EO and radar fusion for fine-grained target classification with a strong few-shot learning baseline

Luca Ballan, Jorge G. O. Melo, Sebastiaan P. van den Broek, Jan Baan, Friso G. Heslinga, Wyke Huizinga, Judith Dijk, and Arta Dilo

TNO Intelligent Imaging, Oude Waalsdorperweg 63, 2597AK The Hague (NL)

ABSTRACT

Combining data from multiple sensors to improve the overall robustness and reliability of a classification system has become crucial in many applications, from military surveillance and decision support, to autonomous driving, robotics, and medical imaging. This so-called sensor fusion is especially interesting for fine-grained target classification, in which very specific sub-categories (e.g. ship types) need to be distinguished, a task that can be challenging with data from a single modality. Typical modalities are electro-optical (EO) image sensors, that can provide rich visual details of an object of interest, and radar, that can yield additional spatial information. Defined by the approach used to combine data from these sensors, several fusion techniques exist. For example, late fusion can merge class probabilities outputted by separate processing pipelines dedicated to each of the individual sensor data. In particular, deep learning (DL) has been widely leveraged for EO image analysis, but typically requires a lot of data to adapt to the nuances of a fine-grained classification task. Recent advances in DL on foundation models have shown a high potential when dealing with in-domain data scarcity, especially in combination with few-shot learning. This paper presents a framework to effectively combine EO and radar sensor data, and shows how this method outperforms stand-alone single sensor methods for fine-grained target classification. We adopt a strong few-shot image classification baseline based on foundation models, which robustly handles the lack of in-domain data and exploits rich visual features. In addition, we investigate a weighted and a Bayesian fusion approach to combine target class probabilities outputted by the image classification model and radar kinematic features. Experiments with data acquired in a measurement campaign at the port of Rotterdam show that our fusion method improves on the classification performance of individual modalities.

Keywords: deep learning, sensor fusion, image classification, few-shot learning, foundation models, EO and radar, decision support

1. INTRODUCTION

Situational awareness (SA) plays a vital role in contemporary military operations, where fine-grained target classification (FGTC) is an important capability. Deep learning (DL) has significantly improved the performance of automated classification of electro-optical (EO) video data, emerging as a powerful tool in defense applications. However, despite these advancements, tackling and automating FGTC with deep learning models remains challenging for state-of-the-art image analysis systems, especially when real-world variability and unseen situations need to be considered. A deep learning model has learned features from training examples to distinguish classes based on visual features extracted from the images. Nevertheless, when classes are visually similar, or when there is a large variety within a class, the learned features are only weakly discriminative for class separation. When the model is applied to unseen situations, these features are even less discriminative. In other terms, the task complexity is given by the high inter-class similarity and intra-class variation occurring in such scenarios. Examples are the classification of different vessel types, and the recognition of behavior in a military scene, where the inherently high variability make it necessary to focus on fine-grained details. A potential solution for this challenge is the integration of EO sensor data with other sources through sensor fusion. That such a fusion of data from multiple sensors can improve the performance, robustness and reliability of automated classification significantly, has been shown in various domains already, including autonomous driving,[1] robotics,[2] and medical imaging.[3]

Corresponding author: Luca Ballan, E-mail: luca.ballan@tno.nl

Signal Processing, Sensor/Information Fusion, and Target Recognition XXXIII, edited by Ivan Kadar, Erik P. Blasch, Lynne L. Grewe, Proc. of SPIE Vol. 13057, 130570K · © 2024 SPIE · 0277-786X · doi: 10.1117/12.3013756

A major challenge in employing deep learning for FGTC tasks in realistic military applications is the substantial amount of data required. Collection and annotation of a sufficiently large dataset to tackle a new situation are typically impractical and often impossible, especially in a military scenario where the data for an FGTC task is often very scarcely available. Recent advancements by the development of foundation models offer a potential solution by demonstrating an ability to compensate for data scarcity and adapt easily to new situations. Foundation models are by definition trained on vast amounts of diverse data, and are known for their effectiveness as image encoders across a variety of domains in zero-shot settings. In other words, it is possible to adapt such general-purpose technologies to fit the task at-hand without further retraining. This "off-the-shelf" performance can be further improved with few-shot learning techniques, providing a way to develop fine-grained target recognition systems that require a relatively small amount of data.

Even with the application of advanced DL models to analyze images, the presence of weak discriminative features within the image data may hinder accurate classification, especially in the FGTC setting. In addressing this limitation, the improvement of the image data involves incorporating additional data modalities such as object kinematics and object dimensions. Information about motion can help resolve ambiguities and contribute to more accurate classifications by distinguishing objects based on their dynamic behavior rather than their static visual appearance. Similarly, when objects share visual features, their physical dimensions become crucial for discrimination.

This study addresses the challenge of constructing a fine-grained target recognition system optimized for scenarios with limited data availability by proposing the combination of a sensor fusion method and the capabilities of foundation models. To validate the applicability of our approach, we apply it to a ship classification task, utilizing visual images acquired with an EO sensor together with both kinematics and dimensional data, acquired with radar. The image processing pipeline is described, incorporating reference examples and explored few-shot learning strategies. The radar kinematics pipeline is also outlined, including feature computation, feature selection and classification. Our approach employs a late-fusion scheme grounded in probabilistic reasoning, representing a realistic use case with a small dataset. The recognition performance of our method is compared with that of the individual sensors to evaluate its efficacy. Notably, the combination of visual and kinematics data outperforms the individual sensors in a realistic scenario, showing the potential of sensor fusion in enhancing fine-grained target recognition in military applications. This paper is organized as follows. Section 2 gives an overview of recent literature. In Section 3, we describe the data used in this work, together with the developed methods. Experiments, with quantitative and qualitative results, are reported in Section 4. In Section 5 we conclude with a discussion on the results, and highlight opportunities for future work.

2. RELATED WORK

The task of deep learning-based image classification has witnessed the development and optimization of two prominent model architectures: Transformers and Convolutional Neural Networks (CNNs). While CNNs have been the conventional choice for image-related tasks, recent developments have introduced transformer-based models as compelling contenders. Transformer-based models, originally designed for sequential data, have demonstrated a notable advantage in handling spatial relationships within images. This adaptability to non-sequential data has sparked interest in exploring their efficacy for image classification. Recent studies have reported instances where transformer-based models exhibit superior performance compared to traditional CNN architectures in most image classification tasks.[4]

It is important to point out that much of the research on novel deep learning-based approaches tends to evaluate their efficacy using generic datasets like ImageNet. While ImageNet has a vast and diverse array of classes, it does not capture the intricacies of fine-grained image classification tasks present in specialized datasets, such as Stanford Cars[5] and Caltech UCSD Birds.[6] Thus, the reliance on generic datasets does not allow to adequately assess the discriminative capabilities required for nuanced tasks, where subtle differences between closely related classes pose a challenge for the required model sensitivity.

Most of the work done on FGTC from image data focuses on automatically finding and highlighting the most discriminative features between a set of classes while reducing the influence of any class-invariant features and background noise. Chou et al.[7] propose a method for multi-scale feature extractors, which selects and combines

the most discriminative feature scales by refining the feature map at different scales. At the same time, it applies background suppression by splitting the features map into foreground and background features. A different approach is taken by Chou et al.[8] with a plug-in module that learns pixel-level feature maps highlighting the most discriminative features in the image space. Do et al.[9] introduce a "Self-Assessment Classifier", which iteratively identifies the class invariant features in the image space and removes them from the image by masking those regions. Despite the promising results shown by the different methods for FGTC in the literature, none of them addresses the issues present in the few-shot learning setting. Some of them even propose whole new neural networks to aid the discriminative features discovery task, which are unfeasible to train in the few-shot setting.

Few-shot learning methods generally fall into two categories: data-centric approaches, involving the augmentation of training data,[10,11] and knowledge-centric approaches, where a model leverages information acquired from a sizable dataset as a foundation for the new task.[12,13] While data-centric approaches have been the most popular in the literature, recent developments resulting in more robust models, such as foundation DL models, have stimulated the development of knowledge-centric approaches in recent years. Foundation DL models have disrupted several fields of research focusing on DL technologies. These are models that are typically trained on broad data such that they can be applied across a wide range of use cases. Leveraging on unsupervised or semi-supervised learning paradigms, these models have proved to successfully learn rich and robust feature representations. In DINOv2,[14] the authors introduce a novel unsupervised transformer-based method, yielding a robust image encoder that generates rich image representations. The authors demonstrate its ability to generalize across diverse datasets, including fine-grained classification datasets like iNaturalist[15] and Places205[16] by training a linear classifier on the image representations generated by the image encoder, which resembles the most naïve few-shot learning strategies.

In some of the knowledge-centric methods, existing model parameters are utilized as a foundation, and an additional set of parameters is trained specifically for the new domain. It involves leveraging the knowledge encoded in the pre-trained model to facilitate learning in a novel task. Hoffman et al.[17] describe the training of a linear classifier atop feature embeddings produced by a CNN model. The CNN serves as the pre-trained model, and its learned feature representations act as a rich knowledge base for the subsequent task. The linear classifier, introduced on top of these embeddings, is then trained to adapt to the characteristics of the new domain, allowing the model to generalize effectively with limited labeled examples. This approach harnesses the transferable knowledge embedded in the pre-trained model while tailoring specific parameters to the nuances of the target task.

Despite the ability of the previously described methods to optimize the information extracted from visual cues, accurate classification may still be hindered in real-world scenarios, especially in the FGTC setting. Leveraging data from additional sensors has been attempted in recent literature, and a division is seen on the chosen fusion techniques. Farahnakian et al.[18] categorize the latter in early, middle and late fusion, according to the level at which an architecture carries out the merger. Late (also referred to as "decision-level") fusion methods generally combine information or probabilities at the output level of individual models. Helgesen et al.[19] argue that different environments have distinct challenges precluding the use of only a single sensor, and express the need for flexibility when combining valuable information. They propose a multi-target tracker where association probabilities are calibrated given separate radar, lidar, electro-optical and infrared measurements. Magnant et al.[20] perform tracking and classification through a recursive Bayesian algorithm based on a Multiple-Model Gaussian Mixture Probability Hypothesis Density (MM-GMPHD).

Multiple papers tackle the problem in the maritime and harbor surveillance scenario. Qu et al.[21] include AIS data as direction, speed and location of vessels through a measurement-based Hungarian algorithm to consider multiple feature similarities for occlusion-robust vessel tracking. The same authors also propose a new multi-sensor benchmark dataset for detection and tracking.[22] The approach developed by Debaque et al.[23] combines several DL classifiers using evidential reasoning based on Dempster-Shafer theory, to better take into account the uncertainty at the last layer of the models. They compare their approach to a baseline Bayesian classifier. Baekkegaard et al.,[24] and Ginoulhac et al.[25] exploit kinematic features for classification. In their approach, time series data is fed to a Recurrent Neural Network (RNN) or to another statistical approach like a Gradient Boosting classifier to process the aggregated temporal variables in a supervised training manner. Reviews on

recent advances in multi-sensor fusion for recognition, situational awareness and understanding are given by Qiao et al.[26] and Samaras et al.[27]

In a large harbor surveillance experiment, Dijk et al.[28] performed daylight and infrared recordings with multiple cameras, and demonstrated how AIS information can be combined with imagery to enrich the information on the objects. The data has been processed further in this work to obtain a suitable dataset for the experiments. Out-of-domain classification on these recordings has been previously investigated, where a DL model trained on the MARVEL dataset is tested according to real-world constraints.[29] Also, small object detection (SOD) leveraging temporal context,[30] and adaptation between multiple sensors[31] have been performed on the aforementioned data.

3. DATA AND METHODS

In this section, we explain describe the data and our proposed solution for fine-grained target classification. Firstly, we present a detailed description of the data collection and processing steps. Then, the individual approaches are introduced, i.e. the EO image-based classifier and the kinematic/dimensional feature classifier. The latter includes the numerical feature computation from radar data. After that, we present our few-shot learning strategy, which is applied in the same manner on both stand-alone methods. Finally, we present our a late-fusion scheme that combines the single sensor probabilities to obtain classification, based on both EO sensor and radar data.

3.1 Data Preparation

The dataset used in our experiments is based on a set of recordings in the harbor of Rotterdam.[28] Data from multiple sources was collected during February-April 2021. Among these sources was EO camera data, tracks originating from radar and the automatic identification system (AIS) data. AIS is used by vessel tracking services and complements the radar tracks with a vessel type and unique identification number, the Maritime Mobile Service Identity (MMSI). The radar tracks contain location and speed. The track data was used to derive the kinematic features, at a certain point in time and space. The MMSI was used to verify ground truth classes for the FGTC task, as subsequently described. The ship dimensions in length and width were obtained from the track data, based from AIS information. Highlights of the dataset collection are shown in Figure 1.

To accurately estimate the dimensions of the vessel in meters, the projection of camera coordinates to world coordinates was computed. Using the taxonomies of the benchmark dataset MARVEL[32] and the MMSI numbers it was possible to categorize the ships in the harbor into multiple classes. The dataset (which we refer to as "Rotterdam 2021") comprises five classes, namely Cargo, Dredger, Passenger, Pleasure and Tanker. Given the unbalanced distribution of these classes in the original recordings, a best effort was made to retrieve a sufficient and diverse number of samples in these five classes. Few extra classes (Tug, Military) have been annotated in the process, but are not considered, due to the insufficient number of samples. The number of samples per class, and distribution of them on the unique MMSIs are highlighted in Figure 2. The number of samples is 445, belonging to 144 unique MMSIs, therefore with an average of roughly 3 samples per MMSI. Nevertheless, given the realistic nature of the dataset collection, some vessels appear less or more frequently in the harbor, as for example 2 Dredgers and a Passenger boat with more than 10 samples per MMSI.

The annotation step was done semi-automatically. Frames were selected according to multiple thresholds defined by using the distance to the camera on the ground plain, as shown on the top-right of Figure 1. Also, since the dataset has a time window of multiple months, it was possible to select samples from different tracks, even those with same MMSI. Therefore the drawn images show a large variety in weather conditions, background, vessel orientation, distance (hence resolution) etc. We used a standard COCO-pretrained YOLOv8[33] model to automatically extract bounding boxes from the wide harbor viewpoint, at all the frames selected from the original recordings. The bounding boxes were cross-validated with the UTM coordinates provided by the AIS data (taking a small error margin into account). Next, it was verified visually that each sample crop belonged to the correct MMSI number (and therefore the correct class). Finally, each bounding box was further refined manually if necessary, or it was removed. Removal was done when e.g. multiple vessel types were visible in the same crop, or bounding box location differences were above the threshold. This procedure, together with some extracted samples, is visualized in Figure 1.

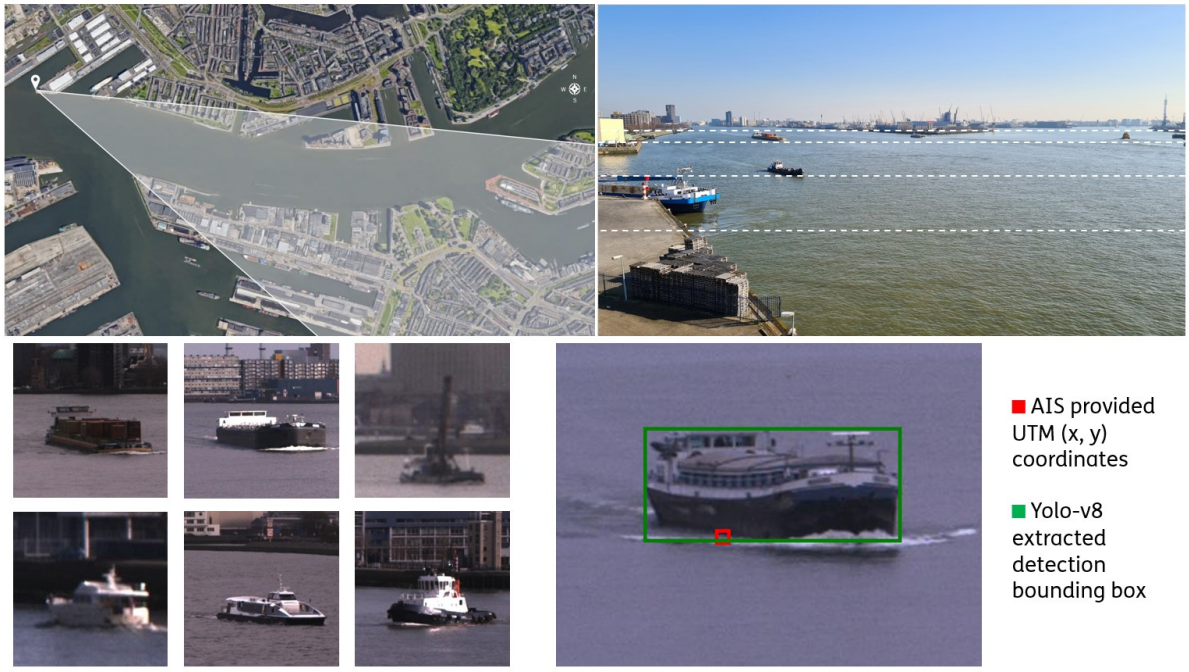

Figure 1. Highlights of the initial data acquisition and dataset preparation steps. Top-left: horizontal FOV of the EO camera setup; top-right: sample frame of the recorded harbor scene; bottom-left: sample crops of a number of classes; bottom-right: image-level alignment between UTM coordinates and YOLOv8 detections.

3.2 EO Classification

Our study utilizes the DINOv2 image encoder as the backbone of the EO classification model. Since it was pre-trained in an unsupervised manner on a carefully curated dataset containing 142 million images, the image encoder learned a thorough understanding of diverse visual patterns. The choice of DINOv2 was based on its proven effectiveness in producing semantically meaningful image representations. The encoder establishes a strong foundation for our subsequent few-shot learning experiments, as it yields visual knowledge ready for fine-tuning on downstream tasks with limited labeled examples. The selected DINOv2 backbone architecture is ViT-g/14. This is the largest available ViT architecture with 2B parameters and a patch resolution of 14x14. Its model weights are downloaded from the official repository*. The model produces a feature embedding of size 1536 for each input image and its weights remain unchanged throughout the experiments. A single-layer Perceptron classifier with 100 neurons was selected for the classification task. The classifier receives a feature embedding generated by the image encoder as input and outputs class probabilities associated with the instance obtained via a Softmax operation.

3.3 Radar Classification

The estimation of class probabilities from kinematic/dimensional features comprises multiple steps. Initially, a set of kinematic features is computed per sample on track level and merged with AIS-based length and width features. Then, a number N of most relevant features is selected through forward Sequential Feature Selection (SFS).[34] SFS identifies in a greedy fashion the most relevant features from a dataset for a particular predictive modeling task. These features are fed into XGBoost, a machine learning algorithm based on boosting decision trees ensembles, which is suitable to handle high-dimensional data and missing values.[35] The dimensional features include width (w) and length (l) of each vessel at sample level. These could be estimated from the distance and orientation of the objects in the scene, given their motion information and the camera-to-world coordinate transformation. Nevertheless, being the corresponding features available in the AIS data, we chose to

*https://github.com/facebookresearch/dinov2

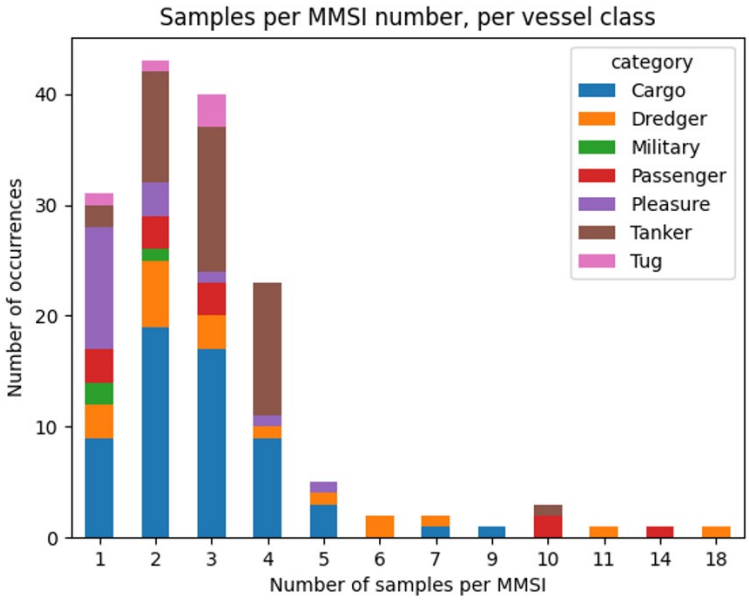

Figure 2. Distribution of classes and MMSIs in the Rotterdam 2021 dataset.

use these values instead. For kinematics, 15 features are computed, belonging to three groups, namely speed (v), tangential acceleration (a^t) and normal acceleration (a^n). We adopt the formulas as defined by Baekkegaard et al.[24] to compute v_i, a_i^t and a_i^n, given a set of observations at any timestamp t_i. For each feature group we store five aggregations over each track in the dataset, for the corresponding samples, to obtain features at track level. These aggregations are the mean (μ), the standard deviation (σ), and the 0.05, 0.50 (median) and 0.95 quantiles (q) of the computed metric distribution. For example, feature q_v^5 of a sample indicates the 0.05 quantile of the speed distribution of the track belonging to that sample. The total of 17 features considered are therefore the following: $\{l,\ w,\ \mu_v,\ \sigma_v,\ q_v^5,\ q_v^{50},\ q_v^{95},\ \mu_{a^t},\ \sigma_{a^t},\ q_{a^t}^5,\ q_{a^t}^{50},\ q_{a^t}^{95},\ \sigma_{a^n},\ q_{a^n}^5,\ q_{a^n}^{50},\ q_{a^n}^{95}\}$.

3.4 Few-shot Learning

The few-shot learning strategy consists of selecting k unique objects from each class in the dataset to use as training data. To ensure that the k selected samples per class do not belong to the same vessel, we check that the selected samples have all different MMSI numbers. The training set is therefore composed by the k selected samples per class, while the evaluation set corresponds to the remainder of the dataset, i.e. all available images of vessels from different MMSI numbers not included in the training set. In the case of the EO classifier, features are extracted from DINOv2 for the whole dataset, and the training set is used to train the final classifier. Similarly, the kinematics/dimensional feature classifier XGBoost is trained on the sample-level and track-level features of the k samples per class (training set), and tested on the corresponding evaluation set. Experiments where done for $k \in \{1, \ldots, 5\}$. We report the mean and standard deviation of the classification accuracy over 100 repetitions ("runs") for each value of k. Each of these runs has a different random split of the dataset for training and evaluation.

3.5 Late-fusion Scheme

The joint classifier is built according to a late-fusion scheme, which requires each of the stand-alone models to output a class probability per sample, normalized over the classes in the dataset. We explored both a weighted and a Bayesian fusion approach.[36] The weighted fusion method takes into account the class-wise accuracy of each model, while the Bayesian fusion method considers priors of encountering each class in the dataset scenario. The prior of each class is estimated as the probability of finding that class in the Rotterdam 2021 dataset. Figure 3 shows the details of the fusion-strategy and the proposed overall multi-sensor fine-grained target classification.

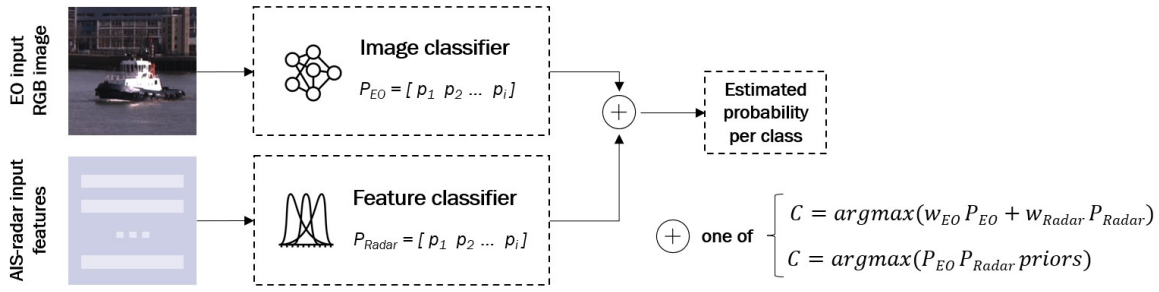

Figure 3. The proposed late-fusion approach for EO+Radar fine-grained target classification. Image data and kinematic/dimensional features are fed as input to the respective models. The output probabilities are then merged according to the adopted late-fusion strategy, and the final class is estimated for each sample in a test setting.

4. EXPERIMENTAL RESULTS

4.1 EO Classification

An initial qualitative analysis is performed on the feature representations generated by the image encoder, as presented in Figure 4. The feature representations are obtained after applying principal component analysis (PCA) on the patch features extracted by the image encoder. The first component is thresholded to isolate the foreground from the background, and a second PCA step is applied to the foreground patches. The first three components of the PCA are colored with three different colors (RGB) and are plotted for visualization. The results suggest that the model is able to recognize a vessel, since it is possible to isolate it from the background, and distinguish its different constituents such as the hull, the bridge and the deck, since different parts of the ship are highlighted with different colors, meaning that different features are obtained.

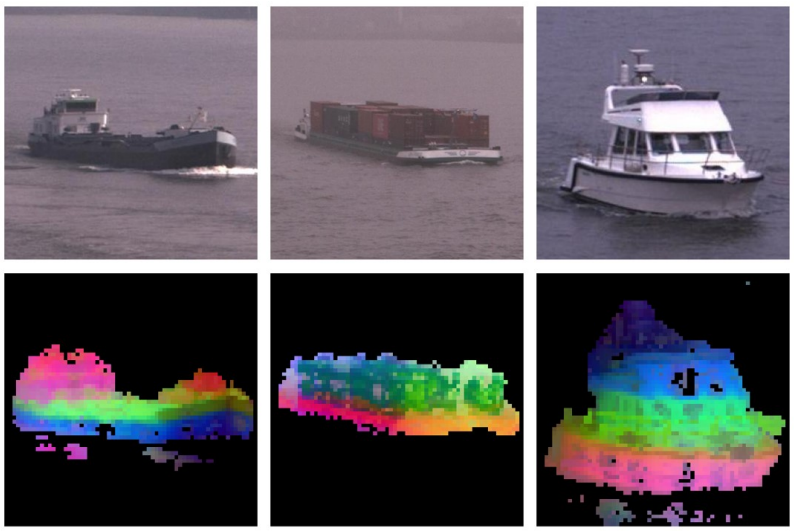

Figure 4. Visualization of the first PCA components. We compute a PCA between the patches of each image in the top row and show their first 3 components below. Each component is matched to a different color channel. Components don't necessarily match between images, because the PCA was computed for each image independently to enable the visualization of more details.

The classifier is trained with Adam optimizer[37] for a maximum of 500 epochs until convergence while trying to minimize the cross-entropy loss. The initial learning rate was set to 0.001 and decreased 5-fold at epoch end, when no improvement in the training loss is observed. The few-shot classification results in Figure 5 show that including a single vessel per class results in an accuracy of $40.5 \pm 8.2\%$, increasing to $57.4 \pm 7.4\%$ when 5 images

of unique vessels per class are used to train the classifier.

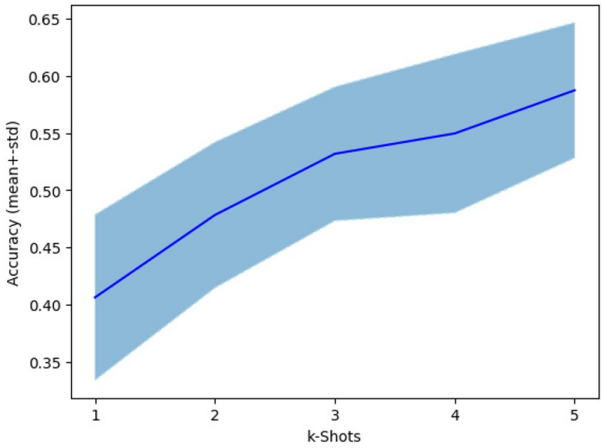

Figure 5. Model performance (average accuracy and standard deviation) for different number of training images per class $k \in \{1, \ldots, 5\}$. For each value of k, 100 experiments are executed, with a new random selection of train instances for each repetition.

To analyze the performance of the classifier in more detail, the value of k was set to 5 and a single experiment was executed. The obtained confusion matrix is shown in Figure 6. The results suggest an increased difficulty in correctly classifying vessels of the class Pleasure, with several instances being classified as Cargo and Dredger, a confusion that a human observer would probably not make. Additionally, the model shows a pattern of misclassification within the classes Tanker, Cargo and Dredger, which relates to the fact instances of these classes often share visual characteristics. This example underscores the challenge of distinguishing between such sets of classes based solely on visual cues, emphasizing the need for nuanced features or enhanced methodologies to improve discrimination accuracy.

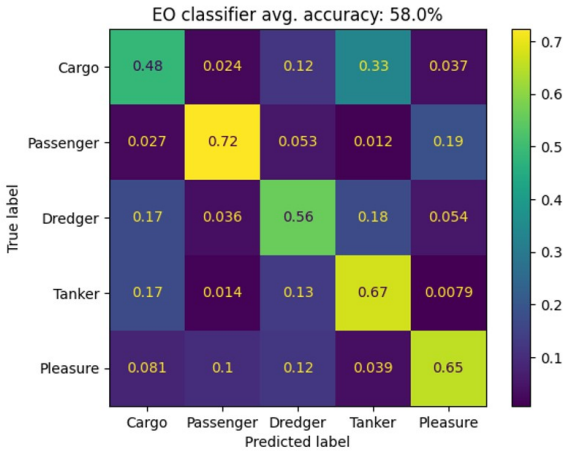

Figure 6. Confusion matrix for the EO classifier on $k = 5$. Results are averaged over 100 runs and normalized for the number of instances in each class (rows sum up to 1). E.g. 0.72 on Pleasure means that 72% of samples of the corresponding class are classified correctly.

While the results presented in this experiment fall slightly below the initial expectations, they provide valuable insights into the problem of training a classifier for FGTC with few example images. Several limitations related to the realistic image acquisition setting, such as low contrast/lightning and low target resolution, may have

influenced these outcomes. These findings serve as a valuable starting point for the proposed methodology that suggests the inclusion of additional data modalities to provide a clearer image of the target objects that ultimately results in better classification performance, even in the few-shot domain.

4.2 Radar features

Dimensional (length, width) and kinematic (speed, tangential and normal acceleration-based) features are computed for the 445 samples in the dataset. The SFS feature selection method is fit with a variable number of features $N \in \{2, \ldots, 17\}$. The estimated optimal performance is given by $N = 10$, with the following set of features: $\{l,\ w,\ \mu_v,\ \sigma_v,\ q_v^5,\ q_v^{50},\ \mu_{a^t},\ \sigma_{a^t},\ q_{a^t}^{50},\ q_{a^n}^5\}$. The usage of additional or different features increased computation time without bringing relevant improvements in the discrimination capability of the model. For visualization purposes, the t-SNE technique[38] (sklearn library implementation[†]) is used to visualize the data points in 2D space in Figure 7. t-SNE is a nonlinear dimensionality reduction technique that models similar objects by nearby points and dissimilar objects by distant points. Larger colored points represent the centers of the distributions for each class. The graph indicates that in a multi-dimensional feature space - here simplified on a flat 2D surface - kinematic/dimensional features can be used to distinguish the various vessel types. The higher overlap between Cargo and Tanker, for example, generally indicates closer characteristics of the two classes, hence the need to use multiple sources of information.

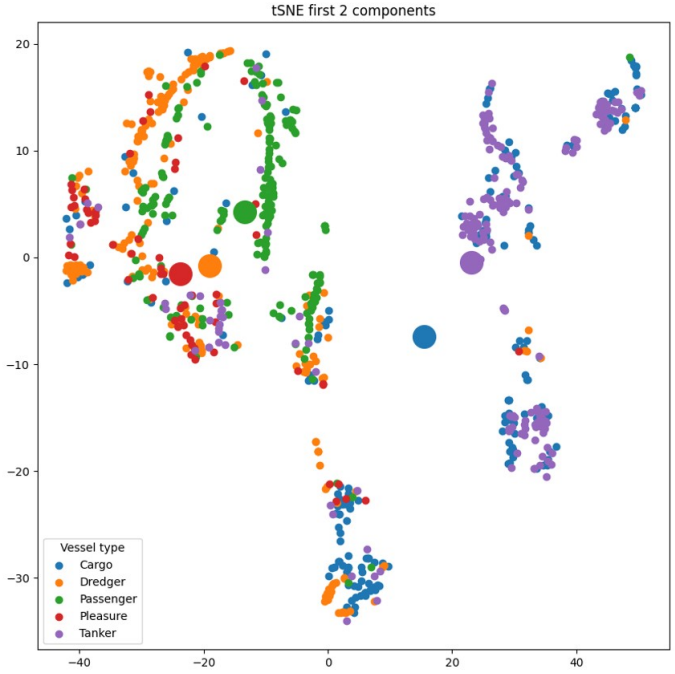

Figure 7. tSNE visualization of the dimensional/kinematic features. Data points are color-clustered by vessel type, and bigger dots represent the class centers. Intuitively, Cargo and Tanker vessel types show a stronger overlap and are more distant apart from the other classes.

4.3 Radar Classification

The stand-alone XGBoost classifier is fit and tested on the kinematic/dimensional features. For the radar feature classification, we used the same experimental setup as for the EO classifier: the results of 100 runs are averaged, where for every run five random MMSIs are selected per class and used to fit the model. During testing, samples belonging to the same MMSIs as those used for training are discarded. Per-class and overall accuracy are shown in Figure 8.

[†] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

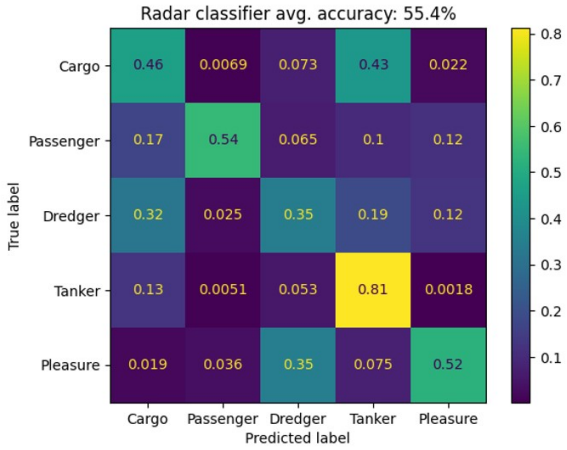

Figure 8. Confusion matrix for the kinematic/dimensional feature classifier on $k = 5$. Results are averaged over 100 runs and normalized for the number of instances in each class (rows sum up to 1).

4.4 Late-fusion Scheme

The Softmax layer outputs of the EO classifier are merged with the estimated output probabilities of the Radar classifier, following the proposed late-fusion scheme. Figure 9 reports per-class and overall accuracy for both the weighted and Bayesian fusion strategies.

In order to qualitatively evaluate the performance of the late-fusion scheme, we report examples of misclassified vessels in two opposite situations, for which the fusion scheme is able to correct the final prediction. Firstly, Figure 10 shows images under challenging conditions, which make the EO classification unreliable; the Radar model instead outputs the right predictions, which ultimately allows the joint model to classify correctly. Figure 11 shows vessel samples belonging to the complementary scenario, i.e. classified wrongly by the Radar model, but showing an EO classifier output with a high probability for the correct class, resulting in a correct joint prediction.

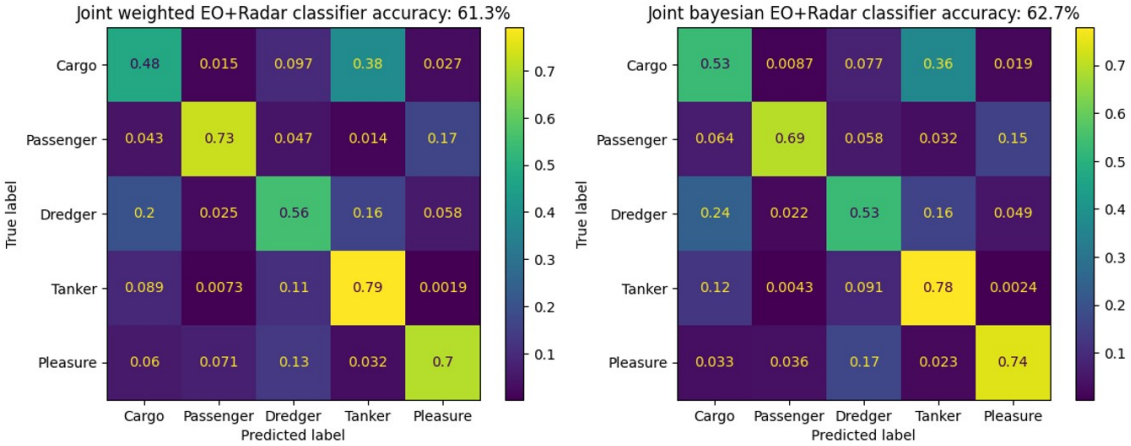

Figure 9. Confusion matrices for the joint EO+Radar classifier. Left: class-wise accuracy weighted fusion. Right: Bayesian-based fusion.

5. DISCUSSION AND FUTURE WORK

The shown qualitative results suggest that the improvement is in some cases limited, as the sensors may capture imprecise and sometimes conflicting data, and an effective combination of diverse sources still remains an open

Figure 10. Examples of "corrected" classification. For the samples above, the EO classifier predicts the wrong vessel type. The joint prediction with radar features, according to the weighted late-fusion scheme, allows for a final correct prediction. In each line, the percentage indicates the highest among the 5 dataset classes.

Figure 11. Examples of "corrected" classification. For the samples above, the Radar classifier predicts the wrong vessel type. The joint prediction with EO features, according to the weighted late-fusion scheme, allows for a final correct prediction. In each line, the percentage indicates the highest among the 5 dataset classes.

challenge. We defined possible alternatives to the explored methodology: 1) the adoption of an early or mid-level fusion scheme, with corresponding pros and cons in terms of simplicity and efficiency; 2) improved EO encoding / learning strategies, as SVM classifier, visual prompt tuning, or others; 3) optimized late-fusion strategy, through better calibration of probabilities; 4) evaluation on a diverse set of few-shot learning scenarios.

6. CONCLUSION

In this study, we have developed an approach for fine-grained classification of EO and radar. We have shown results using real-world data, performing EO and Radar sensor fusion for FGTC on a maritime use case. In particular, we showed that visual, dimensional and kinematic features contain valuable complementary information, and the combination of these in a late-fusion scheme improves the accuracy of an automated recognition system. As real world scenarios often lack data for training FGTC models, we developed a few-shot learning approach that works effectively for both the EO classifier and the kinematic/dimensional feature classifier. The classification performance of our method is compared with that of the individual sensors to evaluate its efficacy. Notably, the combination of visual and kinematics/dimensional data outperforms the stand-alone sensors in the realistic vessel classification scenario, underscoring the potential of sensor fusion in enhancing FGTC in military applications.

ACKNOWLEDGMENTS

The work presented in this paper was performed within V2318 PANOPTES research program, funded by the Dutch Ministry of Defense. The study was carried out within the framework of the Dutch Radar Centre of Expertise (D-RACE), a strategic alliance of Thales Netherlands B.V. and TNO. We gratefully acknowledge the Rotterdam Port Authority for their support in the data acquisition trial.

REFERENCES

[1] Chang, S., Zhang, Y., Zhang, F., Zhao, X., Huang, S., Feng, Z., and Wei, Z., "Spatial attention fusion for obstacle detection using mmWave radar and vision sensor," *Sensors* **20**(4), 956 (2020).

[2] Singh, A., "Vision-radar fusion for robotics BEV detections: A survey," in [*2023 IEEE Intelligent Vehicles Symposium (IV)*], 1–7, IEEE (2023).

[3] Azam, M. A., Khan, K. B., Salahuddin, S., Rehman, E., Khan, S. A., Khan, M. A., Kadry, S., and Gandomi, A. H., "A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics," *Computers in biology and medicine* **144**, 105253 (2022).

[4] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M., "Transformers in vision: A survey," *ACM computing surveys (CSUR)* **54**(10s), 1–41 (2022).

[5] Krause, J., Stark, M., Deng, J., and Fei-Fei, L., "3D Object Representations for Fine-Grained Categorization," in [*2013 IEEE International Conference on Computer Vision Workshops*], 554–561 (2013).

[6] Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P., "Caltech-UCSD birds 200," (2010).

[7] Chou, P.-Y., Kao, Y.-Y., and Lin, C.-H., "Fine-grained visual classification with high-temperature refinement and background suppression," *arXiv preprint arXiv:2303.06442* (2023).

[8] Chou, P.-Y., Lin, C.-H., and Kao, W.-C., "A novel plug-in module for fine-grained visual classification," *arXiv preprint arXiv:2202.03822* (2022).

[9] Do, T., Tran, H., Tjiputra, E., Tran, Q. D., and Nguyen, A., "Fine-grained visual classification using self assessment classifier," *arXiv preprint arXiv:2205.10529* (2022).

[10] Zhang, R., Che, T., Ghahramani, Z., Bengio, Y., and Song, Y., "MetaGAN: An adversarial approach to few-shot learning," *Advances in neural information processing systems* **31** (2018).

[11] Zhou, C., Zhang, Z., Zhou, S., Xing, J., Wu, Q., and Song, J., "Grape leaf spot identification under limited samples by fine grained-GAN," *Ieee Access* **9**, 100480–100489 (2021).

[12] Xu, H., Wang, J., Li, H., Ouyang, D., and Shao, J., "Unsupervised meta-learning for few-shot learning," *Pattern Recognition* **116**, 107951 (2021).

[13] Käppeler, M., Petek, K., Vödisch, N., Burgard, W., and Valada, A., "Few-shot panoptic segmentation with foundation models," *arXiv preprint arXiv:2309.10726* (2023).

[14] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193* (2023).

[15] Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S., "The inaturalist species classification and detection dataset," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 8769–8778 (2018).

[16] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A., "Learning deep features for scene recognition using Places database," *Advances in neural information processing systems* **27** (2014).

[17] Hoffman, J., Tzeng, E., Donahue, J., Jia, Y., Saenko, K., and Darrell, T., "One-shot adaptation of supervised deep convolutional models," *arXiv preprint arXiv:1312.6204* (2013).

[18] Farahnakian, F. and Heikkonen, J., "Deep learning based multi-modal fusion architectures for maritime vessel detection," *Remote Sensing* **12**(16), 2509 (2020).

[19] Helgesen, Ø. K., Vasstein, K., Brekke, E. F., and Stahl, A., "Heterogeneous multi-sensor tracking for an autonomous surface vehicle in a littoral environment," *Ocean Engineering* **252**, 111168 (2022).

[20] Magnant, C., Giremus, A., Grivel, E., Ratton, L., and Joseph, B., "Multi-target tracking using a PHD-based joint tracking and classification algorithm," in [*2016 IEEE Radar Conference (RadarConf)*], 1–6, IEEE (2016).

[21] Qu, J., Liu, R. W., Guo, Y., Lu, Y., Su, J., and Li, P., "Improving maritime traffic surveillance in inland waterways using the robust fusion of AIS and visual data," *Ocean Engineering* **275**, 114198 (2023).

[22] Guo, Y., Liu, R. W., Qu, J., Lu, Y., Zhu, F., and Lv, Y., "Asynchronous trajectory matching-based multimodal maritime data fusion for vessel traffic surveillance in inland waterways," *IEEE Transactions on Intelligent Transportation Systems* (2023).

[23] Debaque, B., Florea, M. C., Duclos-Hindié, N., and Boury-Brisset, A.-C., "Evidential reasoning for ship classification: Fusion of deep learning classifiers," in [*2019 22th International Conference on Information Fusion (FUSION)*], 1–8, IEEE (2019).

[24] Bækkegaard, S., Blixenkrone-Møller, J., Larsen, J. J., and Jochumsen, L., "Target classification using kinematic data and a recurrent neural network," in [*2018 19th International Radar Symposium (IRS)*], 1–10, IEEE (2018).

[25] Ginoulhac, R., Barbaresco, F., Schneider, J.-Y., Pannier, J.-M., and Savary, S., "Target classification based on kinematic data from AIS/ADS-B, using statistical features extraction and boosting," in [*2019 20th International Radar Symposium (IRS)*], 1–10, IEEE (2019).

[26] Qiao, Y., Yin, J., Wang, W., Duarte, F., Yang, J., and Ratti, C., "Survey of deep learning for autonomous surface vehicles in marine environments," *IEEE Transactions on Intelligent Transportation Systems* (2023).

[27] Samaras, S., Diamantidou, E., Ataloglou, D., Sakellariou, N., Vafeiadis, A., Magoulianitis, V., Lalas, A., Dimou, A., Zarpalas, D., Votis, K., et al., "Deep learning on multi sensor data for counter UAV applications—A systematic review," *Sensors* **19**(22), 4837 (2019).

[28] Dijk, J., van den Broek, S. P., den Hollander, R. J., Baan, J., ten Hove, J.-M., and Oorbeek, D., "Multi-sensor information extraction and combination in a large harbor surveillance experiment," in [*Artificial Intelligence and Machine Learning in Defense Applications III*], **11870**, 12–21, SPIE (2021).

[29] den Hollander, R. J., van Rooij, S. B., van den Broek, S. P., and Dijk, J., "Vessel classification for naval operations," in [*Artificial Intelligence and Machine Learning in Defense Applications III*], **11870**, 115–132, SPIE (2021).

[30] Heslinga, F. G., Ruis, F., Ballan, L., van Leeuwen, M. C., Masini, B., van Woerden, J. E., den Hollander, R. J., Berndsen, M., Baan, J., Dijk, J., et al., "Leveraging temporal context in deep learning methodology for small object detection," in [*Artificial Intelligence for Security and Defence Applications*], **12742**, 134–145, SPIE (2023).

[31] Melo, J. G., Ballan, L., van den Broek, B. S., Baan, J., Dijk, J., Huizinga, W., and Dilo, A., "Ship detection in thermal infrared using paired visible light images and domain adaptation via knowledge distillation," in [*Artificial Intelligence for Security and Defence Applications*], **12742**, 165–173, SPIE (2023).

[32] Gundogdu, E., Solmaz, B., Yücesoy, V., and Koc, A., "MARVEL: A large-scale image dataset for maritime vessels," in [*Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V 13*], 165–180, Springer (2017).

[33] Jocher, G., Chaurasia, A., and Qiu, J., "Ultralytics YOLO," (Jan. 2023).

[34] Aha, D. W. and Bankert, R. L., "A comparative evaluation of sequential feature selection algorithms," in [*Pre-proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*], 1–7, PMLR (1995).

[35] Chen, T. and Guestrin, C., "Xgboost: A scalable tree boosting system," in [*Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*], 785–794 (2016).

[36] Tax, D. M., Van Breukelen, M., Duin, R. P., and Kittler, J., "Combining multiple classifiers by averaging or by multiplying?," *Pattern recognition* **33**(9), 1475–1485 (2000).

[37] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014).

[38] Van der Maaten, L. and Hinton, G., "Visualizing data using t-sne.," *Journal of machine learning research* **9**(11) (2008).