook sizes $V \in \{64, 1024, 4096\}$. Smaller V leads to better noise suppression and clearer separation of the objects.

By exploiting limited frequency distributions and reducing entry size, ACC suppresses high-frequency artifacts with rapid fluctuations while preserving primary signal components (Fig. 5, red boxes), resulting in enhanced artifact suppression, improved compression efficiency, and better retention of small, background-blending objects (Fig. 6).

4.3. Frequency-Aware Multiscale Segmentation

Fig.5 shows a pixel-wise change heatmap across frames in real-world underwater imaging sonar video. Brighter areas (7.73% of total), from white to yellow, indicate high-frequency object motions like fish and prawns (red box in Fig.3), while darker areas, from red to black, correspond to low-frequency static structures (blue box). The 2nd and 4th waveforms represent artifacts such as speckle noise, motion blur, and acoustic shadows. Comparing these waveforms reveals clear distinctions between high-frequency signals H_i, low-frequency static signals S, and artifacts.

FAMS decomposes signals into high-frequency temporal components (red box in Fig. 5) and low-frequency structural signals while suppressing rapidly fluctuating artifacts. This leverages image scaling, which shifts spatial frequency components. The transform of an image $f(x, y)$ is given by:

$$F(u, v) = \int \int f(x, y) e^{-j2\pi(ux+vy)} \, dx \, dy \tag{2}$$

Scaling by factor s modifies the transform to:

$$F_s(u, v) = F\left(\frac{u}{s}, \frac{v}{s}\right) \tag{3}$$

This scaling suppresses high-frequency components and enhances low-frequency ones. Unlike natural images, sonar images exhibit distinct frequency signatures (Fig. 5), making this operation effective for segmentation. Learned entry vectors $\mathbf{b}_k$ then transform these components, mapping spatial frequencies $F_s(u, v)$ to temporal domain vectors:

$$f = \text{lookup}(B, F_s) \tag{4}$$

where B is the codebook. The clean output is reconstructed by decoding the aggregated frequency components:

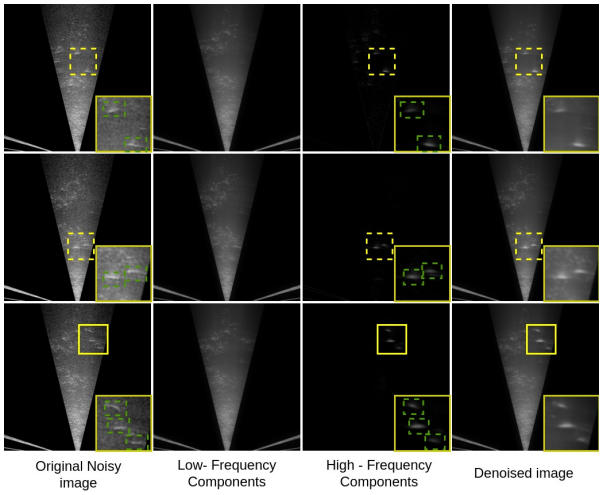


Figure 7. Effectiveness of our FAMS. The green boxes emphasize high-frequency dynamic signals (e.g., marine life).

$$f_{\text{denoised}}(x, y) = \text{Decoder}\left(\sum_k f_k\right) \tag{5}$$

Fig. 7 demonstrates that FAMS, guided by the hedging strategy (Sec. 4.4), preserves high-frequency details H_i (highlighted by red boxes in Fig. 5 and yellow boxes in Fig. 7) in the final layer, while isolating low-frequency structural details with reduced artifacts in earlier layers. This enables effective frequency segmentation and artifact reduction, preserving small, concealed objects blending into the background. The method requires no sequential or temporal input, ensuring efficient and robust inference.

4.4. Hedging Training Strategy

To train SCOPE without clean-noise pairs, synthetic data, or noise assumptions, we propose a self-supervised hedging strategy that uses *low-pass–filtered proxy pairs* to guide artifact correction and frequency-aware learning.

Low-pass proxy pair. As shown in Fig. 8(b)(c), we generate low-pass proxy pairs by first applying the Mixture of Gaussians (MOG) method for change detection to highlight dynamic elements (e.g., fish), followed by a guided filter acting as a low-pass filter. This process coarsely segments temporal frequency components, preserving low-frequency structures and edges while blurring static backgrounds, thereby emphasizing dynamic regions for segmentation guidance. This approach suits underwater imaging sonar data, where dynamic objects show higher temporal variance captured efficiently by MOG, and the guided filter preserves low-frequency spatial structures for robust feature learning. Combining temporal and spatial filtering provides a clear supervisory signal that emphasizes salient motion patterns while suppressing noise and static clutter, enabling effective learning of meaningful frequency representations.

However, low-pass proxy pairs are not directly applicable to artifact correction, reconstruction, or compression

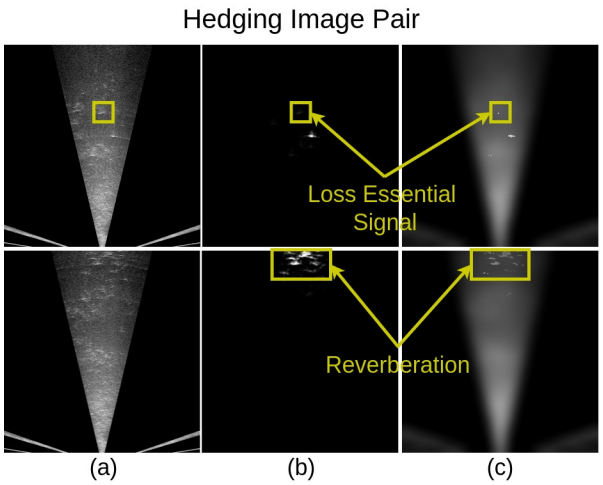


Figure 8. Low-pass proxy pair used in self-supervised learning, not directly applicable to denoising, reconstruction, or compression due to signal loss and reverberation.

in streaming sonar. First, low-frequency regions are often distorted. Second, MOG fails to preserve high-frequency signals from slow or static objects (Fig. 8, top). Third, strong artifacts can introduce reverberation (Fig. 8, bottom). Lastly, generating low-pass proxy pairs requires sequential input and incurs high time complexity due to pixel-wise processing, making real-time edge deployment impractical.

Hedging loss. To enable both artifact correction and frequency-awareness, we compute the loss between the clean output, the noisy input, and the low-pass proxy pair:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1(\hat{I}, I_{noisy}) + \lambda_2 \mathcal{L}_2(\hat{I}, I_{hedging})$$

Here, $\hat{I}$ is the clean output, I_{noisy} is the noisy input, and $I_{hedging}$ is the low-pass proxy pair. The first term, $\mathcal{L}_1(\hat{I}, I_{noisy})$, is a reconstruction loss that enforces structural consistency with the input to prevent over-smoothing and distortion. The second term, $\mathcal{L}_2(\hat{I}, I_{hedging})$, is a hedging loss that guides frequency-aware segmentation by leveraging the low-pass proxy pair, which highlights dynamic features and suppresses static background. This facilitates the removal of high-frequency artifacts and enables the decomposition of each frame into low-frequency structural components and high-frequency dynamic signals.

The weighting factors λ_1 and λ_2 balance the contributions of the two loss terms, ensuring accurate restoration and effectively guiding frequency-aware segmentation.

5. Experiment Results

5.1. Experimental Settings

System setup and data collection. We deployed SCOPE (Fig. 1; more in supplementary) and a continuous underwater monitoring system in three Pacific Northwest rivers to enable real-time salmon tracking, counting, and monitoring, using the ARIS Explorer 1800 and ARIS Explorer 1200 Sonar systems. A Jetson ORIN Nano served as edge device, while data transmission and remote control were managed via Starlink. The underwater monitoring system ran for six months, producing 197,037 images and 1,346 videos with 454,941 annotated bounding boxes labeled by trained technicians under expert and biologist guidance.

Implementation details. Our model employs the vanilla VQ-VAE architecture [43] with a shared codebook size $V = 64$ as described in Section 4.2. Instead of directly quantizing feature map values, the model is encouraged to learn frequency information through the hedging training strategy (Section 4.4). The proposed FAMS uses a hierarchical architecture inspired by VAR [56]. Specifically, both ACC and FAMS are learning-based methods without preset rules or thresholds. They are implemented by generating layers through multiple guided VQ-VAE runs, each run producing a separate layer (Fig. 4, blue region). For generating low-pass proxy pairs used in the loss computation (Section 4.4), we apply a guided filter [32], an edge-preserving smoothing method that uses a guidance image for filtering. Specifically, two hedging images are created by applying the guided filter to the noisy image with MOG-based pixel-wise change detection as guidance, and vice versa. In the hedging loss, we set $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$ to balance preserving structural similarity to the noisy input while effectively removing high-frequency artifacts and segmenting temporal frequency components into low-frequency (static background) and high-frequency (dynamic features).

Inference time. On average, encoding takes 3.06 ms on the edge, while total decoding time for all layers averages 97 ms on the server. Since only the last layer ($\sim$ 26 ms) is typically needed, inference speed meets real-time requirements and supports efficient practical use. The effectiveness of the system in real-world scenarios is further demonstrated through downstream tasks detailed in Section 5.4.

Evaluation. As shown in Fig. 1, we carried out an in-field deployment of the proposed SCOPE across three rivers of the Pacific Northwest. We use SSIM and bpp as primary metrics since our model performs both compression and denoising. Since the output involves reconstruction and denoising, the original noisy input is not a suitable reference. Moreover, noise-free frames are hard to obtain due to device limitations and uncontrollable water conditions like flow speed and suspended particles. Considering these two factors, calculating metrics that require reference images such as PSNR is unsuitable. To further assess fidelity and denoising, we rely on downstream evaluation (Section 5.4).

5.2. Evaluation of Artifact Correction

We target diverse artifact types rather than treating them as general noise, positioning denoising as a coarse correction step tailored to imaging sonar data. Our framework,

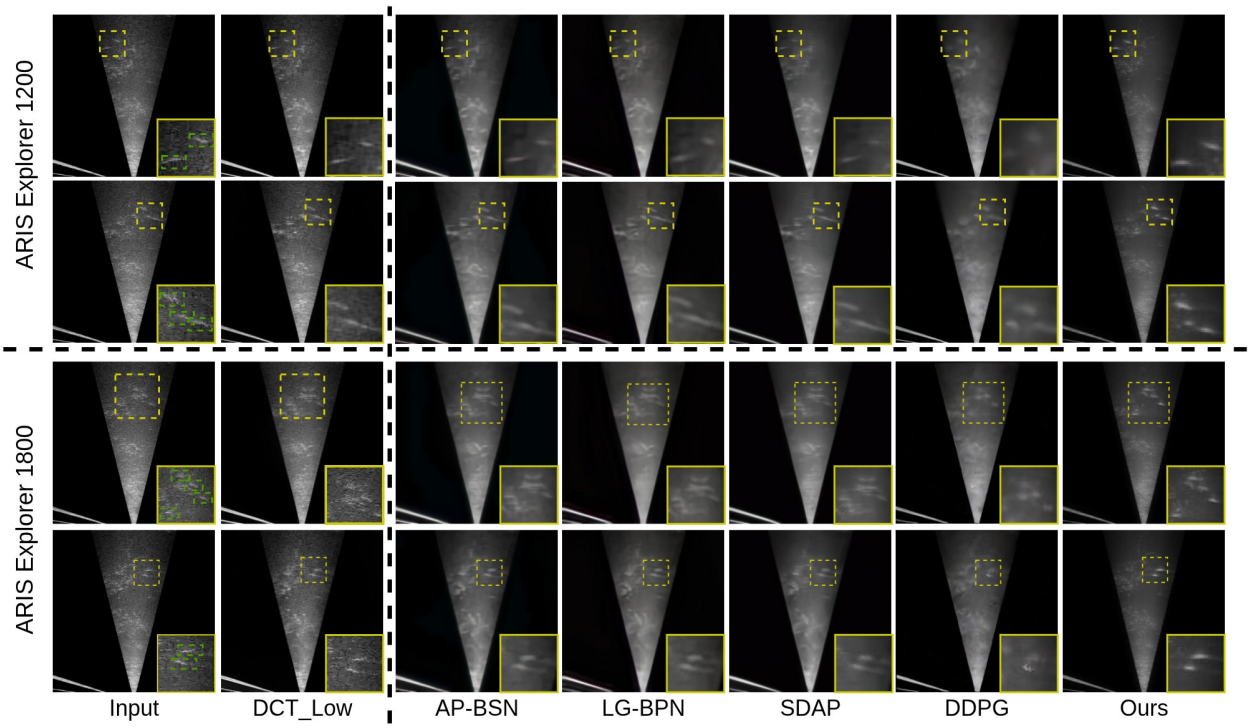


Figure 9. Visual comparison of self-supervised denoising methods on Underwater Imaging Sonar Dataset.

Table 1. Denoising methods comparison on real-world sonar data.

	SSIM [61] ↑	FSIM[62] ↑	BRISQUE [63] ↓
AP-BSN [57]	0.5484	0.8882	76.0703
LG-BPN [58]	0.5440	0.8936	72.8588
SDAP [59]	0.5478	0.9158	78.2003
DDPG [60]	0.5434	0.9168	81.5875
SCOPE(Ours)	**0.7711**	**0.9208**	**70.8516**

Table 2. Compression methods comparison on imaging sonar data.

	BPP ↓	SSIM [61] ↑	BRISQUE [63] ↓
H.264 / H.265 / AV1 (CRF:25)	0.7611 / 0.7396 / 0.6421	-	-
HybridFlow [14]	0.025	-	-
STCM [17]	0.08	-	-
MLIC+ [20]	0.0975	0.7337	69.6311
Continual-Compression [18]	0.05	0.7781	75.2767
SCOPE(Ours) (measured without entropy coding)	$\leq$ **0.0118**	0.7711	70.8516

SCOPE, is compared with state-of-the-art self-supervised denoising methods for real-world scenarios, including AP-BSN [57], LG-BPN [58], SDAP [59], and DDPG [60].

Quantitative measure. Table 1 reports SSIM [61] scores to assess image quality relative to the noisy input. SSIM considers luminance, contrast, and structure, aligning better with human perception than PSNR. As a self-supervised framework, SCOPE achieves the highest SSIM (0.7711), significantly outperforming AP-BSN (0.5484), LG-BPN (0.5440), SDAP (0.5478), and DDPG (0.5434), demonstrating superior structural preservation and visual quality.

Qualitative measure. Fig. 9 present qualitative comparisons in the image and frequency domains. DCT_Low is reconstructed by applying inverse DCT to the 60×60 low-frequency region of the original 600×600 DCT from the noisy input, highlighting the importance of low-frequency information (further discussed in Sec. 5.4). As shown in Fig. 9, SCOPE preserves fine details and more effectively recovers small, concealed objects blended into the background, which is critical for underwater sonar. Moreover, SCOPE suppresses high-frequency artifacts while retaining informative low-frequency signals (see Suppl.: Evaluation).

5.3. Evaluation of Compression

SCOPE is compared with existing VQ-VAE-based image compression methods, including HybridFlow [14], STCM [17], MLIC+ [20], and Continual-Compression [18]. HybridFlow reduces bitrate via a masked codebook, while STCM and MLIC+ focus on entropy optimization. We evaluate SCOPE by comparing its bits per pixel (bpp) against the lowest bpp achieved by these methods. Unlike standard compression benchmarks, our model jointly performs reconstruction and denoising, making PSNR relative to the original image inapplicable.

Table 2 reports the bpp performance. SCOPE achieves the lowest bpp (0.0118), measured without entropy coding. Compared with traditional codecs (H.264/H.265/AV1), it reduces bpp significantly from 0.64 to 0.0118. Since SCOPE outputs are predominantly dark, traditional codecs could further reduce redundancy, while SCOPE alone already cuts bitrate by at least 50% compared to HybridFlow and over 87% compared to MLIC+. This significant reduc-

Table 3. Downstream object detection results on underwater sonar comparing raw source, compression, and denoising methods.

Method	AP50	AP50-95	Precision	Recall	box loss
Raw Image	0.6199	0.21521	0.73917	0.5669	2.26966
DCT_Low	0.61761	0.21476	0.73045	0.57402	2.27095
MLIC+ [20]	0.57479	0.19126	0.69165	0.54103	2.33566
Continual-Compression [18]	0.57852	0.19575	0.70124	0.53802	2.32256
AP-BSN [57]	0.61382	0.21362	0.74366	0.56556	2.26588
LG-BPN [58]	0.61065	0.21177	0.73521	0.56488	2.28128
SDAP [59]	0.58852	0.20141	0.71599	0.54662	2.31268
SCOPE(Ours)	**0.62947**	**0.22232**	**0.75391**	**0.57675**	**2.24997**

tion demonstrates SCOPE's efficiency in preserving critical features while minimizing bandwidth, making it highly suitable for resource-constrained streaming scenarios.

5.4. Downstream task evaluation

Image compression and denoising aim to reduce transmission cost and enhance image quality for perception and downstream applications, such as object detection. However, lossy compression may negatively impact downstream performance. To assess fidelity, we use detection on the original image as a baseline and compare it with compression methods, denoising methods and DCT_Low (Sec. 5.2).

Using an 80/20 training/validation split, Table 3 shows that SCOPE achieves the highest AP50 (0.62947) and AP50-95 (0.22232), along with the best precision (0.75391), recall (0.57675), and lowest val/box loss (2.24997). DCT_Low performs similarly to raw images, emphasizing the importance of preserving low-frequency components, while high-frequency details are less critical. Other denoising models perform poorly on sonar data, highlighting their limitations underwater. Under strict bandwidth constraints (∼4.7 Mbps uplinks), SCOPE outperforms other compression approaches, achieving a 5% improvement with lower bpp.

5.5. Ablation Study

Adaptive Codebook Compression (ACC). We experiment with different codebook sizes $V \in \{64, 1024, 4096\}$ to show that reducing the codebook size not only achieves extremely low bitrates, but also encourages the codebook entries $\mathbf{b}_k$ to learn frequency components from temporal and spatial structures, rather than directly mapping feature values. As shown in Fig. 6, smaller V leads to better artifacts suppression and clearer separation of the objects.

Frequency-Aware Multiscale Segmentation (FAMS). The effect of proposed FAMS is illustrated in Figs. 4 and 7. It suppresses high-frequency artifacts while segmenting low-frequency background structures and high-frequency object signals (e.g., fish). This method extends beyond imaging sonar, showing potential applicability in imaging domains such as MRI and remote sensing. In MRI, it could contribute to improved separation of tissue types. In remote sensing, it could contribute to more accurate change detection and geospatial monitoring.

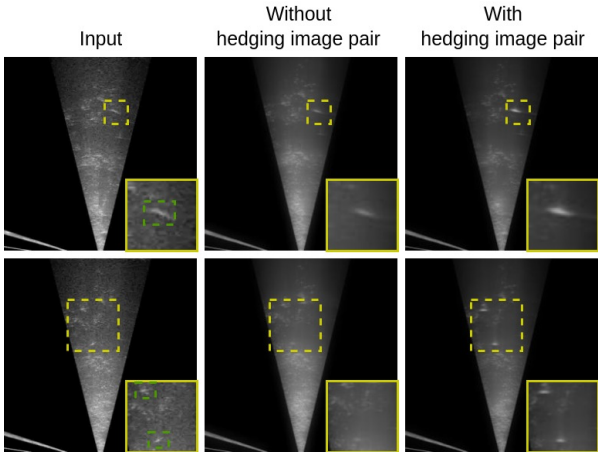


Figure 10. Effectiveness of our Hedging Training Strategy.

Hedging Training Strategy. We evaluate training with and without low-pass proxy pairs combined with different loss functions. Fig. 10 shows that the hedging strategy improves results and visual perception. Without hedging, small, concealed objects tend to blend into the background, causing erosion (row 1) and distortion (row 2) as seen in Fig. 10.

6. Conclusion

In this paper, we presented a self-supervised SCOPE framework for underwater imaging sonar compression and artifact correction. By integrating ACC and FAMS with hedging training, it suppressed artifacts while preserving critical high-frequency signals. Experiments showed superior compression (≤ 0.0118 bpp) suitable for 4.7 Mbps uplinks and artifact correction (SSIM: 0.7711), which improved downstream task accuracy. With 3.1 ms encoding and 97 ms total decoding for all layers, SCOPE enables real-time processing and has been deployed at three Pacific Northwest river sites for salmon and environmental monitoring, enabling practical sonar streaming. The method also shows promise for other imaging fields such as MRI and remote sensing.

Acknowledgment

This research is supported by an NSERC Discovery Grant, a British Columbia Salmon Recovery and Innovation Fund (BCSRIF_2022_401), and a Mitacs Accelerate Cluster Grant. It also received additional funding from Experiment.com. Chi Xu's work is supported by an NSERC Canada Graduate Scholarship–Doctoral (CGS-D) and a Mitacs Globalink Research Award. We are grateful for the trust and collaboration of the Pacific Salmon Foundation, Wild Salmon Center, the Heiltsuk, Haida, Kitasoo Xai'xais, Taku River Tlingit, and Gitga'at First Nations, as well as the Skeena Fisheries Commission and Gitanyow Fisheries Authority for their ongoing partnership in this work.

References

[1] R. E. Jones, R. A. Griffin, and R. K. Unsworth, "Adaptive resolution imaging sonar (aris) as a tool for marine fish identification," *Fisheries Research*, vol. 243, p. 106092, 2021. 1

[2] Y. Wei, Y. Duan, and D. An, "Monitoring fish using imaging sonar: Capacity, challenges and future perspective," *Fish and Fisheries*, vol. 23, no. 6, pp. 1347–1370, 2022. 1

[3] S. Hu and T. Liu, "Underwater rescue target detection based on acoustic images," *Sensors*, vol. 24, no. 6, p. 1780, 2024. 1

[4] B. Shi, T. Cao, Q. Ge, Y. Lin, and Z. Wang, "Sonar image intelligent processing in seabed pipeline detection: review and application," *Measurement Science and Technology*, vol. 35, no. 4, p. 045405, 2024. 1

[5] C. Xu, R. Qian, H. Fang, X. Ma, W. I. Atlas, J. Liu, and M. A. Spoljaric, "Salina: Towards sustainable live sonar analytics in wild ecosystems," in *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 68–81. [Online]. Available: https://doi.org/10.1145/3666025.3699323 1, 2, 3

[6] J. Kay, P. Kulits, S. Stathatos, S. Deng, E. Young, S. Beery, G. V. Horn, and P. Perona, "The caltech fish counting dataset: A benchmark for multiple-object tracking and counting," 2022. [Online]. Available: https://arxiv.org/abs/2207.09295 1

[7] M. Jahanbakht, W. Xiang, L. Hanzo, and M. R. Azghadi, "Internet of underwater things and big marine data analytics—a comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 904–956, 2021. 1

[8] S. Ma, Y. C. Chou, H. Zhao, L. Chen, X. Ma, and J. Liu, "Network characteristics of leo satellite constellations: A starlink-based measurement from end users," in *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, 2023, pp. 1–10. 1

[9] C. Kwan, J. Larkin, B. Budavari, E. Shang, and T. Tran, "Perceptually lossless compression with error concealment for periscope and sonar videos," *Signal & Image Processing An International Journal*, 04 2019. 1

[10] Y. Yu, J. Zhao, C. Huang, and X. Zhao, "Treat noise as domain shift: Noise feature disentanglement for underwater perception and maritime surveys in side-scan sonar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023. 1, 3

[11] T. Yang, T. Zhang, and Y. Yao, "Simnfnd: A forward-looking sonar denoising model trained on simulated noise-free and noisy data," *Remote Sensing*, vol. 16, no. 15, 2024. [Online]. Available: https://www.mdpi.com/2072-4292/16/15/2815 1, 3

[12] F. Zhou, X. Huang, P. Zhang, M. Wang, Z. Wang, Y. Zhou, and H. Yin, "Enhanced screen content image compression: A synergistic approach for structural fidelity and text integrity preservation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 7900–7908. [Online]. Available: https://doi.org/10.1145/3664647.3681354 2, 3

[13] Z. Jia, J. Li, B. Li, H. Li, and Y. Lu, "Generative latent coding for ultra-low bitrate image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 26 088–26 098. 3

[14] L. Lu, Y. Xie, W. Jiang, W. Wang, X. Lin, and Y. Wang, "Hybridflow: Infusing continuity into masked codebook for extreme low-bitrate image compression," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 3010–3018. [Online]. Available: https://doi.org/10.1145/3664647.3681008 3, 7

[15] Z. Zhang, H. Wang, Z. Chen, and S. Liu, "Learned lossless image compression based on bit plane slicing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 27 579–27 588. 3

[16] Y. Mao, P. Chen, S. Wang, S. Wang, and D. Wu, "Peering into the sketch: Ultra-low bitrate face compression for joint human and machine perception," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 2564–2572. [Online]. Available: https://doi.org/10.1145/3581783.3613799

[17] J. Zhang, R. Dong, J. Zheng, M. Chen, L. Zhang, Y. Zhao, and H. Fu, "Spatial-temporal context model for remote sensing imagery compression," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 6063–6072. [Online]. Available: https://doi.org/10.1145/3664647.3681242 3, 7

[18] Z. Duan, M. Lu, J. Yang, J. He, Z. Ma, and F. Zhu, "Towards backward-compatible continual learning of image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 25 564–25 573. 2, 3, 7, 8

[19] H. Kuang, Y. Ma, W. Yang, Z. Guo, and J. Liu, "Consistency guided diffusion model with neural syntax for perceptual image compression," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 1622–1631. [Online]. Available: https://doi.org/10.1145/3664647.3681336 2, 3

[20] W. Jiang, J. Yang, Y. Zhai, P. Ning, F. Gao, and R. Wang, "Mlic: Multi-reference entropy model for learned image compression," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. ACM, Oct. 2023, p. 7618–7627. [Online]. Available: http://dx.doi.org/10.1145/3581783.3611694 3, 7, 8

[21] S. Guo, Z. Chen, Y. Zhao, N. Zhang, X. Li, and L. Duan, "Toward scalable image feature compression: A content-adaptive and diffusion-based approach," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. ACM, Oct. 2023, p. 1431–1442. [Online]. Available: http://dx.doi.org/10.1145/3581783.3611851 3

[22] Z. Chen, L. Relic, R. Azevedo, Y. Zhang, M. Gross, D. Xu, L. Zhou, and C. Schroers, "Neural video compression with spatio-temporal cross-covariance transformers," in *Proceedings of the 31st ACM International Conference on Multi-*

media, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 8543–8551. [Online]. Available: https://doi.org/10.1145/3581783.3611960 2, 3

[23] H. Chihaoui and P. Favaro, "Masked and shuffled blind spot denoising for real-world images," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 3025–3034. 2, 3

[24] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2noise: Learning image restoration without clean data," 2018. [Online]. Available: https://arxiv.org/abs/1803.04189 3

[25] S. Chen, J. Zhang, Z. Yu, and T. Huang, "Exploring efficient asymmetric blind-spots for self-supervised denoising in real-world scenarios," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 2814–2823. 3

[26] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2void - learning denoising from single noisy images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2124–2132. 2, 3

[27] H. Zeng, J. Cao, K. Zhang, Y. Chen, H. Luong, and W. Philips, "Unmixing diffusion for self-supervised hyperspectral image denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 27 820–27 830. 2, 3

[28] J. Cheng, D. Liang, and S. Tan, "Transfer clip for generalizable image denoising," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 25 974–25 984. 2

[29] M. Gong, H. Yang, and P. Zhang, "Feature learning and change feature classification based on deep learning for ternary change detection in sar images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 129, pp. 212–225, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271616301459 2

[30] J. Geng, X. Ma, X. Zhou, and H. Wang, "Saliency-guided deep neural networks for sar image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7365–7377, 2019. 2

[31] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 2, 1999, pp. 246–252 Vol. 2. 2

[32] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013. 2, 6

[33] B. Zhao, Q. Zhou, L. Huang, and Q. Zhang, "Unpaired sonar image denoising with simultaneous contrastive learning," *Computer Vision and Image Understanding*, vol. 235, p. 103783, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314223001637 2

[34] J. Li, S. Shi, L. Lin, Q. Yuan, H. Shen, and L. Zhang, "A multi-task learning framework for dual-polarization sar imagery despeckling in temporal change detection scenarios," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 221, pp. 155–178, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271625000358 2

[35] E. Dalsasso, C. Rambour, N. Trouvé, and N. Thome, "Merlin-seg: Self-supervised despeckling for label-efficient semantic segmentation," *Computer Vision and Image Understanding*, vol. 241, p. 103940, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314224000213 2

[36] A. B. Molini, D. Valsesia, G. Fracastoro, and E. Magli, "Speckle2void: Deep self-supervised sar despeckling with blind-spot convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022. 2

[37] H. Lin, K. Jin, J. Yin, J. Yang, T. Zhang, F. Xu, and Y.-Q. Jin, "Residual in residual scaling networks for polarimetric sar image despeckling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023. 2

[38] J. Ballé, V. Laparra, and E. P. Simoncelli, "Density modeling of images using a generalized normalization transformation," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2016. 3

[39] ——, "End-to-end optimized image compression," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2017.

[40] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2018. 3

[41] F. Mentzer, G. Toderici, D. Minnen, S. J. Hwang, S. Caelles, M. Lucic, and E. Agustsson, "Vct: a video compression transformer," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022. 3

[42] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online]. Available: https://arxiv.org/abs/1312.6114 3

[43] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6309–6318. 3, 4, 6

[44] J. Lucas, G. Tucker, R. B. Grosse, and M. Norouzi, "Don't blame the elbo! a linear vae perspective on posterior collapse," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 3, 4

[45] S. Laine, T. Karras, J. Lehtinen, and T. Aila, "High-quality self-supervised deep image denoising," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/2119b8d43eafcf353e07d7cb5554170b-Paper.pdf 3

[46] C. Kim, T. H. Kim, and S. Baik, "Lan: Learning to adapt noise for image denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 193–25 202. 3

[47] T. Huang, S. Li, X. Jia, H. Lu, and J. Liu, "Neighbor2neighbor: Self-supervised denoising from single noisy

images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 781–14 790. 3

[48] S. Stathatos, M. Hobley, M. Marks, and P. Perona, "Saved: Learning to denoise low-snr video for improved downstream performance," 2025. [Online]. Available: https://arxiv.org/abs/2504.00161 3

[49] Field and Stream, "Best side imaging fish finders," 2025, https://www.fieldandstream.com/gear/best-side-imaging-fish-finders/. 3

[50] J. Joslin, "Imaging sonar review for marine environmental monitoring around tidal turbines," *Report by University of Washington. Report for Offshore Energy Research Association of Nova Scotia (OERA)*, 2019. 3

[51] B. Bulletin, "A.i. search & rescue sonar device comes to ocean city fire dept." https://www.chesapeakebaymagazine.com/a-i-search-rescue-sonar-device-comes-to-ocean-city-fire-dept/, 2024. 3

[52] T. IoT, "The evolution of non-terrestrial networks," https://talkingiot.io/the-evolution-of-non-terrestrial-networks-ntns/, 2025. 3

[53] M. Zhang, J. Li, H. Zhao, L. Shen, and J. Liu, "Starstream: Live video analytics over space networking," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 7909–7917. [Online]. Available: https://doi.org/10.1145/3664647.3680785 3

[54] S. Metrics, "Aris explorer 3000 specifications," http://www.soundmetrics.com/products/aris-sonars/aris-explorer-3000/, 2025. 3

[55] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 868–12 878. 4

[56] K. Tian, Y. Jiang, Z. Yuan, B. PENG, and L. Wang, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 84 839–84 865. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/9a24e284b187f662681440ba15c416fb-Paper-Conference.pdf 4, 6

[57] W. Lee, S. Son, and K. M. Lee, "Ap-bsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 704–17 713. 7, 8

[58] Z. Wang, Y. Fu, J. Liu, and Y. Zhang, "Lg-bpn: Local and global blind-patch network for self-supervised real-world denoising," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18 156–18 165. 7, 8

[59] Y. Pan, X. Liu, X. Liao, Y. Cao, and C. Ren, "Random subsamples generation for self-supervised real image denoising," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 12 116–12 125. 7, 8

[60] T. Garber and T. Tirer, "Image restoration by denoising diffusion models with iteratively preconditioned guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 25 245–25 254. 7

[61] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. 7

[62] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011. 7

[63] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012. 7