Summarise and Identify Large Datasets using Latent Representation UDT 2026

Nisa Andersson, Systems Engineer, Saab, Linköping, Sweden

This work investigates a framework for summarising large-scale seafloor imagery to support efficient survey analysis and increase situational awareness during survey operations. The foundation of the framework focuses on using a self-supervised convolutional autoencoder to learn a compact latent representation of the images. The latent representations enable clustering of images with similar structural characteristics, organising the dataset into coherent groups from which representative overviews can be generated. This reduces redundancy and the number of images requiring manual inspection. The same representation supports similarity-based retrieval through closest search, enabling identification of related imagery. Previous work demonstrated the feasibility of this framework using optical imagery. Here, the approach is extended to Synthetic Aperture Sonar (SAS) imagery, which is based on coherent acoustic backscatter and presents different imaging characteristics. The experimental results show that clustering the latent representations produces groups of structurally similar images, indicating that the embedding space captures relevant feature structure for dataset organisation. Meanwhile, similar image retrieval in the same latent space demonstrates limited effectiveness and requires additional refinement and systematic evaluation to achieve better results.

1 Introduction

Recent multinational exercises Robotic Experimentation and Prototyping Maritime Unmanned Systems (REPMUS) have demonstrated the ability of modern AUV platforms, including the Saab AUV62-MR, to conduct large-area, high-resolution seabed surveys with reliable navigation accuracy. In particular, advances in Synthetic Aperture Sonar (SAS) enable centimetre-scale seabed imaging across long mission durations. However, the data generated during the mission often exceeds the rate at which it can be transmitted through available communication links. Consequently, the datasets are typically retrieved only after the vehicle has been recovered. This creates a bottleneck that restricts timely feedback to the operators and limits the operator's ability to detect, assess, re-task and respond to time-critical events during mission execution. Even when real-time transmission is not crucial, post-mission still imposes a large burden on operators, who must manually review and interpret large volumes of imagery.

One approach to mitigating this challenge is to deploy a supervised classification model onboard the vehicle [1, 2, 3] to perform detection and classify. Such an approach can reduce data volume by transmitting only detected events and improve autonomy. However, their performance often degrades under environmental variability, including changes in seabed composition [4, 5] or gazing angle. Achieving robust generalization requires extensive and diverse labelled datasets, which are costly to obtain and rarely cover all operational scenarios.

An alternative approach is to change the objective from categorical prediction to representation learning. Rather than committing to predefined labels during deployment, the system learns a compact latent embedding that captures the structure of the imagery. Thornton et al. [6] demonstrated that seafloor imagery can be encoded into compact latent embeddings using a learned encoder while preserving spatial relationships. Rather than transmitting full imagery collections or discrete classification outputs, the framework groups visually similar observations in latent space and transmits only the representative latent embeddings along with selected exemplar images for each cluster. This approach separates data collection from final interpretation, allowing imagery to be analysed and labelled after the mission instead of committing to predefined categories during the mission. Moreover, it embeds large mission datasets into a structured latent space that supports distance-based retrieval and cluster-based summarisation. Clustering within this space groups similar observations, enabling operators to examine representative exemplars from each cluster instead of conducting an exhaustive sequential review.

This study extends the framework presented by Thornton et al. to Synthetic Aperture Sonar (SAS) imagery, a sensing modality that differs from optical seafloor data in its underlying characteristics. Given this domain shift, this study qualitatively examines whether an autoencoder-based latent representation preserves sufficient structural information to support unsupervised clustering and similarity-image retrieval for efficient operator analysis.

2 Experiment

Based on the proposed framework [6], this study consists mainly of four parts.

First, a georeferenced convolutional autoencoder is trained using SAS imagery collected during REPMUS 2025. The objective of this stage is to learn a compact latent representation that preserves both structural characteristics of the imagery and geographic continuity between spatially proximate observations.

Second, the dataset is embedded into the latent space, transforming high-dimensional image data into a reduced-dimensional representation.

Third, unsupervised clustering is applied within the latent space to group structurally similar images. This enables cluster-based dataset summarisation, where representative exemplars can be selected from each cluster to reduce redundancy and support potential bandwidth-constrained transmission and the operators for post-processing.

Finally, similarity-based retrieval is performed using distance metrics defined over the latent embeddings. This step evaluates whether the learned representation preserves meaningful neighbourhood structure, allowing query images to retrieve related observations for operator-driven analysis.

2.1 Representation Learning Objective

The first stage of the framework is to learn the latent representations of the images, in which the distance between the representations will serve as a quantitative measure of similarity.

To achieve this, a convolutional autoencoder is trained to map each SAS image into a latent embedding. An autoencoder is a type of artificial neural network that consists of two parts, an encoder and a decoder. The encoder projects the input image into a latent vector, and the decoder reconstructs the image from this representation. The encoder–decoder is trained by minimising the difference between the input and the reconstructed image. This reconstruction-driven optimisation constrains the encoder to preserve the information required to reproduce each image. Consequently, images with similar visual characteristics tend to produce latent embeddings that are close in the latent space.

Formally, let $x_i \in R^{H \times W}$ denote a SAS image acquired at geographic location $p_i \in R^2$. Given a dataset $\{(x_i, p_i)\}_{i=1}^{N}$, the objective is to learn a mapping:

$$\boldsymbol{z_i} = f_\theta(x_i), \quad \boldsymbol{z_i} \in R^d$$

such that images that are similar in appearance, and geographically proximate, are mapped to nearby points in latent space, i.e., $|z_i - z_j|$ small when $x_i \approx x_j$ or $|p_i - p_j|$ small. The desired properties of the latent space must be enforced through the training objective. In practice, this is achieved by defining a loss function; the following subsection formulates this optimisation objective.

2.1.1 Learning Objective

To implement the objective mentioned earlier, the autoencoder is trained under a composite loss function designed to enforce both reconstruction accuracy and spatial consistency in the latent space. While standard autoencoders are trained solely to minimise reconstruction error, the desired embedding in this study must also preserve geographic relationships between observations.

Seafloor geological and ecological features, such as sand ripples, sediment transitions, and pipelines, typically extend over spatial scales larger than a single image. Consequently, neighbouring images often capture related or continuous seabed characteristics. To incorporate this spatial continuity into the learned representation, a georeferenced regularisation term is introduced into the loss function. This term encourages embeddings of geographically proximate images to remain close in latent space, thereby aligning spatial closeness in the physical space, where the image was taken, with neighbourhood relationships in embedding space.

Prior studies have shown that incorporating geolocation-based regularisation improves the structural organisation of the learned embedding compared to reconstruction-only training. Building on this principle, the learning objective (loss function) is defined as:

$$L_{total} = L_{rec} + \lambda L_{geo} = L_{rec} + \lambda KL(P' || Q')$$

where L_{rec} denotes the reconstruction loss, L_{geo} denotes the georeferenced regularisation term, and λ is a hyperparameter balancing between L_{rec} and L_{geo}.

Reconstruction Term

The reconstruction loss L_{rec} is defined as the mean absolute error between the input image x_i and its reconstruction $\hat{x_i}$:

$$L_{rec} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \hat{x_i}|$$

Georeferenced Regularisation Term

The georeferenced loss L_{rec} is defined as the Kullback–Leibler (KL) divergence between two normalised affinity matrices P' and Q'. P' presents pairwise similarity in geographic space, and Q'represents pairwise similarity in latent space. The geographic affinity matrix $P'$is defined as:

$$p'_{ij} = \frac{\left(1 + d(\boldsymbol{p_i}, \boldsymbol{p_j})\right)^{-1}}{\sum_{i'}\sum_{j'}\left(1 + d(\boldsymbol{p_{i'}}, \boldsymbol{p_{j'}})\right)^{-1}}$$

where $d(\boldsymbol{p_i}, \boldsymbol{p_j})$ denotes the geographic distance between coordinates $\boldsymbol{p_i}$ and $\boldsymbol{p_j}$. To incorporate spatial locality constraints:

$$d(\boldsymbol{p_i}, \boldsymbol{p_j}) = min\left(\|\boldsymbol{p_i}, \boldsymbol{p_j}\|_2^2, d_{max}^2\right)$$

where d_{max}^2 limits the influence of samples far away and prevents over-regularisation across unrelated geographic regions.

Latent Affinity Matrix

The latent affinity matrix Q' is defined as the pairwise distances of the latent embeddings:

$$q'_{ij} = \frac{\left(1 + \|\boldsymbol{z_i} - \boldsymbol{z_j}\|_2^2\right)^{-1}}{\sum_{i'}\sum_{j'}\left(1 + \|\boldsymbol{z_{i'}} - \boldsymbol{z_{j'}}\|_2^2\right)^{-1}}$$

where $\boldsymbol{z_i}$ and $\boldsymbol{z_j}$ denote latent representations. Both P' and Q' are normalised such that their entries sum to one, which is required for computing KL.

2.1.2 Proximity-Aware Batch

The effectiveness of the georeferenced regularisation depends on the availability of meaningful geographic relationships within each mini-batch. Since the affinity matrices are computed only from samples within each mini-batch, the geographic relationships inside a batch must contain both nearby and distant images. If samples are drawn randomly from across the survey area, most image pairs will be geographically far apart. In this case, the computed geographic affinities become similar for most pairs, providing little distinction between neighbouring and distant samples. As a result, the KL regularisation has limited ability to enforce meaningful spatial structure that

extends across multiple images in the latent space.

To address this, a proximity-aware sampling strategy was used. For each mini-batch, three anchor images are randomly selected. Around each anchor, approximately 65 geographically close samples are drawn using precomputed spatial metadata. The remaining batch positions are filled with randomly selected samples from the dataset. The batch size is 256.

This multi-anchor construction ensures that each batch contains localised groups of spatially adjacent images as well as more distant samples. Consequently, the resulting geographic affinity matrix P' shows meaningful variation, allowing the KL regularisation term to enforce alignment between physical proximity and distances in latent space.

2.1.3 Training Configuration

The encoder is based on a ResNet-50 architecture without pretrained weights, with the final classification layer removed. The decoder consists of a symmetric upsampling architecture implemented using transposed convolutional layers with batch normalisation to reconstruct the original input

Input images are normalised prior to training. To improve generalization and reduce orientation bias, random rotations and horizontal/vertical flips are applied with probability 0.5.

The geographic distance threshold d^2_{max} is set to 8 meters. The latent dimension was set to 16, to balance compression and reconstruction accuracy based on preliminary empirical tuning [7]. The balancing parameter λ is set to 1000 based on empirical tuning. The autoencoder is trained for 2650 epochs using the PyTorch framework on an NVIDIA L40S GPU.

2.2 Clustering

Clustering is used to examine whether similar SAS images are naturally grouped in the learned latent space. Since distances in latent space are intended to reflect similarity between SAS images, samples that share similar visual characteristics should naturally form groups in latent space.

Various clustering paradigms exist, including centroid-based, density-based, and hierarchical approaches. In this study, a centroid-based method, k-means clustering, is employed due to its simplicity.

The algorithm partitions N latent representations $\{z_i\}_{i=1}^N$ into k clusters by minimising within-cluster variance. The number of clusters k is a user-defined parameter. In this study, k=8 is selected to provide an interpretable grouping for qualitative analysis rather than to achieve optimal partitioning. The objective is structural exploration of the embedding space. The idea is that cluster centroids or representative samples from each cluster can subsequently be used for dataset summarisation and bandwidth-efficient transmission.

2.3 Similarity-Based Image Retrieval

In addition to clustering, similarity-based image retrieval is used to assess whether local relationships in the latent space reflect similarity between SAS images. Given a query image, retrieval is performed by comparing its latent embedding to all other embeddings in the dataset using a distance-based similarity metric.

Let z_q denote the embedding of a query image and $\{z_i\}_{i=1}^N$ the set of latent representations. Images are ranked according to a similarity measure computed between z_q and z_i. Common similarity measures include Euclidean distance and cosine similarity.

In this study, cosine similarity is used, defined as:

$$\mathrm{sim}(z_q, z_i) = \frac{z_q^\top z_i}{|z_q|_2 \, |z_i|_2}$$

Cosine similarity is chosen because it measures angular similarity between embeddings, making it less sensitive to differences in embedding magnitude and more focused on directional alignment in latent space.

3 Results and Discussions

This section presents the results corresponding to the three objectives of the framework: learning a spatially coherent latent representation, organising the dataset through clustering, and enabling similarity-based image retrieval.

3.1 Representation Learning

This subsection evaluates whether the learned latent representation preserves sufficient information for accurate image reconstruction and maintains geographic consistency by embedding spatially proximate images close together.

Learning Objective

Figure 1 shows the total loss over 2650 training epochs. Both reconstruction and the georeferenced loss decrease mostly during the initial phase of training. After this stage, all loss components stabilise and exhibit smooth convergence without oscillatory behaviour.

The total loss closely follows the reconstruction loss after approximately 500 epochs, indicating that the reconstruction term dominates the magnitude of the training objective during later training. The georeferenced loss decreases sharply in the early epochs and subsequently stabilises at a small value, suggesting that the alignment between geographic and latent affinities is established early in training and maintained thereafter.

No evidence of conflict between the two objectives is observed. In particular, the reconstruction loss does not increase when the georeferenced loss decreases, which would indicate opposing gradients. Instead, both terms decrease jointly, implying stable multi-objective optimisation under the selected value of λ.

The relatively small magnitude of the georeferenced loss at convergence suggests that either the geographic and latent affinity distributions are already closely aligned, maybe because the number of neighbours in section 2.1.2 was too high.

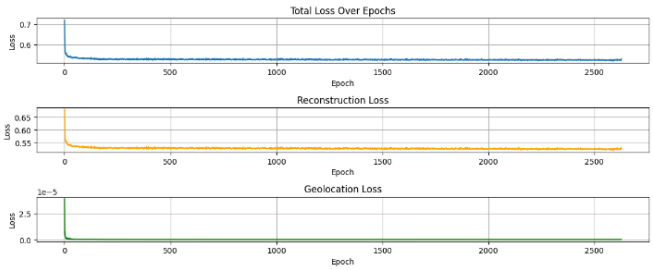

Figure 1 Total loss and loss components during autoencoder training

Affinity Matrix Analysis

To evaluate whether geographic relationships are preserved in the latent space, geographic and latent affinity matrices were computed and examined. Figure 2 and Figure 4 show the sampled images from two different mini-batches (left panels) and their corresponding geographic distributions (right panels), constructed using the proximity-aware sampling strategy described in 2.1.2.

For each batch, two affinity matrices are computed. The geographic affinity matrix is derived from pairwise distances between image coordinates, while the latent affinity matrix is computed from pairwise distances between the corresponding embeddings obtained from the encoder. Figure 3 and Figure 5 present side-by-side comparisons of the geographic affinity matrix (left) and the latent affinity matrix (right) for two different batches.

In both matrices, high-affinity values are shown in yellow, indicating pairs of images that are geographically or latently close. Low-affinity values appear in dark blue or purple, representing distant pairs. In Figure *3*, the latent affinity matrix exhibits structural patterns that closely align with those of the geographic affinity matrix.

Figure *5* presents an alternative batch example. In this case, the geographic affinity matrix exhibits a more fragmented structure, with smaller and more dispersed high-affinity regions. The corresponding latent affinity matrix reflects this distribution, however, not as clearly as in Figure 3.

Across both examples, the structural consistency between geographic and latent affinity patterns suggests that the georeferenced regularisation transferred geological neighbourhood relationships into the learned embedding.

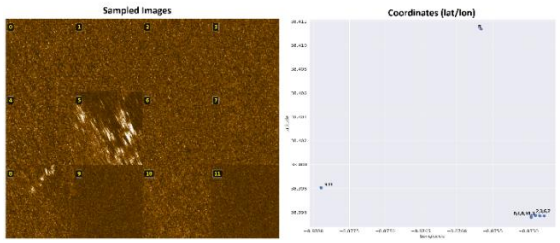

Figure 2 Sample images with corresponding geocoordinates

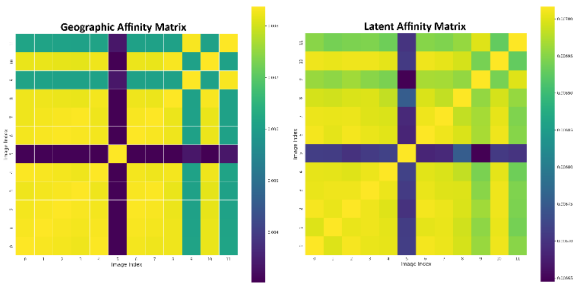

Figure 3 A comparison between the geographic and latent affinity

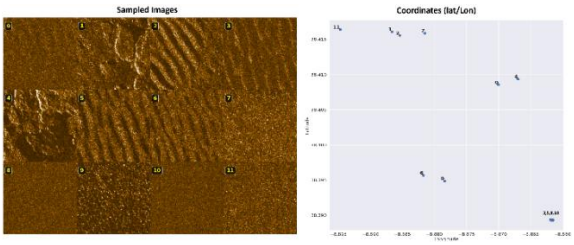

Figure 4 Sample images with corresponding geocoordinates

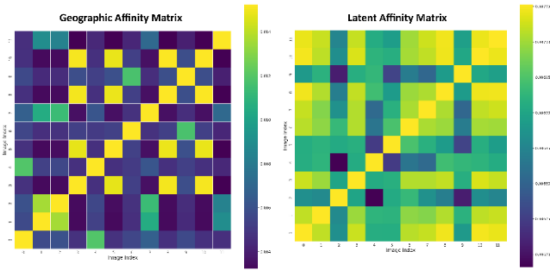

Figure 5 A comparison between the geographic and latent affinity

Reconstruction Accuracy

While the affinity analysis demonstrates that geographic relationships are preserved in latent space, reconstruction performance determines whether the embedding retains sufficient image information. Figure *6* presents representative examples of original input images (top row) and their corresponding reconstructions produced by the decoder (bottom row).

The reconstructed images preserve dominant structural patterns, including distinct seabed textures such as ripples. Large-scale spatial structures and overall intensity distributions are maintained, indicating that the latent representation retains the primary visual characteristics of the input data.

Fine-grained variations appear moderately smoothed in the reconstructions, which is consistent with the dimensionality reduction imposed by the bottleneck architecture of autoencoder.

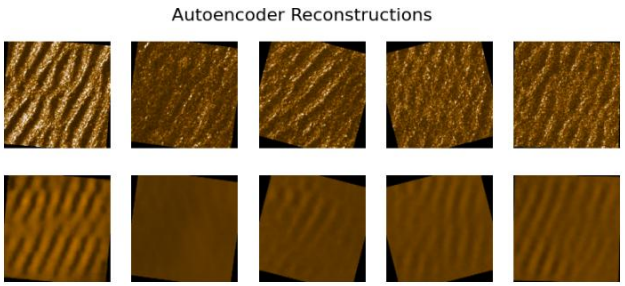

Figure 6 A comparison between the original inputs (upper row) and the reconstructed outputs (lower row)

3.2 Clustering

The qualitative assessment of reconstruction fidelity indicates that the latent representation preserves the dominant visual characteristics of the input imagery. The next stage examines whether the latent representation of the images can be used to organise the large-scale dataset into groups of visually similar images. Such grouping enables cluster-based summarisation, where representative samples from each cluster can be selected to approximate the diversity of the full dataset.

Figure 7 presents a three-dimensional projection of the 16-dimensional latent embeddings. Data points are coloured according to their k-means cluster assignments

(Cluster 0–7). Although the embedding is learned in a 16-dimensional space, the three-dimensional projection provides a qualitative visualisation of the global latent structure and cluster relationships. The visualisation indicates that some clusters exhibit partial separation, while others display overlapping or intertwined boundaries. For example, Cluster 1 and Cluster 4 (shown in orange and brown) appear to occupy adjacent regions in the embedded space. This behaviour is consistent with the representative samples in Figure 8 (rows 2 and 5), where images from these clusters exhibit similar seabed textures and intensity patterns.

The observed overlap may reflect the gradual variation of seabed textures rather than discrete categorical boundaries.

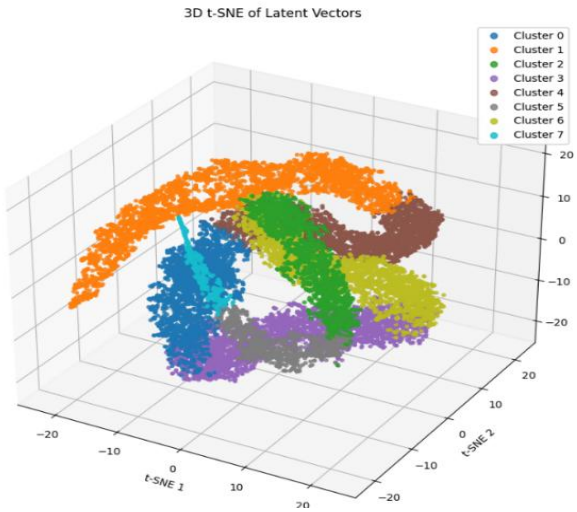

Figure 7 3D Visualisation of 16-dimensional latent vectors clustered into eight groups in 3-dimentsions

3.3 Image Search

Clustering examined whether the latent representations enable meaningful dataset division, whereas similarity-based retrieval assesses whether individual samples are surrounded by visually similar neighbours (e.g., similar images) in latent space.

This section presents the results of similar image retrieval using cosine similarity in latent space. For each query image, the top nine most similar samples are retrieved based on cosine similarity in latent space. Each result figure (Figure 9, 10, and 11) displays the query image in the top-left corner, followed by the nine closest neighbours ranked by similarity.

In Figure 9, the query image contains a mine located within a highly textured seabed background. The retrieved images share similar textural characteristics, although the neighbours do not contain a mine, indicating that object presence is not strongly encoded in the latent representation.

In Figure 10, the query image primarily contains rocky seabed structures. The retrieved results predominantly consist of rock formations, although some samples include partial sand ripple patterns. This suggests that the embedding captures dominant texture characteristics but does not strictly differentiate between closely related seabed types. It is also possible that geographic regularisation influences similarity in latent space; however, quantitative analysis of geographic proximity between query and retrieved samples would be required to confirm this effect.

In the final example, the query image shows a mine on a smooth sandy bottom. The retrieved images include a mixture of rocky scenes and sand ripple environments, suggesting that similarity is influenced by overall texture and intensity distribution rather than the specific presence of an isolated object.

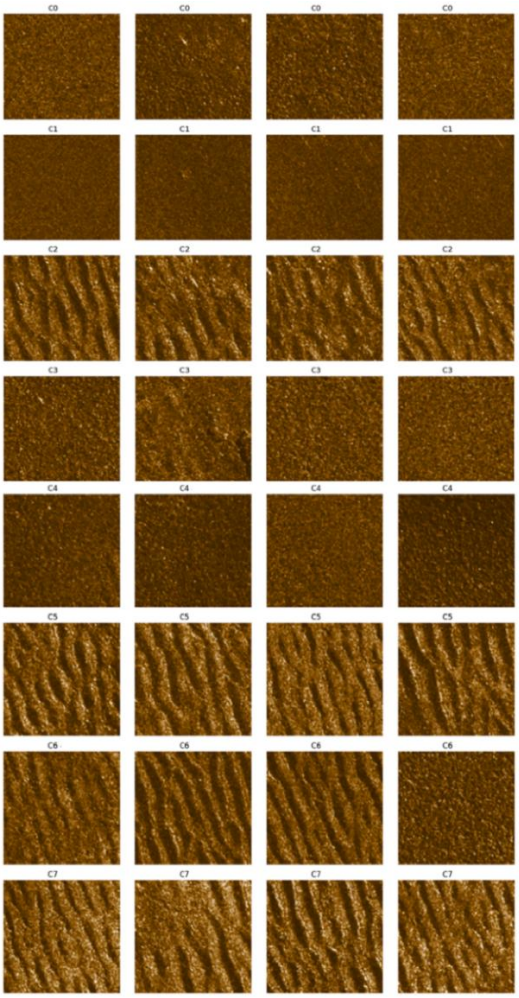

Figure 8 Representative images from the clusters. Each row shows four images from a cluster, with the first row corresponding to cluster 0.

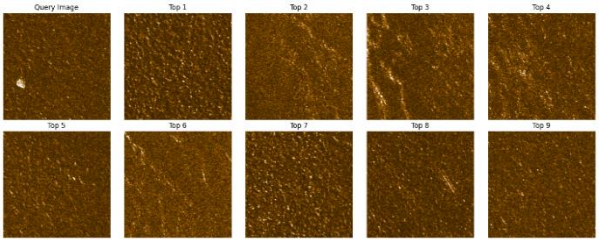

Figure 9 Query image and the top nine most similar images using cosine similarity

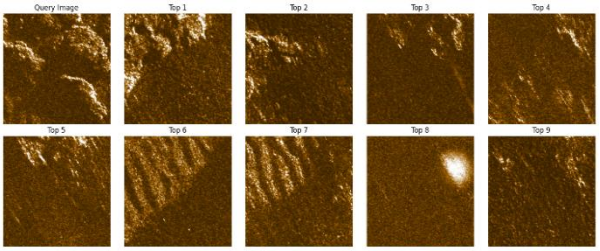

Figure 10 Query image and the top nine most similar images using cosine similarity

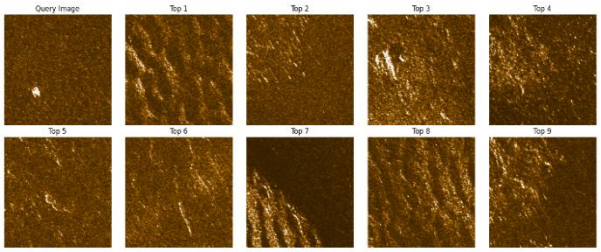

Figure 11 Query image and the top nine most similar images using cosine similarity

4 Conclusions

This study evaluated the feasibility of applying a compression framework to Synthetic Aperture Sonar (SAS) imagery to improve remote situational awareness in bandwidth-constrained AUV operations. Given the high-resolution SAS acoustic imagery, this work examined whether an autoencoder could learn a latent representation that captures informative structure in the data and supports compact summarisation of large datasets for remote interpretation without transmitting full-resolution imagery. This capability is relevant to autonomous survey operations, as summarisation reduces human supervisory load by mitigating exhaustive manual inspection.

Additionally, transmission-efficient monitoring shortens the perception–decision feedback loop, allowing operators to assess survey progress in near real time rather than relying solely on post-mission analysis.

Qualitative evaluation of the results suggests that the autoencoder demonstrates potential for learning latent representations that capture informative structural characteristics in SAS imagery. Visual assessment of reconstructed samples indicates that dominant seabed textures are mostly preserved. In contrast, images containing mixed or transitional characteristics, such as combinations of ripple-patterned and smooth sand bottom, are more challenging to reconstruct.

Clustering results show that embeddings group mainly by texture structure, reflecting meaningful scene characteristics. Notably, despite the presence of stone imagery in the dataset (see Figure 10), no dedicated stone cluster is observed, with these samples seemingly merged into sand-texture clusters.

Similarity-based retrieval using latent embeddings revealed a limited capability, as the returned images were only loosely consistent with query features compared to the study done by [7]. This suggests that the learned representation captures coarse global structure and remains insufficient for fine-grained similarity matching. The potential improvements can be through architectural refinement, training strategies such as balanced seabed sampling, and alternative loss formulations.

Another improvement would be evaluation using quantitative metrics and realistic mission workflows, which would better characterise the framework's operational impact in autonomous underwater surveys.

Overall, the findings highlight both the potential and limitations of autoencoder-based compression and representation learning for SAS imagery, which can be used in autonomous survey workflows.

5. Bibliography

[1] Leonardo Zacchini, Alessandro Ridolfi, Alberto Topini, Nicola Secciani, Alessandro Bucci, Edoardo Topini, Benedetto Allotta, "Deep Learning for on-board AUV Automatic Target Recognition for Optical and Acoustic imagery," in *IFAC-PapersOnLine*, 2020.

[2] Wang Xun, Cao Yanpeng, Wu Shijun, Yang Canjun, "Real-time detection of deep-sea hydrothermal plume based on machine vision and deep learning," *Frontiers in Marine Science*, vol. 10, 2023.

[3] Zhao L, Yun Q, Yuan F, Ren X, Jin J, Zhu X., "YOLOv7-CHS: An Emerging Model for Underwater Object Detection," *Marin Science and Engineering*, vol. 11, 2023.

[4] Lukas Folkman, Kylie A. Pitt, Bela Stantic, "A data-centric framework for combating domain shift in underwater object detection with image enhancement," in *Applied Intelligence*, 2025.

[5] J. L. Walker, Z. Zeng, C. L. Wu, J. S. Jaffe, K. E. Frasier and S. S. Sandin, "Underwater Object Detection Under Domain Shift," *IEEE Journal of Oceanic Engineering*, vol. 49, 2024.

[6] Blair Thornton, Miguel Massot Campos, Adrian Bodenmann, Samuel Simmons, Chihiro Hirai, "Rapid Over-Horizon Awareness of AUV Image Datasets over Low Communication Bandwidths," in *IEEE*, 2024.

[7] Yamada Takaki, Prügel-Bennett Adam, Thornton Blair, "Learning features from georeferenced seafloor imagery with location guided autoencoders," *Journal of Field Robotics*, vol. 38, pp. 52-67, 2021.

[8] Arjuna Balasuriya, Ura T., "Vision-based underwater cable detection and following using AUVs," in *USA*, 2002.

Author/Speaker Biographies

With a master's degree in Computer Science and Engineering, Nisa Andersson is a Systems Engineer at SAAB Underwater Systems specialising in autonomous systems. Her work focuses on bridging the gap between situational awareness and tactical autonomy by developing algorithms that enable robotic platforms to reason and respond to their environment autonomously.

Acknowledgements

This work builds extensively on research conducted by Thornton et al. [6] and I would like to thank Per Abrahamsson (SAAB Underwater Systems) for granting permission to use the SAS dataset. The manuscript was prepared with assistance from OpenAI's ChatGPT 5.2 for grammar and logical clarity.