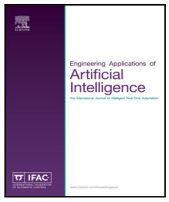
Engineering Applications of Artificial Intelligence 114 (2022) 105157

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Survey paper

Survey on deep learning based computer vision for sonar imagery

Yannik Steiniger [a,*], Dieter Kraus [b], Tobias Meisen [c]

[a] German Aerospace Center (DLR), Institute for the Protection of Maritime Infrastructures, Fischkai 1, Bremerhaven, 27572, Germany
[b] Institute of Water-Acoustics, Sonar-Engineering and Signal-Theory, City University of Applied Sciences Bremen, Neustadtswall 30, Bremen, 28199, Germany
[c] Institute of Technology and Management of the Digital Transformation, University of Wuppertal, Rainer-Gruenter-Straße 21, Wuppertal, 42119, Germany

ARTICLE INFO

Keywords:
Deep learning
Sonar imagery
Computer vision
Automatic target recognition
Status quo review

ABSTRACT

Research on the automatic analysis of sonar images has focused on classical, i.e. non deep learning based, approaches for a long time. Over the past 15 years, however, the application of deep learning in this research field has constantly grown. This paper gives a broad overview of past and current research involving deep learning for feature extraction, classification, detection and segmentation of sidescan and synthetic aperture sonar imagery. Most research in this field has been directed towards the investigation of convolutional neural networks (CNN) for feature extraction and classification tasks, with the result that even small CNNs with up to four layers outperform conventional methods.

The purpose of this work is twofold. On one hand, due to the quick development of deep learning it serves as an introduction for researchers, either just starting their work in this specific field or working on classical methods for the past years, and helps them to learn about the recent achievements. On the other hand, our main goal is to guide further research in this field by identifying main research gaps to bridge. We propose to leverage the research in this field by combining available data into an open source dataset as well as carrying out comparative studies on developed deep learning methods.

1. Introduction

Sidescan sonar (SSS) and synthetic aperture sonar (SAS) systems are among the most prominent sensors when investigating the sea floor. Common applications are the search for unexploded ordnances (UXO), wrecks or segmenting different bottom types. These sensors, which are mounted on a towfish or autonomous underwater vehicle (AUV), emit an acoustic ping and receive the backscattered signal. The recorded time-signals are further processed, e.g. by stacking consecutive pings on top of each other, in order to form an image. During sea floor scanning missions, which can extend over several days, a large number of images are captured. Manually inspecting the data is cumbersome and time consuming, since objects of interest are very rare (e.g. for the object detection task) or because large parts of the images need to be annotated (e.g. for bottom type segmentation). Thus, an automatic analysis of sonar imagery is crucial and has been researched extensively for many years (Johnson and Deaett, 1994; Nelson and Tuovila, 1995; Langner et al., 2009). Typically, computer vision methods for analyzing sonar images are combined under the term automatic target recognition (ATR).

Since the success of AlexNet (Krizhevsky et al., 2012) in the 2012 ImageNet Large Scale Video Recognition Challenge (ILSVRC) convolutional neural networks (CNN) have become state-of-the-art in computer vision tasks. It took a few more years for CNNs to be applied to SSS and SAS images (see Fig. 1). Most recent, Vision Transformer (ViT) (Dosovitskiy et al., 2021) are more and more often applied in vision tasks, which surpass CNNs on standard classification benchmarks. However, they have not been applied to sonar imagery yet. In general the application of deep learning methods to the sonar imagery domain is several years behind the state-of-the-art as shown by selected methods in Fig. 1. Their breakthrough started around 2016. Since then the number of research publications with respect to deep learning applied to SSS and SAS imagery for computer vision tasks has constantly grown. In most cases the developed deep learning methods outperform classical approaches which shows the transformative character of this technology also for the sonar imagery domain. Researchers who have worked on traditional ATR methods for the automatic analysis of sonar images now need to adapt to this change of the state-of-the-art and get accustomed to deep learning.

Due to the importance of deep learning as a tool for ATR, the increased number of research papers and the quick development in deep learning research, a survey on deep learning based computer vision for sonar imagery is necessary. For this, we systematically review the past and current research in this field and present the main findings. With over 60 publications considered, our status quo review serves researchers who are new to this field as well as sonar engineers who are familiar with the conventional processing but not with deep learning as

* Corresponding author.
E-mail addresses: yannik.steiniger@dlr.de (Y. Steiniger), dieter.kraus@hs-bremen.de (D. Kraus), meisen@uni-wuppertal.de (T. Meisen).

https://doi.org/10.1016/j.engappai.2022.105157
Received 28 March 2022; Received in revised form 20 June 2022; Accepted 29 June 2022
Available online 8 July 2022

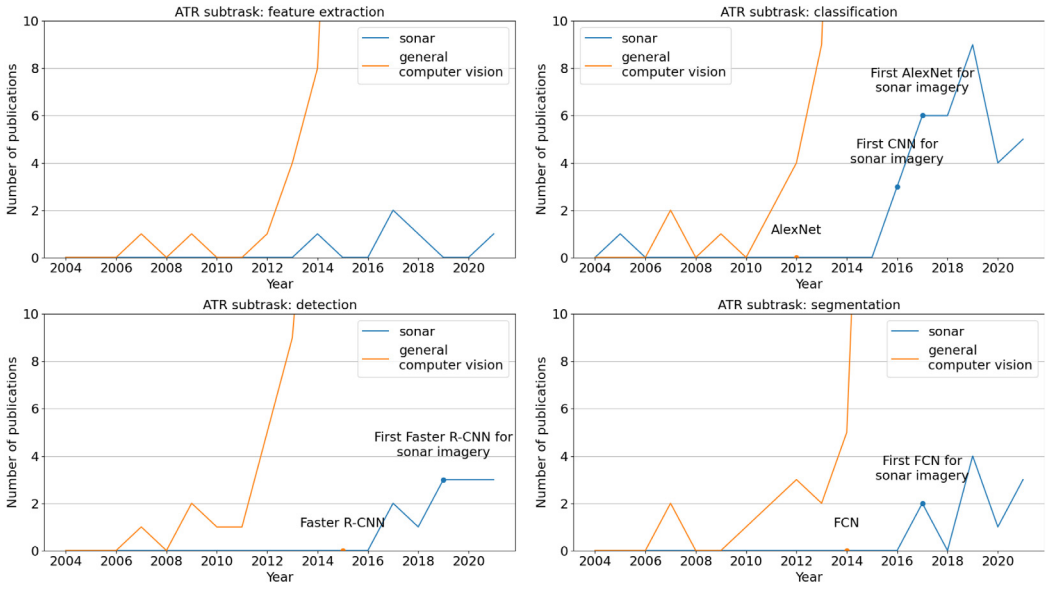


Fig. 1. Number of relevant publications per year for the considered computer vision tasks. For comparison the orange curve shows the number of general deep learning publications with respect to computer vision and the corresponding subtask.

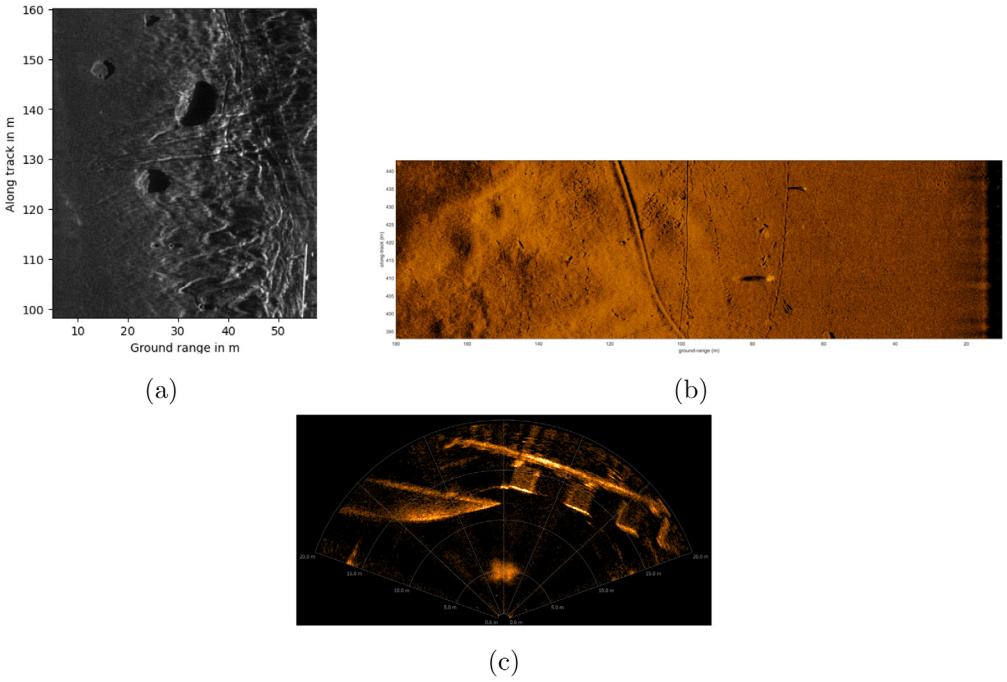


Fig. 2. Examples of sonar images from different sensors. (a) A SSS image. (b) A SAS image[1]. (c) A FLS image.

a starting point. The holistic overview and discussions help researchers who are already working on deep learning solutions in the field of SSS and SAS imagery to see the development over the past few years as well as the current trends. Furthermore, by compressing the work done so far, our main goal is to guide further research in this field by revealing the current state-of-the-art as well as identifying main research gaps to bridge. We focus especially on SSS and SAS imagery and exclude work done with data from forward looking sonar (FLS) since the type of images captured by this sensor differs from the other two, see Fig. 2 where one example image from each sensor is displayed. Nevertheless, it should be noted that deep learning applications have also been investigated for FLS images, like Valdenegro-Toro (2016), Fuchs et al. (2018), Jin et al. (2019) and Fan et al. (2021) to mention a few research works.

From the example sonar images in Fig. 2 the difference compared with natural RGB images becomes clear. As sonar images represent acoustic intensities, features that utilize color information are useless. Moreover, the resolution of sonar images is smaller compared to natural RGB images. High resolution SAS systems are able to achieve a resolution of a few centimeters. Still, the computer vision methods have to deal with less details compared to conventional images. Due to the large domain gap between sonar and natural RGB images a simple application of state-of-the-art deep learning methods will fail. In addition, speckle noise, multipath returns, sea surface reflections

[1] Image captured with a SeaCat Vision Mk2 SAS, provided by ATLAS ELEKTRONIK GmbH.

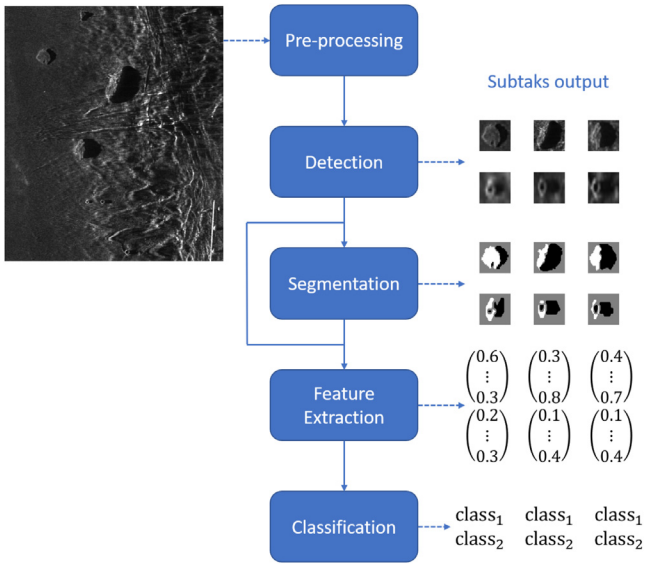


Fig. 3. Typical conventional ATR processing chain. In a first step the sonar image undergoes some pre-processing. From this image ROIs are detected. From the ROIs the highlight and shadow regions are segmented and features are calculated. Based on the extracted features the ROIs are classified into given classes, e.g. target or clutter.

and other effects related to the underwater environment cause the interpretation of sonar images to be challenging.

Fig. 3 shows the typical conventional ATR processing chain for SSS and SAS images (Fei et al., 2015) together with the main information captured at each step. From this we derive detection, segmentation, feature extraction and classification as the computer vision subtasks that are considered in this survey. In the conventional ATR processing chain, the detection step serves mainly for extracting regions of interest (ROI). Afterwards, within the segmentation subtask, the object highlight and the acoustic shadow are segmented. In the feature extraction step, hand-crafted features are extracted from the ROI as well as the segmentation results, e.g. size of the acoustic shadow. A classifier, e.g. a support vector machine (SVM), is finally used to predict a class based on these features for each ROI. In our survey we use the ATR processing chain as a guideline and investigate what deep learning methods have been applied in the individual subtasks (except the pre-processing). Note however, that we will consider a different ordering of subtasks compared with the ATR processing chain in the structure of our survey, as we move from the less complex to the most complex task. In the deep learning context feature extraction is most of the time directly included in the classification step. Furthermore, a deep learning detector simultaneously predicts the location, size and class of an object and thus includes the classification. Deep learning based segmentation adds even more complexity because the objects are detected on a pixel level. Thus, we consider the deep learning approaches to the ATR subtasks in the order: feature extraction, classification, detection and segmentation.

We searched for relevant publications using Clarivates *Web of Science* with a combination of the keywords: deep learning, side scan sonar, sidescan sonar, synthetic aperture sonar, automatic target recognition, feature extraction, classification, detection, segmentation. Our search results in 5 papers regarding feature extraction, 35 for classification, 12 for detection and 10 regarding segmentation. An overview on the deep learning methods applied in the individual ATR subtasks is depicted by the flowchart in Fig. 4. The number associated with an arrows towards a method indicates the number of times it was applied. We marked the most popular methods in each subtask with a bold arrow. Different architectures of one deep learning method are summarized under the name of the method, e.g. ResNet covers ResNet-18, ResNet-50 and other types of this architecture. For a faster look up on the abbreviations in Fig. 4 we refer to the list of abbreviations in Appendix A of the appendix. We investigate the contributions of the individual papers in terms of the designed or applied deep learning methods, their main findings and their relation to other works. From this we derive current research gaps with respect to the three viewpoints:

- *Method*: What are the most promising methods?
- *Completeness*: Are there state-of-the-art deep learning methods for computer vision that have not been applied to the sonar imagery domain yet?
- *Data*: Deep leaning has profited from large publicly available datasets. How is the situation for SSS and SAS images?

We are aware of the existing review by Neupane and Seok (2020), who considered deep learning approaches for ATR in FLS, SSS, SAS and passive sonar systems. However, in our survey we focus particularly on work done using SSS and SAS data which is only lightly covered in their review. This focus on the two sensors as well as the explicit separation by ATR subtasks helps the reader to easily get the information he or she needs. Moreover, the publications regarding SSS and SAS images considered in Neupane and Seok (2020) are mainly about generating synthetic images and are not directly ATR related. This becomes evident when looking at the intersection between the considered papers, as only three publications considered in our work are listed in Neupane and Seok (2020). Furthermore, due to the quick development in the field of deep learning, our survey gives an update on the current state-of-the-art in ATR for SSS and SAS images. Although there is only one year between the two surveys, 20% of the papers covering this range of topics are published in 2021. Domingos et al. presented a survey on deep learning methods for underwater shoreline surveillance (Domingos et al., 2022). Their main focus lies on methods for processing data from passive sonar systems. Approaches for SSS and SAS data are only lightly covered. Another review on literature regarding ATR for sonar imagery was done by Hożyń (2021). He focuses on the classification and detection mines in SSS and SAS images. Traditional as well as deep learning methods are reviewed. In contrast to his work, we focus not only on methods designed for mine classification and detection but consider more applications like the detection of wrecks or the segmentation of different sea bottom types. Another survey (Teng and Zhao, 2020) covers literature relating to underwater target recognition based on optical sensors. To the best of our knowledge, no survey provides a holistic view on deep learning based computer vision for SSS and SAS imagery.

It should be noted that CNNs, besides being successfully applied to various computer vision tasks, still some general challenges, like overfitting, slow convergence, getting stuck in local minima or poor performance on small datasets exist. Since a detailed discussion on CNNs is beyond the scope of this work we refer the reader to the current literature such as Alzubaidi et al. (2021) or Chai et al. (2021).

This paper is organized as follows: We divide the main part of our survey into the aforementioned computer vision tasks and consider feature extraction in Section 2, classification in Section 3, detection in Section 4 and segmentation in Section 5. After analyzing the existing research for the individual ATR subtasks, we derive current challenges in deep learning based computer vision for sonar imagery in Section 6. We close our paper with a summary in Section 7.

2. Feature extraction

2.1. Subtask explanation

In order to classify ROIs in the conventional ATR processing chain, features need to be calculated in order to train a classifier. This subtask of feature extraction is depicted in Fig. 5. Before the rise of deep learning and especially CNNs, those features were based on hand-crafted

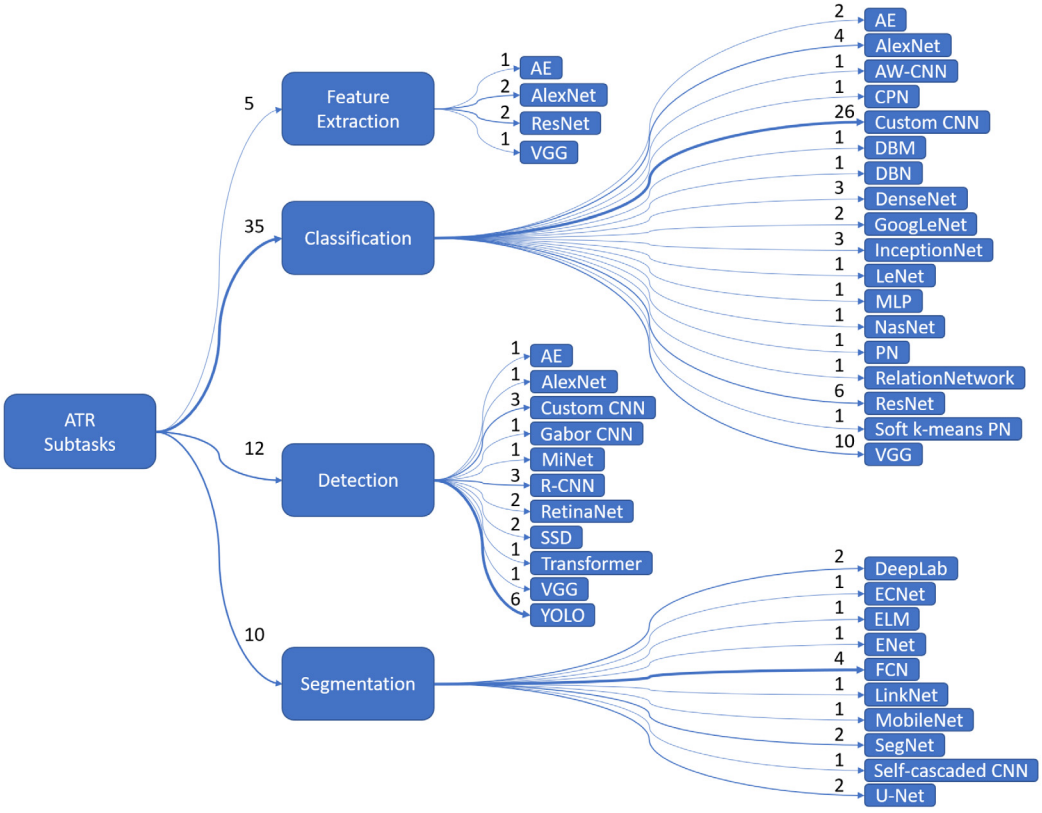


Fig. 4. Flowchart of the ATR subtasks and applied deep learning methods.

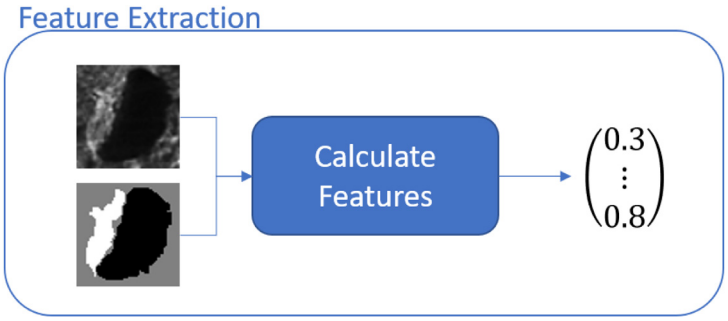


Fig. 5. Scheme of the feature extraction subtask. Based on mathematically formulated rules a feature vector is extracted from the ROI.

Table 1
Publications on deep learning for feature extraction from sonar imagery.

Paper	Task	Method
Isaacs (2014)	Representation learning for MLO	AE
Zhu et al. (2017)	Feature extraction for object recognition	AlexNet
McKay et al. (2017)	Mine recognition	VGG-16, VGG-19, VGG-f, AlexNet
Rutledge et al. (2018)	Archaeological site detection	ResNet-50
Divyabarathi et al. (2021)	Shipwreck, plain wreck	ResNet-50

engineering. Since in SSS and SAS images the acoustical shadow contains important information, features are typically based on the pixel intensities as well as the segmented highlight and shadow areas. Some examples of hand-crafted features are: object area, object mean intensity, difference between object and background mean intensity (Perry and Guan, 2004). From literature, several deep learnings methods used to extract feature from SSS and SAS images are investigated. We present in the following section how these methods improve the performance of an ATR processing chain.

2.2. Deep learning applications

Table 1 lists papers which investigate deep learning for the feature extraction step. The column *Task* summarizes the main objective of each paper while *Method* lists the used deep learning methods. Because abbreviations are used in this and the following tables, a list of abbreviations is given in the appendix for a faster look up. We sorted the papers in the table by their year of publication.

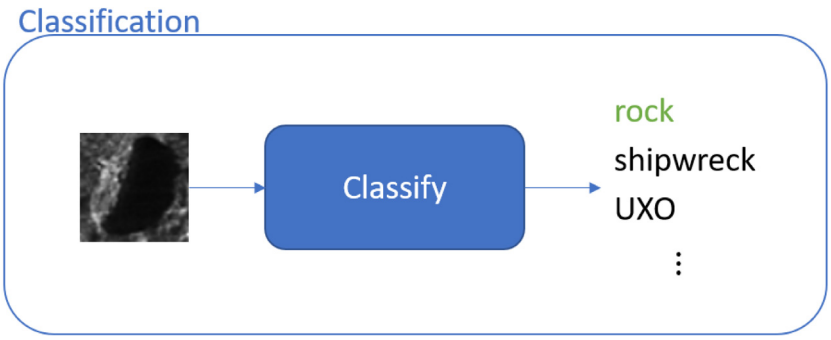


Fig. 6. Scheme of the classification subtask. A ROI needs to be assigned to a specific class, typically based on extracted features.

In four of the five papers listed in Table 1 standard CNN models like AlexNet, VGG or ResNet were implemented. The fifth paper used an auto-encoder (AE) to generate features from a sonar image. All papers covered a different task, ranging from mine recognition to archaeological site detection.

One of the first applications of deep learning to SSS or SAS data was the usage of CNNs for the calculation of abstract features (Zhu et al., 2017; McKay et al., 2017). Prior to this, Isaacs used latent Dirichlet allocations (LDA) and AE with a single hidden layer to extract features of objects in SAS images (Isaacs, 2014). The purpose of LDA and AE was to learn a representation of the snippet containing the object in an unsupervised manner. The AE was trained on 600 snippets from a SAS image database from the NATO Centre for Maritime Research and Experimentation (CMRE) MUSCLE AUV system. However, in this work the quality of the learned representations was only assessed empirically. Whether these representations are useful for further classification is still open for investigation.

Multiple works have shown that using a pre-trained CNN for feature extraction in combination with an SVM for classification improves the classification performance compared to the usage of simpler features (Zhu et al., 2017; McKay et al., 2017; Rutledge et al., 2018). In Zhu et al. (2017) a pre-trained AlexNet was used for this purpose. The features were extracted without fine-tuning from the last fully connected layer. For comparison, the two conventional feature extraction methods local binary patterns and histogram of oriented gradients were selected. The SVM trained with features from AlexNet outperformed the ones trained with the conventional features. In McKay et al. (2017) different CNNs were analyzed for the task of extracting features in an ATR processing chain. The authors considered VGG-16, VGG-19, VGG-f and AlexNet and compared them to a scale invariant feature transform bag-of-words model and a sparse reconstruction-based classifier. An SVM was used to classify the sonar images based on the extracted features. Again, the main result underlined that the CNNs serve as a better feature extractor, with the deeper VGG-19 and AlexNet being the best methods. In their work the authors also fine-tuned some layers of VGG-f, which consists only of the first three convolutional layers and a fully convolutional layer for the classification. With only 20 training samples per class they found this CNN to perform worse than the classical feature extractors combined with an SVM when classifying blocks, cones, spheres and cylinders. A slightly different application was considered in Rutledge et al. (2018) where ROIs were ranked in order to generate an input sequence for a path planning module of an AUV. The ranking was done using a ranking SVM which was trained with hand-crafted features as well as features extracted from a ResNet-50 pre-trained on the ImageNet dataset. Again, the features from the deep CNN showed a better performance and generalization ability. Recently, features from a ResNet-50 were used in combination with an ensemble of four different classifiers, namely logistic regression, random forest, naive Bayes and SVM (Divyabarathi et al., 2021). The authors showed that the ensemble improves the performance compared to the individual classifiers. However, no analysis of the classification performance of the ResNet-50 was carried out.

All available work showed that features extracted from CNNs are better suited for further processing with an SVM than conventional or hand-crafted features. No publication has investigated the combination of deep learning and conventional features. Nevertheless, we will show in the next section that some hand-crafted features have been included in a CNN for classification with a positive effect on the performance. Features can also be extracted by an AE but their usability for the ATR process is still unclear. However, nowadays in general computer vision task the feature extraction step is not carried out explicitly anymore. Instead CNNs have shown to be powerful enough to classify images without the need of a complementary algorithm like an SVM. With only a few papers regarding the feature extraction subtask being published, this also indicates that a separate feature extraction is not part of a deep learning based ATR processing chain today.

3. Classification

3.1. Subtask explanation

A primary objective of an ATR system is to tell an operator what can be seen in a sonar image. For this, ROIs are assigned to predefined classes based on features extracted from the image. One purpose is for example to distinguish between an UXO and a ROI generated due to clutter. The classification subtask is shown in Fig. 6. Based on features either gained hand-crafted or learned by a CNN, classical machine-learning methods like an SVM can be used to classify ROIs. The following section will investigate the recent advances in the classification subtask using deep learning methods like multilayer perceptrons (MLP) or CNNs, which are in this case not exclusively used for feature extraction.

3.2. Deep learning applications

Publications which tackle the classification problem with a deep learning approach are listed in Table 2. The column *Classes* summarizes the different objects that are intended to be classified in the individual papers. As before, the deep learning methods used are listed under *Method*. The term CNN in this column refers to a custom network designed for that particular work. When the specific architecture of a network is not stated (e.g. VGG instead of VGG-16) it was not clear from the paper which architecture was used.

The papers listed in Table 2 are grouped by the classes considered for classification. From this grouping it is seen that most works (11 out of the 35 considered publications) deal with the binary classification between a target and clutter. Note that the classification of target and clutter can also be seen as a refinement of the detection since clutter may be considered as background. Since the detection step leading to the ROIs that are classified in these 11 publications is not based on deep learning, we assigned them to the classification subtask. Two further common use cases are the differentiation between a mine-like object (MLO) and a false alarm as well as the classification of different

Table 2
Publications on deep learning for classification of sonar imagery.

Paper	Classes	Method
Chen and Summers (2016)	Sandy ridges, sand wave, rocky, rocky sand, flat sand	AE, CNN
Berthold et al. (2017)	Coarse, mixed, sand, fine sediment	GoogLeNet
Luo et al. (2019)	Sand wave, mud, reef	CNN
Qin et al. (2021)	Sand wave, mud, reef	AlexNet, LeNet, DenseNet-52, DenseNet-100, DenseNet-151, ResNet-3-2,[a] ResNet-3-3, ResNet-4-2, VGG-13, VGG-16
Ye et al. (2018)	Ship-&plane wreck vs. other	VGG-11, ResNet-18
Wang et al. (2019)	Shipwreck, plain wreck, stone, tire, shoal, sand ripple	AW-CNN, CNN, DBN
Huo et al. (2020)	Shipwreck, plane wreck, corpse, mine, background	CNN, VGG
Xu et al. (2020)	Shipwreck, plane wreck, sand, stone	CNN
Li et al. (2021)	Shipwreck, plane wreck, others	AE, ResNet-34
Nayak et al. (2021)	Archaeological sites vs. background	CNN
Cheng et al. (2022)	Shipwreck, plane wreck, see floor	VGG19
Karjalainen et al. (2019)	Rock vs. cylinder	CNN
Ochal et al. (2020)	18 objects[b]	PN, Relation Network, Soft k-means PN, CPN, CNN, ResNet-18, ResNet-50
Williams and Dugelay (2016)	Target vs. clutter	DBM
Williams (2017)	Target vs. clutter	CNN
Williams (2018b)	Target vs. clutter	CNN
Williams (2018a)	Target vs. clutter	CNN
Gerg and Williams (2018)	Target vs. clutter	CNN, VGG
Galusha et al. (2019)	Target vs. clutter	CNN
d'Alès de Corbet et al. (2019)	Target vs. clutter	CNN
Williams et al. (2019)	Target vs. clutter	CNN, VGG-16
Berthomier et al. (2020)	Target vs. clutter	CNN
Williams (2021)	Target vs. clutter	CNN
Gerg and Monga (2022)	Target vs. clutter	CNN, DensNet-121, ResNet-18
Williams (2019)	UXO vs. Non-UXO	CNN
Dzieciuch et al. (2017)	MLO vs. background	CNN
Chapple et al. (2017)	MLO, NMLO, FAO, seabed	InceptionNet
Gebhardt et al. (2017)	MLO vs. Non-MLO	CNN
Phung et al. (2019)	MLO, NMLO, FAO	CNN, AlexNet, VGG-16, VGG-19, ResNet-50, Inception-v3, GoogLeNet
Bouzerdoum et al. (2019)	MLO, NMLO, FAO	CNN, VGG
Quidu et al. (2005)	Mines	MLP
Williams (2016)	Mines	CNN
McKay et al. (2017)	Mines	VGG-16, VGG-19, VGG-f, AlexNet
Zhu et al. (2018)	Mines	CNN
Warakagoda and Midtgaard (2018)	Mines	AlexNet, VGG-16, DenseNet-161, Inception-ResNet-v2, NasNet-large, CNN

[a]This notation refers to 3 residual blocks with 2 residual connections.

[b]Only the classes anchor, cube, plane, boar and pyramid are explicitly mentioned in the paper.

types of mines (five publications each). Seven publications considered ship and plane wrecks beside other classes. Additional information about the datasets used in the papers is summarized in Table A.1 in the appendix. Nearly 50% of the investigated research papers implemented a custom CNN architecture. The second most used architecture is VGG which was considered in 10 out of the 35 works.

Because it is not clear that a method which performs well on one type of object (e.g. shipwrecks) will also perform well on another type (e.g. mines) and whether results from the former apply to the latter, the following discussion is grouped by the class of objects as given in Table 2. By discussing and comparing the individual papers we will derive common findings as well as research gaps.

Chen and Summers, Berthold et al. Luo et al. and Qin et al. used CNNs for the classification of different sediment types (Chen and Summers, 2016; Berthold et al., 2017; Luo et al., 2019; Qin et al., 2021). In an early work from 2016, Chen and Summers showed that pre-training of an AE improves the classification performance of the encoder CNN after fine-tuning. Less than 100 labeled samples per class in the fine-tuning step were sufficient to achieve a classification accuracy of up to 88.2%. In addition, they showed that a generative adversarial network (GAN) can generate realistic images of different seafloor types. Berthold et al. used GoogLeNet to classify the four types coarse sediment, mixed sediment, sand and fine sediment (Berthold et al., 2017). The network was trained from scratch using patches from a sidescan mosaic image. It showed a classification accuracy of 83% for the class sand but failed to classify the fine sediment correctly. Luo et al. compared a shallow CNN based on LeNet-5 with a deeper CNN based on AlexNet and found that the former performed better when dealing with a small dataset (<250 samples per class) (Luo et al., 2019). For the class sand wave it achieved an accuracy of 93.43%. The work of Qin et al. (2021) extended their previous work (Luo et al., 2019) by considering pre-training on grayscale CIFAR-10 images and augmentation of the dataset using a GAN which generates synthetic SSS snippets. They analyzed a large number of different CNNs for the classification task (see Table 2). Deeper networks especially benefit from the pre-training with DenseNet-151 and ResNet-4-2 having the lowest error rate. Note that the notation ResNet-4-2 here means four residual blocks with two residual connections. The augmentation using a GAN was only applied to some subset of the data but showed a slight improvement. Because different datasets and metrics were used in the four papers, a direct comparison between the results is difficult. Nevertheless, the benefit of using a pre-training dataset was also observed in many other computer vision problems (Valverde et al., 2021; Mensink et al., 2021). Identifying different seabed types is also a classical segmentation problem in the underwater domain which is further covered in Section 5.

Researchers from the Harbin Engineering University, China focused their work on classifying shipwrecks, plane wrecks and other objects like stones (Ye et al., 2018; Wang et al., 2019; Xu et al., 2020; Li et al., 2021). In Ye et al. (2018) transfer learning from ImageNet to SSS images was analyzed using a VGG-11 and a ResNet-18. The authors found that training the whole network from scratch as well as fine-tuning the whole network results in overfitting. Fine-tuning of the network's last layer led to a model with good performance. Deep belief networks (DBN) and a method called adaptive weights convolutional neural network (AW-CNN) were investigated in Wang et al. (2019). AW-CNN uses the weights from a DBN to initialize the CNN which results in a better performance compared to a CNN trained from scratch. Xu et al. combined the classification output of a CNN and an SVM which was trained with features from the CNN (Xu et al., 2020). This incorporation of the SVM led to an increased classification accuracy. The authors also augmented their dataset with synthetic images generated from a Wasserstein GAN. Augmenting the dataset in this way increased the performance even further. However, no analysis of the image quality or diversity was made. A zero-shot classification approach based on style transfer was developed in Li et al. (2021). In zero-shot learning not a single real sonar image is directly used for the training of a CNN. Using an AE architecture, the style of SSS images was transferred to natural RGB images of objects similar to the considered classes (e.g. ships for the class shipwrecks). A ResNet-34 was trained on these synthetic images and tested on real sidescan sonar data, achieving a mean accuracy of 75%. Huo et al. analyzed transfer learning of CNNs for classifying SSS images of shipwrecks, plane wrecks, corpses, mines and background samples (Huo et al., 2020). They found that pre-training a VGG-19 on ImageNet and fine-tune it on the sonar images performs better than using an SVM with hand-crafted features or training a CNN with two convolutional layers from scratch. In Nayak et al. (2021) additional hand-crafted features were incorporated into a CNN with three convolutional layers in order to improve the classification of sidescan sonar snippets into the classes archaeological sites (e.g. shipwrecks) and background. An additional improvement was made by pre-processing all snippets using an edge detector and feeding these images to the CNN. Comparability of the individual results and derivation of a best approach is hard, just as with feature extraction, since the datasets used differ in many dimensions, e.g. number of images as shown in Table A.1, pre-processing or number of classes. Nevertheless, one common finding so far is that pre-training or initialization of a CNN using DBN is to be preferred over the training from scratch. Recently, Cheng et al. improved the classification performance of a VGG19 by incorporating mulit-domain pre-training and attention modules to the network (Cheng et al., 2022). Synthetic aperture radar (SAR) images, which have a small domain gap to sonar images, were used to pre-train the first convolutional layers for the VGG19. Grayscaled optical images of ships, airplanes and the sea surface were used to pre-train the fully connected layers. This type of transfer-learning combines the benefits from the low-level similarity between SAR and SSS as well as from the high-level similarity between the three classes in sonar and optical images. In addition, a combination of channel and spatial attention was used to help the network to focus on important parts of the input images.

One paper considered a CNN for the classification of sonar images into the class rock or cylinder. Karjalainen et al. trained a CycleGAN and showed that a CNN trained on the synthetic images performs as good as a CNN trained on real data (Karjalainen et al., 2019). Enlarging the training dataset with synthetic images generated by a GAN is an approach which is more and more often adopted in the sonar imagery domain Reed et al. (2019), Steiniger et al. (2020), Jegorova et al. (2020), Xu et al. (2020) and Qin et al. (2021).

In Ochal et al. (2020) few-shot learning was investigated for the classification of SSS images and compared to transfer learning. More precisely, Prototypical Network (PN), Relation Network, Soft k-means PN and Consistent Prototypical Network (CPN) were used as few-shot learning methods while a custom CNN, ResNet-18 and ResNet-50 were used for transfer-learning. On a dataset consisting of simulated SSS images from 18 different objects the few-shot learning method performed slightly better than the pure CNNs. Contrary to the findings of other works (Ye et al., 2018; Huo et al., 2020) training both ResNets from scratch rather than using the pre-trained weights obtained using the ImageNet database performed better. However, the number of simulated images per class was not mentioned in Ochal et al. (2020) making a fair comparison between the different works more difficult.

A lot of work in the field of classification of sonar imagery has been done by the NATO CMRE (Williams and Dugelay, 2016; Williams, 2017, 2018b,a; Gerg and Williams, 2018; d'Alès de Corbet et al., 2019; Williams et al., 2019; Berthomier et al., 2020; Williams, 2021, 2019, 2016). Most of the work deals with the differentiation between target and clutter snippets from a SAS, where the class target typically includes mine-like and other man-made objects. In their first work, Williams et al. trained a deep Boltzmann machine (DBM) in a multi-view scenario where images of one object from different perspectives are available (Williams and Dugelay, 2016). Using multiple views increased the performance of the DBM. The benefit was greater if the images were fused prior to the training rather than fusing the

predictions of the individual views. Combining multiple views of an object was further investigated in d'Alès de Corbet et al. (2019). Here two different CNN architectures were considered. One averaged over the predictions of the CNN applied to each view individually. The second one had a separate branch for each view which were then concatenated prior to a fully connected layer. The authors showed that the second approach gives a slight improvement over averaging of the individual predictions. In Williams (2017) different custom CNN architectures with the number of convolutional layers ranging from two to five were studied and compared to a relevant vector machine (RVM). The results showed that all CNNs outperform the RVM and an ensemble of CNNs led to an additional improvement. Their best performing single model was a CNN with four convolutional layers. Some publications deal with additional input information and how it can be used to improve the performance of a CNN (Williams, 2018b; Gerg and Williams, 2018; Williams et al., 2019). The phase information of the complex SAS data rather than the amplitude was successfully used in Williams (2018b) to classify target and clutter snippets. However, the CNN architectures are not comparable with the ones from Williams (2017) and thus no direct statement about which type of input image is better suited can be made. Nevertheless, in most recent publications only the images containing the amplitude are used, indicating that this representation severs as the better input for CNNs. The usages of additional input information for the classification of SAS snippets was further analyzed by Gerg and Williams (2018). The amplitude, phase and 2D power spectral density (PSD) were considered individually and combined as separate inputs to a CNN. Phase-only and PSD-only performed worse but the PSD in addition to the amplitude gave a slight boost over amplitude-only. Additionally, a VGG pre-trained on ImageNet and fine-tuned on SAS amplitude images achieved the best performance. In Williams et al. (2019) different representations of SAS snippets were considered for the classification. This includes translation and horizontal flipping at test time as well as the phase image and frequency spectrum as two additional branches in the CNN. All three branches were concatenated prior to the fully connected layer. The augmentation at test-time and the additional input information improved the classification performance. Furthermore, an ensemble of four small CNNs which used multiple representations achieved the same performance as a VGG-16 pre-trained on ImageNet and fine-tuned on the amplitude images. Target-concept transfer and sensor transfer were studied in Williams (2018a). In the first case a CNN was pre-trained to classify between mines and clutter and then transferred to classify between UXO and clutter which now contained mines as well. For sensor transfer the CNN was pre-trained on data from one SAS system and transferred to data from another SAS which operated at a different frequency band. On both transfer learning tasks, the performance after fine-tuning was better if the amplitude information was used rather than the phase information. Berthomier et al. added auxiliary information like image quality or target shape to be predicted by a CNN (Berthomier et al., 2020). This resulted in a slight drop in classification performance considering target vs. clutter but at the same time led to a model which outputs more information about an image. Several different CNN architectures were recently studied in Williams (2021). In total eight CNNs with the number of convolutional layers ranging from four to twenty were designed and compared. One finding from this work was that CNNs with smaller kernel sizes but larger pooling factors tended to perform better than CNNs with larger kernels and smaller pooling factors. Their best performing CNN had twelve convolutional layers with mostly 4×4 and 5×5 kernels. In conjunction with previous work (Williams, 2017; Williams et al., 2019) they found that test-time augmentations using translation and horizontal flipping as well as an ensemble of the eight CNNs improved the performance. Other researchers also investigated CNNs for the classification of SAS snippets (Galusha et al., 2019). They considered a dual-frequency SAS and designed a CNN with two input channels, one for each frequency. Data augmentation using horizontal flipping and a slight rotation of + /- 10 degrees during training improved the performance of this CNN. The authors provided no analysis regarding the benefit of using two frequencies. But since the previously discussed publications have shown the benefit of multiple representations and because the two images for different frequencies provide the network with more information an improvement over the single-frequency approach is expected. Gerg and Monga showed that incorporating domain knowledge about SAS images and the detection step leading to the ROIs into the neural network improves the classification performance (Gerg and Monga, 2022). Their first domain knowledge prior stated that SAS images contain speckle noise. A U-Net was trained in order to learn a denoising of the sonar image prior to the feature extraction with a DenseNet-121. Additionally, the overall network did not only predict the class of the object but also its location inside the ROI. This was based on the second domain knowledge prior that the detector extracts the snippet with the object centered in the middle. Thus, augmenting the dataset through translation and learning this offset directly led to a translation invariant CNN. Their architecture achieved a better classification performance than a DenseNet-121 without priors, a ResNet-18 and the CNN described by Galusha et al. (2019) but with only a single input. So far, research work on the classification between target and clutter snippets has shown that multiple views from an object as well as different representations are beneficial for the use of CNNs. An optimal network architecture is not found, as one work suggested four convolutional layers while a more recent work proposed twelve layers. However, these custom CNNs were not compared to state-of-the-art networks like EfficientNet or ViT. Comparisons between a fine-tuned VGG and a custom CNN showed a similar performance for the two networks.

In Williams (2019) the previous work on target-concept transfer learning (Williams, 2018a) was expanded and applied to the classification between UXO and non-UXO. Four CNNs with four to twelve convolutional layers each were trained on two datasets captured by different SAS systems. The first training dataset contained nearly 3000 target snippets while the second dataset only contained 65 UXO images. Similar to Williams (2018a) pre-training on the first dataset and fine-tuning on the second led to a better performance than training only on the second dataset.

Another prominent use case of deep learning is the classification of sonar snippets into the classes mine-like object (MLO), non-mine-like object (NMLO) and false alarm object (FAO) (Dzieciuch et al., 2017; Chapple et al., 2017; Gebhardt et al., 2017; Phung et al., 2019; Bouzerdoum et al., 2019). Dzieciuch et al. used a CNN with only one convolutional layer and analyzed the effect of the number of training epochs and batch size on the classification accuracy (Dzieciuch et al., 2017). Using 250 training images they achieved an accuracy of 99% when classifying between MLO and background snippets which were considered as FAO in this case. As is commonly known, training converged faster using a larger batch size. In Chapple et al. (2017) an InceptionNet pre-trained on ImageNet was fine-tuned using sonar images to improve the performance of an ATR system. Similar to Williams (2021) Gebhardt et al. performed a study on different CNN architecture with a varying number of convolutional layers ranging from one to nine (Gebhardt et al., 2017). More than five convolutional layers led to a slight improvement in accuracy at the cost of a longer inference time. Their best model was a seven layer CNN and all CNNs with more than one convolutional layer outperformed an SVM. A study about deep pre-trained CNNs was carried out in Phung et al. (2019). In their analysis VGG-19 performed best, while ResNet-50 only achieved an accuracy of 55.76%. A second finding from their comparison was that in most cases replacing the last layer of a deep pre-trained CNN with an SVM performs equal or better than fine-tuning the whole CNN. An explanation for this was not given in the publication but the small training dataset of only 199 samples give rise to the argumentation that more data is required for successful fine-tuning. The authors also proposed a hierarchical Gaussian process (HGP) classifier which used the features from the fully connected layer of a CNN as an input. This

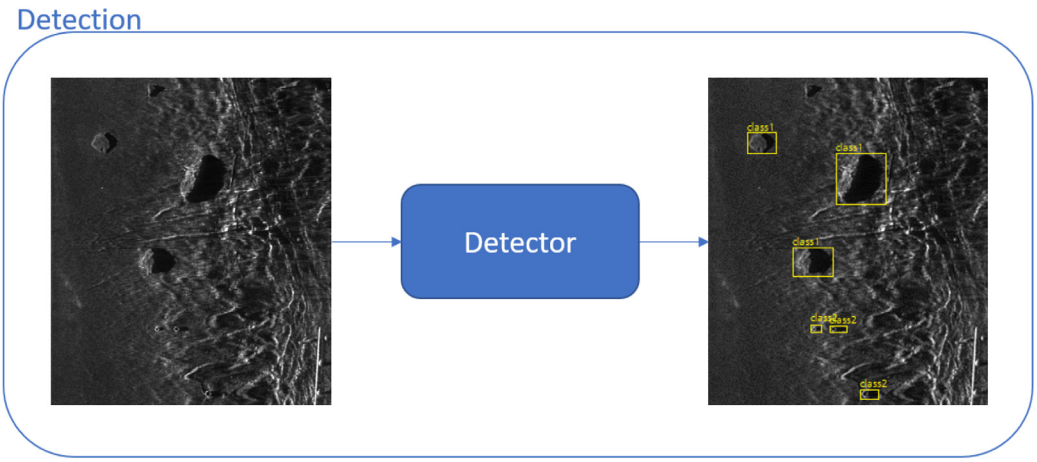


Fig. 7. Scheme of the detection subtask. The detector localizes the object by predicting a bounding box and simultaneously classifies the containing object.

combination of CNN and HGP reached a higher classification accuracy than the VGG-19 at the cost of higher computational complexity. In Bouzerdoum et al. (2019) transfer learning and data augmentation were investigated. For transfer learning they considered a VGG network and found that, similar to Phung et al. (2019), replacing the last fully connected layer with an SVM performs better than fine-tuning the whole network. The augmentation methods found to be useful are horizontal flipping, scaling and extracting the object and placing it onto a different background snippet. Furthermore, a shallow CNN consisting of two convolutional layers was designed, which outperformed the combination of VGG and SVM (which in turn had outperformed the fine-tuned VGG). This finding seems contrary to some other work like Gerg and Williams (2018) and Williams et al. (2019) were the VGG showed the best performance. One reason for these two different findings is most likely the dataset size. Bouzerdoum et al. (2019) used 176 MLO, 40 NMLO and 196 FAO snippets, while in Gerg and Williams (2018) and Williams et al. (2019) 2912 target and 29,280 clutter snippets were used for training.

The first work dealing with a deep learning based approach for sonar image analysis was done by Quidu et al. (2005). They used an MLP to classify different types of mines based on hand-crafted features extracted from the objects shadow region. More than ten years later, David Williams was the first to apply CNNs to sonar imagery for the task of classifying mines in a binary manner (Williams, 2016). In his work he showed that, similar to his findings in Williams and Dugelay (2016), the CNN outperforms a RVM based on hand-crafted features. As already mentioned in Section 2 the authors of McKay et al. (2017) were not able to fine-tune a VGG network without employing an SVM on the classification task due the lack of data. In Zhu et al. (2018) the authors used an AlexNet-like CNN for the classification of different types of mines. Due to a lack of data they pre-trained the CNN with images simulated by a ray tracer and showed that this increases the classification accuracy. The fine-tuned CNN as well as the one trained from scratch outperformed an SVM. Transfer learning in the context of mine classification was also investigated by Warakagoda and Midtgaard (2018). Different ways of fine-tuning AlexNet and VGG-16 were analyzed by freezing specific parts of the networks. They found that fine-tuning the whole network performs better than fine-tuning only some layers at the input or output of the network. Furthermore, the authors compared several deep CNNs pre-trained on ImageNet and fine-tuned on their SAS image data. Networks with a better performance on ImageNet also performed better on the sonar data with NasNet-large being the best model on both datasets.

This overview has shown that deep learning methods have led to improvements on several different sonar image classification tasks over the past years. Starting from MLPs which used hand-crafted features, the design of shallow custom CNNs as well as the transfer of deeper networks to the sonar domain have emerged as main research fields. One obstacle when applying deep learning to sonar image classification is the lack of (publicly) available data. This results in rather small networks, the development of specific augmentation methods (e.g. using GANs or ray tracing), few- or zero-shot approaches and special improvements which utilize the properties of the sonar domain. Multiple views of an object, the combination of different data types (e.g. the phase information) and domain knowledge improves the classification of sonar images. In deep learning the quality and amount of data used for training and testing has a high impact on the reported performance of a method. Nevertheless, all investigated papers used a different dataset (see also Table A.1 in the appendix). To make further progress in applying deep learning to sonar imagery, a fair comparison in terms of performance, training and inference speed and stability between developed methods is necessary. For this, we suggest to create at least one publicly available benchmark dataset.

4. Detection

4.1. Subtask explanation

In the conventional ATR processing chain for SSS and SAS images the detection serves mainly as a method for locating possible ROIs. The objective of this approach is to favor a high recall over a high precision as false alarms are filtered out by a subsequent classification step (see Fig. 3). In classical ATR processing chains the detection step is typically implemented as a basic template matching procedure (Fei et al., 2015). Considering detection from a deep learning perspective, it combines the localization, classification and estimation of the extent of an object into one step. This subtask is displayed in Fig. 7. The algorithms directly output the location of an object in the form of a bounding box enclosing it. Here a high recall and precision should be achieved. Deep learning detection methods that are applied and especially designed for object detection in sonar images are considered in the following section.

4.2. Deep learning applications

The research papers dealing with the detection of objects in SSS or SAS images are listed in Table 3. In this table the column *Object* specifies the type of object that is considered for detection in the respective work while *Method*, as before, summarizes the deep learning approaches used. Most works considered the detection of MLOs (5 out of the 12 publications). In 50% of the publications a version of YOLO

Table 3
Publications on deep learning for detection in sonar imagery.

Paper	Object	Method
Xu et al. (2019)	Shipwrecks	YOLOv1, Faster R-CNN
Jiang et al. (2020)	Shipwrecks, plane wrecks, corpse	Faster R-CNN, YOLOv1, SSD
Yu et al. (2021)	Shipwrecks, container	YOLOv5, Transformer
Einsidler et al. (2018)	Anomaly, rock	YOLOv2
Feldens et al. (2019)	Rock	RetinaNet
Feldens (2020)	Rock	RetinaNet
Denos et al. (2017)	MLO	AE, CNN
McKay et al. (2017)	Mines	VGG-16, VGG-19, VGG-f, AlexNet
Berthomier et al. (2019)	MLO	CNN
Le et al. (2020)	MLO	Gabor CNN, R-CNN, Fast R-CNN, Faster R-CNN, Tiny YOLOv3, YOLOv3, SSD300
Topple and Fawcett (2021)	MLO, NMLO	MiNet
Steiniger et al. (2021a)	Target	YOLOv2, YOLOv3, CNN

was used for detection. The second most utilized architecture is Faster R-CNN which was considered in three works. The individual papers are grouped by common objects that should be detected.

A typical object to be detected in sonar images is a shipwreck. In Xu et al. (2019) YOLOv1 and Faster R-CNN were compared for detecting shipwrecks. Due to the lack of own data the authors took SSS, SAS, multibeam echosounder and optical images from the internet as training and test data. A GAN-based approach was used to further augment the dataset. With the additional data the performance of YOLOv1 improved, although the generated images show artifacts which make them easy to distinguish from real sonar images. The Faster R-CNN was not trained with the augmented dataset but achieved a better performance than YOLOv1 when trained on the baseline dataset. An active learning approach for the detection of shipwrecks and other objects was discussed in Jiang et al. (2020). The three methods uncertainty sampling, uncertainty and diversity sampling and local information selection were compared. Faster R-CNN was used as a detector. With both uncertainty sampling methods 1500 samples were selected to achieve the same performance as when 3000 samples were selected randomly. This shows a higher efficiency and reduces the labeling effort. The framework was extended to SSD and YOLOv1 with Faster R-CNN and SSD performing slightly better than YOLOv1. Very recently, Yu et al. were the first to investigate the self-attention mechanism in the context of object detection in sonar images (Yu et al., 2021). Adding a multi-head self-attention module to YOLOv5s led to an increase in performance with nearly no computational overhead. The authors also found that pre-training is preferable over training from scratch, however the used pre-training dataset is not specified.

The first group to apply a standard deep learning detector to SSS images was Einsidler et al. (2018). They showed that YOLOv2 can already be trained with less than 150 images to detect rocks and other objects on the seafloor. However, no performance evaluation in terms of a calculated metric was provided. Feldens et al. also considered the detection of rocks (Feldens et al., 2019). In their work they trained a RetinaNet and showed that a better performance is achieved if smaller patches from the SSS image are used. Specifically, patches of size 25 m^2 and 225 m^2 with a resolution of 0.25 m per pixel were considered prior to upscaling to the input size of the RetinaNet. Further analyses showed that a rock needs to encompass at least 3×3 pixels in the image to be detected reliably. In Feldens (2020) this work was extended by using a single-stage residual network for super-resolution. Increasing the resolution of the sonar image through super-resolution resulted in an improvement of the detection especially for small rocks. Besides the classical application of ATR, findings like these also serves for studies on the maritime ecosystem or for geoengineering purposes.

In Denos et al. (2017) an AE was used in a first stage to learn features from synthetic SAS snippets containing a simulated MLO. The reconstruction error of the AE was then used to generate a heatmap of a larger, real SAS image and to extract snippets of background without an object. A VGG-like CNN was trained on the simulated target and real background snippets. However, this method led to a large number of false alarms since other objects were also filtered by the AE and were thus not contained in the background class. The method from McKay et al. (2017), which was already discussed in Section 2, can be extended to perform detection. Here the combination of AlexNet as a feature extractor and an SVM was used to calculate a score for small patches of a large sonar image indicating the presence of an object. The resulting heatmap was thresholded to obtain the final detections. This approach was further analyzed by Berthomier et al. (2019). A CNN with four convolutional layers, which had been trained for the classification between MLO and clutter, was used instead of the AlexNet and SVM combination. For the detection the input size of the CNN was increased in order to receive a large SAS image. This method performed well for a rather flat seafloor but still needs to be improved for more complex seafloor types, e.g. the presence of sand ripples. In Le et al. (2020) a Gabor filter based neural network was designed and compared to a large range of deep learning detectors. The architecture of the Gabor CNN was similar to YOLOv3 and performed detection at multiple scales. By using Gabor filters instead of classical convolution kernels the detection performance was improved compared to YOLOv3. However, this improvement came at cost of a longer inference time. Other detectors considered in the paper (see Table 3) performed worse. The Gabor CNN was also compared to the previously mentioned approach from McKay et al. and showed a higher detection rate as well as a lower false alarm rate while running at a higher speed. A so-called MiNet was developed for on-board detection of MLOs on an AUV in Topple and Fawcett (2021). MiNet is a one stage detector similar to YOLO with a smaller backbone. The authors suggested an incremental training procedure. First the network was trained on synthetic images. In a second step, synthetic images of objects were combined with real background samples. Finally, real sonar images were used for training. The network fulfilled the on-board requirements in terms of memory

consumption (9.9 MB) and processing power (0.0122 GFLOPS) and was able to detect objects in a sonar image within minutes. Unfortunately, no quantitative analysis or comparisons with other methods were done in the publication.

A comparison between YOLOv2, YOLOv3 and a small CNN used for detection similar to Berthomier et al. (2019) was carried out in Steiniger et al. (2021a). YOLOv2 and YOLOv3 were pre-trained on the MS COCO dataset, while the CNN was trained from scratch. In the experiments YOLOv3 showed the best performance leading the authors to the conclusion that fine-tuning a standard deep learning detector is preferable over training a custom network from scratch.

Reviewing the work on the detection subtask has shown that basic CNNs applied to a whole sonar image as well as standard deep learning detection methods like YOLO have been investigated in the literature. Custom methods involve the usage of Gabor filters which have proven to be suitable for sonar images as well as custom backbones which are smaller compared to standard models. Both observations are a consequence of sonar images having fewer details compared with natural RGB images, and that only a limited amount of training data is available. However, recently developed methods like DetectoRS or DETR have not yet been considered for SSS and SAS images. Finally, as for the classification subtask, the lack of a common dataset makes it hard to answer which detection algorithm works best on SSS or SAS images. In addition, no consistent metric is reported in the investigated papers. Future results should always be reported in a standard metric like average precision to allow an easy and fair comparison.

5. Segmentation

5.1. Subtask explanation

The segmentation of a sonar image results in a masked image where specific regions in the image are assigned a certain label. In the conventional ATR processing chain segmentation is used to extract the highlight and shadow areas caused by an object in order to further extract features from these areas, e.g. the size or shape. In the context of computer vision the segmentation is used for a larger variety of goals. Three types of segmentation are distinguished: semantic segmentation, instance segmentation and panoptic segmentation. An example of each type of segmentation is shown in Fig. 8. Semantic segmentation assigns a class label to each pixel, e.g. to identify different seafloor types in a sonar image. In contrast to this, instance segmentation has the goal of assigning an object ID and a class label to all pixels belonging to one object. Pixels which do not belong to an object would not be masked in instance segmentation. Panoptic segmentation combines both semantic and instance segmentation and assigns a semantic label to each pixel as well as an object ID to all objects. Research on investigating deep learning methods to solve the segmentation task on SSS and SAS images is reviewed in the following section.

5.2. Deep learning applications

Table 4 lists the papers that we identified using the keywords and that deal with the task of sonar image segmentation. The column *Task* specifies what should be segmented for that particular work and is considered for the grouping of the investigated papers. *Method* again summarizes the deep learning methods that are used.

Common segmentation tasks are the segmentation of sand waves (3 of the 10 research works), segmentation of highlight, shadow and background regions (2 of 10) and segmentation of prominent linear structures (2 of 10). In four of the ten research works a fully convolutional network (FCN) was used. Besides this, DeepLab, U-Net and SegNet were used in two works each. All of the identified research papers considered only semantic segmentation for sonar images. For instance and panoptic segmentation objects need to be present in the images which are then assigned an object ID. However, only the work from Chen and Summers (2017) considered the segmentation of objects. Segmenting objects can be seen as an extension of the detection step, since in addition to a bounding box, a mask must also be predicted for each object. The delay between the development of new deep learning methods or tasks and their application to the sonar imagery domain (see Fig. 1) can thus also been observed here. We will now investigate the segmentation approaches of the individual tasks in more detail.

Zheng et al. considered the task of differentiating between water column and seafloor as a segmentation problem and applied DeepLabV3+ to solve it (Zheng et al., 2021). They also proposed a preprocessing module which combined the original image, a horizontally flipped version and the mean of these two into three channels. This preprocessing helped the segmentation network to learn the symmetrical behavior of the sea bottom line on the port and starboard side and improved the mean intersection over union (IoU) slightly. Furthermore, based on a first coarse segmentation the image was cropped and used for a fine segmentation. The overall method achieved a mean pixel error when comparing the extracted sea bottom line with the ground truth of 1.1 pixel. Note that the authors only provided the mean pixel error and not an average of the absolute error which can be misleading since large positive and negative errors are evened out. Nevertheless, the comparison with a conventional method showed a qualitatively better result for the deep learning approach. It should be noted that other deep learning approaches for determining the sea bottom line exist which process the one dimensional time signals of a sonar (Yan et al., 2020, 2021; Qin et al., 2021). Because this is not considered as sonar imagery the papers are not listed in Table 4. Promising approaches for this particular use case are based on a 1D U-Net (Yan et al., 2021; Qin et al., 2021).

Wu et al. as well as Wang et al. segmented prominent lines in sonar images (Wu et al., 2019; Wang et al., 2020). Wu et al. developed a method which they call efficient convolutional network (ECNet) and which is based on an encoder–decoder architecture (Wu et al., 2019). The final output was the average of the output from the first three encoders and the last decoder. To deal with the imbalance between highlight and background pixels the authors implemented a weighted loss function. Their method was compared to a U-Net, SegNet and LinkNet and showed a better performance with a mean IoU of 66.18%. Wang et al. used a simpler approach based on a VGG-16 transformed into a FCN (Wang et al., 2020). Skip connections and batch normalization were added to the architecture to deal with exploding gradients during training. To solve the problems provoked by the imbalanced dataset the authors used a weighted loss function too. The FCN outperformed traditional methods like fuzzy c-means and the Canny edge detector, reaching a mean IoU of 83.05%. However, one needs to be careful when comparing this result with the work of Wu et al. because Wang et al. augmented a base dataset of 50 sonar images to obtain 2000 images and split this augmented dataset into training and test set. This causes similar images to be contained in both sets.

In Song et al. (2017) the authors used a FCN to segment sonar images into the typical regions highlight, shadow and background. A Markov random field (MRF) was applied to post-process the segmented image. The combination of FCN and MRF showed a better performance than both methods individually. In Song et al. (2021) the authors extended their work by considering self-cascaded CNNs for the segmentation task. The self-cascaded CNN was compared to a normal CNN of the same size and the conventional methods fuzzy c-means and MRF. As in Song et al. (2017) the segmentation result was post-processed by a MRF. The performance of the self-cascaded CNN was the best in their comparison. However, the FCN from Song et al. (2017) was not taken into account. When comparing the results of the two papers one needs to be careful since the dataset as well as the reported metric (accuracy in Song et al. (2017) and mean IoU in Song et al. (2021)) are different. Nevertheless, the qualitative segmentation results shown in the papers indicate a good performance for both methods but the self-cascaded CNN was trained with less data.

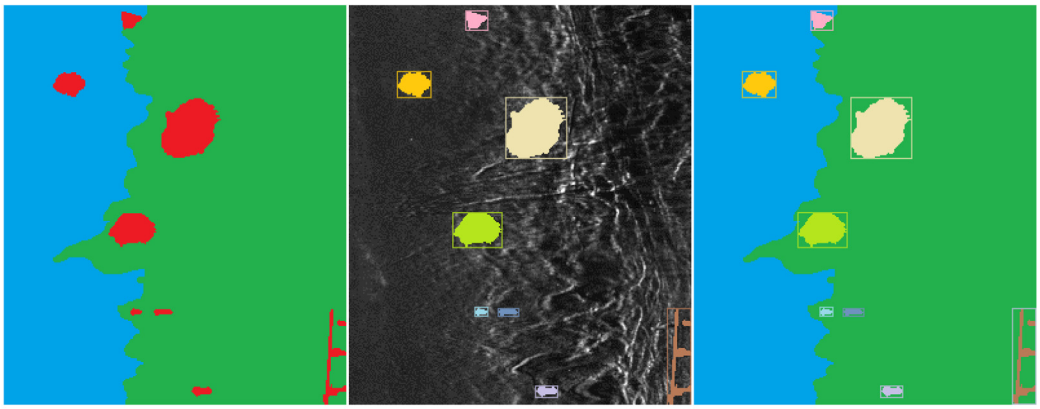


Fig. 8. Different types of segmentation of a sidescan sonar image. (a) Semantic segmentation. Each pixel is assigned a semantic label, e.g. flat, wave, object. (b) Instance segmentation. All pixel belonging to an object are assigned a unique ID. (c) Panoptic segmentation. Each pixel is assigned a semantic label and pixels belonging to an object are assigned a ID.

Table 4
Publications on deep learning for segmentation of sonar imagery.

Paper	Task	Method
Zheng et al. (2021)	Water column	DeepLabV3+
Wu et al. (2019)	Lines	ECNet, U-Net, SegNet, LinkNet
Wang et al. (2020)	Lines	FCN
Song et al. (2017)	Highlight, shadow, background	FCN
Song et al. (2021)	Highlight, shadow, background	self-cascaded CNN
Yu et al. (2019)	Sand waves	DeepLab
Li et al. (2019)	Sand waves	ENet, SegNet
Nian et al. (2021)	Sand waves	ELM, MobileNetV1
Rahnemoonfar and Dobbs (2019)	Seagrass, pothole	U-Net, FCN
Chen and Summers (2017)	Target	FCN

Another application of segmentation is the determination of different seafloor structures or sediment types. Yu et al. used DeepLab in order to segment sand waves from the remaining seafloor (Yu et al., 2019). They improved the performance by augmenting the dataset and increasing the image resolution using a CNN for super-resolution. Adding a MRF in a post-processing step again showed a slight improvement. In Li et al. (2019) the same authors compared ENet and SegNet for the segmentation of sand waves. For ENet a MRF was added for post-processing purposes. In their comparison ENet was faster than SegNet and showed a better performance. Again the datasets used for training and testing were different in Yu et al. (2019) and Li et al. (2019) which needs to be mentioned when comparing the performance of ENet or SegNet with the method of Yu et al. Nevertheless, the large difference in mean IoU between ENet with 0.90 and the extended DeepLab with 0.59 is most likely not only dataset related. Segmentation of sand waves was also considered by Nian et al. (2021). Here the authors used an extreme learning machine (ELM) to classify sub-sequences of the individual pings of SSS data into sand wave or no sand wave. The predicted label was then mapped to the sonar image at the location of the considered sub-sequence to form the segmentation result. Their method was compared to MobileNetV1 which classified not the sub-sequences but patches into the two categories. Because the ELM only deals with one dimensional data the network was smaller and thus faster at test time. ELM-based segmentation had a slightly higher performance in terms of accuracy and F1-score compared to MobileNetV1. Rahnemoonfar and Dobbs segmented potholes in a seagrass bed (Rahnemoonfar and Dobbs, 2019). They developed a U-Net like architecture with dense blocks on the encoder side and so-called inception-deconvolutions on the decoder side. Inception-deconvolutions are based on transposed convolutional layers with rectangular kernels as suggested by the InceptionNet architecture. Their model was outperforming an FCN on the segmentation task and used less parameters.

A slightly different approach of segmenting a sonar image was considered in Chen and Summers (2017). The authors employed a fully convolutional ladder network, which simultaneously segmented the image, reconstructed it from an internal representation and classified it. Using the two additional tasks, data which is not labeled for segmentation but for classification can improve the training of the network. The method was only evaluated qualitatively but generalized well on the test images shown in the work.

Although different types of segmentation exist, all research on segmenting SSS or SAS images has focused solely on semantic segmentation. The considered deep learning segmentation methods outperformed simple conventional baselines like fuzzy c-means. Comparisons with a stronger baselines, e.g. snake based approaches, are necessary in order to assess the potential of deep learning methods for the segmentation of sonar images. Finally, as for the classification subtask, the lack of a common dataset makes it hard to compare the results between different papers and to give an answer to the question which algorithm is the best to segment a sonar image.

6. Research directions

Based on the abstraction of all 62 papers considered here, we will derive further research directions from three different viewpoints: method, completeness, data. The viewpoint *method* recaptures for each subtask the most promising deep learning methods which are treated as state-of-the-art. They need be considered by researchers starting their work in the field of SSS and SAS image processing. In addition to this, *completeness* shows current gaps between state-of-the-art methods in more common computer vision applications and sonar imagery. Potential further research directions and promising methods are derived from that. Finally, *data* lists relevant concerns as well as research directions with respect to the datasets used for ATR in SSS and SAS images.

6.1. Methods

Despite the lack of a consistent test dataset, there are still some findings regarding an optimal deep learning method which are supported by multiple papers. Suggestions for which deep learning method to use in the individual subtasks based on the previous discussions are summarized in Fig. 9. Note however, that a final decision on which single method is best suited for the respective ATR subtask cannot be deduced due to the lack of comparability of the individual publications. A broad comparison on a benchmark dataset is necessary in order to close this research gap.

For the feature extraction subtask CNNs have shown to provide better features for a subsequent classification than hand-crafted engineering. However, since CNNs also outperform conventional classification algorithms like SVM the feature extraction task has lost its relevance and is mostly included directly in the classification. Nevertheless, hand-crafted features have shown good performance for more than 20 years (Dobeck et al., 1997). An investigation into what extent a CNN learns features similarly to those methods may give both insight to the CNN and the opportunity to combine beneficial hand-crafted features into a CNN.

When considering classification, the methods can be divided into three groups depending on the amount of available training data. Less than 200 images are not sufficient to fine-tune a VGG network or to train a CNN from scratch. However, replacing the fully connected layers with an SVM should lead to good results. Fine-tuning a VGG architecture or training a shallow CNN with up to 4 convolutional layers is a good choice when having a dataset of medium size. If more data is available larger networks like DenseNet can be fine-tuned.

Only few comparisons were made between different deep learning detection algorithms in the considered research papers. Faster R-CNN seems to perform better than YOLOv1 but the updated version YOLOv3 outperforms the former. A very promising approach are Gabor CNNs which are very suited for SSS and SAS images. In general, fine-tuning a pre-trained deep learning detector leads to better results than those obtained when training a custom CNN from scratch and applying it for detection. For semantic segmentation a post-processing using MRF has shown to lead to better results in every paper where it was applied. The typical deep learning methods which are applied for this subtask are FCN or encoder–decoder architectures.

6.2. Completeness

Looking at the state-of-the-art on the standard computer vision benchmark ImageNet self-attention based models are currently at the top. However, such methods, e.g. VisionTransformers (Dosovitskiy et al., 2021), SWIN Transformers (Liu et al., 2021) or very recently CoAtNet (Dai et al., 2021), have not yet been applied to the sonar domain. Transformers generally show their full potential when a very large dataset is available. Whether they can successfully be trained with a limited amount of available SSS or SAS images needs to be investigated. The most recent network architecture that is considered in one of the consulted research papers is NasNet which was published in 2018. Since then, more architectures, e.g. ResNeXt or EfficientNet, have been proposed and show better performance on ImageNet than NasNet. Transfer-learning of larger models has proven to be beneficial when only a medium sized dataset is available. Thus, a study on transfer-learning of modern CNN architectures and a comparison to a custom CNN has to be considered in one of the next steps. Research on augmenting and generating synthetic sonar images needs to be continued to ensure that larger models can be trained using limited data.

As with the classification subtask transformers have currently been applied for detection with promising results. The only paper which considered an attention mechanism for sonar imagery are Yu et al. (2021) and Cheng et al. (2022). Other recently proposed deep learning detectors with attention like DETR (Carion et al., 2020) or without attention like CenterNet2 (Zhou et al., 2021) or DetectoRS (Qiao et al., 2021) need be considered for detecting objects in sonar images. In all research papers regarding detection that are considered in this work a comparison with a conventional method, e.g. template matching, is missing. Without this comparison it is not proven that deep learning methods perform better than conventional methods for the detection of objects in SSS and SAS images. Both research gaps, the lack of investigation of most recent deep learning detectors as well as the comparison with conventional methods, can be closed by carrying out a study about deep learning detection algorithms for SSS and SAS images which considers both types of methods.

As shown in Section 5 only semantic segmentation methods are investigated for SSS and SAS images so far. Very recently the task of panoptic segmentation was proposed, which combines semantic and instance segmentation. For sonar images this type of segmentation is relevant because different seafloor types as well as objects lying on the seafloor would be segmented at the same time. This combines multiple different tasks that are relevant in the sonar domain and leads to a more complete analysis of a sonar image. Thus, we propose the investigation of panoptic segmentation for SSS and SAS images. Current state-of-the-art method for panoptic segmentation on MS COCO is Mask2Former, which again uses attention and should also be considered for sonar images.

Two evolving fields in deep learning which are close related are the uncertainty estimation and the calibration of computer vision models (Shen et al., 2021; Rajaraman et al., 2022). In critical applications like the classification between MLO and NMLO knowledge about the uncertainty of the prediction is essential. Model calibration is especially necessary if the used dataset is unbalanced, which is often the case for sonar image datasets. So far, to the best of our knowledge, no work has been done on applying such methods in the field of sonar imagery. This leaves uncertainty estimation and model calibration as another important research direction. For uncertainty estimation, Abbaszadeh Shahri et al. (2022) recently proposed a method that uses an ensemble of networks created by dropping connections between neurons. Closely related to uncertainty estimation is sensitivity analysis, where the influence of the input on the output is analyzed in order to get more inside to a model. In computer vision applications, those analyses are typically carried out through adversarial examples (Linardatos et al., 2020). Getting more insides to a neural network and removing its black box character is part of research on explainable artificial intelligence. Because the automatic processing of sonar images on an AUV is one main application of the research (Li et al., 2019; Song et al., 2021), an explanation of the network's predictions is essential. Some of the investigated papers consider such explanations, e.g. by interpreting the learned feature maps of a CNN (Williams, 2021) or by determining which part of the input image is relevant for the prediction (Williams, 2017). Nevertheless, more work needs to be done in order to better explain why a network makes certain decisions. We refer the reader to Linardatos et al. (2020) and Samek et al. (2019) for a broad overview of methods for explainable artificial intelligence.

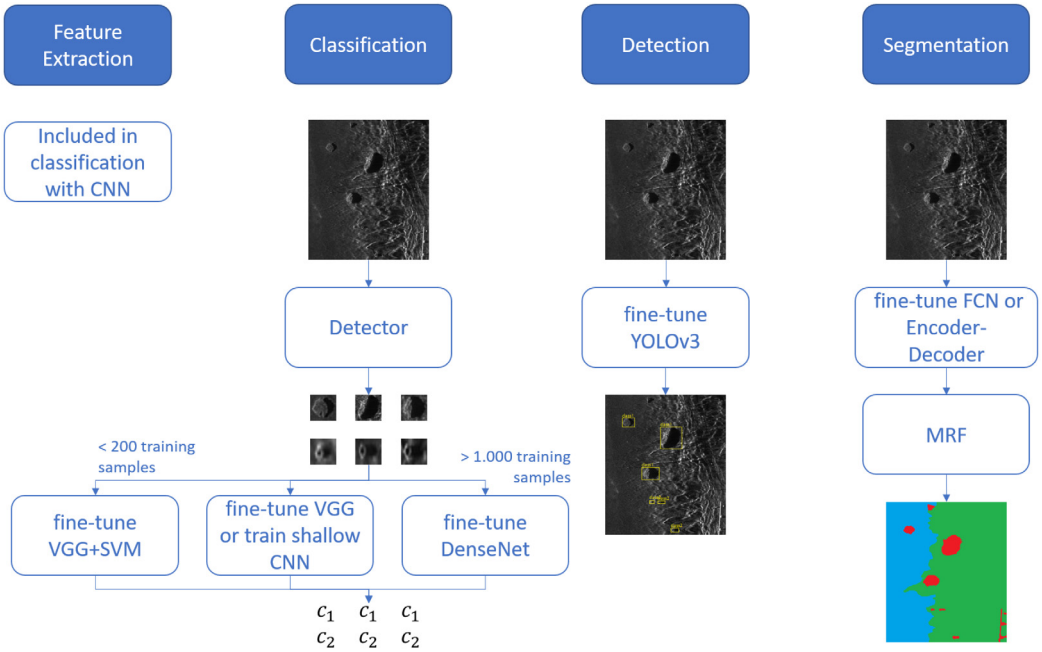


Fig. 9. Suggested deep learning methods for the individual subtasks in the ATR processing chain.

Training of CNNs involves tuning of several hyperparameter and can get stuck in local optima. To overcome these problems, metaheuristic algorithms, like genetic programming or grey wolf optimizer, have been designed. A combination of CNNs and metaheuristic algorithms have recently been applied to medical computer vision problems (Oyelade and Ezugwu, 2022). Such hybrid models have not been considered for sonar imagery yet but should be investigated for a further improvement in this field.

6.3. Data

Our survey discovered that the most challenging problem when applying deep learning to any of the ATR subtasks is the lack of a public dataset for training and testing of the methods. The military nature of the data, e.g., the classification of mines, is one reason why the data has not been made public. From the scientific perspective, this lack of a common benchmark dataset is especially relevant since it makes the comparison of different papers nearly impossible. To leverage the progress of deep learning for sonar imagery a common baseline dataset has to be created. Since a large sonar image dataset is typically not existing at only one institute or company an open source solution gives all researchers the possibility to contribute to this baseline dataset and thereby to enlarge it.

Additionally, the dataset needs to cover different challenging scenarios for the different subtasks. For classification the survey shows a broad range of applications. A dataset for classification thus needs to consider many different classes (e.g. seafloor types and mine types). In addition, a major challenge when classifying sonar images is the variable nature of highlight and shadow appearance when the angle from which the object is viewed changes. A cylinder from the front may look more like a stone than a typical cylinder. The detection is also expected to depend on the type of seafloor, since more complex surfaces such as sand ripples or rock fields generate a large number of false alarms. A dataset for this subtask needs to capture different objects as well as different seafloors. As previously stated panoptic segmentation is a promising application for segmenting sonar images. Not much work has been done so far on labeling seafloor and objects on a pixel level. Thus, an additional challenge here is to annotate the data. Especially the speckle noise in sonar images will be challenging for generating high quality annotations.

Besides comparability, an open source dataset helps researchers who only have a small dataset available. Current ways to deal with limited data are to enlarge the dataset by standard augmentation (Galusha et al., 2019; Bouzerdoum et al., 2019) or using GANs to generate synthetic data (Jegorova et al., 2020; Steiniger et al., 2021b). Few-shot learning seems also to be suited for this case but the only paper applying it to SSS images deals with synthetic data. Improving ways to generate synthetic data and few- or even zero-shot learning are research directions which should be followed in order to improve current algorithms.

Finally, fine-tuning standard deep learning models on sonar data has shown to be beneficial. However, no analysis has yet been carried out as to which pre-training dataset is most suited for dealing with sonar images. Other works on transfer-learning in computer vision have shown that the selection of the pre-training dataset can have a large impact on the final performance (Mensink et al., 2021). Especially the large domain gap between the common pre-training datasets like ImageNet or MS COCO on the one hand and SSS or SAS data on the other hand motivates further research toward the most suited pre-training dataset. The recent results from Cheng et al. that pre-training on SAR image leads to an improvement over the standard usage of ImageNet further indicates that a deeper analysis of the used pre-training dataset can lead to better networks.

7. Conclusion

This survey has reviewed research papers which consider deep learning approaches for the typical ATR subtask feature extraction, classification, detection and segmentation. We have shown that not only in standard computer vision but also in the sonar domain deep learning is a quickly developing field. CNNs are outperforming classical methods when it comes to feature extraction and classification. For detection and segmentation, a comparison with state-of-the-art conventional methods is still missing. In terms of applying state-of-the-art deep learning methods to the sonar domain, the research community is a few years behind. Transformers are currently the main research topic in computer vision but have only been considered in one out of the 62

research works. Because of the large domain gap between sonar images and natural RGB images as well as the limited amount of training data, ways to utilize the full potential of state-of-the-art deep learning model for sonar images need to be investigated. Multiple works have used GANs to generate synthetic images which is a promising approach to deal with limitations due to a small dataset. Other contributions that should be highlighted are the usage of multiple representation for sonar image classification, Gabor-CNNs for detection in sonar images and incorporating attention modules to classification and detection models.

Another crucial finding is the lack of a publicly available benchmark dataset which harms the comparability of developed methods. Such a sonar image dataset would give a boost to the development of better deep learning methods for the classification, detection and segmentation of SSS or SAS images. Finally, we propose the following five main research directions:

- Studying state-of-the-art deep learning detection algorithms applied to sonar images.
- Applying panoptic segmentation of SSS and SAS images.
- Employing different pre-training datasets when fine-tuning deep learning models.
- Building an open source SSS and SAS benchmark dataset.
- Improve the generation of synthetic sonar images.

CRediT authorship contribution statement

Yannik Steiniger: Conceptualization, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing. **Dieter Kraus:** Conceptualization, Analysis and/or interpretation of data, Writing – original draft, Supervision. **Tobias Meisen:** Conceptualization, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yannik Steiniger reports financial support was provided by ATLAS ELEKTRONIK GmbH.

Acknowledgment

The authors would like to thank ATLAS ELEKTRONIK GmbH for their support of this work.

Table A.1
Information about the datasets used in the classification tasks.

Paper	Sonar	Number of training samples	Number of test samples
Chen and Summers (2017)	SAS	<500[a]	1000
Berthold et al. (2017)	SSS	?	?
Luo et al. (2019)	SSS	545	144
Qin et al. (2021)	SSS	900	300
Ye et al. (2018)	SSS	235	100
Wang et al. (2019)	SSS	1.545*	515*
Huo et al. (2020)	SSS	833	257
Xu et al. (2020)	SSS	291	
Li et al. (2021)	SSS	365	274
Nayak et al. (2021)	SSS	7.148*	
Cheng et al. (2022)	SSS	725	341
Karjalainen et al. (2019)	SSS	207	?
Ochal et al. (2020)	SSS (simulated)	?	?
Williams and Dugelay (2016)	SAS	6.950	204
Williams (2017)	SAS	764	764
Williams (2018b)	SAS	?	196.252[b]
Williams (2018a)	SAS	?	?
Gerg and Williams (2018)	SAS	32.192	24.726
Galusha et al. (2019)	SAS	?	?
d'Alès de Corbet et al. (2019)	SAS	52	104
Williams et al. (2019)	SAS	32.192	24.133
Berthomier et al. (2020)	SAS	655.438[c]	24.725
Williams (2021)	SAS	655.426	24.133
Gerg and Monga (2022)	SAS	27.748	21181
Williams (2019)	SAS	46.003	15398
Dzieciuch et al. (2017)	SSS	250	250
Chapple et al. (2017)	SSS, SAS	?	?
Gebhardt et al. (2017)	SSS	4.326	5.138
Phung et al. (2019)	SSS	199	198
Bouzerdoum et al. (2019)	SSS, SAS	330	82
Quidu et al. (2005)	SAS	?	4
Williams (2016)	SAS	659	659
McKay et al. (2017)	SAS	80	40
Zhu et al. (2018)	SAS	123	41
Warakagoda and Midtgaard (2018)	SAS	4.380	420

[a]The number of unlabeled training data is not stated.
[b]927 target and 195 325 clutter samples.
[c]2924 target and 652 514 clutter samples.

Appendix A. List of abbreviations

AE	Auto-encoder
ATR	Automatic target recognition
AUV	Autonomous underwater vehicle
AW-CNN	Adaptive weights convolutional neural network
CMRE	Centre for Maritime Research and Experimentation
CNN	Convolutional neural network
CPN	Consistent prototypical network
DBN	Deep belief network
DBM	Deep Boltzmann machine
ECNet	Efficient convolutional network
ELM	Extreme learning machine
FAO	False alarm object
FCN	Fully convolutional network
FLOP	Floating-point operations per second
FLS	Forward looking sonar
GAN	Generative adversarial network
HGP	Hierarchical Gaussian process
ILSVRC	ImageNet Large Scale Video Recognition Challenge
IoU	Intersection over union
LDA	Latent Dirichlet allocations
MLO	Mine-like object
MLP	Multi layer perceptron
MRF	Markov random field
NMLO	Non-mine-like object
PN	Prototypical network
PSD	Power spectral density
ROI	Region of interest
RVM	Relevant vector machine
SAR	Synthetic aperture radar
SAS	Synthetic aperture sonar
SSS	Sidescan sonar
SVM	Support vector machine
UXO	Unexploded ordnance
ViT	Vision transformer

Appendix B. Datasets used for classification

Table A.1 lists the number of sonar images that are used to train the classification methods in the respective works. Some authors have not specified this number which is indicated by a question mark. If only the number of training samples is stated, no information about the train/test split was given in the paper. In some cases only the amount of training data after augmentation is given, indicated by a star in the table. Note that the datasets might not be balanced, e.g., the dataset in Williams and Dugelay (2016) contains 2526 target and 4424 clutter snippets.

References

Abbaszadeh Shahri, A., Shan, C., Larsson, S., 2022. A novel approach to uncertainty quantification in groundwater table modeling by automated predictive deep learning. Nat. Resour. Res. 31 (3), 1351–1373. http://dx.doi.org/10.1007/s11053-022-10051-w.

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L., 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J. Big Data 8 (1), 53. http://dx.doi.org/10.1186/s40537-021-00444-8.

Berthold, T., Leichter, A., Rosenhahn, B., Berkhahn, V., Valerius, J., 2017. Seabed sediment classification of side-scan sonar data using convolutional neural networks. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, pp. 1–8. http://dx.doi.org/10.1109/SSCI.2017.8285220.

Berthomier, T., Williams, D.P., d'Alès de Corbet, B., Dugelay, S., 2020. Exploiting auxiliary information for improved underwater target classification with convolutional neural networks. In: Global Oceans 2020: Singapore — U.S. Gulf Coast. IEEE, pp. 1–10. http://dx.doi.org/10.1109/IEEECONF38699.2020.9389138.

Berthomier, T., Williams, D.P., Dugelay, S., 2019. Target localization in synthetic aperture sonar imagery using convolutional neural networks. In: OCEANS 2019 MTS/IEEE Seattle. IEEE, pp. 1–9. http://dx.doi.org/10.23919/OCEANS40490.2019.8962774.

Bouzerdoum, A., Chapple, P.B., Dras, M., Guo, Y., Hamey, L., Hassanzadeh, T., Le, H.T., Nezami, O., Orgun, M., Phung, S.L., Ritz, C.H., Shahpasand, M., 2019. Improved deep-learning-based classification of mine-like contacts in sonar images from autonomous underwater vehicle. In: Proceedings of the 5th Underwater Acoustics Conference and Exhibition (UACE). pp. 179–186.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision — ECCV 2020, Vol. 12346. In: Springer eBook Collection, Springer International Publishing and Imprint Springer, Cham, pp. 213–229. http://dx.doi.org/10.1007/978-3-030-58452-8_13.

Chai, J., Zeng, H., Li, A., Ngai, E.W., 2021. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Mach. Learn. Appl. 6, 100134. http://dx.doi.org/10.1016/j.mlwa.2021.100134.

Chapple, P.B., Dell, T., Bongiorno, D., 2017. Enhanced detection and classification of mine-like objects using situational awareness and deep learning. In: Proceedings of the 4th Underwater Acoustics Conference and Exhibition (UACE). pp. 529–536.

Chen, J.L., Summers, J.E., 2016. Deep neural networks for learning classification features and generative models from synthetic aperture sonar big data. In: 172nd Meeting of the Acoustical Society of America, Vol. 29. Acoustical Society of America, 032001. http://dx.doi.org/10.1121/2.0000458.

Chen, J.L., Summers, J.E., 2017. Deep convolutional neural networks for semi-supervised learning from synthetic aperture sonar (SAS) images. In: 173rd Meeting of Acoustical Society of America and 8th Forum Acusticum. Acoustical Society of America, 055018. http://dx.doi.org/10.1121/2.0001018.

Cheng, Z., Huo, G., Li, H., 2022. A multi-domain collaborative transfer learning method with multi-scale repeated attention mechanism for underwater side-scan sonar image classification. Remote Sens. 14 (2), 355. http://dx.doi.org/10.3390/rs14020355.

Dai, Z., Liu, H., Le V., Q., Tan, M., 2021. CoAtNet: Marrying convolution and attention for all data sizes. http://dx.doi.org/10.48550/arXiv.2106.04803, arXiv:2106.04803.

d'Alès de Corbet, B., Williams, D.P., Dugelay, S., 2019. Target classification using multi-view synthetic aperture sonar Imagery. In: Proceedings of the 5th Underwater Acoustics Conference and Exhibition (UACE). pp. 227–233.

Denos, K., Ravaut, M., Fagette, A., Lim, H.-S., 2017. Deep learning applied to underwater mine warfare. In: OCEANS 2017 MTS/IEEE Aberdeen. IEEE, pp. 1–7. http://dx.doi.org/10.1109/OCEANSE.2017.8084910.

Divyabarathi, G., Shailesh, S., Judy, M.V., 2021. Object classification in underwater SONAR images using transfer learning based ensemble model. In: 2021 International Conference on Advances in Computing and Communications (ICACC). IEEE, pp. 1–4. http://dx.doi.org/10.1109/ICACC-202152719.2021.9708373.

Dobeck, G.J., Hyland, J.C., Smedley, L., 1997. Automated detection and classification of sea mines in sonar imagery. In: Dubey, A.C., Barnard, R.L. (Eds.), Detection and Remediation Technologies for Mines and Minelike Targets II. SPIE, pp. 90–110. http://dx.doi.org/10.1117/12.280846.

Domingos, L.C.F., Santos, P.E., Skelton, P.S.M., Brinkworth, R.S.A., Sammut, K., 2022. A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance. Sensors 22 (6), http://dx.doi.org/10.3390/s22062181.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. http://dx.doi.org/10.48550/arXiv.2010.11929.

Dzieciuch, I., Gebhardt, D., Barngrover, C., Parikh, K., 2017. Non-linear convolutional neural network for automatic detection of mine-like objects in sonar imagery. In: Visarath, L., Patrick, P.A. (Eds.), Proceedings of the 4th International Conference on Applications in Nonlinear Dynamics (ICAND), Vol. 6. In: Lecture Notes in Networks and Systems, Springer, Cham, pp. 309–314. http://dx.doi.org/10.1007/978-3-319-52621-8_27.

Einsidler, D., Dhanak, M., Beaujean, P.-P., 2018. A deep learning approach to target recognition in side-scan sonar imagery. In: OCEANS 2018 MTS/IEEE Charleston. IEEE, pp. 1–4. http://dx.doi.org/10.1109/OCEANS.2018.8604879.

Fan, Z., Xia, W., Liu, X., Li, H., 2021. Detection and segmentation of underwater objects from forward-looking sonar based on a modified mask RCNN. Signal, Image Video Process. 15, 1135–1143. http://dx.doi.org/10.1007/s11760-020-01841-x.

Fei, T., Kraus, D., Zoubir, A.M., 2015. Contributions to automatic target recognition systems for underwater mine classification. IEEE Trans. Geosci. Remote Sens. 53 (1), 505–518. http://dx.doi.org/10.1109/TGRS.2014.2324971.

Feldens, P., 2020. Super resolution by deep learning improves boulder detection in side scan sonar backscatter mosaics. Remote Sens. 12 (14), 2284. http://dx.doi.org/10.3390/rs12142284.

Feldens, P., Darr, A., Feldens, A., Tauber, F., 2019. Detection of boulders in side scan sonar mosaics by a neural network. Geosciences 9 (4), 159. http://dx.doi.org/10.3390/geosciences9040159.

Fuchs, L.R., Gallstrom, A., Folkesson, J., 2018. Object recognition in forward looking sonar images using transfer learning. In: 2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV). IEEE, pp. 1–6. http://dx.doi.org/10.1109/AUV.2018.8729686.

Galusha, A.P., Dale, J., Keller, J., Zare, A., 2019. Deep convolutional neural network target classification for underwater synthetic aperture sonar imagery. In: Bishop, S.S., Isaacs, J.C. (Eds.), Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIV. SPIE, pp. 18–28. http://dx.doi.org/10.1117/12.2519521.

Gebhardt, D., Parikh, K., Dzieciuch, I., Walton, M., Hoang Vo, N.A., 2017. Hunting for naval mines with deep neural networks. In: OCEANS 2017 MTS/IEEE Anchorage. IEEE, pp. 1–5.

Gerg, I.D., Monga, V., 2022. Structural prior driven regularized deep learning for sonar image classification. IEEE Trans. Geosci. Remote Sens. 60, 1–16. http://dx.doi.org/10.1109/TGRS.2020.3045649.

Gerg, I.D., Williams, D.P., 2018. Additional representations for improving synthetic aperture sonar classification using convolutional neural networks. In: Proceedings of the 4th International Conference on Synthetic Aperture Sonar Synthetic Aperture Radar. pp. 11–22. http://dx.doi.org/10.48550/arXiv.1808.02868.

Hożyń, S., 2021. A review of underwater mine detection and classification in sonar imagery. Electronics 10 (23), 2943. http://dx.doi.org/10.3390/electronics10232943.

Huo, G., Wu, Z., Li, J., 2020. Underwater object classification in sidescan sonar images using deep transfer learning and semisynthetic training data. IEEE Access 8, 47407–47418. http://dx.doi.org/10.1109/ACCESS.2020.2978880.

Isaacs, J.C., 2014. Representational learning for sonar ATR. In: Bishop, S.S., Isaacs, J.C. (Eds.), Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XIX. SPIE, pp. 1–9. http://dx.doi.org/10.1117/12.2053057.

Jegorova, M., Karjalainen, A.I., Vazquez, J., Hospedales, T.M., 2020. Unlimited resolution image generation with R2D2-GANs. In: Global Oceans 2020: Singapore — U.S. Gulf Coast. IEEE, pp. 1–5. http://dx.doi.org/10.1109/IEEECONF38699.2020.9389260.

Jiang, L., Cai, T., Ma, Q., Xu, F., Wang, S., 2020. Active object detection in sonar images. IEEE Access 8, 102540–102553. http://dx.doi.org/10.1109/ACCESS.2020.2999341.

Jin, L., Liang, H., Yang, C., 2019. Accurate underwater ATR in forward-looking sonar imagery using deep convolutional neural networks. IEEE Access 7, 125522–125531. http://dx.doi.org/10.1109/ACCESS.2019.2939005.

Johnson, S.G., Deaett, M.A., 1994. The application of automated recognition techniques to side-scan sonar imagery. IEEE J. Ocean. Eng. 19 (1), 138–144. http://dx.doi.org/10.1109/48.289460.

Karjalainen, A.I., Mitchell, R., Vazquez, J., 2019. Training and validation of automatic target recognition systems using generative adversarial networks. In: 2019 Sensor Signal Processing for Defence Conference (SSPD). IEEE, pp. 1–5. http://dx.doi.org/10.1109/SSPD.2019.8751666.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 25. MIT Press, pp. 1097–1105.

Langner, F., Knauer, C., Jans, W., Ebert, A., 2009. Side scan sonar image resolution and automatic object detection, classification and identification. In: OCEANS 2009—Europe. IEEE, pp. 1–8. http://dx.doi.org/10.1109/OCEANSE.2009.5278183.

Le, H.T., Phung, S.L., Chapple, P.B., Bouzerdoum, A., Ritz, C.H., Tran, L.C., 2020. Deep gabor neural network for automatic detection of mine-like objects in sonar imagery. IEEE Access 8, 94126–94139. http://dx.doi.org/10.1109/ACCESS.2020.2995390.

Li, C., Ye, X., Cao, D., Hou, J., Yang, H., 2021. Zero shot objects classification method of side scan sonar image based on synthesis of pseudo samples. Appl. Acoust. 173, 107691. http://dx.doi.org/10.1016/j.apacoust.2020.107691.

Li, K., Yu, F., Wang, Q., Wu, M., Li, G., Yan, T., He, B., 2019. Real-time segmentation of side scan sonar imagery for AUVs. In: 2019 IEEE Underwater Technology. IEEE, pp. 1–5. http://dx.doi.org/10.1109/UT.2019.8734319.

Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2020. Explainable AI: A review of machine learning interpretability methods. Entropy 23 (1), http://dx.doi.org/10.3390/e23010018.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, pp. 9992–10002. http://dx.doi.org/10.1109/ICCV48922.2021.00986.

Luo, X., Qin, X., Wu, Z., Yang, F., Wang, M., Shang, J., 2019. Sediment classification of small-size seabed acoustic images using convolutional neural networks. IEEE Access 7, 98331–98339. http://dx.doi.org/10.1109/ACCESS.2019.2927366.

McKay, J., Gerg, I., Monga, V., Raj, R.G., 2017. What's mine is yours: Pretrained CNNs for limited training sonar ATR. In: OCEANS 2017 MTS/IEEE Anchorage. IEEE, pp. 1–7.

Mensink, T., Uijlings, J., Kuznetsova, A., Gygli, M., Ferrari, V., 2021. Factors of influence for transfer learning across diverse appearance domains and task types. http://dx.doi.org/10.48550/arXiv.2103.13318, arXiv:2103.13318.

Nayak, N., Nara, M., Gambin, T., Wood, Z., Clark, C.M., 2021. Machine learning techniques for AUV side-scan sonar data feature extraction as applied to intelligent search for underwater archaeological sites. In: Ishigami, G., Yoshida, K. (Eds.), Field and Service Robotics, Vol. 16. In: Springer Proceedings in Advanced Robotics, Springer, Singapore, pp. 219–233. http://dx.doi.org/10.1007/978-981-15-9460-1_16.

Nelson, S.R., Tuovila, S.M., 1995. Fractal-based image processing for mine detection. In: Dubey, A.C., Cindrich, I., Ralston, J.M., Rigano, K.A. (Eds.), Detection Technologies for Mines and Minelike Targets. SPIE, pp. 454–465. http://dx.doi.org/10.1117/12.211342.

Neupane, D., Seok, J., 2020. A review on deep learning-based approaches for automatic sonar target recognition. Electronics 9 (11), 1972. http://dx.doi.org/10.3390/electronics9111972.

Nian, R., Zang, L., Geng, X., Yu, F., Ren, S., He, B., Li, X., 2021. Towards characterizing and developing formation and migration cues in seafloor sand waves on topology, morphology, evolution from high-resolution mapping via side-scan sonar in autonomous underwater vehicles. Sensors 21 (9), 3283. http://dx.doi.org/10.3390/s21093283.

Ochal, M., Vazquez, J., Petillot, Y., Wang, S., 2020. A comparison of few-shot learning methods for underwater optical and sonar image classification. In: Global Oceans 2020: Singapore — U.S. Gulf Coast. IEEE, pp. 1–10. http://dx.doi.org/10.1109/IEEECONF38699.2020.9389475.

Oyelade, O.N., Ezugwu, A.E., 2022. Characterization of abnormalities in breast cancer images using nature–inspired metaheuristic optimized convolutional neural networks model. Concurr. Comput.: Pract. Exper. 34 (4), http://dx.doi.org/10.1002/cpe.6629.

Perry, S.W., Guan, L., 2004. Pulse-length-tolerant features and detectors for sector-scan sonar imagery. IEEE J. Ocean. Eng. 29 (1), 138–156. http://dx.doi.org/10.1109/JOE.2003.819312.

Phung, S.L., Nguyen, T.N.A., Le, H.T., Chapple, P.B., Ritz, C.H., Bouzerdoum, A., Tran, L.C., 2019. Mine-like object sensing in sonar imagery with a compact deep learning architecture for scarce data. In: 2019 Digital Image Computing: Techniques and Applications (DICTA). IEEE, pp. 1–7. http://dx.doi.org/10.1109/DICTA47822.2019.8945982.

Qiao, S., Chen, L.-C., Yuille, A., 2021. DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 10208–10219. http://dx.doi.org/10.1109/CVPR46437.2021.01008.

Qin, X., Luo, X., Wu, Z., Shang, J., 2021. Optimizing the sediment classification of small side-scan sonar images based on deep learning. IEEE Access 9, 29416–29428. http://dx.doi.org/10.1109/ACCESS.2021.3052206.

Quidu, I., Burlet, N., Malkasse, J.-P., Florin, F., 2005. Automatic classification for MCM systems. In: OCEANS 2005 — Europe. IEEE, pp. 844–847. http://dx.doi.org/10.1109/OCEANSE.2005.1513166.

Rahnemoonfar, M., Dobbs, D., 2019. Semantic segmentation of underwater sonar imagery with deep learning. In: 2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 9455–9458. http://dx.doi.org/10.1109/IGARSS.2019.8898742.

Rajaraman, S., Ganesan, P., Antani, S., 2022. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. PLoS One 17 (1), e0262838. http://dx.doi.org/10.1371/journal.pone.0262838.

Reed, A., Gerg, I.D., McKay, J.D., Brown, D.C., Williams, D.P., Jayasuriya, S., 2019. Coupling rendering and generative adversarial networks for artificial SAS image generation. In: OCEANS 2019 MTS/IEEE Seattle. IEEE, pp. 1–10. http://dx.doi.org/10.23919/OCEANS40490.2019.8962733.

Rutledge, J., Yuan, W., Wu, J., Freed, S., Lewis, A., Wood, Z., Gambin, T., Clark, C.M., 2018. Intelligent shipwreck search using autonomous underwater vehicles. In: Lynch, K. (Ed.), 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 1–8. http://dx.doi.org/10.1109/ICRA.2018.8460548.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R., 2019. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Vol. 11700, 1st ed. 2019 In: Springer eBooks Computer Science, Springer, Cham, http://dx.doi.org/10.1007/978-3-030-28954-6.

Shen, Y., Zhang, Z., Sabuncu, M.R., Sun, L., 2021. Real-time uncertainty estimation in computer vision via uncertainty-aware distribution distillation. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 707–716. http://dx.doi.org/10.1109/WACV48630.2021.00075.

Song, Y., He, B., Liu, P., 2021. Real-time object detection for AUVs using self-cascaded convolutional neural networks. IEEE J. Ocean. Eng. 46 (1), 56–67. http://dx.doi.org/10.1109/JOE.2019.2950974.

Song, Y., Zhu, Y., Li, G., Feng, C., He, B., Yan, T., 2017. Side scan sonar segmentation using deep convolutional neural network. In: OCEANS 2017 MTS/IEEE Anchorage. IEEE, pp. 1–4.

Steiniger, Y., Groen, J., Stoppe, J., Kraus, D., Meisen, T., 2021a. A study on modern deep learning detection algorithms for automatic target recognition in sidescan sonar images. In: Proceedings of the 6th Underwater Acoustics Conference and Exhibition (UACE). Acoustical Society of America, 070010. http://dx.doi.org/10.1121/2.0001470.

Steiniger, Y., Kraus, D., Meisen, T., 2021b. Generating synthetic sidescan sonar snippets using transfer-learning in generative adversarial networks. J. Mar. Sci. Eng. 9 (3), 239. http://dx.doi.org/10.3390/jmse9030239.

Steiniger, Y., Stoppe, J., Meisen, T., Kraus, D., 2020. Dealing with highly unbalanced sidescan sonar image datasets for deep learning classification tasks. In: Global Oceans 2020: Singapore — U.S. Gulf Coast. IEEE, pp. 1–7. http://dx.doi.org/10.1109/IEEECONF38699.2020.9389373.

Teng, B., Zhao, H., 2020. Underwater target recognition methods based on the framework of deep learning: A survey. Int. J. Adv. Robot. Syst. 17 (6), 172988142097630. http://dx.doi.org/10.1177/1729881420976307.

Topple, J.M., Fawcett, J.A., 2021. MiNet: Efficient deep learning automatic target recognition for small autonomous vehicles. IEEE Geosci. Remote Sens. Lett. 18 (6), 1014–1018. http://dx.doi.org/10.1109/LGRS.2020.2993652.

Valdenegro-Toro, M., 2016. Object recognition in forward-looking sonar images with convolutional neural networks. In: OCEANS 2016 MTS/IEEE Monterey. IEEE, pp. 1–6. http://dx.doi.org/10.1109/OCEANS.2016.7761140.

Valverde, J.M., Imani, V., Abdollahzadeh, A., de Feo, R., Prakash, M., Ciszek, R., Tohka, J., 2021. Transfer learning in magnetic resonance brain imaging: A systematic review. J. Imaging 7 (4), 66. http://dx.doi.org/10.3390/jimaging7040066.

Wang, H., Gao, N., Xiao, Y., Tang, Y., 2020. Image feature extraction based on improved FCN for UUV side-scan sonar. Mar. Geophys. Res. 41 (4), 18. http://dx.doi.org/10.1007/s11001-020-09417-7.

Wang, X., Jiao, J., Yin, J., Zhao, W., Han, X., Sun, B., 2019. Underwater sonar image classification using adaptive weights convolutional neural network. Appl. Acoust. 146, 145–154. http://dx.doi.org/10.1016/j.apacoust.2018.11.003.

Warakagoda, N.D., Midtgaard, Ø., 2018. Transfer-learning with deep neural networks for mine recognition in sonar images. In: Proceedings of the 4th International Conference on Synthetic Aperture Sonar Synthetic Aperture Radar, 40, pp. 115–122.

Williams, D.P., 2016. Underwater target classification in synthetic aperture sonar imagery using deep convolutional neural networks. In: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, pp. 2497–2502. http://dx.doi.org/10.1109/ICPR.2016.7900011.

Williams, D.P., 2017. Demystifying deep convolutional neural networks for sonar image classification. In: Proceedings of the 4th Underwater Acoustics Conference and Exhibition (UACE). pp. 513–520.

Williams, D.P., 2018a. Convolutional neural network transfer learning for underwater object classification. In: Proceedings of the 4th International Conference on Synthetic Aperture Sonar Synthetic Aperture Radar. pp. 123–131.

Williams, D.P., 2018b. Exploiting phase information in synthetic aperture sonar images for target classification. In: OCEANS 2018 MTS/IEEE Kobe. IEEE, pp. 1–6. http://dx.doi.org/10.1109/OCEANSKOBE.2018.8559255.

Williams, D.P., 2019. Transfer learning with SAS-image convolutional neural networks for improved underwater target classification. In: 2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 78–81. http://dx.doi.org/10.1109/IGARSS.2019.8898611.

Williams, D.P., 2021. On the use of tiny convolutional neural networks for human-expert-level classification performance in sonar imagery. IEEE J. Ocean. Eng. 46 (1), 236–260. http://dx.doi.org/10.1109/JOE.2019.2963041.

Williams, D.P., Dugelay, S., 2016. Multi-view SAS image classification using deep learning. In: OCEANS 2016 MTS/IEEE Monterey. IEEE, pp. 1–9. http://dx.doi.org/10.1109/OCEANS.2016.7761334.

Williams, D.P., Hamon, R., Gerg, I., 2019. On the benefit of multiple representations with convolutional neural networks for improved target classification using sonar data. In: Proceedings of the 5th Underwater Acoustics Conference and Exhibition (UACE). pp. 187–194.

Wu, M., Wang, Q., Rigall, E., Li, K., Zhu, W., He, B., Yan, T., 2019. ECNet: Efficient convolutional networks for side scan sonar image segmentation. Sensors 19 (9), 2009. http://dx.doi.org/10.3390/s19092009.

Xu, L., Wang, X., Wang, X., 2019. Shipwrecks detection based on deep generation network and transfer learning with small amount of sonar images. In: 2019 IEEE 8th Data Driven Control and Learning Systems Conference (DDCLS). IEEE, pp. 638–643. http://dx.doi.org/10.1109/DDCLS.2019.8909011.

Xu, Y., Wang, X., Wang, K., Shi, J., Sun, W., 2020. Underwater sonar image classification using generative adversarial network and convolutional neural network. IET Image Process. 14 (12), 2819–2825. http://dx.doi.org/10.1049/iet-ipr.2019.1735.

Yan, J., Meng, J., Zhao, J., 2020. Real-time bottom tracking using side scan sonar data through one-dimensional convolutional neural networks. Remote Sens. 12 (1), 37. http://dx.doi.org/10.3390/rs12010037.

Yan, J., Meng, J., Zhao, J., 2021. Bottom detection from backscatter data of conventional side scan sonars through 1D-UNet. Remote Sens. 13 (5), 1024. http://dx.doi.org/10.3390/rs13051024.

Ye, X., Li, C., Zhang, S., Yang, P., Li, X., 2018. Research on side-scan sonar image target classification method based on transfer learning. In: OCEANS 2018 MTS/IEEE Charleston. IEEE, pp. 1–6. http://dx.doi.org/10.1109/OCEANS.2018.8604691.

Yu, Y., Zhao, J., Gong, Q., Huang, C., Zheng, G., Ma, J., 2021. Real-time underwater maritime object detection in side-scan sonar images based on transformer-YOLOv5. Remote Sens. 13 (18), 3555. http://dx.doi.org/10.3390/rs13183555.

Yu, F., Zhu, Y., Wang, Q., Li, K., Wu, M., Li, G., Yan, T., He, B., 2019. Segmentation of side scan sonar images on AUV. In: 2019 IEEE Underwater Technology. IEEE, pp. 1–4. http://dx.doi.org/10.1109/UT.2019.8734433.

Zheng, G., Zhang, H., Li, Y., Zhao, J., 2021. A universal automatic bottom tracking method of side scan sonar data based on semantic segmentation. Remote Sens. 13 (10), 1945. http://dx.doi.org/10.3390/rs13101945.

Zhou, X., Koltun, V., Krähenbühl, P., 2021. Probabilistic two-stage detection. http://dx.doi.org/10.48550/arXiv.2103.07461, arXiv:2103.07461.

Zhu, P., Isaacs, J.C., Fu, B., Ferrari, S., 2017. Deep learning feature extraction for target recognition and classification in underwater sonar images. In: 2017 IEEE 56th Annual Conference on Decision and Control (CDC). IEEE, pp. 2724–2731. http://dx.doi.org/10.1109/CDC.2017.8264055.

Zhu, K., Tian, J., Huang, H., 2018. Underwater object images classification based on convolutional neural network. In: 2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP). IEEE, pp. 301–305. http://dx.doi.org/10.1109/SIPROCESS.2018.8600472.