Transformers and Ensembles for Object Detection in Sidescan Sonar Images

Yannik Steiniger[1]

[1]German Aerospace Center, Institute for the Protection of Maritime Infrastructures, 27572 Bremerhaven, Germany

Contact author: Yannik Steiniger, Fischkai 1, 27572 Bremerhaven, Germany, yannik.steiniger@dlr.de

Abstract: *Transformer models have evolved as the state-of-the-art for computer vision tasks, such as image classification and object detection. For sonar images, they recently have been applied for classifying snippets of synthetic aperture sonar images. In this work, deep learning models for object detection based on standard convolutional neural networks (CNN) and on transformers are investigated and compared. The Retina-SWIN transformer model achieves a true-positive rate of up to 99.5%. However, all investigated deep learning detectors suffer from a high number of false alarms. Thus, the models are incorporated into two-step pipelines and ensembles. With both setups the number of false alarms is reduced. However, the detection ensembles show a lower true-positive rate than the two-step pipeline. The best overall model is Retina-SWIN, which is trained for localisation of objects, combined with a custom CNN, which classifies the extracted snippets. It reaches a maximum true-positive rate of 94.8% at 33.4 false alarms, which is a reduction of over 80% compared to the baseline Retina-SWIN.*

Keywords: *Deep Learning, Automatic Target Recognition, Sonar Imagery, Sidescan Sonar, Transformer*

1. INTRODUCTION

Being able to observe a scene under water is crucial for the security of maritime infrastructures. Due to physical limitations of optical systems, imaging sonar systems, such as sidescan or synthetic aperture sonar, which utilise acoustic instead of electro-magnetic waves, are commonly used to survey large seafloor areas. These sensors are typically mounted on an autonomous underwater vehicle, which follows a predefined path to scan the requested area. Analysing the captured data automatically is essential to increase the autonomy of a holistic surveying system. Nowadays, deep learning methods like convolutional neural networks (CNN) are primarily used for computer vision tasks relevant for the analysis of sonar images, such as classification, detection and segmentation [1, 2, 3, 4]. Recently, vision transformers have emerged as the state-of-the-art in many computer vision tasks and their application to the classification of sonar images has been investigated as well [5, 6]. However, transformer-based models for detecting objects in sonar images have not been analysed yet.

In this work, CNNs and transformer models are compared for the detection of objects in sidescan sonar images. As representatives for the CNN methods one-stage as well as two-stage detectors are selected. The transformer-based methods investigated here are Retina-SWIN [7] and Deformable DETR (DDETR) [8], which use the attention mechanism in the network backbone or detection head, respectively. In addition, all detectors are used in two different two-step pipelines, which further process predictions with another CNN to reduce the number of false alarms [9]. Furthermore, the combination of multiple different object detectors into an ensemble is studied. For classification, such ensembles of CNNs have shown to improve the overall performance [1]. The results show that the transformer-based method Retina-SWIN can reach a true-positive rate of 99.5% but at the cost of a high false alarm rate. Processing the detections with a CNN in the two-step pipeline significantly reduces the number of false alarms per image but also negatively impacts the true-positive rate. The best trade-off is achieved with a Retina-SWIN model, which is trained for localisation and a subsequent CNN for classification, reaching a maximum true-positive rate of around 95% at slightly over 30 false alarms per image. An ensemble of different object detectors can increase the detection performance but at the cost of an increased number of false alarms.

The remainder of the paper is organised as follows. First, Section 2 provides a brief introduction into CNN- and transformer-based object detection. Afterwards, the experimental setup is described in Section 3. In Section 4 the results of the experiments are presented and discussed. Finally, the paper closes with a summary in Section 5.

2. BACKGROUND

2.1. CNN-BASED DETECTORS

Generally, CNN-based detectors can be divided into two categories: one-stage and two-stage detectors. A two-stage detector, like CenterNet2 [10], predicts in the first stage a large number of region of interests (ROI) and filters, classifies and refines these ROIs in the second stage. In contrast to this, a one-stage detector, like YOLOv8 [11], directly predicts bounding boxes and object classes from an input image without intermediate ROIs. This typically makes one-stage detectors faster than two-stage detectors but also less accurate.

2.2. TRANSFORMER-BASED DETECTORS

A transformer is a deep learning architecture that has originated from natural language processing and that translates an input sequence into an output sequence via an encoder-decoder architecture and an attention mechanism [12]. The components of the attention mechanism are a set of key-value pairs $\mathbf{K} \in \mathcal{R}^{N_k \times N_l}$ and $\mathbf{V} \in \mathcal{R}^{N_v \times N_l}$ as well as queries $\mathbf{q} \in \mathcal{R}^{N_k \times 1}$, where N_l is the number of key-value pairs, e.g., the number of words in a sentence and N_k and N_v are the number of features of the keys, queries and values, respectively. $\mathbf{K}$, $\mathbf{V}$, and $\mathbf{q}$ are realized through a linear combination of the inputs to the attention module with learnable weight matrices and the attention is calculated as

$$\text{attention}(\mathbf{K}, \mathbf{V}, \mathbf{q}) = \text{softmax}\left(\frac{\mathbf{q}^{\mathrm{T}}\mathbf{K}}{\sqrt{N_k}}\right)\mathbf{V}^{\mathrm{T}}. \tag{1}$$

In computer vision models, this attention mechanism can either be used in the backbone of the model, which learns to extract general features from the input image, or in the head of the model, which performs the specific task based on these features, e.g., the prediction of bounding boxes and object classes. In the backbone, in order to be applied to an image, the input is split into patches, which form the input sequence of the encoder. The encoder applies the attention mechanism and outputs learned image features. When used in the detection head, attention is used in an encoder and decoder, where the encoder further transforms features from a backbone model and the decoder maps an input query sequence into predictions using these features.

3. EXPERIMENTAL SETUP

3.1. SIDESCAN SONAR DATASET

To train the object detectors as well as the classification models for the two-step approach sidescan sonar images from several sea trials are used [13, 9]. In these images objects from the four classes *Tire*, *Rock*, *Cylinder* and *Wreck* were annotated with manually defined bounding boxes tightly enclosing the object highlight and acoustic shadow. The detection dataset consists of the whole sidescan sonar images as input and the bounding box coordinates and class labels as target variables. For the classification dataset the snippets defined by the bounding boxes are extracted from the sonar images. In addition, random background snippets are selected to form a fifth *Background* class. The final classification dataset consists of the extracted snippets as input and the class labels as the target variable.

Images from the classes *Tire*, *Cylinder* and *Wreck* are very limited. Thus, the training and test split for both datasets is done such that these classes are roughly split 50:50. However, multiple images of the same objects, e.g., from different viewing angels, exist in the dataset. Thus, the data is split such that images from the same object are either in the training or test set. Additionally, if multiple objects are in the same sidescan sonar image, their associated snippets are all assigned to either the training or test set. This ensures that the same objects are used for training and testing the detectors and the classifier. These restrictions result in the number of samples in the training and test set for each class reported in Table 1.

Table 1: Overview about the datasets. Sidescan images are used for detection. Snippets from the classes Tire, Rock, Cylinder, Wreck and Background are used for classification.

Dataset	Number of sidescan images	tires	rocks	cylinders	wrecks	background
Training	769	24	2288	15	10	1390
Test	128	12	167	22	10	719

3.2. DETECTION PIPELINES

Besides the direct application of the deep learning detectors two additional pipelines previously introduced in [9] are used and shown in Figure 1. The purpose of these two-step pipeline is to reduce the number of false alarms. In the first two-step approach the deep learning detection models are used for pure localisation of the objects. To train these models all objects are considered as one general class *Target*. The classification is carried out in the second step by an additional CNN. In the second two-step pipeline the detections from the deep learning models are filtered by another CNN. This classifies the extracted snippets either into *Target* or *Background*. Detections whose corresponding snippets are classified as background will be filtered.

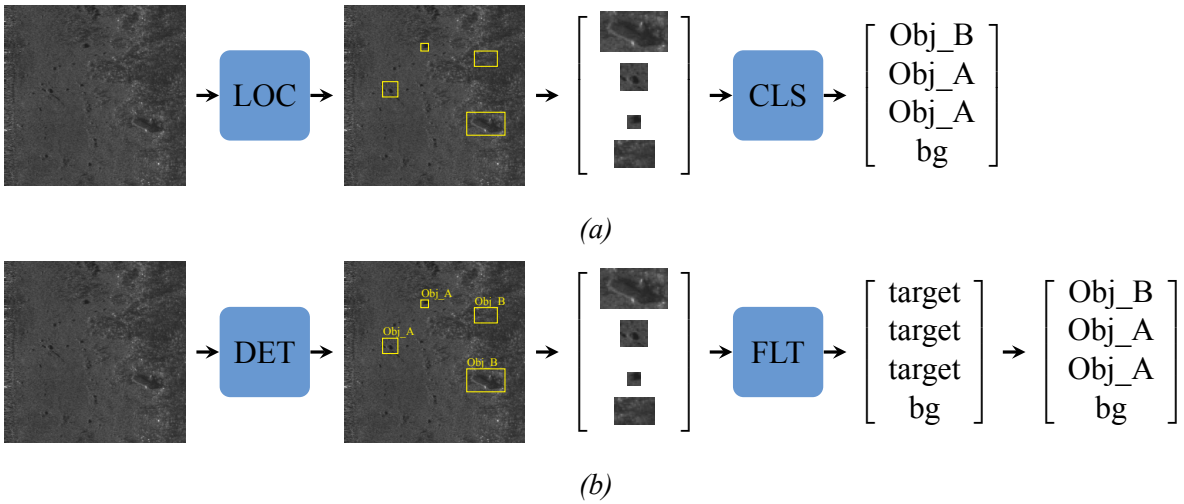

Figure 1: Two-step detection pipelines. (a) The detector is only used for localization (LOC) while a CNN classifies the extracted snippets in a subsequent step (CLS). (c) The detector (DET) predicts bounding boxes and object classes and another CNN carries out a binary classification (FLT) to reduce the number of false alarms by filtering background snippets.

In this study, the deep learning models YOLOv3 [14], YOLOv8 [11], CenterNet2 [10], Retina-SWIN [7] and DDETR [8] are considered. The models are initialized with pre-trained weights based on the MS COCO dataset. They are trained on the sonar data with their standard configuration using the MMDetection toolbox [15]. All models are trained for 100 epochs except Retina-SWIN which is trained for only 5 epochs to prevent performance degradation. As classification and filtering networks the same shallow architecture with three convolutional layers and training parameter as in [9] is used.

3.3. DETECTOR ENSEMBLING

Combining the output of multiple models into one ensemble for a final decision has shown to be beneficial for sonar image classification [1]. In this work an ensemble is build by combining the output of multiple different detectors. Based on the previous sections we consider three types of ensembles DET, LOC+CLS and DET+FLT, meaning that only outputs from within one detection pipeline are combined. In addition, the detections of the DET ensemble after ensembling are filtered with the CNN used in the filtering step of the DET+FLT pipeline. The combination of the individual predictions of the models in an ensemble is based on the number of detectors that predict the same object class at a similar position. Two detections are considered to belong to the same object if the predicted class matches and the intersection over union (IoU) of the bounding boxes exceeds a value of 0.5. If a pre-defined number of models n_{det} detects the same object it is counted as a detection of the ensemble. For the final prediction of the ensemble the bounding box with the highest prediction score is considered.

4. RESULTS

The detection performance in all setups is compared based on receiver operating characteristic (ROC)-like curves. In contrast to standard ROC curves for classification performances, here the average number of false alarms per image is given on the abscissa. In addition to the IoU, the pixel-wise Euclidean distance d between the center pixel of the true and predicted bounding box is considered to determine true-positive detections. Figure 2 displays the ROC-like curves of the five models for the three detection pipelines. As expected the two-step pipelines produce less false alarms but also reduce the number true-positive detections. Based on the IoU for determining true-positives, DDETR shows the worst performance of the five deep learning detectors. The performance is better if d is considered, which shows that the shape of the bounding boxes predicted by DDETR does not match the true ones. This is also the reason for the bad performance of DDETR in the DET+FLT pipeline, where snippets are extracted from the sonar image and feed to the filtering CNN based on the predicted bounding boxes. With badly defined bounding boxes parts of the object are missing in the snippet which leads to a wrong classification result. Retina-SWIN shows the best performance in this comparison achieving a maximum true-positive rate of 99.5%.

For a further quantitative analysis, Table 2 lists the maximum true-positive rate TPR_{max}, the false alarm rate corresponding to this value as well as the false alarm rate for true-positive rates of 0.9 and 0.5 of the individual detectors and pipelines. The deep learning model trained for localisation in the LOC+CLS pipeline is indicated with an -L. The CNN-based models show a higher TPR_{max} in the DET+FLT pipeline while the both transformer-based models perform better in the LOC+CLS setup. Retina-SWIN for localization and the CNN for classification achieves the best trade-off of a high TPR and a low number of false alarms with a TPR_{max} of 94.8% and 33.4 false alarms per image. If a low false alarm rate is preferable over a high TPR, YOLOv8 for localization and the CNN for classification gives the best result with only 0.45 false alarms per image at a TPR of 50%.

With the five deep learning models available, five different ensembles E1-E5 are investigated by successively combining more detectors. Based on their individual performance in the previous analysis, starting with only Retina-SWIN in E1, YOLOv3, YOLOv8, CenterNet2 and DDETR are subsequently added to form E2-E5. Figure 3 shows the performance metrics ex-

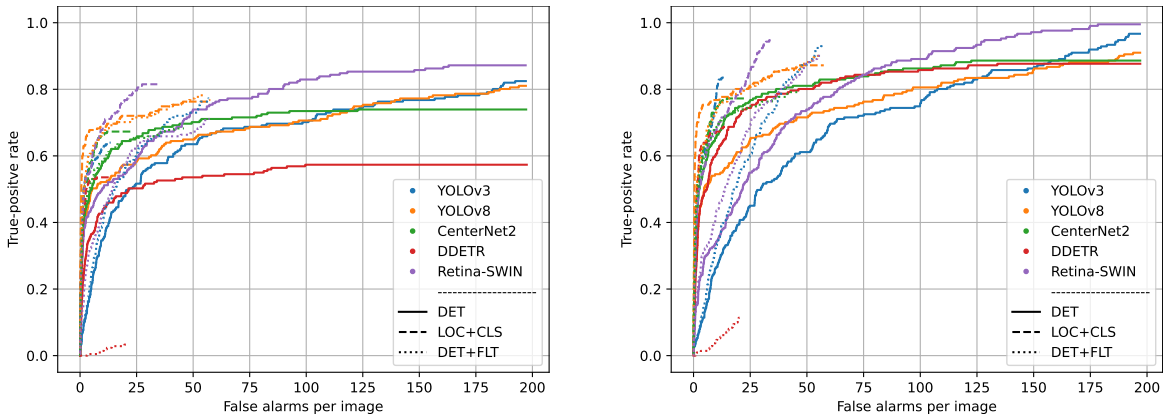

Figure 2: ROC-like curves of the detectors with (a) IoU=0.5 and (b) d=10 px.

tracted from the ROC-like curves for the ensembles. For each ensemble the minimum number of detectors that have to predict a specific detection in order to count as a detection of the ensemble n_{det} is varied from one to four. For example, with E3 and $n_{det} = 2$ at least two models of Retina-SWIN, YOLOv3 and YOLOv8 have to predict an object of the same class at the same position to generate a detection of this ensemble. The ROC-like curves with $d = 10$ px are used to extract the metrics in Figure 3. However, the plot for IoU=0.5 show the same behaviour. Figure 3 additionally shows the performance metrics of Retina-SWIN-L+CNN as a reference.

Increasing n_{det} reduces the maximum true-positive rate but also the number of false alarms. Enlarging the ensemble set while keeping n_{det} fixed increases the maximum true-positive rate but also the number of false alarms. Since Retina-SWIN already has a high TPR_{max}, increasing the set of models to form the ensemble while counting a detection if only one model outputs a prediction only increases the number of false alarms. An additional filtering of the detections from the ensemble slightly reduces the TPR_{max} but also drastically the number of false alarms. Creating ensembles from the LOC+CLS detection pipeline leads to a worse performance compared to the Retina-SWIN-L+CNN baseline. The number of false alarms can be reduced in this ensemble from 33.4 to 6.3 but the TPR_{max} drops below 75%. The same behaviour can be seen for the DET+FLT pipeline. However, here the best TPR_{max} for an ensemble with at least two matched detections is 78.2% at a false alarm rate of 13. Compared to the baseline, no ensemble configuration can increase the maximum true-positive rate and at the same time reduce the number of false alarms. The best ensemble is E3 with the DET+FLT pipeline and $n_{det} = 3$ which achieves a TPR_{max} of 87.7% at 31.4 false alarms per image.

5. SUMMARY

This work has compared CNN- and transformer-based models for object detection in side-scan sonar images. Retina-SWIN, which uses the attention mechanism in the backbone model to extract general features, achieves the highest true-positive rate of 99.5%. A two-step detection pipeline in which the detector is trained for localization only and the classification is carried out by an additional CNN significantly reduced the number of false alarms per image from 178 to 33 while keeping a high maximum true-positive rate of 94.8%. Furthermore, ensembles of the detectors are constructed by combining the predictions of different models. These ensem-

Method	TPR_{max}	**False alarms**		
		@ TPR_{max}	@ TPR=0.9	@ TPR=0.5
YOLOv3	0.967	192.016	163.891	29.698
YOLOv8	0.910	193.837	187.651	4.636
CenterNet2	0.886	122.450	-	2.504
DDETR	0.877	133.961	-	5.860
Retina-SWIN	**0.995**	178.295	103.504	21.643
YOLOv3-L+CNN	0.839	12.884	-	2.240
YOLOv8-L+CNN	0.872	45.636	-	**0.450**
CenterNet2-L+CNN	0.773	14.674	-	1.519
DDETR-L+CNN	0.687	**13.054**	-	1.589
Retina-SWIN-L+CNN	0.948	33.403	**27.876**	2.550
YOLOv3+CNN	0.934	56.411	52.860	17.767
YOLOv8+CNN	0.900	53.295	53.295	1.264
CenterNet2+CNN	0.791	42.860	-	2.085
DDETR+CNN	0.114	19.899	-	-
Retina-SWIN+CNN	0.900	55.000	55.000	13.953

Table 2: Maximum true-positive rate and number of false alarms for the detection methods with d=10 px. Best values marked in bold.

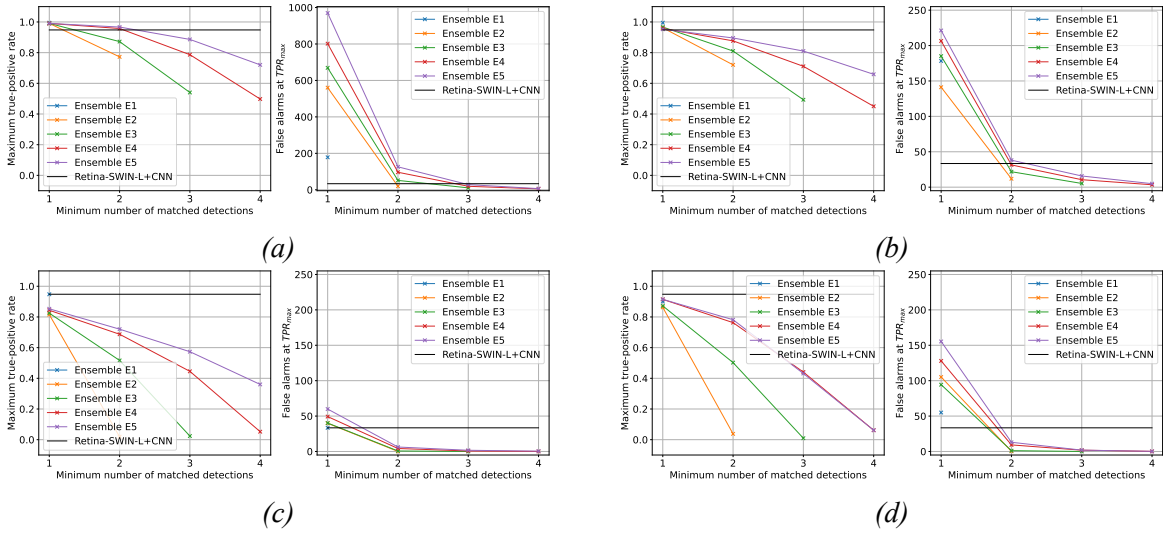

Figure 3: Maximum true-positive rate and number of false alarms of the ensembles with d=10 px. Ensembles are based on the detection pipeline (a) DET, (b) DET with a filtering after ensembeling, (c) LOC+CLS and (e) DET+FLT. Note that the y-axis scale in the right sub-figure in (a) differs from the others.

bles can also reduce the number of false alarms but reduces the maximum true-positive rate too much. The best ensemble combines detections from Retina-SWIN, YOLOv3 and YOLOv8 and achieves a maximum true-positive rate of 87.7% and reduces the false alarms to 31.

REFERENCES

[1] D. P. Williams: "Demystifying Deep Convolutional Neural Networks for Sonar Image Classification" in *Proceedings of the 4th Underwater Acoustics Conference and Exhibition (UACE)*, (Skiathos, 2017).

[2] L. Li, Y. Li, C. Yue, G. Xu, H. Wang, X. Feng: "Real-time underwater target detection for AUV using side scan sonar images based on deep learning", *Applied Ocean Research* **138**, **(2023)**.

[3] I. D. Gerg, V. Monga: "Deep Multi-Look Sequence Processing for Synthetic Aperture Sonar Image Segmentation", *IEEE Transactions on Geoscience and Remote Sensing* **61**, **(2023)**.

[4] Y. Steiniger, D. Kraus, T. Meisen: "Survey on deep learning based computer vision for sonar imagery", *Engineering Applications of Artificial Intelligence* **114**, **(2022)**.

[5] B. W. Sheffield, J. Ellen, B. Witmore: "On vision transformers for classification tasks in side-scan sonar imagery" in *International Conference on Synthetic Aperture Sonar and Synthetic Aperture Radar 2023* (Lerici, 2023).

[6] N. Waraiagoda, Ø. Midtgaard: "Vision Transformers for Sonar Image Classifaction" in *International Conference on Underwater Acoustics 2024* (Bath, 2024).

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo: "Swin transformer: Hierarchical vision transformer using shifted windows" in *2021 IEEE/CVF International Conference on Computer Vision* (virtual, 2021).

[8] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai: "Deformable DETR: Deformable Transformers for End-to-End Object Detection" in *9th International Conference on Learning Representations (ICLR)* (virtual, 2021).

[9] Y. Steiniger, J. Stoppe, D. Kraus, T. Meisen: "On the detection and classification of objects in scarce sidescan sonar image dataset with deep learning methods" in *7th Underwater Acoustic Conference and Exhibition (UACE)* (Kalamata, 2023).

[10] X. Zhou, V. Koltun, P. Krähenbühl: "Probabilistic two-stage detection", *arXiv:2103.07461*, 2018, [online] Available: https://arxiv.org/abs/2103.07461.

[11] G. Jocher, A. Chaurasia, J. Qiu: "YOLO by Ultralytics" (8.0.0), 2023, [Software] https://github.com/ultralytics/ultralytics.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin: "Attention Is All You Need" in *Advances in Neural Information Processing Systems 30* (Long Beach, 2017).

[13] Y. Steiniger, J. Groen, J. Stoppe, D. Kraus, T. Meisen: "A study on modern deep learning detection algorithms for automatic target recognition in sidescan sonar images", in *6th Underwater Acoustics Conference and Exhibition* (virtual, 2021).

[14] J. Redmon, A. Farhadi: "YOLOv3: An Incremental Improvement", *arXiv:1804.02767*, 2018, [online] Available: https://arxiv.org/abs/1804.02767.

[15] K. Chen *et al.*: "MMDetection", *arXiv:1906.07155*, 2019, [online] Available: https://arxiv.org/abs/1906.07155.