Journal of
Marine Science and Engineering

MDPI

Article

Side-Scan Sonar Image Matching Method Based on Topology Representation

Dianyu Yang [1], Jingfeng Yu [2,3], Can Wang [1], Chensheng Cheng [1], Guang Pan [1], Xin Wen [1] and Feihu Zhang [1,*]

[1] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China; 2019100496@mail.nwpu.edu.cn (D.Y.); wangcan2017@mail.nwpu.edu.cn (C.W.); chengchensheng@163.com (C.C.); guang.pan@nwpu.edu.cn (G.P.); wenxin666@mail.nwpu.edu.cn (X.W.)
[2] Ganjiang Innovation Academy, Chinese Academy of Sciences, Ganzhou 341000, China; jfy@gia.cas.cn
[3] Department of Automation, Tsinghua University, Beijing 100190, China
[*] Correspondence: feihu.zhang@nwpu.edu.cn; Tel.: +86-155-9661-1656

Abstract: In the realm of underwater environment detection, achieving information matching stands as a pivotal step, forming an indispensable component for collaborative detection and research in areas such as distributed mapping. Nevertheless, the progress in studying the matching of underwater side-scan sonar images has been hindered by challenges including low image quality, intricate features, and susceptibility to distortion in commonly used side-scan sonar images. This article presents a comprehensive overview of the advancements in underwater sonar image processing. Building upon the novel SchemaNet image topological structure extraction model, we introduce a feature matching model grounded in side-scan sonar images. The proposed approach employs a semantic segmentation network as a teacher model to distill the DeiT model during training, extracting the attention matrix of intermediate layer outputs. This emulates SchemaNet's transformation method, enabling the acquisition of high-dimensional topological structure features from the image. Subsequently, utilizing a real side-scan sonar dataset and augmenting data, we formulate a matching dataset and train the model using a graph neural network. The resulting model demonstrates effective performance in side-scan sonar image matching tasks. These research findings bear significance for underwater detection and target recognition and can offer valuable insights and references for image processing in diverse domains.

Keywords: sidescan sonar; semantic segmentation; topological features; attention mechanism; image matching

check for updates

Citation: Yang, D.; Yu, J.; Wang, C.; Cheng, C.; Pan, G.; Wen, X.; Zhang, F. Side-Scan Sonar Image Matching Method Based on Topology Representation. *J. Mar. Sci. Eng.* **2024**, *12*, 782. https://doi.org/10.3390/jmse12050782

Academic Editor: Rouseff Daniel

Received: 6 March 2024
Revised: 30 April 2024
Accepted: 4 May 2024
Published: 7 May 2024

1. Introduction

In recent decades, side-scan sonar technology has emerged as an indispensable tool in marine science research owing to its efficient application in ocean exploration and underwater environmental monitoring. Side-scan sonar is capable of generating high-resolution images of underwater topography, which plays a pivotal role in mapping benthic habitats, classifying underwater objects, and conducting detailed studies of seafloor geomorphology [1–3]. However, the complexity of side-scan sonar images has historically posed challenges in extracting and interpreting their information [4]. The rapid advancement of information processing and image feature extraction technologies in recent years has offered new avenues for addressing this challenge. Numerous algorithms have demonstrated outstanding performance in image recognition and matching [5–7], opening up new possibilities for the automatic processing and analysis of side-scan sonar images [8].

In the realm of side-scan sonar image analysis, the task of matching images, specifically aligning side-scan sonar images of the same location from diverse angles, has perennially posed a formidable challenge. This challenge stems from many factors, encompassing the inherent low quality of side-scan sonar images, pronounced distortion, and the intricate nature of the features depicted [9,10]. These complexities render the extraction of features and subsequent image matching an exceptionally demanding endeavor for side-scan sonar images. Contemporary research predominantly revolves around feature extraction and matching using traditional feature-matching algorithms, employing clustering for feature extraction and matching, and harnessing machine learning for the extraction and matching of features [11].

Feature extraction and matching based on traditional algorithms have been among the earliest research endeavors in related fields, demonstrating superior performance in relatively simplistic environments with distinct landmarks. However, their efficacy in complex settings heavily depends on the design of the matching algorithms. Among the most widely employed are feature descriptors such as Scale-Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF), and the more recently developed KAZE. These algorithms furnish robust tools for image matching, and are particularly adept at handling images in challenging environmental conditions. They facilitate the identification of similar feature points across disparate images, thereby enabling precise image registration and object recognition. Ghate and Nikose (2021) [12] analyzed the repeatability of SIFT and SURF descriptors with the aim of optimizing preprocessing methods for underwater images by refining the extraction process of key feature points. Additionally, Pourfard et al. (2021) [13] proposed a method that amalgamates the KAZE algorithm with a modified SURF descriptor for the registration of Synthetic Aperture Radar (SAR) images, specifically targeting the issue of speckle noise. In the realm of 2D object recognition, Bansal et al. (2021) [14] conducted a comparative study, scrutinizing SIFT, SURF, and ORB feature descriptors to evaluate their performance across diverse scenarios. Lozano-Vázquez et al. (2022) [15] explored an array of image enhancement and feature extraction methods, with a focus on reliably extracting image features such as SIFT and SURF under low-light conditions. Finally, Shaharom et al. (2023) [16] presented a comprehensive review of the application of the SIFT and SURF algorithms in multispectral image matching, delving into various research strategies aimed at augmenting the performance of these algorithms in processing multispectral images. It is evident that the concerted efforts employing different feature descriptors enable the extraction of image features to a considerable extent. However, in underwater environments significant manual optimization is necessary to ensure a requisite degree of accuracy.

Clustering algorithms have demonstrated significant value and efficiency in sonar image processing and target recognition. Mainstream clustering algorithms categorize data points into multiple clusters, each consisting of data points with similar characteristics, thereby simplifying subsequent data analysis. In sonar image processing, K-means clustering proves effective in classifying and recognizing features within images, facilitating the identification and categorization of underwater objects and regions. Anilkumar et al. (2019) [17] proposed an algorithm that incorporates an enhancement technique based on Contrast Limited Adaptive Histogram Equalization (CLAHE) coupled with clustering algorithms to track underwater cables and enhance image quality, consequently enhancing the accuracy of line detection. This approach enhances the visibility of underwater images, presenting a viable technical solution for cable tracking. Additionally, Vikas (2017) [18] introduced an image processing algorithm utilizing fuzzy C-means clustering to identify and eliminate shadows in side-scan sonar images. Through cluster analysis, this technique effectively distinguishes shadows from actual objects in the images, thereby enhancing the efficiency and accuracy of sonar image processing [19].

The underwater target detection algorithm proposed by Li et al. (2022) [20] utilizes an improved YOLOv4 model. It incorporates K-means++ clustering for anchor boxes to adapt to underwater scene characteristics and employs image enhancement techniques to enhance detection performance. This method demonstrates significant advantages in improving the accuracy and real-time performance of underwater target detection. In another study, Xinyu et al. (2017) [21] investigated a sonar image processing method that integrates k-means clustering-based image segmentation for denoising and segmenting sonar images. This approach effectively distinguishes targets, background, and noise within the images, providing an effective tool for sonar image analysis. Additionally, Rajput et al. (2022) [22] examined the utilization of spider search algorithms for noisy sonar image segmentation coupled with clustering methods to differentiate objects, shadows, and background. By simulating the principles of natural selection and genetic algorithms, this method optimizes the clustering process, thereby enhancing the efficiency and accuracy of sonar image segmentation.

An overview of the literature reveals the extensive application and significance of the k-means clustering algorithm in sonar image processing. Nevertheless, these studies predominantly employ clustering directly on the original sonar images, often following noise reduction at most. Consequently, the outputs frequently exhibit numerous outliers, lack robustness, and demonstrate significant variability when there are changes in the viewpoint of the images, rendering them ineffective in complex underwater environments.

With the advancement of data science, the application of deep learning technologies in sonar image processing and underwater environment monitoring has become a key driving force in the development of this research field. Due to their excellent performance in image recognition, classification, and feature extraction, deep learning models are widely used for automatic identification, tracking, and detection of sonar targets. Neupane and Seok (2020) [23] reviewed the latest advancements in deep learning algorithms for automatic sonar target recognition, detailing key information on data acquisition, datasets, and hyperparameters, providing a reference for research in this field. Moreover, Li et al. (2020) [24] explored the effectiveness of deep learning frameworks in marine remote sensing image classification, demonstrating how these frameworks can extract information from marine remote sensing images through eight typical applications. Misiuk et al. (2023) [25] reviewed the relevant developments of underwater surveying and mapping technology in the past 30 years and pointed out the future development directions of related technologies. In underwater object classification, Gav et al. (2015) [26] focused on using ultrasonic technology for classification, highlighting the effectiveness of different methods such as Bayesian and Principal Component Analysis (PCA) in achieving underwater object classification. These studies showcase the application of deep learning in sonar data analysis and highlight its significant role in marine science research and underwater environment monitoring.

The study by Gašparović et al. (2022) [27] proposed an automated deep learning approach for detecting underwater pipelines and compared the performance of six different CNN detectors for underwater object detection. Singh and Bhat (2021) [28] conducted a review of deep learning-based algorithms for enhancing underwater images, focusing on the latest methodologies, datasets, and evaluation criteria, then compared the outcomes of these algorithms. These investigations underscore the potential of deep learning technologies in ameliorating the quality of underwater imaging and image enhancement. While the utilization of deep learning technologies in the underwater domain has been steadily increasing in recent years, its prevalence still lags behind that in terrestrial and aerial image processing. This discrepancy can be attributed to the relative difficulty of acquiring sonar data, which results in a smaller total volume of available data, posing challenges in meeting the data requirements of models. Consequently, more research has been directed towards local feature extraction and image quality enhancement, with fewer studies focusing on extracting global features from sonar images and performing image matching.

Summarizing the above, research on feature extraction and image matching in the underwater sonar image domain remains scarce and is in a relatively nascent stage. Additionally, due to the complexity of underwater environments, single feature extraction methods struggle to achieve very good results, and a combination of multiple methods is the mainstream direction for future development. Currently, feature extraction and matching methods have emerged based on the bag-of-words model combined with clustering and feature descriptors [29], along with image matching methods that utilize machine learning and semantic information to extract image topological representations [30], both of which have achieved relatively good results. Although these studies have not been applied in the underwater domain, they can provide inspiration.

While delving into research in related fields, we encountered a particularly enlightening article by H. Zhang, M. Xue, et al. (2023) [31]. Their paper introduces SchemaNet, a novel reasoning paradigm characterized by pattern reasoning, with the aim of producing interpretable predictions through reconstructing the forward pass of a Deep Neural Networks (DNN). A salient aspect of the paper lies in its utilization of a diverse array of methodologies, including deep neural networks, clustering algorithms, topological feature extraction, and graph neural networks, to facilitate multidimensional feature extraction from images. This encompasses spatial, semantic, and attentional features. The integration of features from various dimensions significantly enhances the efficacy of image matching tasks.

Their study centered on evaluating the efficacy of interpretable models for classification tasks. Expanding upon this, our research applies these models to the domain of underwater sonar image matching. We utilize raw sonar images as inputs for our custom semantic segmentation model, yielding segmentation results that incorporate both semantic and spatial information. Subsequently, we extract image topological structures from these results. Finally, we feed these data into a graph neural network and compute the correlations among the network outputs to accomplish image matching.

The primary contributions of this article can be summarized as follows:

1. We employ a self-designed CNN-based semantic segmentation model as the teacher model for training the DeiT model and conducting knowledge distillation. This process involves obtaining intermediate layer outputs and self-attention matrices.
2. Based on SchemaNet, we utilize clustering methods to cluster data from intermediate layers into multiple visual words. These words serve as nodes in the image's topological structure. Additionally, we integrate self-attention matrices to calculate the weights of nodes and their interaction coefficients, which act as the edges in the image's topological structure.
3. We extract feature vectors from the image's topological structure using a graph neural network and transform the classification problem into an image matching problem. This approach allows the graph neural network to be trained to output the topological similarity between different images as the criterion for image matching.

2. Methodology and Method

2.1. DeiT and Knowledge Distillation

Introduced by A. Vaswani, the transformer [32] model has emerged as the predominant model in deep learning owing to its outstanding performance. Initially designed for natural language processing tasks, subsequent efforts by researchers led to the development of the Vision Transformer (ViT) [33] model, effectively extending the transformer architecture to the realm of image processing. The transformer architecture departs from the conventional convolutional structures prevalent in traditional deep neural networks, opting instead for an abundance of attention matrices to extract image features. Essentially, this approach can be considered equivalent to convolutional kernels of variable sizes, resulting in a notably high performance ceiling. However, it is essential to note that the transformer model's effectiveness is contingent on extensive training with substantial datasets. For instance, the ViT model relies on the proprietary JFT-300M dataset, comprising 300 million images, an impractical scale for general tasks, thereby posing challenges to the widespread application of the model.

DeiT (Data-efficient Image Transformer) [34] was developed to mitigate the challenge of extensive data pretraining requirements encountered by traditional transformer-based visual models such as ViT. DeiT introduces a technique called "knowledge distillation", wherein a proficient high-performance model (referred to as the teacher model) guides the training of a smaller model (known as the student model). Through this approach, DeiT, acting as the student model, can assimilate richer and more refined information from the teacher model, typically a larger well-trained CNN, without the necessity of directly assimilating vast datasets.

In Figure 1, the class token is used to represent the semantic category of input information, which is input into the model in parallel with the embedded patch sequence. This segmentation is necessary in order to align with the sequence-form input required by the transformer structure. The distillation token is a concept introduced during the training process for knowledge distillation. It is utilized to capture knowledge from the teacher model to assist in training the student model. By incorporating a special token into the input sequence, the model can effectively concentrate on essential information from the teacher model.

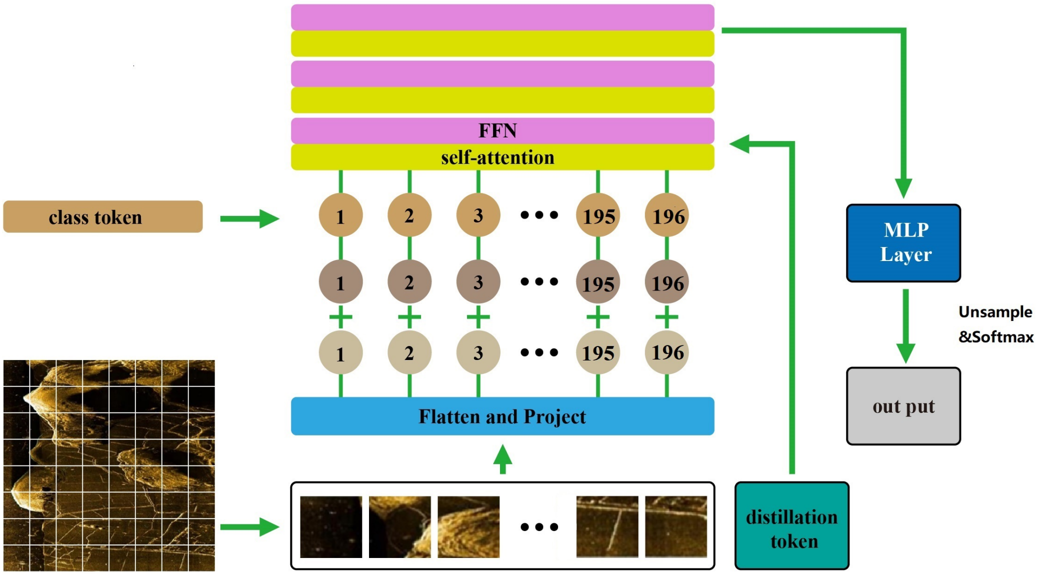

Figure 1. The DeiT model architecture applied to our task.

The Feed Forward Network (FFN) constitutes a feedforward neural network comprising two linear layers intertwined with a nonlinear activation function. Typically, a residual connection and layer normalization are incorporated between these linear layers to facilitate nonlinear transformation and feature mapping at each position. Self-attention stands as a fundamental component within the transformer model, employed for computing attention weights across distinct positions within the input sequence. It encompasses three linear transformations (query, key, and value) followed by the calculation of attention scores through the softmax function, culminating in the multiplication of these scores by the value to derive the final representation.

The total loss incurred during the training of the DeiT model comprises two components, which correspond to self-supervised learning and knowledge distillation, respectively. These are delineated as follows:

(1) Self-supervised learning loss aims to enable the model to learn image feature representations in an unsupervised manner, denoted as $\mathcal{L}_{\text{CE}}$ in the figure. Here, Lce corresponds to the cross-entropy loss. If we assume the model's predicted distribution for image patches as p and the true label distribution as q, then the formula for $\mathcal{L}_{\text{CE}}$ is

$$\mathcal{L}_{\text{CE}} = -\sum_{i} q_i \log(p_i) \tag{1}$$

(2) Knowledge distillation loss is integral to the process of knowledge distillation, playing a pivotal role in shaping the model's learning.

$$\mathcal{L}_{\text{teacher}} = \tau^2 \text{KL}\big(\psi(Z_s/\tau), \psi(Z_t/\tau)\big) \tag{2}$$

The temperature parameter, denoted as τ, plays a crucial role in regulating the smoothness of the probability distribution during the distillation process. The calculation of the Kullback–Leibler (KL) divergence is represented by KL. The output probability distributions of the student model and the teacher model, denoted as $\psi(Z_s/\tau)$ and $\psi(Z_t/\tau)$, respectively, undergo a temperature adjustment. This adjustment involves dividing the original Z_s and Z_t by the temperature parameter τ and subsequently transforming them into probability distributions using the softmax function, denoted as ψ.

The combined total loss incorporating both parts is as follows, with λ representing the weight parameter:

$$\mathcal{L}_{\text{global}} = (1-\lambda)\mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{teacher}}. \tag{3}$$

In light of the limited volume of side-scan sonar data, which typically only reaches into the thousands, we employed a self-designed semantic segmentation network architecture rooted in Convolutional Neural Networks (CNNs) [35] as the reference model for the Data-efficient Image Transformer (DeiT). By means of distillation training, we attained superior results compared to those obtained with the transformer architecture.

The configuration of our model architecture, depicted in Figure 2, comprises an encoder and a decoder each consisting of four layers, akin to the U-Net architecture. The encoder of the model is segmented into upper and lower channels. In light of the relatively large size of the utilized sonar images, the primary channel at the bottom employs large convolution kernels to accommodate a broad receptive field, while the upper channel employs smaller convolution kernels to compensate for this. Additionally, the main channel integrates SE-blocks to modulate attention weights across diverse channels, thereby augmenting segmentation accuracy. Subsequently, the decoder amalgamates the output features from both channels and employs transposed convolution for downsampling, ultimately yielding pixel-level classification results. We employ this model as the teacher model, not exclusively for enhancing the semantic segmentation performance of DeiT but rather to extract the self-attention matrices from its intermediate layers. Post-distillation, the DeiT model exhibits pixel-level classification capabilities analogous to those of the teacher model. This approach additionally ensures that the self-attention matrices generated from its intermediate layers encompass pertinent information regarding the semantic features.

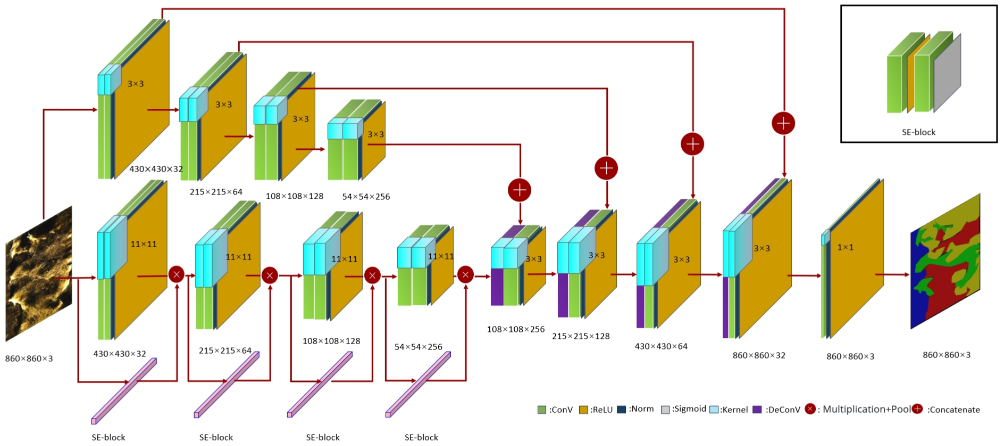

Figure 2. Semantic segmentation model.

2.2. Extraction of Image Topological Structure Based on Attention

Traditional feature matching methods, exemplified by SIFT and SURF, exhibit suboptimal robustness when confronted with sonar images, which frequently manifest substantial distortion. Similarly, matching methods grounded in the bag-of-words model prove inadequate in underwater environments, as they lack the diverse and semantically rich visual words prevalent in urban settings. Consequently, an adept matching method for underwater sonar images requires robustness and the ability to assimilate information pertaining to location, semantics, and interactions within the image for achieving heightened accuracy. This semantic information is encapsulated in the output features of a semantic segmentation network. Through clustering, a sequence of feature nodes with distinct characteristics can be derived. By leveraging these feature nodes as the vertices of a topological structure and the inter-node relationships as edges, an effective topological structure for image matching can be constructed.

When the training of the DeiT network concludes, the subsequent step involves extracting the output features from the layer immediately preceding the activation function. These features then undergo discretization. This discretization is specifically achieved by applying k-means clustering to a set of visual tokens extracted from the dataset. The goal is to construct a visual vocabulary with a predefined size M. In this context, M is deliberately set to 128. This process effectively quantizes the continuous feature space into a discrete set of tokens. This transformation facilitates a more structured and interpretable representation of the image content, offering potential benefits for a variety of tasks, including image matching and retrieval.

$$\text{Ingredient}(x) = \underset{i \in \{1,\dots,M\}}{\text{argmin}} \parallel x - \omega_i \parallel_2 \tag{4}$$

The term Ingredient (x) denotes the index of the visual word ω, with x representing the extracted deep output features, excluding both the class token and the distillation token.

After discretizing the features, the resulting discretized sequence is denoted as $\tilde{X} = (\tilde{x}_1, \dots, \tilde{x}_n)$. Simultaneously, the sequence of nodes, referred to as visual words ω and represented as V when functioning as nodes, is unique as well: $V = \text{Unique}(\tilde{X})$. Each node contains both its index (i.e., the vertex's ordinal number) and weight, where the weight is determined by two criteria, namely, its contribution to the prediction probability of the network model and its proportion among all visual words. The formula for calculating the weight is as follows:

$$\lambda_\nu = \alpha_1 \sum_{i \in \Xi(\nu|\tilde{X})} \psi_i^{\text{CLS}} + \alpha_2 |\Xi(\nu|\tilde{X})|. \tag{5}$$

The symbol $|\Xi(\nu|\tilde{X})|$ signifies the frequency of occurrence of the node indexed by v within all nodes in the image. The set $\Xi(\nu|\tilde{X})$ denotes the collection of nodes. The notation ψ_i^{CLS} represents the attention matrix corresponding to the i-th patch concerning the class token. Additionally, $\alpha_{1,2} > 0$ denotes predefined weights designed to balance these two terms. The component preceding λ_ν signifies the contribution to the prediction probability of the network model, while the latter component indicates its proportion among all visual words.

The connections among vertices, representing the edges of the topology, are established through the calculation of interactions between these vertices. For any pair of vertices $u, v \in V$ with distinct indices, the edge weight is determined based on two criteria: the mean attention among pairs of diverse nodes and their spatial adjacency relationship. The formula for computing the average attention between pairs of distinct nodes is as follows:

$$e_{u,v}^{\text{attn}} = \frac{1}{|\Pi[(u,v)|\tilde{X}]|} \sum_{(i,j) \in \Pi[(u,v)|\tilde{X}]} \Psi_{i,j}^V. \tag{6}$$

In this context, $\Pi[(u,v)|\tilde{X}]$ denotes the Cartesian product of the sets $\Xi(\nu|\tilde{X})$ and $\Xi(u|\tilde{X})$. Additionally, $\Psi_{i,j}^V$ represents the attention matrix between distinct patches i and j relative to the patch tokens. It is important to note that the model's output attention matrix comprises submatrices for various tokens.

Similarly, the adjacency is defined as

$$e_{u,v}^{\text{adj}} = \frac{1}{|\Pi[(u,v)|\widetilde{\mathrm{X}}]|} \sum_{(i,j)\in\Pi[(u,v)|\widetilde{\mathrm{X}}]} \frac{1}{\epsilon + \|\mathrm{Pos}(i) - \mathrm{Pos}(j)\|_2}. \tag{7}$$

In this context, the function $Pos(i)$ is defined to return the original 2D coordinates of the input visual tokens relative to the model output feature plane after flattening. The interaction between vertices u and v is ultimately represented as a weighted sum of two components, with β_1 and β_2 serving as predefined weights.

$$e_{u,v} = \beta_1 e_{u,v}^{\text{attn}} + \beta_2 e_{u,v}^{\text{adj}} \tag{8}$$

After processing, the characteristics extracted from the intermediate layers of the model are converted into a sequence of nodes comprising indices and weights, accompanied by edges with weights that reflect the interactions between various pairs of nodes. Subsequently, these topological features are employed for matching.

2.3. Image Matching Based on Graph Neural Networks

We employ Graph Convolutional Network (GCN) [36] as the feature extractor for the graph (as shown in Figure 3), representing its topological structure. To facilitate the integration of graph information into the GCN module, we assign a trainable embedding vector to each node in the topological graph. Each vector is initialized independently, drawn from a multivariate Gaussian distribution $N(0, I_d)$.

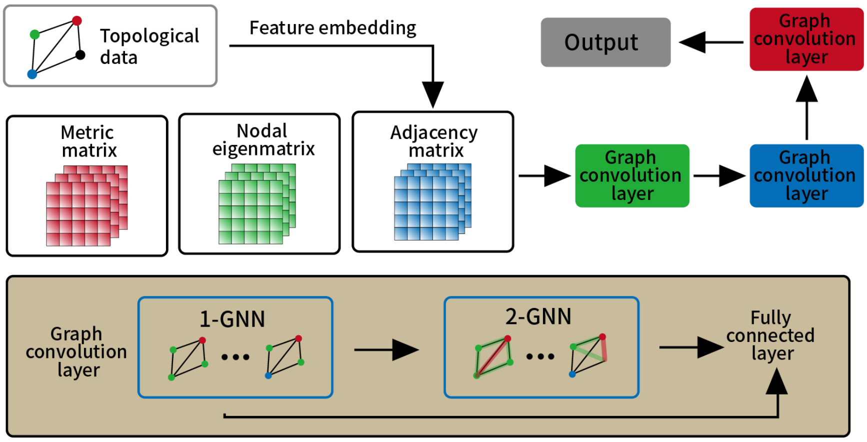

Figure 3. Graph neural network model. Each graph convolutional network layer contains two layers of information aggregation structure, the first aggregating node information and the second aggregating edge information.

The network output of the GCN is

$$\mathrm{OutPut}(F) = \mathrm{N}(\sigma((I_d + E)FW)), \tag{9}$$

where F represents the topological features used as input to the Graph Neural Network (with its dimension being $|V| \times d$, where V represents the number of nodes and d represents the dimension of the embedding vectors), σ denotes a nonlinear activation function, N stands for feature normalization, W is a trainable parameter matrix, and E represents edge weights. After passing through all the GNN layers, the features are aggregated through weighted average pooling. Unlike the GCN in the diagram, the final output is not classified through the results output by an MLP; rather, a feature vector z is obtained. When data are input into the network model, a series of vector outputs $Z = (z_1, \ldots, z_n)$ are obtained, with n being the number of samples in the dataset. When an image is processed entirely to obtain its output z_i ($i \in n$), the final prediction logic is defined by computing the inner product similarity, with $y = z_{gt} z_i^{\top}$ and with z_{gt} representing the output of the ground truth.

3. Experiment

All experiments were carried out on an Intel Core i9-10900F CPU @ 2.8 GHz × 20 with 64 GB of RAM. Additionally, an NVIDIA GeForce 3090 GPU with 24 GB of VRAM was utilized. The experimental setup included CUDA Toolkit 11.3, CUDNN V8.2.1, Python 3.9, and PyTorch-GPU 1.10.1, all running on the Ubuntu 20.04 operating system.

3.1. Dataset Collection

The sonar image dataset was collected by the HYDRO 3060 side-scan sonar in Qiandao Lake, Jiande, Hangzhou, China. Part of the publicly available data has been uploaded to this GitHub repository (https://github.com/YDY-andy/Sonar-dataset, accessed on 6 July 2023). The collected sonar data underwent processing using software, converting the data from XTF format to waterfall video and then obtaining sonar data in image format through frame-by-frame sampling. Our team developed and manufactured the sonar equipment, which was equipped on an underwater robot. Additionally, the Autonomous Underwater Vehicle (AUV) was outfitted with inertial navigation, Doppler, global positioning systems, and ultra-short baseline positioning systems, enabling manual cruising by remote control or following a pre-established program trajectory (Figure 4). The AUV collected all data for this paper while employing a slow and constant speed cruising at an altitude of 10 m underwater, with a maximum distance of 50 m from the bottom. Figure 4 depicts the underwater robot used for data collection. The original image data size was 960 × 900 pixels, with an example of the original sonar image presented in Figure 5.

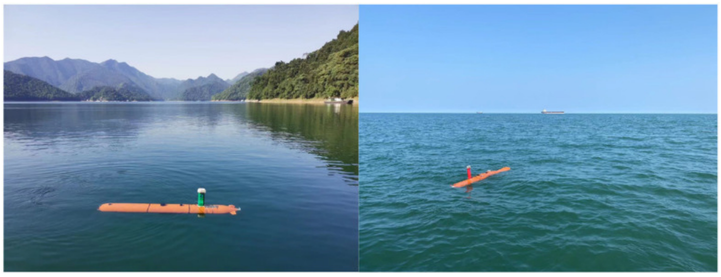

Figure 4. Self-developed AUV for data collection.

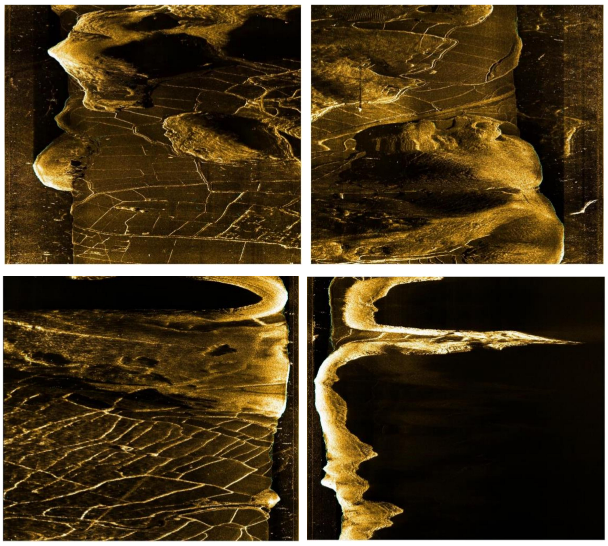

Figure 5. The original sonar image (each sonar image is cropped down the middle into two images in order to increase data volume and reduce image size).

The training of neural network models necessitates labeled data serving as ground truth; thus, annotating the sonar data becomes imperative. The LabelMe labeling software (version 5.01) developed by the Massachusetts Institute of Technology served as the tool for image labeling. A total of five categories were assigned to the sonar data labels: (1) water; (2) seamounts; (3) ground (formerly cultivated land); (4) shadowed areas; and (5) unlabeled areas (background). The term "unlabeled areas" primarily denotes the fragmented regions remaining after labeling the images with the first four categories. The labeled images are depicted in Figure 6.

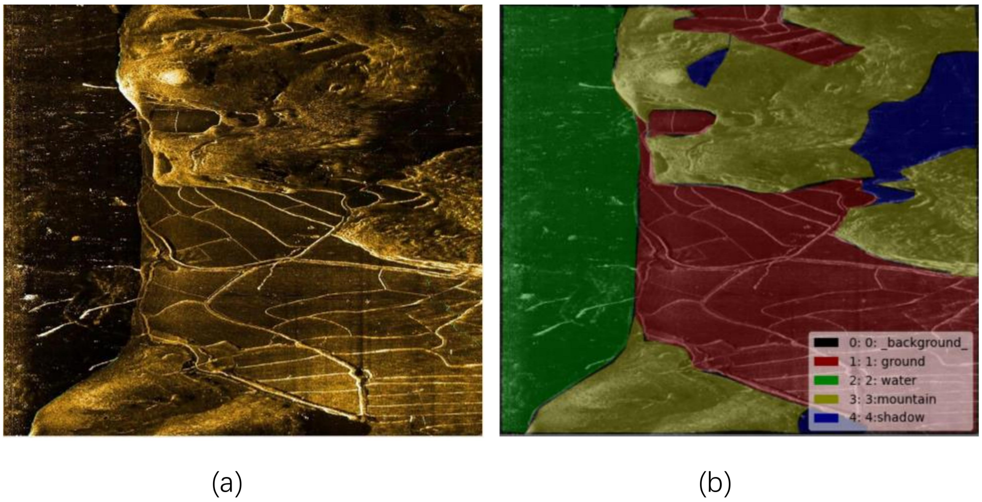

Figure 6. (**a**) Original image and (**b**) label.

To facilitate the image matching process, which necessitates generating images akin to the ground truth but with specific distortions for training, we employed four augmentation techniques: random cropping, random rotation, random erasing, random expansion, and random contrast adjustments. Figure 7 showcases the augmented training images.

Random cropping, random rotation, random erasing, and random contrast adjustments are commonly used methods in machine learning data augmentation. Random expansion, our original method, randomly selects a random number of contours from a semantic block using edge detection method and performs dilation processing on them. This method can create distorted images more naturally.

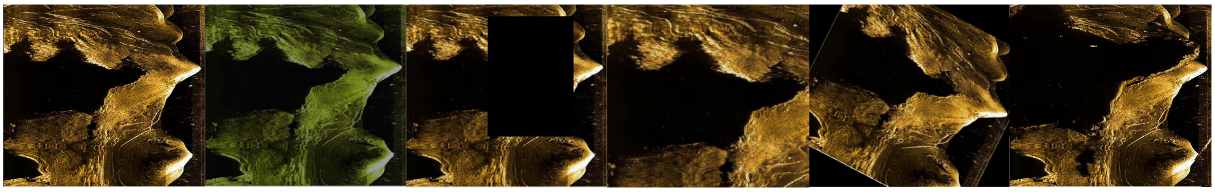

Figure 7. From left to right: the original image, contrast change, random erasure, random cropping, random rotation, and random expansion.

3.2. *Algorithm Architecture and Network Training*

The algorithm's architecture involves the training of two network models (DeiT and GCN), the generation of topological structures for samples serving as ground truth and those used in the training set, and the final matching process, as illustrated below (Algorithm 1).

Algorithm 1 Topological feature construction and matching

```
Require: $D = \{(x_i, y_i)\}_{i=1}^{D}$: the training dataset with $D$ samples;
    $Backbone(\cdot)$: DeiT distilled by our model (remove activation function layer);
    $Matcher(\cdot,\cdot)$: the graph matcher; $M$: the amount of visual vocabulary; $\Theta$: the set
    of all trainable parameters; $\delta_t$: the sparsification threshold.
procedure INITIALIZATION
    Training $Backbone(D)$
    Sample a subset $\hat{D} \subset D$ as the ground truth
    Amplify data on $\hat{D}$ to obtain $\tilde{D}$
    for $(x, y) \in \hat{D}$ do
        $\Omega \leftarrow Discretize(M, (Backbone(x)))$        ▷ $\Omega$:isual words (tpology nodes set)
        $\hat{G} \leftarrow$ Topology$(Backbone(x)[attn], \Omega)$        ▷ $\hat{G}$:Topology of ground truth
        for $\hat{e}_{i,j} \in \hat{E}$ do        ▷ Remove edges with weights below the threshold
            if $\hat{\lambda}_i < \delta_t$ or $\hat{\lambda}_j < \delta_t$ then
                $\hat{e}_{i,j} \leftarrow$ Null
            end if
        end for
    end for
end procedure
procedure TRAINING($Matcher(\tilde{G}, \hat{G})$)
    for $(x, y) \in \tilde{D}$ do
        $\Omega \leftarrow Discretize(M, (Backbone(x)))$
        $\tilde{G} \leftarrow$ Topology$(Backbone(x)[attn], \Omega)$        ▷ $\tilde{G}$:Topology
        for $\hat{e}_{i,j} \in \hat{E}$ do
            if $\hat{\lambda}_i < \delta_t$ or $\hat{\lambda}_j < \delta_t$ then
                $\hat{e}_{i,j} \leftarrow$ Null
            end if
        end for
        $\hat{y} \leftarrow Matcher(\tilde{G}, \hat{G})$
        Compute the final loss and gradient $\nabla_\Theta$ w.r.t. parameters $\Theta$
        Update parameters $\Theta$ with AdamW optimizer
    end for
end procedure
$\hat{G} \leftarrow$ Initialization
Training$(Matcher(\tilde{G}, \hat{G}))$
```

The training hyperparameters for all the networks are presented in Table 1. Because the matching process relies on topological structures generated from deep network features, maintaining the original images' high resolution is unnecessary. Hence, the images input to the DeiT and GCN were resized to 224 × 224, facilitating a higher batch size.

Table 1. Training hyperparameters.

Hyperparameters	Segmentation	DeiT	GCN
Num of workers	8	8	8
Batch size	3	64	64
Oprimizer	SGD	AdamW	AdamW
Learning rate	0.01	0.001	0.001
Learning policy	poly	cosine annealing	cosine annealing
Train epoch	50	50	50
number of hidden layer	/	12	2
Multi-heads	/	3	/

The network training process is visualized in the following figures (Figures 8 and 9).

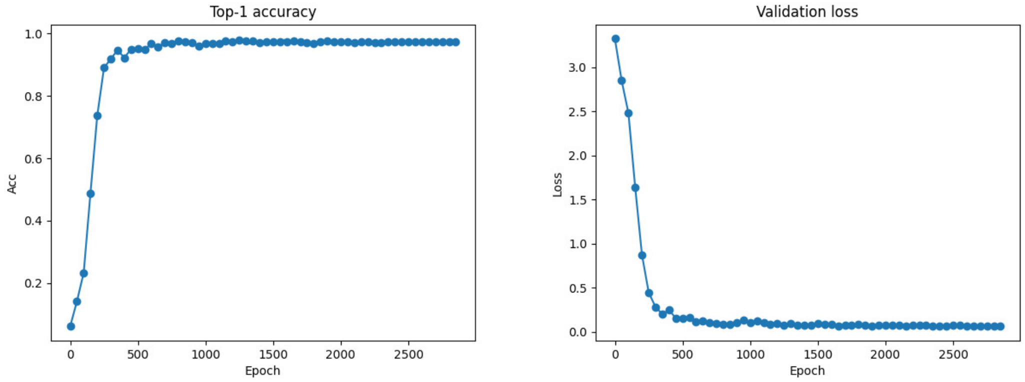

Figure 8. The change curve of the accuracy (Acc) and training loss function for the DeiT network.

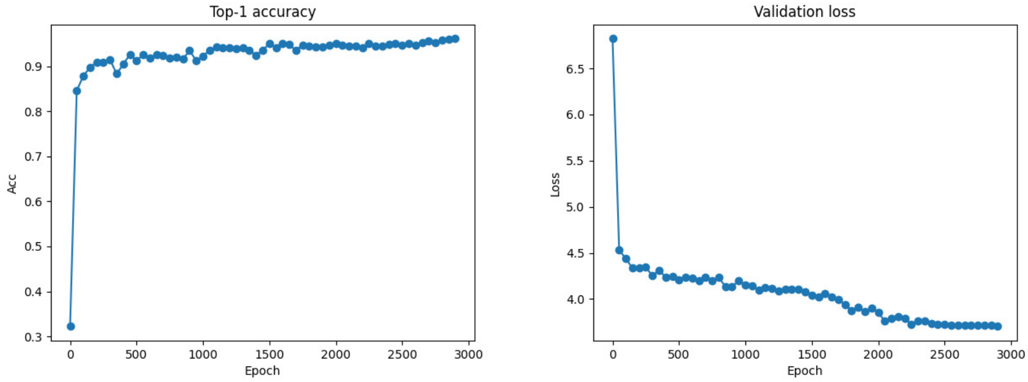

Figure 9. The change curve of the accuracy (Acc) and training loss function for the matcher.

The DeiT network essentially converges at 1500 epochs, while the matcher only converges around 2500 epochs. Moreover, at the point of convergence, the validation loss of the matcher is significantly higher than that of the DeiT network. This phenomenon could be attributed to the distillation training method employed by the DeiT network. Distillation training allows the DeiT network to acquire knowledge from an already well-trained teacher network, thereby enhancing its generalization capability and resulting in a faster convergence rate. On the other hand, the classification task faced by the matcher is more complex, as it requires integrating a broader range of modal information, which likely increases the difficulty of convergence.

4. Results and Analysis

In order to compare the extraction results of different sonar images, we selected two images from the same location but with different side-scan sonar perspectives as examples, as shown in Figure 10. The node numbers and distributions extracted by the model are displayed in Figure 11, while the topology structure containing the weights is shown in Figure 12. Additionally, we randomly selected two images unrelated to these two as negative samples for comparison, with the node distribution and topological structure of the negative samples presented in Figures 13 and 14. Because the original images were cropped into 14 × 14 = 196 patches when input into the DeiT network, these patches correspond to the dimensions of the input data. After clustering, each dimension was assigned a node index, which is reflected in the images as shown, with a total of 196 squares, each containing the corresponding node index.

From the figures, it can be observed that the basic node distributions of the two images are quite similar, with the highest proportion of nodes being 8, 40, and 99, among others. However, certain regions differ. Additionally, the distribution of nodes does not completely align with the semantic regions, as the node distribution depends not only on semantic features but on certain latent features, such as the connectivity and shape between regions, which are difficult to describe explicitly. This represents an effective extraction and summarization of the high-dimensional features of the images.

In the process of integrating nodes and edge weights to construct a more concrete representation of the topological structure, we exclude nodes and edges with weights less than 0.02 to ensure the sparsity of the structure, retaining only the relatively important nodes and edges in the figure. The circles in the figure represent nodes, while the remaining values are weights. It can be seen that the topological structures are highly similar, as they represent different perspectives of the same location, with identical node indices and only a few differences in edges and weights.

Overall, there is almost no evident similarity in the weights of nodes and the composition of edges, indicating that the similarity between sonar images from different locations is low. However, some patterns can still be observed in the figures. For example, all four images contain nodes 26, 40, and 99 and the weights of these three nodes are generally relatively high, suggesting that the model considers these three nodes to represent common features in sonar images. Moreover, in all four images, there are edges between nodes 26 and 40 as well as between nodes 40 and 99, indicating that the model recognizes a correlation between these important nodes. Despite the likelihood that the topological structures of unrelated images do not match, there is still some invariance in the local structures that reflect inherent features of sonar images.

We compared our method with other models in the field of image matching (or those capable of performing image matching) using a sonar dataset. The results are presented in Table 2.

Table 2. Model performance comparison.

Method	ACC	FLOPs
SIFTs	18.682	/
BagNet	68.343	1.10 G
SchemaNet	92.456	1.20 G
Our model	96.140	1.18 G

Here, SIFT refers to image matching using the classic SIFT descriptor, BagNet [37] is a model that employs deep networks to extract a visual bag-of-words for image matching and, SchemaNet is the original model used in [31], utilizing VIT as a teacher model for the classification network, which we have adapted for the task of image matching on a sonar dataset.

It is evident that the performance of classic SIFT is limited, and that the effectiveness of BagNet, which solely relies on visual bag-of-words matching, is also unsatisfactory. While SchemaNet performs better, the use of ViT as a teacher model is not as effective as employing our CNN-based semantic segmentation model, likely due to the limited amount of sonar image data.

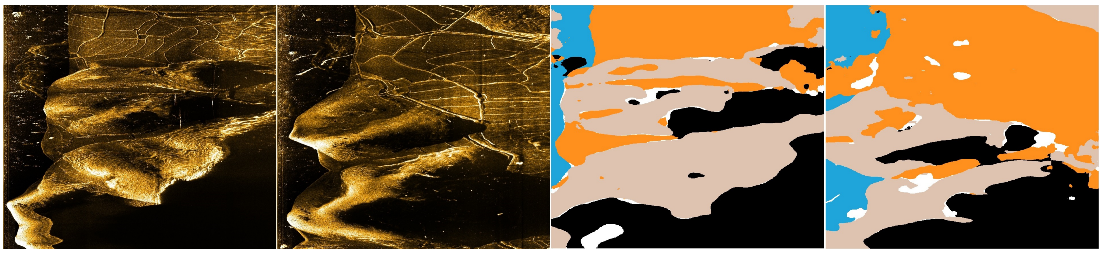

Figure 10. Left, middle, and right show the original images, the annotation results of the original images, and the output of the semantic segmentation network, respectively.

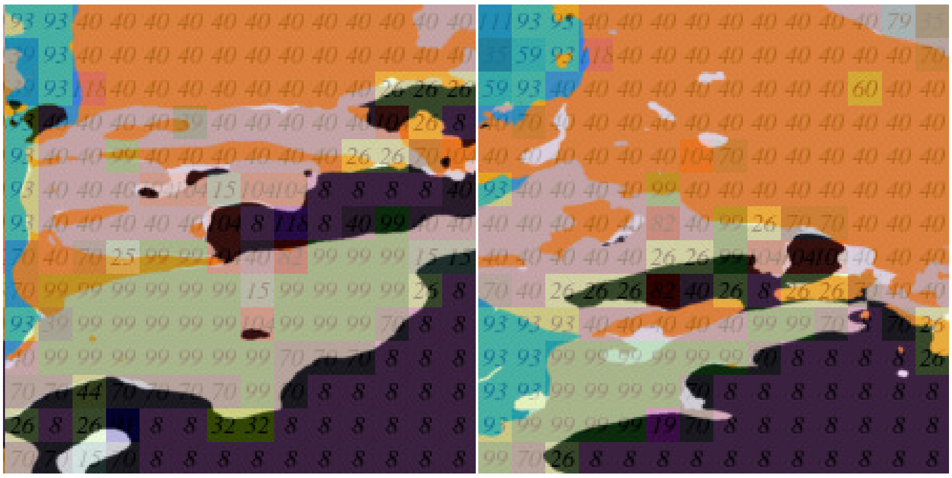

Figure 11. Node distribution of positive samples (images from the same location but different perspectivesm, the numbers in the figure represent the node index).

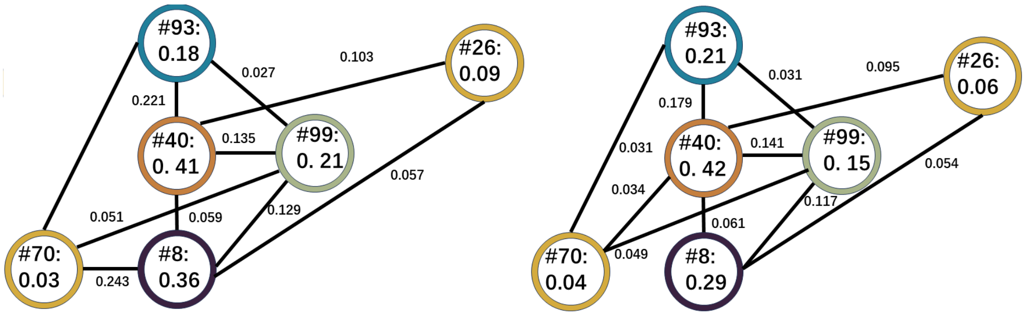

Figure 12. Positive sample topology structure; points and edges with weights below 0.02 will be discarded.

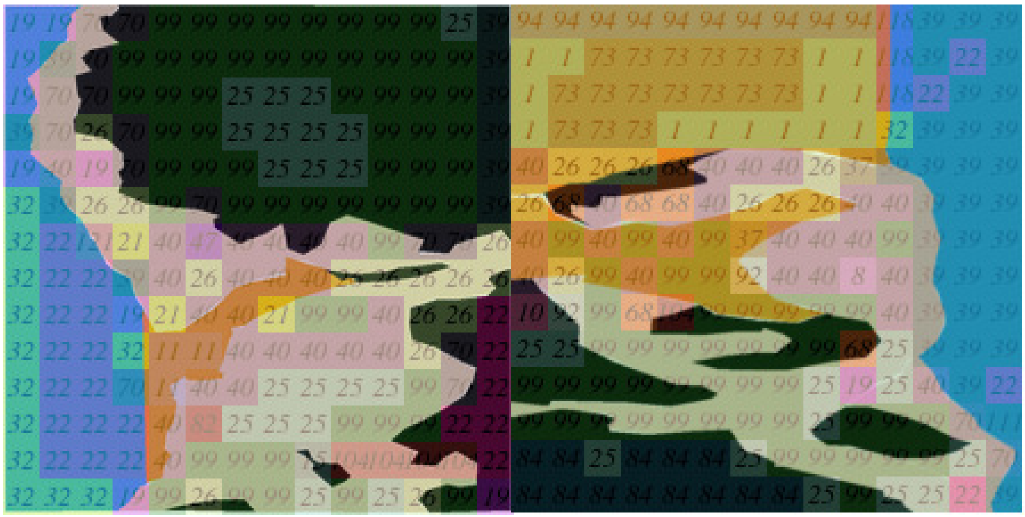

Figure 13. Node distribution of negative samples (unrelated images, the numbers in the figure represent the node index).

Table 3 lists the image (topology) matching normalization results for four images.To increase the discrimination of the output results, we normalized the self-matching score of the same image to 2.

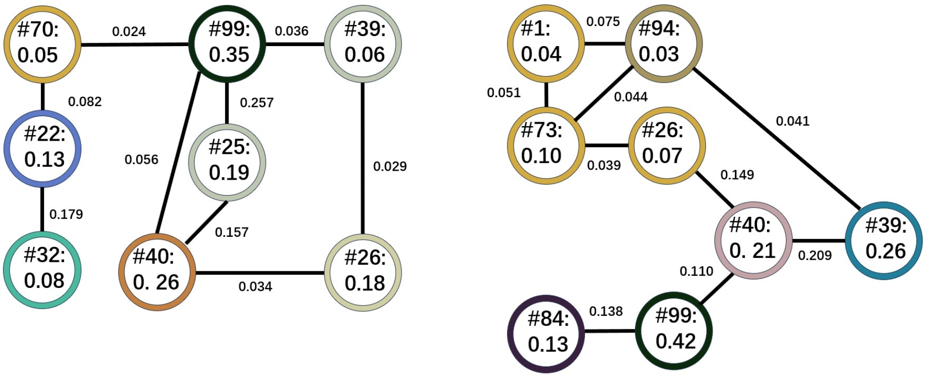

Figure 14. The topological structure of negative samples.

Table 3. Different image matching scores.

	Positive Sample B	**Negative Sample A**	**Negative Sample B**
Positive sample A	0.6257	0.2025	0.1917
Positive sample B	/	0.2463	0.2138
Negative sample A	/	/	0.1779

The data presented in Table 3 reveal that images from the same location but with different perspectives achieved the highest matching scores, whereas the scores for matching between unrelated images were comparatively lower. Moreover, a comprehensive matching test conducted on the dataset revealed that the highest score for matches between images with no relation was 0.4324, which remains significantly lower than the scores for related images. This indicates that our model effectively utilizes topological relevance to match images based on their shared location in sonar imagery.

5. Conclusions

This study investigates the application of topological matching for analyzing underwater side-scan sonar images, offering novel insights and solutions for underwater environment detection and image processing. Feature matching in sonar imagery poses persistent challenges in underwater settings due to image quality limitations, complex features, and susceptibility to distortion. However, by leveraging SchemaNet and integrating deep learning and knowledge distillation techniques, the method proposed in this paper effectively addresses these challenges and offers a robust solution for feature matching in underwater side-scan sonar images. Our model can extract the topological structure of sonar images containing various high-dimensional feature information and use it as a benchmark for image matching. The image topology structure, which includes semantic and regional relationships, is very effective in dealing with low-quality images and perspective distortion.

The paper introduces a self-designed semantic segmentation module as the teacher model, which trains the DeiT network to generate attention matrices. These matrices serve as the foundation for constructing the topological representation of images. A curated set of representative sonar images is then utilized for matching, with data augmentation techniques employed to create diverse images for the training set. This approach enables the network to perform image matching even in the presence of geometric and perspective distortions, as evidenced by the output of the final model.

6. Discussion

This research introduces a novel solution and technological framework for feature matching in underwater sonar images, with significant theoretical and practical implications. Future research endeavors could focus on further refining and optimizing the topological feature-based matching model, thereby expanding its applicability to a broader spectrum of underwater scenarios and contributing to advancements in underwater environment detection and monitoring. Meanwhile, this study has a significant limitation, namely, that while using deep learning models to extract the topological structure of images for image matching can achieve high accuracy in matching results, it cannot achieve fine matching between the pixels of two images. At present, the main methods for achieving fine matching are descriptor methods such as SIFT. Thus, combining these two approaches to achieve high accuracy with fine pixel-level matching is a research direction that needs to be considered in the future.

Author Contributions: D.Y. and F.Z. conceived the study and put forward the methodology; J.Y. arranged the testing and data collection site and designed the experimental method; C.C. and C.W. performed the data collection and preprocessing; D.Y. carried out the software for the experiments and wrote the first draft of the manuscript; X.W. provided assistance on hardware devices; F.Z. and G.P. reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Key Research and Development Program of China (2023YFC2808400).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Acknowledgments: The authors would like to thank Xujia Hou, Tingfeng Tan, Wei Zhang, and Xinyu Guo et al. We thank them for their help during the experiment. We also thank all the authors of [31] for their inspiring work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Brown, C.; Smith, S.; Lawton, P.; Anderson, J. Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques. *Estuar. Coast. Shelf Sci.* **2011**, *92*, 502–520. [CrossRef]
2. Riyait, V.; Lawlor, M.; Adams, A.; Hinton, O.; Sharif, B. A Review of the ACID Synthetic Aperture Sonar and other Sidescan Sonar Systems. *Int. Hydrogr. Rev.* **1995**, *72*, 115–123.
3. Kasatkin, B. Anomalous phenomena in sound propagation near the sea floor: A review. *Acoust. Phys.* **2002**, *48*, 379–387. [CrossRef]
4. Blondel, P. *The Handbook of Sidescan Sonar*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2010.
5. Jayanthi, N.; Indu, S. Comparison of Image Matching Techniques. *Int. J. Latest Trends Eng. Technol.* **2016**, *7*, 396–401. [CrossRef]
6. Alam, M.; Morshidi, M.; Gunawan, T.; Olanrewaju, R. A Comparative Analysis of Feature Extraction Algorithms for Augmented Reality Applications. In Proceedings of the 2021 IEEE 7th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), Bandung, Indonesia, 23–25 August 2021; pp. 59–63. [CrossRef]
7. Ma, J.; Sun, Q. Image Recognition Method based on Artificial Intelligence Technology. In Proceedings of the 2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 27–29 May 2022; pp. 971–974. [CrossRef]
8. Fakiris, E.; Blondel, P.; Papatheodorou, G.; Christodoulou, D.; Dimas, X.; Georgiou, N.; Kordella, S.; Dimitriadis, C.; Rzhanov, Y.; Geraga, M.; et al. Multi-Frequency, Multi-Sonar Mapping of Shallow Habitats—Efficacy and Management Implications in the National Marine Park of Zakynthos, Greece. *Remote Sens.* **2019**, *11*, 461. [CrossRef]
9. Ye, X.; Li, P.; Zhang, J.; Shi, J.; Guo, S. A feature-matching method for side-scan sonar images based on nonlinear scale space. *J. Mar. Sci. Technol.* **2016**, *21*, 38–47. [CrossRef]
10. Zhang, J.; Tao, B.; Liu, H.; Jiang, W.; Gou, Z.; Wen, F. A mosaic method based on feature matching for side scan sonar images. In Proceedings of the 2016 IEEE/OES China Ocean Acoustics (COA), Harbin, China, 9–11 January2016; pp. 1–6. [CrossRef]
11. Trzcinska, K.; Tegowski, J.; Pocwiardowski, P.; Janowski, L.; Zdroik, J.; Kruss, A.; Rucinska, M.; Lubniewski, Z.; Schneider von Deimling, J. Measurement of Seafloor Acoustic Backscatter Angular Dependence at 150 kHz Using a Multibeam Echosounder. *Remote Sens.* **2021**, *13*, 4771. [CrossRef]
12. Ghate, S.; Nikose, D. Analysis of the Repeatability of SIFT and SURF Descriptors Techniques for Underwater Image Preprocessing. *Int. J. Adv. Res. Sci. Commun. Technol.* **2021**, *4*, 24–31. [CrossRef]
13. Pourfard, M.; Hosseinian, T.; Saeidi, R.; Motamedi, S.; Abdollahifard, M.; Mansoori, R.; Safabakhsh, R. KAZE-SAR: SAR Image Registration Using KAZE Detector and Modified SURF Descriptor for Tackling Speckle Noise. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5207612. [CrossRef]

14. Bansal, M.; Kumar, M.; Kumar, M. 2D object recognition: A comparative analysis of SIFT, SURF and ORB feature descriptors. *Multimed. Tools Appl.* **2021**, *80*, 18839–18857. [CrossRef]
15. Lozano-Vázquez, L.; Miura, J.; Rosales-Silva, A.; Luviano-Juárez, A.; Mújica-Vargas, D. Analysis of Different Image Enhancement and Feature Extraction Methods. *Mathematics* **2022**, *10*, 2407. [CrossRef]
16. Shaharom, M.; Tahar, K. Multispectral Image Matching Using SIFT and SURF Algorithm: A Review. *Int. J. Geoinform.* **2023**, *19*, 13–21. [CrossRef]
17. Anilkumar, S.; Dhanya, P.; Balakrishnan, A.; Supriya, M. Algorithm for Underwater Cable Tracking Using CLAHE based Enhancement. In Proceedings of the 2019 International Symposium on Ocean Technology (SYMPOL), Ernakulam, India, 11–13 December 2019; pp. 129–137. [CrossRef]
18. Vikas, A. Detection and Removal of Shadows for Side Scan Sonar Images By Effective Image Processing Algorithms. *Int. J. Adv. Res. Ideas Innov. Technol.* **2017**, *3*, 1561–1565.
19. Alevizos, E.; Schoening, T.; Koeser, K.; Snellen, M.; Greinert, J. Quantification of the fine-scale distribution of Mn-nodules: Insights from AUV multi-beam and optical imagery data fusion. *Biogeosci. Discuss.* **2018**, *2018*, 1–29.
20. Li, B.; Liu, B.; Li, S.; Liu, H. Underwater Target Detection Based on Improved YOLOv4. In Proceedings of the 2022 41st Chinese Control Conference (CCC), Hefei, China, 25–27 July 2022; pp. 7012–7017. [CrossRef]
21. Tang, X.; Zhang, X.; Xu, X.; Sheng, J.; Xiang, Y. Methods for Underwater Sonar Image Processing in Objection Detection. In Proceedings of the 2017 International Conference on Computer Systems, Electronics and Control (ICCSEC), Dalian, China, 25–27 December 2017; pp. 941–944. [CrossRef]
22. Rajput, S.; Chawra, R.; Wani, P.; Nanda, S. Noisy Sonar Image Segmentation using Reptile Search Algorithm. In Proceedings of the 2022 International Conference on Connected Systems & Intelligence (CSI), Trivandrum, India, 31 August–2 September 2022; pp. 1–10. [CrossRef]
23. Neupane, D.; Seok, J. A Review on Deep Learning-Based Approaches for Automatic Sonar Target Recognition. *Electronics* **2020**, *9*, 1972. [CrossRef]
24. Li, X.; Liu, B.; Zheng, G.; Ren, Y.; Zhang, S.; Liu, Y.; Gao, L.; Liu, Y.; Zhang, B.; Wang, F. Deep-learning-based information mining from ocean remote-sensing imagery. *Natl. Sci. Rev.* **2020**, *7*, 1584–1605. [CrossRef]
25. Misiuk, B.; Brown, C.J. Benthic habitat mapping: A review of three decades of mapping biological patterns on the seafloor. *Estuar. Coast. Shelf Sci.* **2023**, *2023*, 108599. [CrossRef]
26. Gav, P.; Kawitkar, R.; Balan, M. Review on Ultrasonic Techniques forUnderwater Object Classification. *Int. J. Adv. Res. Electr. Electron. Instrum. Energy* **2015**, *4*, 879–882.
27. Gašparović, B.; Lerga, J.; Mauša, G.; Ivasic-Kos, M. Deep Learning Approach For Objects Detection in Underwater Pipeline Images. *Appl. Artif. Intell.* **2022**, *36*, 2146853. [CrossRef]
28. Singh, N.; Bhat, A. A Detailed Understanding of Underwater Image Enhancement using Deep Learning. In Proceedings of the 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 22–23 October 2021; pp. 1–6. [CrossRef]
29. Xi, K.; He, J.; Hao, S.; Luo, L. SLAM Loop Detection Algorithm Based on Improved Bag-of-Words Model. In Proceedings of the 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Chengdu, China, 19–21 August 2022; pp. 683–689. [CrossRef]
30. Wang, G.A.; Yang, S.; Liu, H.; Wang, Z.; Yang, Y.; Wang, S.; Sun, J. High-order information matters: Learning relation and topology for occluded person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6449–6458.
31. Zhang, H.; Xue, M.; Liu, X.; Chen, K.; Song, J.; Song, M. Schema Inference for Interpretable Image Classification. *arXiv* **2023**, arXiv:2303.06635.
32. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
33. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
34. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Jégou, H. Training data-efficient image transformers & distillation through attention. *arXiv* **2020**, arXiv:2012.12877.
35. Yang, D.; Cheng, C.; Wang, C.; Pan, G.; Zhang, F. Side-scan sonar image segmentation based on multi-channel CNN for AUV navigation. *Front. Neurorobot.* **2022**, *16*, 928206. [CrossRef]
36. Morris, C.; Ritzert, M.; Fey, M.; Hamilton, W.L.; Grohe, M. Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 4602–4609.
37. Brendel, W.; Bethge, M. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv* **2019**, arXiv:1904.00760.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.