This paper explores change detection in repeat-track side-scan sonar imagery through feature matching. It addresses insufficient matching accuracy and stability in low-contrast, noisy, and geometrically distorted side-scan sonar imagery typically collected from surface vehicles. The experiment included a comparison of classical, convolutional, and transformer-based feature matching methods (SIFT, DISK, SuperPoint, LoFTR, and LightGlue) on two real-world datasets, Atlantic and Baltic. The results were evaluated quantitatively and qualitatively. Quantitative evaluation used displacement, angular stability, and reprojection error metrics, as well as resource consumption metrics like execution time and memory usage. In addition, matching maps and change maps for pairs of images were generated and analyzed qualitatively. All methods produced interpretable change maps for the low-noise Baltic dataset, whereas the wave-affected Atlantic dataset with stripe- and speckle noise only occasionally produced consistent maps. The SuperPoint + LightGlue method demonstrated the highest ratio of inlier correspondences after RANSAC filtering (43.4% and 65.6%) and the lowest mean reprojection error (36.0 and 3.9 px), while LoFTR provided the densest coverage (up to 97%) consuming up to 15× more computational resources. These results confirm the advantage of transformer-based matching methods under challenging conditions due to their global receptive field. In contrast, CNN-based methods performed better in low-noise, well-aligned images. Overall, the findings indicate that deep feature matchers can improve the applicability and reliability of change detection in tasks such as humanitarian demining, autonomous underwater navigation, image mosaicking, and related applications

Keywords: side-scan sonar, feature matching, deep learning, change detection, computer vision

UDC 004.932.4 : 004.89 : 621.45
DOI: 10.15587/1729-4061.2025.346940

CHANGE DETECTION IN SIDE-SCAN SONAR IMAGERY BASED ON DEEP LEARNING FEATURE MATCHING METHODS

Oleksandr Katrusha*
Corresponding author
E-mail: oleksandr.katrusha@gmail.com
Dmytro Prylipko
Master of Science
EvoLogics GmbH
Wagner-Régeny-Straße, 4,
Berlin, Germany, 12489
Kostiantyn Yefremov*
PhD
*Department of Artificial Intelligence
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
Beresteiskyi ave., 37, Kyiv, Ukraine, 03056

Received 03.10.2025
Received in revised form 05.12.2025
Accepted 15.12.2025
Published 30.12.2025

How to Cite: *Katrusha, O., Prylipko, D., Yefremov, K. (2025). Change detection in side-scan sonar imagery based on deep learning feature matching methods. Eastern-European Journal of Enterprise Technologies, 6 (2 (138)), 51–62. https://doi.org/10.15587/1729-4061.2025.346940*

1. Introduction

Side-scan sonar (SSS) is extensively used for wide range of underwater applications. Detecting changes between same-track missions is crucial for mine counter-measures (MCM), rescue operations, underwater and coastal inspection, port security, marine environment studies, pipeline inspection, underwater archeology and other use cases of side-scan sonar imagery.

However, analysis of SSS images over long missions is significantly more challenging than optical images. This complexity is caused by low contrast, a small number of visible objects, speckle noise, geometric distortions caused by platform motion, and the presence of prominent acoustic shadows and highlights. These factors make it difficult to extract and reliably match keypoints between different missions, which is essential for change detection tasks, mosaicking and simultaneous localization and mapping (SLAM). The number of images obtained during a single survey can reach thousands. Manual processing of such information volumes requires significant human resources and time, and makes it impossible to make timely decisions in military, rescue, or engineering operations.

The key area of modern research in this field is the application of deep learning and computer vision methods to improve the accuracy and reliability of image matching, change detection, and object localization. Neural networks consider complex patterns of intensity, texture and geometry of the seabed, inaccessible to classical algorithms. However, most existing studies of SSS images matching use either classical methods, such as SIFT, SURF, and ORB, or convolutional neural networks (CNNs), transferred from the optical domain [1]. Classical methods are not effective enough in cases of low contrast, strong speckle noise, and lack of distinctive features, while CNN-based approaches rely on local receptive fields and are therefore limited in capturing global context. The lack of open data for training and fine-tuning also limits the use of convolutional approaches, while the active development and success of transformer-based models make it possible to solve the problems of image matching and change detection assuming the image global context under complex conditions.

The evolution of deep learning methods will help move to automated processing of SSS data. This will enable operator-free work of autonomous systems in real time. From a practical point of view, the research will help in mine counter-measures, environmental monitoring, inspection of marine structures, and the development of autonomous navigation.

For Ukraine, these tasks are of particular importance: the consequences of the war create a need for new approaches to

humanitarian demining and the safe use of coastal waters. Thus, research aimed at applying modern deep learning methods to SSS images matching and change detection is relevant for both science and practical usage.

2. Literature review and problem statement

A detailed comparison of keypoint detection methods performed on sonar images from autonomous underwater vehicles was carried out in [1]. Ten methods (nine classical AKAZE, BRISK, FAST, Harris, ORB, Shi-Tomasi, SIFT, SURF, SAR-SIFT, and the KeyNet deep detector) were compared on a dataset of sixteen SSS and forward-looking sonar (FLS) images pre-processed by filters. The methods were evaluated by the number of points found, coverage density, time spent, and other metrics. However, the dataset was quite limited, and the images contained strong features, which facilitated the identification of keypoints. In addition, the images were collected from underwater vehicles, which are much less affected by noise and waves than surface ones.

The classification of methods for detecting changes in remote sensing images divides them into two main types, as noted in review [2]: pixel-based (PBCD) and object-based (OBCD). The former implements image comparison at the pixel level, which requires precise alignment. The latter aims to compare objects previously detected by other methods. The basis of the PBCD approach is the process of image co-registration, that aligns two or more images using navigation information or matching features within them with subpixel accuracy. The paper confirms the relevance of deep learning methods for matching (in particular, RPC, SIFT, DTM, CACO, and RANSAC) and detecting changes in (radar) remote sensing images but not for SSS images, which have their own specificity. It also does not mention the transformer-based models that have gained popularity in recent years.

After aligning the images by keypoints, a change map suitable for manual or automated analysis can be created by simple subtraction as, for example, in study [3]. It used a convolutional neural network (CNN) to classify potential changes into "object" and "non-object" and reduce the percentage of false classifications. The matching was performed using a third-party library without specifying the algorithm. However, due to the lack of open SSS datasets, the study was mainly based on a synthetic dataset and several real-world SSS images from underwater vehicles. It is known that synthetic data cannot always reproduce the actual environment in sufficient detail. Thus, the problem of matching points using deep methods on sufficiently large real-world datasets remains unsolved, partially due to the lack of large public datasets.

In work [4], it was proposed to use deep segmentation models (Segment Anything Model) to detect changes in synthetic aperture sonar (SAS) images. The model showed better results than traditional log-ratio based change detection. Point detection and matching were performed using classical SIFT and RANSAC methods. SAS images, although similar to SSS ones, are generally of much better quality, so the question of applicability of the methods to SSS remained open.

Image matching can also be performed for simultaneous localization and mapping (SLAM). For example, in [5], the classical A-KAZE point detection method [6] is used for matching images obtained from forward-looking sonar (FLS). However, the application of these methods to SSS images has not been considered.

A comparative study [7] of matching methods' effectiveness on images obtained in same-track SAS missions was conducted. The SIFT algorithm was used for feature matching, and the RANSAC method or its modifications, in particular M-estimator Sample Consensus (MSAC), were used to refine the results by filtering out incorrect matching. Another work [8] also used SIFT and SURF as the main methods for feature matching, along with other approaches. However, deep feature detection and matching methods were not considered in these studies.

Feature matching is also a key element in mosaicking and automatic target recognition (ATR). In [9], the sonar-to-optical style transfer was used to increase the proportion of correct matches and TFeat, a local feature descriptor based on the VGG-19 convolutional neural network was proposed, which exceeded the classical descriptors in quality. However, the experiment was performed on a limited set of six images from an AUV that contained prominent features and no transformer matchers were used.

In [10], SIFT was chosen to detect corner points on the SSS image from an underwater vehicle due to its high matching accuracy. However, there were numerous noticeable lines on the flat seafloor left by anchors, which significantly helped in detecting keypoints in these high-quality images.

In the experimental results published in [11], the SURF algorithm was used as a keypoint detector for SAS image matching. Also, in [12], SIFT was used to match SAS images in a multi-stage registration system using canonical correlation analysis. However, their application to SSS images was not investigated.

In [13], a SONIC convolutional matching model was proposed for FLS images, and an experimental comparison was made with AKAZE and the LightGlue transformer model [14]. However, their application to SSS images was not investigated.

Therefore, it can be concluded that for the tasks of change detection and feature matching in remote sensing images, in particular sonar images, methods based on classical feature detectors, such as SIFT, SURF, A-KAZE, or PatchMatch [15, 16], in combination with descriptor matching and geometric refinement via RANSAC, are widely used. At the same time, there are numerous studies of deep learning methods including CNNs for the above-mentioned tasks. They demonstrate increased robustness to noise, intensity variations and image distortions, and show advantages over classical algorithms in low contrast conditions. However, convolutional methods mostly work with local perception fields and consequently do not always take into account the wider context of the scene, which is a significant drawback for sonar images with a small number of prominent texture features.

New transformer-based approaches offer promising ways to overcome these limitations. Instead of the traditional "detect-describe-match" sequence, modern algorithms such as LoFTR [17] use an attention mechanism that allows for global analysis of correspondences between images. LoFTR generates dense pixel-wise matches without the need for an external detector, while LightGlue dynamically adapts the matching process using descriptors obtained from deep detectors such as SuperPoint [18], DISK [19], or classical SIFT. These methods may be particularly promising for SSS images, where local textures are weakly expressed, but global structural features of the seabed morphology remain stable.

Therefore, the unsolved part of the problem is the lack of a comparative assessment of modern feature matching methods

of different architectures (classical, convolutional and transformer-based) specifically for real-world SSS images, including noisy ones. Existing works are mostly focused on processing optical or SAS images from underwater vehicles with distinct high contrast features, which significantly simplifies the task of keypoint detection and matching, while real-world sonar images collected from surface vehicles have low contrast, a small number of noticeable high contrast objects, shadows, high speckle and stripe noise. The question of methods' effectiveness by the "accuracy-resource consumption" criteria still remains open, which is critically important for their use in autonomous systems with limited computing resources.

In other words, it is advisable to conduct a study to systematically compare the effectiveness and resource utilization of modern classical, convolutional, and transformer-based methods for the tasks of features matching and changes detection. The datasets for the study should be large enough and contain real-world side-scan sonar images collected from a surface vehicle.

3. The aim and objectives of the study

The aim of the study is to determine the suitability of modern deep neural feature matching networks for the task of change detection in side-scan sonar images under real-world conditions. This will help define the influence of network architecture on the accuracy and stability of feature matching, and outline the directions of further adaptation of deep models to the sonar images domain.

To achieve the goal, the following research tasks were defined:

– to generate two real-world SSS data sets obtained from same-track missions of uncrewed surface vessels (USV);

– to experimentally compare and quantitatively evaluate the effectiveness of classical, convolutional, and transformer-based feature matchers (SIFT, DISK, SuperPoint, LoFTR, LightGlue) on the generated data;

– to determine the computational efficiency of the selected methods (execution time, memory usage);

– to conduct a qualitative analysis of matching maps and change maps.

4. The study materials and methods

The object of the study is the process of feature matching between side-scan sonar images with the purpose of change detection in the underwater environment.

The hypothesis of the study can be formulated as follows. Modern deep learning algorithms based on convolutional networks and transformers can provide higher density and geometric accuracy of feature matching on real-world sonar images compared to classical methods. This should allow to improve the quality of change detection even in noisy and distorted images from real SSS missions. At the same time, it is known that transformer-based architectures require much more computational resources, which should be confirmed by research results.

In the process of research, the following assumptions and simplifications were adopted, to focus on the main task of the study without complicating the model and introducing additional restrictions and corrections, such as variations in equipment parameters or changes in environmental conditions.

Assumptions accepted:

1. Image pairs (reference and matching) were obtained with the same SSS parameters – slant range, viewing angle, etc.

2. Changes on the seabed between missions are local and do not affect its overall geometry.

3. The influence of waves and deviations of the carrier vehicle track is insignificant and does not require additional correction.

Simplifications adopted:

1. When detecting keypoints, the water column was excluded (masked), which allowed to avoid unnecessary false matches and skip areas without useful information.

2. Additional intensity equalization after time-value gain (TVG) correction was not applied to evaluate the algorithms on real distorted signal.

3. Homographic image warping was applied, which is assumed to be close to the true three-dimensional geometry of the seabed surface.

Simplified seabed geometry model and exclusion of water column area increases the calculation stability and allows for a more accurate quality assessment of the comparison.

Data.

The experiment was based on field data from two real SSS operations, each consisting two "lawn mower" missions along the same track (reference and matching, respectively). The missions were performed using an EvoLogics Sonobot 5 USV equipped with SSS with carrier frequency of 500 kHz. Each point on the seabed within the area of each mission was covered at least twice. Several objects were placed on the seabed between passes to test the ability of the algorithms to detect changes. The first operation *Atlantic* was carried out off the coast of Portugal in September 2024. The matching mission was carried out under conditions of strong waves and wind, which led to the appearance of noise artifacts in the images – wave patterns and stripe noise. The second operation *Baltic* was carried out close to the pier near the city of Świnoujście (Poland) in November 2024.

Hardware.

All calculations, including inference of deep learning models, were performed on a Lenovo Legion computer with GeForce RTX4070 (8GB) graphics adapter, Intel Core i-7 CPU, and 32 GB of RAM. The code and weights of the models were taken from the corresponding repositories listed in the references.

Method.

The first step in constructing a relevant seafloor change map was to identify keypoints for each pair of reference and matching images. In this paper, the terms *keypoints* and *features* are used interchangeably to denote local image correspondences, with the terminology adapted to the specific matching approach (keypoint-based or dense)

$$\left\{k_1,k_2,\ldots,k_r\right\}\in\mathbb{N}^2,\ \left\{k'_1,k'_2,\ldots,k'_m\right\}\in\mathbb{N}^2. \quad (1)$$

SIFT was chosen as the reference detector due to its popularity in optical and sonar imagery. DISK and SuperPoint methods were chosen for comparison as modern deep learning approaches, whose effectiveness needs to be verified on SSS data. The maximum number of keypoints was set to 4096 to ensure maximum coverage, and the descriptor length was set to 256. All other parameters remained set by default.

In the second stage, the obtained keypoints were matched to form matching pairs

$$M=\left\{\left(k_i,k'_j\right)\middle|\, i\in\left[1,r\right],j\in\left[1,m\right]\right\}. \quad (2)$$

For testing purposes, a modern LightGlue matcher based on the attention mechanism was chosen. The dense LoFTR matcher does not have a separate detector since it forms both keypoints and matches in one pass. LoFTR with the ResNet-FPN backbone was initialized with weights obtained on outdoor datasets; other parameters left standard.

After obtaining pairs of matched keypoints for each image, the homography matrix H = homography(M) was calculated and the matching image was warped to geometrically aligned at the pixel level with reference one. The homography was calculated using the OpenCV implementation [20] with RANSAC filtering (threshold = 5).

Ultimately, a change map was constructed by absolute subtraction of images at the pixel level. The filtered change map was obtained after thresholding the result to highlight the most significant changes and suppress noise. The resulting change map was used for qualitative assessment of the accuracy and adequacy of matching.

Evaluation.

The evaluation was performed both quantitatively and qualitatively. For a balanced quantitative comparison of the selected feature detection and matching methods, two groups of metrics were considered:

1. Quality of keypoint matching.

2. Efficiency of resource usage.

The key metrics included the number of keypoints per image, the number of matches, spatial coverage (percentage of 64×64 pixel blocks containing at least one match), and the percentage of matches that passed the filtering (RANSAC Inlier %). These indicators directly reflect the method's ability to detect and match reliable features in sonar images, noisy and heterogeneous by nature.

Since ground truth matches require laborious manual labeling, indirect quantitative metrics were used to assess the matching geometric quality. The image pairs are of the same size, geo-aligned, and acquired with the same hardware. This implies that true matches should have close coordinates within each image. Deviations of a few pixels are possible due to variations in roll, pitch, vehicle speed, or sea level due to turbulence, tides, or other factors.

Therefore, the following metrics were used to assess the quality of the matching:

1. Displacement – the difference between the coordinates of the corresponding keypoints in two images $d_i = \left\| k_i' - k_i \right\|$. For a perfect match, the expected value and standard deviation of the displacement should be within a few pixels.

2. Matching angle – the angle between the line connecting the matching points in the adjacent images and the "horizon"

$$\theta_i = \arctan\left(\left(y_i' - y_i \right) / \left(x_i' + W - x_i \right) \right), \quad (3)$$

where (x_i, y_i) are the coordinates of the i-th point on the reference image; $\left(x_i', y_i'\right)$ are the coordinates of the corresponding point on the matching image; W is the width of the image; θ_i is the angle of the line connecting the corresponding points. Ideally, the expected value of the angle should be close to zero, and the standard deviation should be within a few degrees. The number of outliers was then estimated from the angle distribution.

3. Reprojection error, defined as

$$e = \frac{1}{N} \sum_{i=1}^{N} \left\| x_i' - proj\left(Hx_i \right) \right\|_2, \quad (4)$$

where N is the number of matches; $x_i \in N^2$ is the keypoint of the reference image; $H \in R^{3 \times 3}$ is the calculated homography matrix; *proj*(.) – projection function converting homogeneous coordinates back to Cartesian; x_i' is the corresponding keypoint in the matching image.

This indicator quantitatively estimated the distance between the projected position of the point and its pair in another image. A smaller value of the standard deviation means higher quality of the match after alignment.

The study also used the resource efficiency indicators – the processing time per one image, the consumption of random access memory (RAM), as well as the amount of allocated and peak memory of the graphics processor (GPU). These indicators are critical for assessing the methods implementation in real underwater missions, where computing resources are often limited. The evaluation of each indicator was carried out separately for each mission.

5. Results of the matching models' comparative research

5. 1. Generated datasets

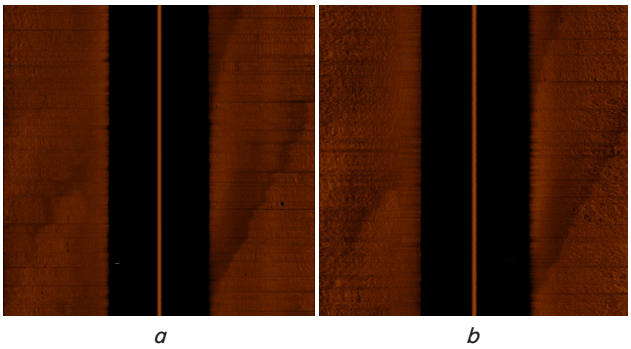
Based on the collected field data, two datasets were generated. When designating specific pairs of images (reference and matching), the notation *Baltic* 1, *Atlantic* 5, and so on was used. The main parameters of both datasets are given in Table 1.

Table 1

Dataset parameters

Dataset	Mission length, m	Time between missions	Altitude, m	Slant range, m	Number of image pairs	Bottom type
Atlantic	2500	7 days	16–18	51	20	Sand
Baltic	700	1 hour	8–10	51	14	Sand and silt

No additional intensity correction was applied after standard TVG correction. The water column was masked to avoid false keypoint matching and skewed distribution. Typical images from Atlantic mission are shown on Fig. 1.



a *b*

Fig. 1. Sample reference and matching images for *Atlantic* 4 pair: *a* – reference image; *b* – matching image

The matching mission images contained distortion and stripe noise (Fig. 1, *b*), which noticeably changed the visible structure of the seabed. Each track from both the reference and repeat missions was divided into relatively straight segments with a course deviation of no more than 12°, determined by the GNSS coordinates of the vessel. The navigation information was converted into "North-East" coordinate

system relative to the initial point of the reference mission. Each matching segment was paired with the nearest reference segment using the cKDTree algorithm [21], where more than 90% pings were within the half-meter threshold. In the absence of matches for a particular ping, the previous ping from the paired image was reused. After that, the sonar image data was sliced and converted into fragments of 1024×1024 pixels with an overlap of 128 pixels. End images of the segments were cropped vertically to approximately 1024×670 pixel size.

5. 2. Experimental comparison and quantitative evaluation of efficiency

Table 2 summarizes the results of keypoint detection and matching efficiency for both datasets. SIFT, DISK, and SuperPoint generated an average of 3,000 to 4,000 keypoints per image; however, the number of matches remaining after LightGlue filtering was low, ranging from 24 (SIFT) to 66 (SuperPoint). At the same time, LoFTR formed a much denser set of points – an average of 472 points per image, almost completely covering the image. However, after RANSAC, the proportion of inlier matches remained low (approximately 8%), indicating geometric inconsistency of a large fraction of the correspondences.

Table 3 further examines the geometric quality of matches. In the Atlantic set, traditional SIFT and DISK exhibited rather large displacement mean (≈ 130 pixels) and reprojection error mean (> 300 pixels), reflecting unstable correspondences. At the same time, the combination of SuperPoint + LightGlue showed better metrics. The displacement mean (≈ 36 pixels) and reprojection error (mean 36 pixels, standard deviation 72) indicates better geometric stability of the matches. LoFTR, with a fairly dense coverage, showed large displacement standard deviation (182 pixels) and moderate reprojection error mean (≈ 142 pixels). Although SuperPoint was the best in terms of metrics, it still failed to provide reliable enough matches comparing to Baltic dataset, which becomes obvious from qualitative analysis (see below). In the less noisy Baltic dataset containing contrast features of a pier, all methods showed significantly better metrics. SIFT, DISK and SuperPoint: reprojection error mean (< 5 pixels), displacement standard deviation (< 5 pixels), while LoFTR produced results with higher variance due to sporadic false matches. In general, classical and sparse methods (SIFT, DISK, SuperPoint) showed high quality, whereas dense transformer-based LoFTR was statistically times worse than its sparse counterparts.

Matching angle distributions for LoFTR and SuperPoint + LightGlue on Fig. 2 help understand the details behind the collected metrics.

In the Baltic set, both LoFTR and SuperPoint + LightGlue produced a set of matching with angles concentrated near zero, as reflected in low angular standard deviations (approximately 1.3° for LoFTR and 0.3° for SuperPoint). LoFTR has a sharp peak with minimal spread, while SuperPoint has a well-centered but broader bell-shaped distribution.

In the Atlantic set, however, LoFTR (*c*) retains a centered bell-shaped distribution but with "heavy tails" due to larger angular deviations (6–7°). This indicates a higher number of outliers and false matches under difficult conditions. At the same time, SuperPoint + LightGlue (*d*) generated very few matches with sparse and irregular angle distributions. This explains its relatively small standard deviation (≈ 1.9°) but indicates the instability of the method with a small number of matches.

Table 2

Keypoints and matches

Dataset	Method	Keypoints per image	Matches per image pair	Spatial coverage, %	RANSAC inliers, %
Atlantic	SIFT + LightGlue	3199	24	48	27.1
	DISK + LightGlue	4096	42	57	18.6
	SuperPoint + LightGlue	3351	66	34	43.4
	LoFTR	472	472	96	8.0
Baltic	SIFT + LightGlue	3252	1088	86	61.3
	DISK + LightGlue	4096	1772	77	57.3
	SuperPoint + LightGlue	3835	1394	83	65.6
	LoFTR	4674	4674	97	64.3

Table 3

Geometric consistency of matching

Dataset	Method	Displacement mean, pixels	Matching angle standard deviation, degrees	Displacement standard deviation, pixels	Reprojection error mean, pixels	Reprojection error standard deviation, pixels
Atlantic	SIFT + LightGlue	131.0	5.71	96.2	506.6	1124.0
	DISK + LightGlue	128.5	8.26	120.9	298.6	487.8
	SuperPoint + LightGlue	35.9	1.91	26.2	36.0	71.8
	LoFTR	135.5	6.78	182.4	141.7	184.2
Baltic	SIFT + LightGlue	6.1	0.35	4.2	4.3	3.3
	DISK + LightGlue	5.9	0.35	4.0	4.8	3.6
	SuperPoint + LightGlue	5.6	0.34	3.8	3.9	2.7
	LoFTR	10.0	1.37	28.7	8.8	28.9

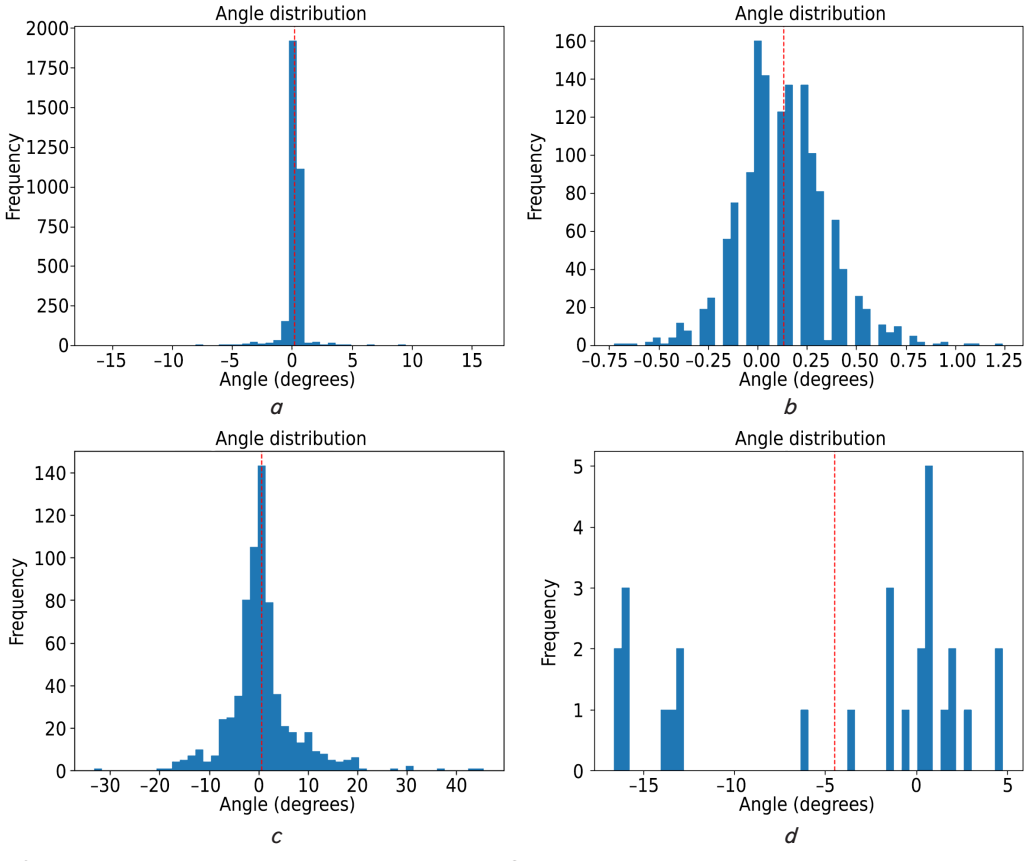


Fig. 2. Histogram of matching angles for LoFTR and SuperPoint of sample image pairs from both datasets: *a* – LoFTR for the *Baltic* 6 pair; *b* – SuperPoint + LightGlue for the *Baltic* 6 pair, *c* – LoFTR for the *Atlantic* 17 pair, *d* – SuperPoint + LightGlue for the *Atlantic* 17 pair; in all images, the *X* axis is the matching angle in degrees, the *Y* axis is the frequency of the angle in the sample

Based on matching angles distribution analysis, it can be concluded that LoFTR provides denser, but less stable keypoint matches. SuperPoint + LightGlue generates fewer, but more accurate correspondences. The methods effectiveness significantly depends on the data set properties. On noisy sets, classical and convolutional methods may lose their ability to generate coherent matches with good scene coverage.

5. 3. Estimating computational efficiency

Computational resources usage is analyzed in Table 4. On both datasets, the SIFT, DISK, and SuperPoint methods demonstrated similarly low memory requirements and high execution speed. RAM consumption was limited to 25 MB, allocated GPU did not exceed 70 MB, and peak – 532 MB per dataset. The average processing time of one image pair ranged from 34 ms to 66 ms, which makes these methods suitable for large-scale and high-load applications.

Transformer-based LoFTR, however, required significantly more resources: for the Atlantic set, the average processing time for a pair of images was approximately 820 ms with a peak GPU memory consumption of about 4 GB per set and an allocated GPU memory of 649 MB per set. Similar results were observed for Baltic – about 856 ms per image pair and up to 3.4 GB of peak GPU memory per set, and up to 1009 MB allocated GPU memory per set.

Table 4

Resource usage efficiency

Dataset	Method	Time per image pair, ms	RAM, MB	Allocated GPU, MB	Peak GPU, MB
Atlantic	SIFT + LightGlue	59	15	57	346
	DISK + LightGlue	34	11	63	532
	SuperPoint + LightGlue	55	9	65	298
	LoFTR	820	541	1009	3995
Baltic	SIFT + LightGlue	57	25	82	319
	DISK + LightGlue	62	18	81	498
	SuperPoint + LightGlue	66	15	107	365
	LoFTR	856	196	649	3348

5. 4. Qualitative evaluation of matching and change maps

Quantitative metrics provide means of objective comparison but do not sufficiently reflect the spatial distribution of matches or the visual quality of the resulting change maps, which is the ultimate methods goal. Therefore, an additional qualitative analysis of the keypoints distribution, the consistency of matches, and the quality of the change maps was conducted on typical pairs of images from both datasets. Four main aspects were taken into account:

1. Overall number of mapped keypoints.
2. Distribution of keypoints across the image.
3. Visual quality of mappings, presence of false matches.
4. Quality of the resulting change maps.

The aim of the analysis was to identify the patterns and individual cases not covered by the metrics and to assess the suitability of the change map for operator's review. The results, summarized in Table 5, demonstrate significant differences between methods and datasets, highlighting the trade-off between density, reliability, and interpretability of keypoint matching mapping in sonar images.

Fig. 3–8 show typical images from each experiment, illustrating the observations described above and demonstrating the methods performance.

Qualitative analysis confirmed that the matching efficiency depends on the detector-matcher combination, as well as on the data set. In the Atlantic set, with a homogeneous seabed texture, low contrast and noise, SIFT + LightGlue and DISK + LightGlue generated only sparse, often false matches. SuperPoint + LightGlue provided more consistent matches, but still with insufficient and skewed coherence. LoFTR provided the highest density and uniformity of matches, which allowed for high-quality change maps to be obtained on some pairs, although numerous false matches prevented the calculation of a correct homographic transformation.

Table 5

Qualitative observations

Dataset	Method	Number of matched keypoints	Spatial distribution of keypoints	Matching quality	Change map
Atlantic	SIFT + LightGlue	Dozens of sparse points	Distributed evenly, closer to contrast areas. In some images grouped in one corner. Some images with very few points	Many false matches (angles), 100% for some images	Failed to produce interpretable change maps
	DISK + LightGlue	Dozens of points, more than SIFT	More evenly than in SIFT, less bound to contrast areas. All on one side for two images	False matches present	Failed to produce interpretable change maps
	SuperPoint + LightGlue	From several to hundreds of points	Big variations from very dense groups to even distribution. One image without matched points	Highly coherent matches with few false ones	Interpretable change map in a few pairs
	LoFTR	Hundreds of points, evenly distributed	Evenly distributed with occasionally dense groups in nadir or high contrast areas	Visible minority of false matches; often grouped together	Interpretable change map in a few pairs, overall better quality
Baltic	SIFT + LightGlue	Hundreds of points	Evenly distributed. Higher density in high contrast areas (pier, nadir, sandbars, bottom objects)	Almost no false matches	Good change map, minor distortion
	DISK + LightGlue	Hundreds of points, more than SIFT	Grouped closer to nadir area. In some images left or right-most 1/5 of images without keypoints. Very few points near nadir/contrast areas	Few or no false matches	Overall good, few pair with skewed warping
	SuperPoint + LightGlue	Hundreds of points	Evenly distributed, high density in nadir/contrast areas/edges, pier Some well textured areas without matches	Few or no false matches	Overall good, few pairs with skewed warping
	LoFTR	Times more than other methods	Evenly distributed, some small areas without keypoints. Higher density in nadir area	Few or no false matches	Overall good, few pairs with skewed warping

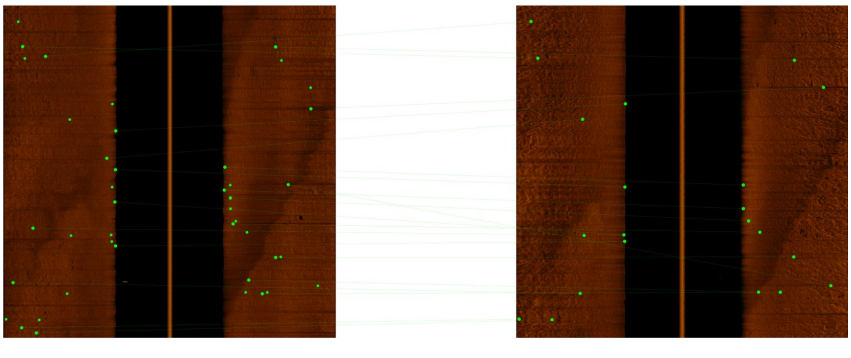


Fig. 3. SIFT + LightGlue matching result for *Atlantic* 4 pair — sparse keypoints with many false matches

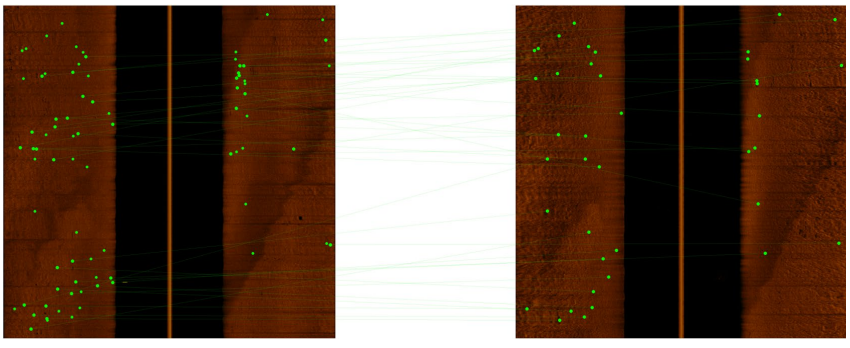


Fig. 4. DISK + LightGlue matching result for *Atlantic* 4 pair — tighter matching, but still with a large number of false matches

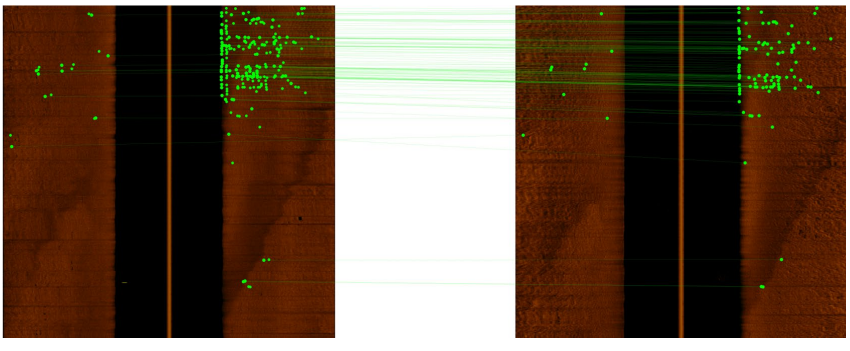


Fig. 5. SuperPoint + LightGlue matching result for *Atlantic* 4 pair — matches concentrated in one corner

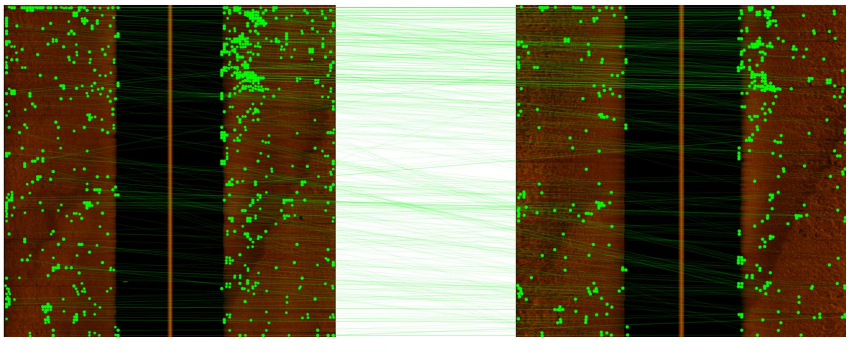
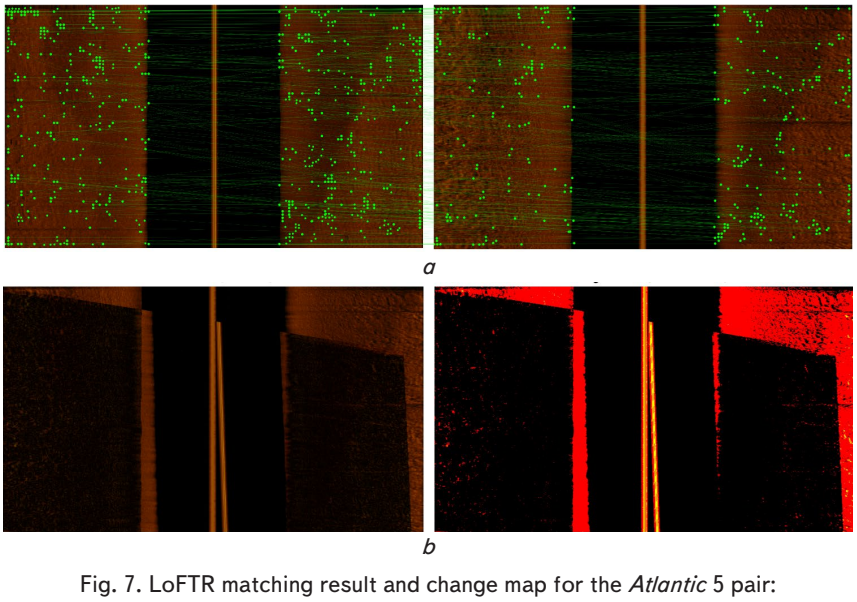


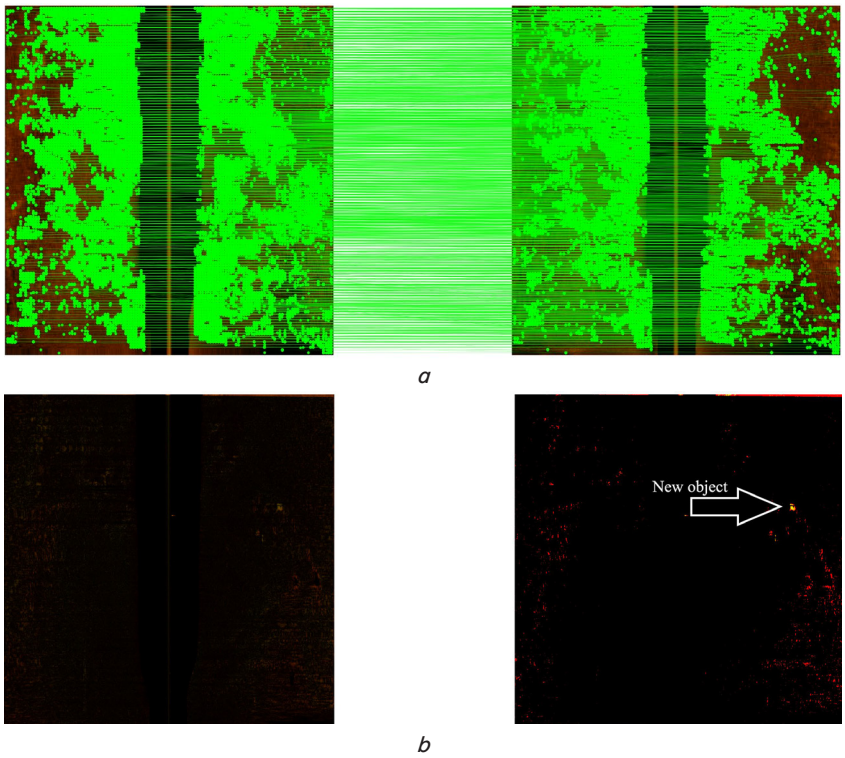
Fig. 6. The result of the LoFTR matching for the *Atlantic* 4 pair — numerous evenly distributed points, but with a significant number of false matches



a

b

Fig. 7. LoFTR matching result and change map for the *Atlantic* 5 pair:
a — matching map; *b* — change maps, original (left), and after thresholding (right)



a

b

Fig. 8. LoFTR matching result and change map for the *Baltic* 9 pair: *a* — highly coherent matches of evenly distributed numerous keypoints; *b* — change maps, original (left), and after thresholding (right); new object is marked with a white arrow

In the Baltic set, with better contrast and the presence of clear structures (pier, sandbars), all methods demonstrated much better results. SIFT + LightGlue and DISK + LightGlue gave high-quality coverage with a minimum of errors; SuperPoint + LightGlue achieved the best balance between density and stability; LoFTR, as in the first set, created the most matches, with few false ones. All methods allowed for the construction of high-quality informative change maps that clearly reflected changes at the bottom (for example, in Fig. 8), achieving the main goal.

6. Discussion of matching models research

Our results (Tables 2–5, Fig. 2–8) confirm the different behavior of the methods by quantitative metrics, matching quality, and resource consumption. The patterns revealed in the study explain the effectiveness of each approach, considering their architectural features, input data and task type.

Combinations of classical (SIFT) and convolutional (DISK, SuperPoint) detectors in combination with LightGlue form sparse but stable matches, which provides small displacement and low reprojection error (Table 3). Such stability can be explained by the limited local perception fields of CNN architectures, where descriptors distinguish intense contrasts even in the presence of noise. However, due to the low density of matches, sparse methods fail to provide full-fledged image matching at the pixel level and create interpretable change maps (Fig. 3–5).

In contrast, the transformer-based LoFTR model provided the highest density of matches (Table 2), which allowed to achieve continuous coverage of the scene (Fig. 6–8). This is presumably the result of the global attention mechanism, which analyzes the dependences between all pixels simultaneously. However, the high number of false matches and large variance of reprojection error (Table 3) may be explained by the fact that the model was trained on optical data and is too sensitive to low-contrast sonar images of a different nature.

Unlike the approach in [3], where CNN is used to reduce false positives in the change map, this study directly applied deep learning models to feature matching, providing more detailed geometric assessment. Compared to the classical gradient-based SIFT method, which provides accuracy under high contrast conditions, the combination of SuperPoint + LightGlue demonstrated better stability even on noisy images. The results of our study indicate the need for adaptation of deep models' parameters to acoustic data, since without such adaptation, the accuracy and quality of matching may be insufficient.

Thus, this study addresses the lack of systematic comparison of detectors and matchers with different architectures (classical, convolutional, and transformer-based) on real-world SSS images. For the first time, a comprehensive quantitative assessment (Tables 2–4) and qualitative analysis of matching and change maps (Fig. 7, 8) for two datasets were performed, which paves the way to practical recommendations. In particular, SuperPoint + LightGlue can be considered the best compromise for real-time tasks, while LoFTR is appropriate for offline analysis with a priority on completeness of coverage and matching. Its resource intensity is explained by its transformer architecture, that performs matching over the entire image grid, not only over a limited set of keypoints.

Thus, our work reasonably achieves the initial objective: the suitability of modern deep learning models for sonar tasks has been assessed and their real limitations have been identified.

It is worth noting that the study was limited by the lack of ground-truth matches, hence the evaluation used indirect metrics. In addition, the test datasets had a relatively small volume (two missions), and did not cover the variety of bottom types. Also, the use of pre-trained models without domain

adaptation to acoustic data reduced their ability to generalize, and the computational complexity of LoFTR (Table 4) limits its application in autonomous vehicles.

The research limitations also include the simplification of the seabed geometry model (homographic approximation), that ignores its three-dimensional structure. Besides, transformer-based LightGlue point matcher was not compared with classical matchers (BFMatcher, FLANN) that would make the analysis more objective.

Further research should cover the domain-specific fine-tuning of the research deep models on actual sonar data, as well as the development of more advanced match filtering algorithms, that could incorporate a priori data (sonar position and orientation, information about the bottom material, etc.). The design of hybrid convolutional-transformer architectures considering the physical properties of the acoustic signal seems promising. Also, it is necessary to compare matching methods on more heterogeneous data sets and bottom types (rocky bottom, vegetation, variable depth), and the use of the such methods and change maps in automatic object detection tasks.

7. Conclusions

1. Within the scope of the study, two real-world SSS image datasets were generated – *Atlantic* and *Baltic*, based on the data collected by the EvoLogics Sonobot 5 USV. Image pairs from the reference and matching missions were aligned with high accuracy: over 90% of ping pairs were within 0.5 m. The final ping mismatch did not exceed 2–3 pixels. The obtained data correspond to real conditions in a typical SSS search missions with significant speckle and stripe noise, homogeneous bottom, and a small number of prominent features. This favorably distinguishes them from synthetic data often used in modern studies and provides a qualitatively relevant assessment of the selected methods.

2. Comparative analysis on the generated datasets revealed significant differences in the behavior of the methods under noisy (Atlantic) and low-noise (Baltic) conditions. On the low-noise dataset, all methods provided high stability and geometric accuracy of the matches: the reprojection error mean was 3.9–8.8 pixels, and the percentage of inliers – from 57% to 65.6% for all models. LoFTR showed a high result of matches (64.3% inliers) with the largest scene coverage (97%). SuperPoint + LightGlue achieved the lowest error mean (3.9 pixels with a standard deviation of 2.7 pixels) and the highest percentage of inliers (65.6%). On the other hand, on the high-noise Atlantic dataset, a significant drop in the quality was observed for all methods. Classical SIFT and convolutional DISK generated only several matches (18–66) on average per image pair, and a low percentage of inliers (18.6–27.1%), which led to unstable alignment. LoFTR produced the highest number of matches (472) and maximum coverage (96%), but the percentage of filtered matches was the lowest (8%), which led to high reprojection error (141.7 ± 184.2 pixels). The most balanced result on noisy data was shown by SuperPoint + LightGlue, with the smallest displacement mean (35.9 pixels), the lowest reprojection error mean (36.0 pixels), and the highest inlier percentage (43.4%). Therefore, SuperPoint + LightGlue was the best method according to the adopted metrics.

3. In terms of resource usage, the classical and convolutional methods remained the most economical (< 0.06 s per image pair and up to 532 MB of peak memory per set). At the same time, DISK + LightGlue was the best on the Atlantic set (0.034 s per frame), and the worst in terms of peak memory on both sets (532 MB and 498 MB GPU, respectively). LoFTR turned out to be the most resource-greedy (up to 0.86 s per frame, up to 4 GB of GPU memory per set). Thus, no fundamental difference between the classical and convolutional algorithms in resource consumption was found. Instead, as expected, the transformer-based LoFTR spends an order of magnitude more memory and time on the same tasks.

4. Qualitative analysis of the maps of matches and changes confirmed that methods' effectiveness depends on the level of image noise. Some edge cases of the methods were also found. On the Baltic set, all algorithms generated sufficiently stable matches and interpreted change maps. This allows to recommend the use of classical and convolutional approaches under simple conditions. On the noisy Atlantic, classical and convolutional methods detected relatively few points and created low density of matches, making it impossible to generate high-quality change maps for most pairs. LoFTR, on the other hand, provided a high number of matches and dense coverage, but most of them were unstable. This can be explained by the high sensitivity of transformers to noise artifacts and the lack of additional filtering of matches for consistency. It can be concluded that for noisy underwater images, the use of LoFTR with adaptation to sonar data and additional filtering of matches by coherence seems promising, and could make it possible to combine its coverage and stability of the results.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

The data will be provided upon reasonable request.

Use of artificial intelligence

The authors declare the use of generative AI in the research and manuscript preparation process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to generative AI tools under full human supervision:

– code optimization;

– proofreading and editing.

Generative AI tool used: Chat-GPT5.

The authors bear full responsibility for the final manuscript.

Generative AI tools are not listed as authors and are not responsible for the final results.

Acknowledgments

The authors would like to thank EvoLogics GmbH for providing equipment, data, and technical support.

Authors' contributions

Oleksandr Katrusha: Conceptualization, Methodology, Software, Data curation, Formal analysis, Visualization, Writing – original draft; **Dmytro Prylipko**: Conceptualization, Data curation, Supervision; **Kostiantyn Yefremov**: Methodology, Project administration, Writing – review & editing.

References

1. Zhou, X., Yuan, S., Yu, C., Li, H., Yuan, X. (2022). Performance Comparison of Feature Detectors on Various Layers of Underwater Acoustic Imagery. Journal of Marine Science and Engineering, 10 (11), 1601. https://doi.org/10.3390/jmse10111601
2. Shafique, A., Cao, G., Khan, Z., Asad, M., Aslam, M. (2022). Deep Learning-Based Change Detection in Remote Sensing Images: A Review. Remote Sensing, 14 (4), 871. https://doi.org/10.3390/rs14040871
3. Steiniger, Y., Schröder, S., Stoppe, J. (2022). Reducing the false alarm rate of a simple sidescan sonar change detection system using deep learning. 22nd International Symposium on Nonlinear Acoustics, 48, 070022. https://doi.org/10.1121/2.0001642
4. Hedlund, W. (2024). Change Detection in Synthetic Aperture Sonar Imagery Using Segment Anything Model. Linköping University. Available at: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1877741&dswid=6083
5. Westman, E., Hinduja, A., Kaess, M. (2018). Feature-Based SLAM for Imaging Sonar with Under-Constrained Landmarks. 2018 IEEE International Conference on Robotics and Automation (ICRA), 3629–3636. https://doi.org/10.1109/icra.2018.8461004
6. Alcantarilla, P., Nuevo, J., Bartoli, A. (2013). Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. Procedings of the British Machine Vision Conference 2013, 13.1-13.11. https://doi.org/10.5244/c.27.13
7. Myers, V., Saebo, T. O., Hansen, R. (2018). Comparison of Co-Registration Techniques for Synthetic Aperture Sonar Images from Repeated Passes. Synthetic Aperture Sonar and Radar, 40 (2).
8. Myers, V., Quidu, I., Zerr, B., Sabo, T. O., Hansen, R. E. (2020). Synthetic Aperture Sonar Track Registration With Motion Compensation for Coherent Change Detection. IEEE Journal of Oceanic Engineering, 45 (3), 1045–1062. https://doi.org/10.1109/joe.2019.2909960
9. Zhou, X., Yu, C., Yuan, X., Luo, C. (2021). Matching Underwater Sonar Images by the Learned Descriptor Based on Style Transfer Method. Journal of Physics: Conference Series, 2029 (1), 012118. https://doi.org/10.1088/1742-6596/2029/1/012118
10. Zhang, J., Xie, Y., Ling, L., Folkesson, J. (2023). A fully-automatic side-scan sonar simultaneous localization and mapping framework. IET Radar, Sonar & Navigation, 18 (5), 674–683. https://doi.org/10.1049/rsn2.12500
11. Midtgaard, O., Hansen, R. E., Saebo, T. O., Myers, V., Dubberley, J. R., Quidu, I. (2011). Change detection using Synthetic Aperture Sonar: Preliminary results from the Larvik trial. OCEANS'11 MTS/IEEE KONA, 1–8. https://doi.org/10.23919/oceans.2011.6107272
12. G-Michael, T., Marchand, B., Tucker, J. D., Marston, T. M., Sternlicht, D. D., Azimi-Sadjadi, M. R. (2016). Image-Based Automated Change Detection for Synthetic Aperture Sonar by Multistage Coregistration and Canonical Correlation Analysis. IEEE Journal of Oceanic Engineering, 1–21. https://doi.org/10.1109/joe.2015.2465631
13. Gode, S., Hinduja, A., Kaess, M. (2024). SONIC: Sonar Image Correspondence using Pose Supervised Learning for Imaging Sonars. 2024 IEEE International Conference on Robotics and Automation (ICRA), 3766–3772. https://doi.org/10.1109/icra57147.2024.10611678
14. Lindenberger, P., Sarlin, P.-E., Pollefeys, M. (2023). LightGlue: Local Feature Matching at Light Speed. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 17581–17592. https://doi.org/10.1109/iccv51070.2023.01616
15. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D. B. (2009). PatchMatch. ACM Transactions on Graphics, 28 (3), 1–11. https://doi.org/10.1145/1531326.1531330
16. Zhang, J., Xie, Y., Ling, L., Folkesson, J. (2025). A Dense Subframe-Based SLAM Framework With Side-Scan Sonar. IEEE Journal of Oceanic Engineering, 50 (2), 1087–1102. https://doi.org/10.1109/joe.2024.3503663
17. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X. (2021). LoFTR: Detector-Free Local Feature Matching with Transformers. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8918–8927. https://doi.org/10.1109/cvpr46437.2021.00881
18. DeTone, D., Malisiewicz, T., Rabinovich, A. (2018). SuperPoint: Self-Supervised Interest Point Detection and Description. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). https://doi.org/10.1109/cvprw.2018.00060
19. Tyszkiewicz, M. J., Fua, P., Trulls, E. (2020). DISK: Learning local features with policy gradient. arXiv. http://doi.org/10.48550/ARXIV.2006.13566
20. Culjak, I., Abram, D., Pribanic, T., Dzapo H., Cifrek, M. (2012). A brief introduction to OpenCV. 2012 Proceedings of the 35th International Convention MIPRO, 1725–1730. Available at: https://ieeexplore.ieee.org/document/6240859/authors#authors
21. Gutiérrez, G., Torres-Avilés, R., Caniupán, M. (2024). cKdtree: a Compact Kdtree for Spatial Data. Alberto Mendelzon Workshop on Foundations of Data Management. Available at: https://www.semanticscholar.org/paper/cKdtree%3A-a-Compact-Kdtree-for-Spatial-Data-Guti%C3%A9rrez-Torres-Avil%C3%A9s/1106acf86126909ec2ad8ffa174cbd6e4dcca329