Towards Explainable Visual Vessel Recognition Using Fine-Grained Classification and Image Retrieval

Heiko Karus[1,2,3], Friedhelm Schwenker[2], Michael Munz[3], Michael Teutsch[1]

[1] Hensoldt Optronics GmbH, Germany, [2] Ulm University, Germany,
[3] Ulm University of Applied Sciences, Germany
{heiko.karus, michael.teutsch}@hensoldt.net

Abstract

The precise recognition of vessel types is critical for applications in maritime surveillance, but manual visual inspection is slow and error-prone. Automated fine-grained object recognition helps to quickly and accurately categorize vessels, as long as they overcome challenges of modern deep and machine learning-based methods such as class imbalance or intransparency. This paper utilizes recent literature to work towards this task considering two different problem formulations: fine-grained classification and image retrieval. We create a novel dataset called Military MARVEL consisting of 15,858 images of 137 military vessel classes to conduct the extensive, survey-like experiments. On two maritime datasets, we demonstrate that end-to-end fine-grained classification leads to slightly higher accuracy for the price of less flexibility compared to image retrieval. Heatmap-based eXplainable Artificial Intelligence methods are combined with the vessel recognition approaches to assess the achieved transparency quantitatively and qualitatively. Notably, quantitative measures are consistent with qualitative evaluation, especially in fine-grained classification. Our Military MARVEL dataset is available at https://github.com/HensoldtOptronicsCV/FineGrainedVesselRecognition.

1. Introduction

Coastal and maritime surveillance systems typically use radar and sonar to detect targets. Enhancing these systems with vision-based and electro-optical sensors can improve the effectiveness [39]. Except for certain remote sensing applications [24–26], however, computer vision in the maritime domain is a niche topic in recent literature [34, 47, 63] with rather few public benchmarking datasets [5, 6, 29, 39]. Among them are just a few works of literature, in which fine-grained vessel recognition based on small appearance differences is targeted [19, 62]. Fine-grained vessel recognition, however, can be a highly relevant function in maritime surveillance applications as it enhances the situational awareness by distinguishing between up to several hundreds of object classes. This enriches the meta data of an object list provided by a maritime object detector that usually comes with only few, coarse-grained object classes.

In this paper, we compare two different problem formulations for the application of vessel recognition: fine-grained classification and image retrieval. While there is a strong relation between these two tasks [58, 61], they are solved with different Deep Neural Network (DNN) architectures and machine learning approaches [48, 56, 64]. We explore the potential of recent state-of-the-art work in both fine-grained classification and image retrieval [11, 17, 27]. As the existing public datasets [19, 62] provide rather poor image quality [43] or linked images in the internet get increasingly lost over time, we collected a novel dataset called Military MARitime VEsseLs (Military MARVEL). This dataset consists of 15,858 images across 137 different military vessel classes. Based on the Military MARVEL and the commercial Janes Fighting Ships (JFS) database, we analyze and evaluate the fine-grained classification and image retrieval approaches. Finally, state-of-the-art methods for heatmap-based eXplainable Artificial Intelligence (XAI) [15, 23, 38, 45] are utilized to increase the transparency of the decision-making process, thereby increasing the acceptance of deep learning and facilitating human understanding of the predictions made.

Our contributions can be summarized as follows: (1) we provide a thorough analysis of two different problem formulations, namely fine-grained classification and image retrieval, for the same task of visual vessel recognition. (2) we create a novel dataset to conduct the experiments. This dataset is called Military MARVEL and will be published on acceptance. (3) we combine state-of-the-art XAI methods with the fine-grained classification and image retrieval

methods analyzing the potential of fostering transparency.

The remainder of this paper is organized as follows: related work is presented in Section 2. The creation of the Military MARVEL dataset is described in Section 3. Experimental results are presented in Section 4. We conclude in Section 5.

2. Related Work

Fine-grained Classification is a type of image classification that focuses on distinguishing between closely related categories that have subtle differences in appearance [31, 56]. Previous studies have addressed this challenge by using Convolutional Neural Networks (CNNs) to extract informative features at multiple levels [18, 22]. They have employed training strategies that accommodate varying levels of granularity, identified discriminative objects or components, and investigated feature interactions through pairwise learning. Recent advancements in Transformer-based methodologies have been used for fine-grained classification [11, 20]. This is achieved by leveraging feature fusion across multiple Transformer layers and selective attention mechanisms for parts delineation [11, 20].

Deep Metric Learning (DML) is a process that learns a feature embedding to quantify the similarity between objects to facilitate image retrieval [64]. The metric learning losses can be divided into two categories: pair-based and proxy-based. Pair-based methods [3, 10, 46] emphasize sample-to-sample relations, while proxy-based methods, such as Proxy-NCA [35] and Proxy-Anchor [27], focus on proxy-to-sample relations. Proxy-based methods provide better generalization with low training complexity, but they may sacrifice the exploration of semantic information in sample-to-sample relations [59].

Explainable Artificial Intelligence (XAI) aims to improve the interpretability of deep learning models [8, 38, 60]. Explanation methods are categorized based on scope and mechanism, with various taxonomies available. Local explanations interpret individual data points, while global explanations summarize models across datasets [23]. Explanations are classified as white-box or black-box, depending on the level of model access required. Black-box methods [15, 38, 40, 60] are model-agnostic, offering broader applicability, whereas white-box methods [8, 45, 50, 65] typically require access to the model's gradient values. Additionally, XAI methods can be split into occlusion-based approaches [15, 38, 40, 60], where the original image is occluded, fed to the model, and results are compared for salience map creation, activation-based approaches [45, 50, 54] that require gradients, and activation-based approaches tailored for ViTs [1, 8].

3. The Military MARVEL Dataset

The use of large-scale, comprehensive benchmark datasets has significantly improved visual object classification [30, 41], especially when training DNNs. For robust generalization in fine-grained recognition, rather domain-specific training data is required, resulting in the publication of datasets for specific object categories [44, 66]. Therefore, we introduce the Military MARVEL dataset. The dataset name and the data collection approach are adopted from the already existing MARVEL dataset for civil ships [19]. Military MARVEL is currently the largest dataset for fine-grained visual military ship categorization. It comprises 15,858 images across 137 classes, constructed through a semi-automated clustering method. We started with 90,000 webscraped images that we carefully filtered removing duplicates, nighttime images, or low-quality images.

Data Collection: The Military MARVEL dataset was created to serve in benchmarking advanced fine-grained visual recognition techniques for military vessel classification and retrieval. It consists of 15,858 annotated images obtained from a shipspotting website[1]. The annotations include ship category, date, photographer, location, and International Maritime Organization (IMO) number for vessel identification. Although many annotated shipspotting images are available online, their use is usually limited to research purposes, requiring explicit permission from photographers for commercial use due to the vastness of the dataset. Military MARVEL comprises around 90,000 images taken by 1,036 photographers, extensively covering various types of military vessels. The use of annotated visual data from hobbyists can significantly benefit research efforts, provided that the necessary permissions for usage are obtained.

Data Cleanup: When constructing image analysis datasets, it is important to prioritize quality over quantity to avoid biased outcomes and reduced performance [14]. To ensure the quality of the dataset, it is important to identify and remove duplicate images, as they can skew the performance metrics of machine learning models trained on the dataset [36]. Prior to detection and removal of duplicates, a banner was removed at the bottom of each image. While image hashing offers an alternative approach [32, 52], it carries the risk of false positives. The computation of the Structural Similarity Index (SSIM) [55] helps to identify duplicate images and ensures a meaningful assessment of perceptual similarity. Automated processes are used to remove small, grayscale, blurred, and nighttime images, which improves the quality of the dataset. Manual curation further enhances the integrity of the dataset by meticulously removing poor-quality or irrelevant images and ensuring that only single-vessel images are included. Iterative maintenance is employed to ensure the accuracy of the dataset and the definition of classes.

[1] www.shipspotting.com

Despite advancements in automation, human intervention remains essential for recognising image quality. This emphasises the need for thorough manual cleaning processes.

Finding Distinctive Classes: The process of collecting a dataset for military vessels poses challenges due to evolving vessel designs and similarities between models. Initially, grouping vessels based on categories from shipspotting websites yielded only eleven classes, which is insufficient for fine-grained recognition. Another method, grouping by vessel identification numbers (IMO numbers), resulted in 380 classes, and by vessel titles, 452 classes, although many had fewer than 30 samples. To achieve more precise recognition, we organised the data using Wikimedia[2] metadata. This was done primarily through manual inspection, aided by identifying model mergers and verifying differences through rigorous examination of sample images, Wikimedia metadata, and manufacturer websites. We merged similar models when visual differences were indistinct, resulting in 137 classes, each with at least 30 samples. This ensured a more detailed dataset for visual recognition.

Diversity Maximization: A potential limitation of relying on a limited number of photographers to generate a dataset is the potential for unwanted correlations to be introduced into the data [33, 53]. While photographers may be active over a period of years, there may be regional dependencies due to factors such as certain ships visiting certain ports more frequently. To address this, a filtering process was applied to maximise the internal diversity of the dataset. Specifically, pairwise similarity scores were calculated for each pair of images within a variant based on photographer, time, vessel and port. A total of at least 30 images per class were then incrementally and greedily selected in order of decreasing diversity to minimise internal correlation. The resulting images were randomly divided into training, validation and test subsets. This approach was effective in reducing internal correlation, as evidenced by a significant decrease in the classification performance of the baseline classifiers. To avoid complex dependencies in the data, the option of isolating different photographers into different subsets was also considered, but ultimately rejected. There are up to 213 different photographers in a class and at least four photographers in a class. On average, there are 37 different photographers.

Dataset Properties: The dataset contains 15,858 images at a resolution of 224×224 pixels with annotated vessel classes belonging to 137 classes and there are 116 samples per class on average, at least 30 samples per class and at most 902 samples per class (see Fig. 1). The vessels in the dataset can be assigned to eleven different categories with at least 48 samples of fast attack craft and at most 4,366 samples of frigates (see Fig. 2). There are 1,442 samples on average per category. The backgrounds of the images in the Military MARVEL dataset vary greatly (see Fig. 3). Images were taken on canals with a lot of vegetation in the background, on the open sea with a lot of blue background, and in harbors with a lot of buildings in the background. The manual cleanup process as described earlier resulted in six dominant perspectives in the images (see Fig. 4). The dataset contains all possible views except the top, front and rear views of the vessel. The Military MARVEL dataset exhibits a fine-grained structure, in which the vessel cate-

[2]https://www.wikimedia.org/

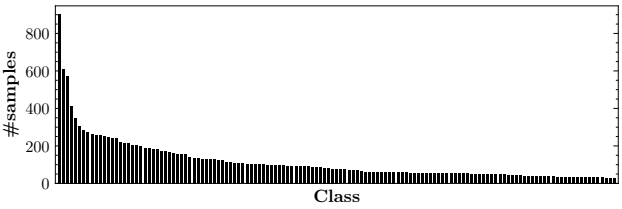


Figure 1. Number of samples per class in Military MARVEL.

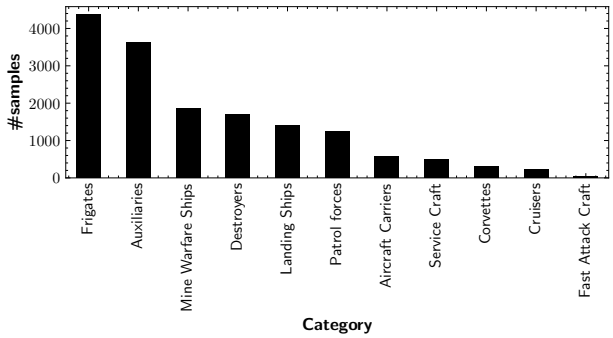
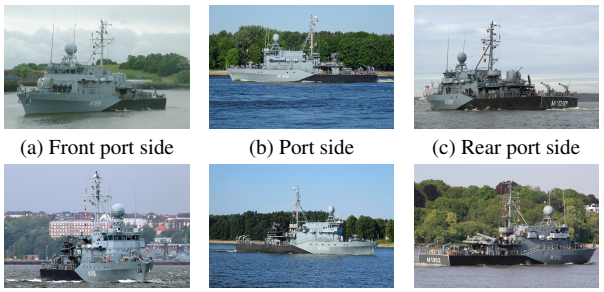


Figure 2. Number of samples per category in Military MARVEL.

(a) Canal with trees (b) Open sea (c) Harbor

Figure 3. Different image backgrounds in Military MARVEL.



(a) Front port side (b) Port side (c) Rear port side

(d) Front starboard side (e) Starboard side (f) Rear starboard side

Figure 4. Different perspectives in Military MARVEL.

(a) Hameln class (b) Kulmbach class

Figure 5. Example for two classes with small appearance difference within the Military MARVEL dataset.

gory is further subdivided into subcategories with nuanced differences in visual features and patterns. Two representative subcategories of the dataset, shown in Fig. 5, are the Hameln class and the Kulmbach class, which are distinguished solely by variations in the configuration of the radar structure on their superstructures. In addition, the intra-class variation of the dataset is substantial, as the classes contain a wide range of perspectives and backgrounds, resulting in significant intra-class variance.

4. Experiments and Results

In extensive experiments, we work towards explainable vessel recognition. We use well-established protocols for training and evaluation [27, 33] to assess and compare methods for image retrieval and fine-grained classification. Besides two vessel recognition datasets, we additionally use two well-established datasets for fine-grained visual recognition [28, 53] as verification datasets. We then explore the current state-of-the-art for this application in (1) visual encoders for feature extraction and clustering, (2) problem formulation via fine-grained image classification or image retrieval, and (3) XAI. We conclude this section with a discussion on the suitability of either fine-grained classification or image retrieval for visual vessel recognition as well as the performance capability of XAI methods.

4.1. Datasets

We obviously utilize the newly created Military MARVEL dataset as presented in Section 3. The commercial JFS database[3] contains 8,460 unique images, including submarines, ships, amphibious vehicles, and naval aircraft. The images are not inherently categorized. However, the information available in the XML files allows the images to be classified into different classes. To utilize this dataset, a subset was created called the JFS_20 dataset, which contains 5,158 images belonging to 20 different classes. In this way, we could guarantee that at least 15 samples exist for each class. Unlike the Military MARVEL dataset, the JFS_20 dataset does not only contain ships. It also contains submarines and military aircraft, resulting in a higher intra-class variance and a lower inter-class variance. The dataset is imbalanced with an average of about 133 samples per class. Furthermore, we use the well-established CUB200 [53] for visual bird recognition and the CARS196 [28] for visual car recognition that have been used for years in related literature to compare fine-grained classification and image retrieval approaches. For better comparability in both fine-grained classification and image retrieval closed-set [4] splits are used, *i.e.* no rejection class is considered. Table 1 shows an overview of the used datasets.

[3]https://www.janes.com/capabilities/defence-equipment-intelligence/naval-combat-systems

Table 1. Overview of the used datasets. Military MARVEL and JFS_20 are datasets for visual vessel recognition. We use the well-established datasets CUB200 and CARS196 for verification.

Dataset	Classes	#images	#training	#test
CUB200 [53]	200	11,788	5,994	5,794
CARS196 [28]	196	16,185	8,144	8,041
Military MARVEL	137	15,858	11,042	4,816
JFS_20	20	5,158	3,461	1,697

4.2. Comparison of Visual Encoders

We evaluate the discriminative abilities of ViT-Small [16], DINO ViT-Small [7], and ResNet50 [21] in fine-grained datasets using the Ranking with Medoids index (RankM) [13]. RankM evaluates the distinctiveness of the extracted features with a number following the principle *the higher the better*. ViT-Small with 22 million parameters is pre-trained on ImageNet-21K [41] producing a 384-dimensional feature vector, while ResNet50 with 23 million parameters is pre-trained on ImageNet-21K, too, producing a 2,048-dimensional vector. ViT-Small with DINO, our third considered approach, is foundation model-based using self-supervised training on ViT-Small using ImageNet-21K without labels. On ImageNet-21K, ViT-Small achieves 78.1 % accuracy, slightly surpassing DINO with ViT-Small with 77.0 % and ResNet50's 76.7 % [7, 9].

RankM was calculated for the four datasets listed in Table 1 and the results are visualized in Table 2. For all four datasets the RankM is higher for features extracted by ViT-Small compared to features extracted by ResNet50. ViT-Small also outperforms ViT-Small with DINO on all datasets, which is consistent with the results of Amir *et al.* [2]: they showed that supervised ViT features are grouped by classes and DINO features are mostly grouped by object parts. As a result, we only consider ViT as visual encoder in the remainder of this section.

Table 2. RankM↑ analysis to measure dataset granularity of pre-trained backbones. The best result for each dataset is written in bold font. ViT-Small clearly outperforms the other approaches.

Feature extractor	CUB200	CARS196	Military	JFS_20
ViT-Small [16]	**0.839**	**0.354**	**0.572**	**0.484**
DINO ViT-Small [7]	0.664	0.217	0.536	0.435
ResNet50 [21]	0.504	0.295	0.399	0.421

4.3. Comparison of Fine-Grained Classification and Image Retrieval

Our experiments compare fine-grained classification and image retrieval methods using RankM and t-SNE plots [51]. The approach-specific quantitative evaluation is conducted with top-1 accuracy [11, 49], for fine-grained image classification and Recall@1 for image retrieval [27, 57]. While RankM, top-1 accuracy, and Recall@1 are quantitative measures that follow the principle *the higher the better*, t-SNE visualizes the clustering capability of an encoder in a dimensionality-reduced 2D representation. Intentionally, we chose top-1 accuracy and Recall@1 as approach-specific metrics measures as they measure similar aspects of performance: matching the correct class with the first hit. Thus, they are well-suited for a direct comparison of fine-grained classification and image retrieval methods. First, we perform ablation studies considering recent literature and then we compare these two different problem formulations for the same task of visual vessel recognition.

Fine-Grained Classification: We consider easy-to-use DNN architectures flavored with recent literature: the original ViT-Small network as well as the novel Plug-In Module (PIM) [11] together with ViT-Small. The utilized original models use pre-trained weights from the PyTorch Image Models repository on ImageNet-21k. We put a linear, fully-connected layer on top of the final hidden state of the ViT's [CLS] token as classification head. Batch size was 32, Stochastic Gradient Descent (SGD) optimizer with momentum 0.9, initial learning rate 0.002, and cosine annealing scheduler were used. To handle imbalanced data, minority classes were oversampled using WeightedRandomSampler [37]. Augmentation included random cropping, flipping, and the AutoAugment policy [12].

Image Retrieval: We utilize the following image retrieval techniques taken from recent literature: ViT-Small with Proxy-Anchor (PA) training [27] and ViT-Small with Hyperbolic Embeddings (HYP) [17]. PyTorch Image Models' pre-trained weights are used with SGD optimizer, cosine annealing scheduler, and specific batch sizes for PA. Hyperbolic embeddings are applied to ViT-Small pre-trained on ImageNet-21K using frozen linear projection for patch embeddings and generating 384-dimensional representations for HYP. The matching between the query image and the database is done via cosine similarity [17, 27].

Table 3. RankM↑ calculation to measure dataset granularity. Fine-tuning on the training sets of each dataset strongly boosts the performance of feature extraction. The visual encoder of Image Retrieval (IR) outperforms Fine-Grained Classification (FGC).

Method	Task	CUB	CARS196	Military	JFS_20
ViT-Small [16]	-	0.839	0.354	0.572	0.484
ViT-Small [16]	FGC	0.875	0.848	0.989	0.657
PA [27]	IR	**0.889**	**0.934**	**0.997**	**0.716**

Table 4. Comparison of top-1 accuracy↑ and Recall@1↑ (in percent) of various methods for Fine-Grained Classification (FGC) and Image Retrieval (IR). Only minor improvement is given by PIM compared to the baseline ViT-Small.

Method	Task	CUB200	CARS196	Military	JFS_20
ViT-Small [16]	FGC	86.9	90.4	99.6	65.3
PIM [11]	FGC	**87.5**	**91.0**	**99.7**	**65.8**
PA [27]	IR	84.1	90.6	99.6	55.4
HYP [17]	IR	85.4	88.7	99.5	51.1

Comparison: We start with an analysis of the RankM measure. Table 3 provides an overview. The ViT-Small model pre-trained on ImageNet-21K only is drastically improved according to the RankM measure by fine-tuning the visual encoder for both approaches fine-grained classification and image retrieval with PA training. Image retrieval slightly outperforms fine-grained classification. This observation is confirmed by the t-SNE plots in Fig. 6, where more distinct clusters appear in the lower row that represents the visual encoder of the image retrieval approach. Finally, Table 4 shows that end-to-end trainable fine-grained classification performs slightly better in top-1 accuracy compared to image retrieval in terms of Recall@1, respectively. The improvement given by task-specific recent literature with PIM [11] and HYP [17] is rather small compared to the additional effort. Our experiments on the benchmark datasets CUB200 [53] and CARS196 [28] show that the published results of HYP [17], PIM [11] and ViT-Small [16] can be reproduced. Results of PA [27] in combination with ViT-Small on CUB200 [53] and CARS196 [28] are plausible.

The top-1 accuracy and the Recall@1 on the proposed Military MARVEL dataset is very high indicating that the dataset is well-engineered and and not too difficult. Furthermore, the JFS_20 dataset seems to be the most challenging dataset. Since this dataset was constructed by humans for human-based database queries of meta information rather than image information, this dataset is not well-annotated for computer vision applications. Delving deeper into the dataset, there are classes with large intra-class variance and rather small inter-class variance leading to difficulties in clustering as visualized in the t-SNE plots of Fig. 6.

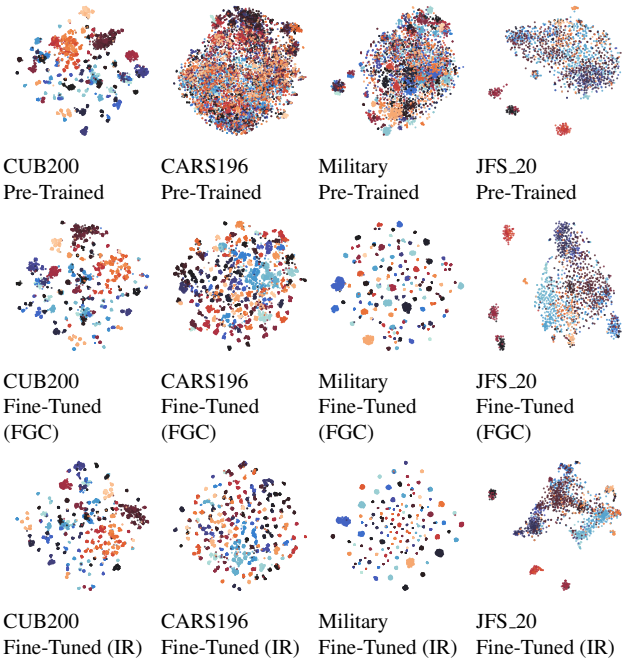
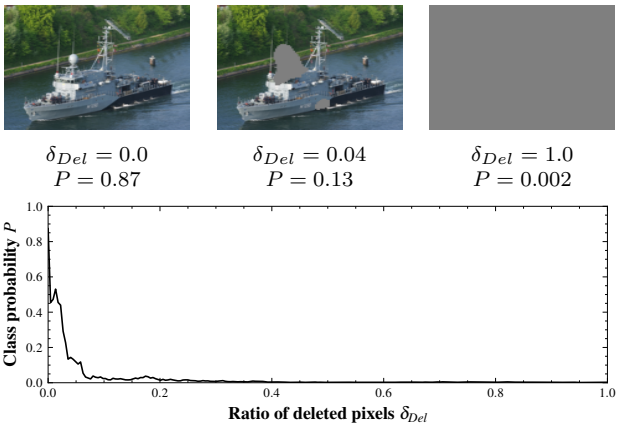


Figure 6. Feature visualization via t-SNE plots of pre-trained encoders (upper row), encoders fine-tuned with Fine-Grained Classification (FGC) in the center row, and encoders fine-tuned with Image Retrieval (IR) in the lower row. Class labels are visualized in different colors. Clustering is expected to work well for all datasets especially after fine-tuning except for JFS_20, which is not well-annotated for computer vision applications.

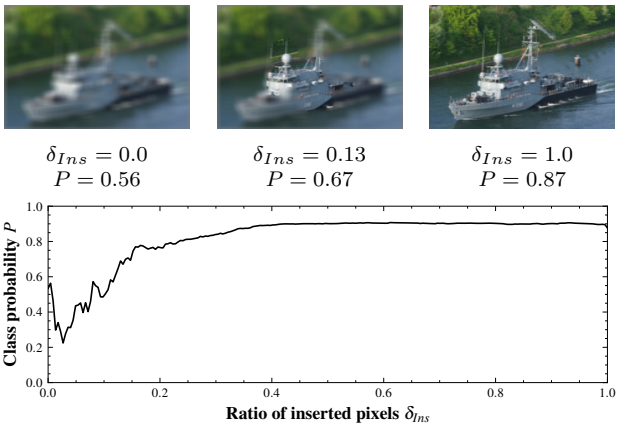
4.4. Explainable Artificial Intelligence

In this section, heatmap-based XAI methods [42] are analyzed for their suitability in our application of visual vessel recognition. Particularly, we evaluate the occlusion-based XAI methods RISE [38], Sliding window [60] and Saliency Based Similarity Maps (SBSM) [15], the activation-based methods Sim-Score [54], Grad-CAM [45] and Sim-CAM [50], and attention-based methods such as Attention rollout [1] and Transformer-IBAV with and without Layer-wise Relevance Propagation (LRP) [8]. They are evaluated quantitatively using the deletion and insertion metric [38]. The deletion metric measures the decrease in the probability of the predicted class as an increasing number of significant pixels are removed. Each pixel's significance is given by the generated saliency map (see Fig. 7). A distinct explanation is indicated by a sharp decrease in the probability curve, resulting in a low Area-Under-Curve (AUC) value. The insertion metric evaluates the increase in probability as more pixels are introduced with a higher AUC indicating a more effective explanation (see Fig. 8). Therefore, a blurred image is generated by smoothing with a Gaussian kernel of size 15. By using both the deletion and insertion metrics, we perform a comprehensive evaluation of XAI methods.



Class probability P over ratio of deleted pixels δ_{Del}.

Figure 7. Deletion metric with ratio of deleted pixels δ_{Del} and class probability P of Kulmbach class. AUC $= 0.025$ for this deletion metric example.



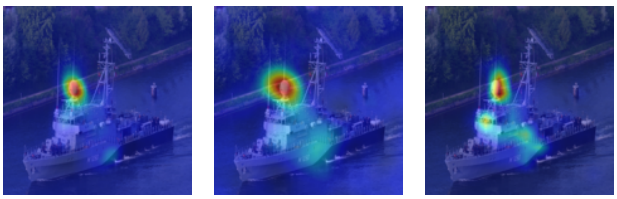
Class probability P over ratio of inserted pixels δ_{Ins}.

Figure 8. Insertion metric with ratio of inserted pixels δ_{Ins} and class probability P of Kulmbach class. AUC $= 0.814$ for this insertion metric example.

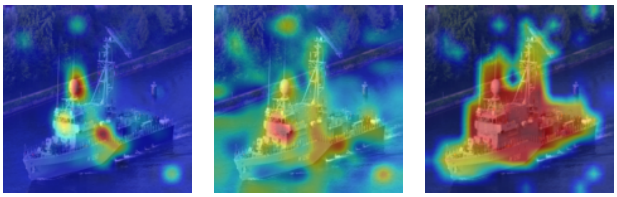
Fine-Grained Classification: We consider the heatmap-based XAI methods for fine-grained classification called Sliding window [60], RISE [38], Transformer-IBAV with and without LRP [8], Attention rollout [1], and Grad-CAM [45] on the Military MARVEL and JFS_20 datasets. For RISE, the binary random masks were stochastically generated with equiprobable values of zero and one. Specifically, a total of 14,400 masks were used, each with dimensions of $h = w = 11$ and consistently applied with $H = W = 224$ in all experimental settings. Binary masks for the Sliding Window method were generated using a window size of 18×18 pixels and a step size of two, resulting in 14,400 masks. These parameters were found by minimizing deletion metric for the image shown in Fig. 7.

Table 5. Quantitative evaluation of heatmaps of XAI methods for fine-grained classification using insertion and deletion metric for the Military MARVEL and the JFS_20 dataset. RISE outperforms the other methods by a good margin.

Data	Method	Deletion ↓	Insertion ↑
Military	RISE [38]	**0.035 ± 0.029**	**0.819 ± 0.109**
	Sliding Window [60]	0.048 ± 0.053	0.798 ± 0.130
	Transformer-IBAV w/ LRP [8]	0.040 ± 0.027	0.785 ± 0.131
	Transformer-IBAV w/o LRP [8]	0.040 ± 0.028	0.783 ± 0.132
	Grad-CAM [45]	0.142 ± 0.089	0.756 ± 0.149
	Attention rollout [1]	0.182 ± 0.091	0.674 ± 0.156
JFS_20	RISE [38]	**0.154 ± 0.133**	**0.745 ± 0.144**
	Sliding Window [60]	0.175 ± 0.153	0.690 ± 0.181
	Transformer-IBAV w/ LRP [8]	0.161 ± 0.096	0.627 ± 0.187
	Transformer-IBAV w/o LRP [8]	0.167 ± 0.097	0.623 ± 0.189
	Grad-CAM [45]	0.192 ± 0.111	0.593 ± 0.196
	Attention rollout [1]	0.243 ± 0.114	0.517 ± 0.191



(a) Sliding Window (b) RISE (c) T-IBAV with LRP



(d) T-IBAV (e) Attention rollout (f) Grad-Cam

Figure 9. Heatmaps of various XAI methods for fine-grained classification. Since the Kulmbach class distinguishes itself from the most similar class Hameln just by the highlighted superstructure (a spherical radar), the findings of Table 5 are confirmed.

Five methods are quantitatively evaluated on both the Military MARVEL and the JFS_20 dataset. Only true positive samples from the test sets are considered. Table 5 shows the results. Occlusion-based methods, especially RISE, achieved the lowest deletion and highest insertion scores on Military MARVEL, while attention-based and activation mapping XAI methods achieved higher deletion and lower insertion scores. The popular Grad-CAM method performs weak on both datasets. Figure 9 presents example images overlayed with heatmaps generated by the XAI methods. Heatmaps of the occlusion-based methods correctly focus on the spherical radar on top of the superstructure. The heatmap of the activation and attention-based methods focuses on the whole vessel with some artifacts in the background. RISE achieves the best XAI performance on both datasets for fine-grained classification.

Table 6. SBSM with random masks outperforms other XAI methods based on the insertion and deletion metrics for image retrieval on the Military MARVEL dataset. The margin is rather low.

Data	Method	Deletion ↓	Insertion ↑
Military	SBSM (random masks) [15]	**0.326 ± 0.041**	**0.885 ± 0.055**
	SBSM (sliding window) [15]	0.361 ± 0.055	0.879 ± 0.058
	Sim-CAM [54]	0.344 ± 0.055	0.868 ± 0.058
	Sim-Score [50]	0.513 ± 0.206	0.746 ± 0.156
JFS_20	SBSM (random masks) [15]	**0.334 ± 0.073**	**0.765 ± 0.100**
	SBSM (sliding window) [15]	0.445 ± 0.127	0.698 ± 0.128
	Sim-CAM [54]	0.423 ± 0.119	0.695 ± 0.107
	Sim-Score [50]	0.497 ± 0.158	0.630 ± 0.151

Image Retrieval: We compare the similarity-based heatmaps generated by four algorithms, namely SBSM with binary masks and SBSM with random masks [15], Sim-CAM [54], and Sim-Score [50] qualitatively and quantitatively. The same setup is used for SBSM as previously described for RISE and Sliding Window. Table 6 shows the results of the quantitative evaluation. The deletion metric values are quite similar. Notably, there is no significant correlation between the AUC and the qualitative analysis as shown in Fig. 10. However, the two occlusion-based methods showed higher AUC compared to Sim-CAM. The insertion metric also provides similar AUC values. We assume that the image retrieval-based cosine similarity as confidence indication is not as meaningful for calculating the deletion/insertion metrics compared to the classifier confidence in fine-grained classification. Another indication is the strong difference of the deletion/insertion values compared to Table 5. However, the two occlusion-based methods perform slightly better compared to activation-based methods. SBSM with random masks has the highest AUC.

Figure 10 presents a example query image and its corresponding top-1 retrieved image overlayed with heatmaps generated by the different algorithms. The retrieved image is correct. We see that the heatmap of the occlusion-based methods is predominantly focused on the spherical radar on top of the superstructure, whereas the heatmap of the activation-based methods focuses on the whole vessel with some artifacts in the background. We assume that qualitative evaluation of heatmaps, which typically neglects failure cases, can be misleading. In our application, however, SBSM seems to be the choice for XAI in image retrieval.

4.5. Visualization of the Decision-Making Process

In this case study, we take two samples from Military MARVEL to visualize model-based decision-making for both fine-grained classification and image retrieval. Heatmaps are used to gain further insight into the models and the data.

A sample of the Military MARVEL dataset from Kulmbach class is presented in Fig. 11a. We generate heatmaps for the top-3 predicted classes shown in Figures 11b to 11d.

(a) Query image (b) SBSM (binary) (c) SBSM (random)

(d) Sim-Score (e) Sim-CAM

Figure 10. Heatmaps of various XAI methods for image retrieval. The spherical radar was correctly detected as most relevant image region, but the rather similar numbers in Table 6 do not correlate with the strong performance difference observed in the figure.

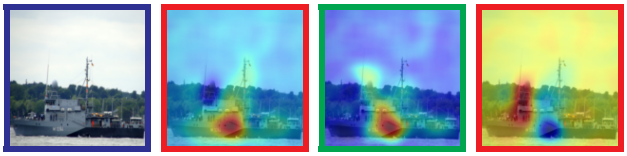
(a) Kulmbach (b) Kulmbach (c) Hameln (d) Blue Ridge

Figure 11. Heatmaps of top-3 predicted classes for Kulmbach class sample.

The Kulmbach class is characterized by a spherical radar located on top of the superstructure, which is effectively captured by the heatmap in Fig. 11a. Note that the Hameln class is similar in appearance to the Kulmbach class just without the spherical radar. The difference between these two classes is highlighted by the blue circle in the heatmap shown in Fig. 11c. The heatmap also shows a high similarity along the diagonal line on the hull. The same sample was used as query image for image retrieval and the heatmaps of the three most similar images are shown in Fig. 12. The heatmaps of Figures 12b to 12d show that the retrieved images are similar to the query image in terms of the spherical radar on top of the superstructure and the diagonal line on the hull. Another sample of the Kulmbach class is shown in Fig. 13a. The sample is incorrectly classified as the Hameln class. The heatmap of Hameln class in Fig. 13b shows the importance of the diagonal line on the hull and that the spherical radar on top of the superstructure is less important for Hameln class. The heatmap in Fig. 13c indicates that the diagonal line on the hull is also important for the Kulmbach class. In addition, the spherical radar contributes to the models' decision. It seems that the model does not detect the spherical radar. In contrast, all three retrieved images

(a) Query image (b) Rank 1 (c) Rank 2 (d) Rank 3

Figure 12. Heatmaps of correctly retrieved images for a Kulmbach class query image.



(a) Kulmbach (b) Hameln (c) Kulmbach (d) Elbe

Figure 13. Heatmaps of top-3 predicted classes for Kulmbach class sample.

(a) Query image (b) Rank 1 (c) Rank 2 (d) Rank 3

Figure 14. Heatmaps of correctly retrieved images for a Kulmbach class query image.

are correct as shown in Fig. 14. Figures 14b to 14d show that the retrieved images and the query image are similar in terms of the spherical radar on the superstructure.

5. Conclusion

In this paper, we introduced a novel dataset for visual vessel recognition, namely the Military MARVEL dataset that consists of 15,858 images arranged in 137 different vessel classes. Based on this dataset and other related datasets, we conducted a systematic analysis of the state-of-the-art in fine-grained image classification, image retrieval, and XAI. Our key findings are: (1) a well-engineered dataset such as Military MARVEL can quite easily be utilized for high-performance visual vessel recognition with state-of-the-art approaches for fine-grained classification and image retrieval. (2) modern Transformer-based visual encoders such as ViT are powerful feature extractors that outperform CNNs. (3) end-to-end fine-grained image classification achieves higher accuracy according to the top-1 accuracy measure compared to image retrieval but image retrieval is more flexible and extendable especially to previously unseen classes. (4) occlusion-/black-box-based XAI methods such as RISE are both powerful and easy-to-implement tools to gain insight into a DNN's decision process.

References

[1] Samira Abnar and Willem Zuidema. Quantifying Attention Flow in Transformers. In *Annual Meeting of the Association for Computational Linguistics*, 2020. 2, 6, 7

[2] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT Features as Dense Visual Descriptors. In *ECCV Workshops*, 2022. 4

[3] Sean Bell and Kavita Bala. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics (TOG)*, 34(4):1–10, 2015. 2

[4] Abhijit Bendale and Terrance E. Boult. Towards Open Set Deep Networks. In *IEEE CVPR*, 2016. 4

[5] Domenico D. Bloisi, Luca Iocchi, Andrea Pennisi, and Luigi Tombolini. ARGOS-Venice Boat Classification. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2015. 1

[6] Borja Bovcon, Jon Muhovič, Janez Perš, and Matej Kristan. The MaSTr1325 dataset for training deep USV obstacle detection models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019. 1

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE CVPR*, 2021. 4, 5

[8] Hila Chefer, Shir Gur, and Lior Wolf. Transformer Interpretability Beyond Attention Visualization. *IEEE CVPR*, 2020. 2, 6, 7

[9] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations. *ICLR*, 2022. 4

[10] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE CVPR*, 2005. 2

[11] Po-Yung Chou, Chu-Hsing Lin, and Wen-Chung Kao. A Novel Plug-in Module for Fine-Grained Visual Classification. *ArXiv*, abs/2202.03822, 2022. 1, 2, 5

[12] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning Augmentation Strategies From Data. *IEEE CVPR*, 2019. 5

[13] Yin Cui, Zeqi Gu, Dhruv Kumar Mahajan, Laurens van der Maaten, Serge J. Belongie, and Ser-Nam Lim. Measuring Dataset Granularity. *ArXiv*, abs/1912.10154, 2019. 4

[14] Samuel Dodge and Lina Karam. Understanding How Image Quality Affects Deep Neural Networks. In *Conference on the Quality of Multimedia Experience (QoMEX)*, 2016. 2

[15] Bo Dong, Roddy Collins, and Anthony Hoogs. Explainability for Content-Based Image Retrieval. In *IEEE CVPR Workshops*, 2019. 1, 2, 6, 7

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 4, 5

[17] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic Vision Transformers: Combining Improvements in Metric Learning. In *IEEE CVPR*, 2022. 1, 5

[18] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly Supervised Complementary Parts Models for Fine-Grained Image Classification From the Bottom Up. *IEEE CVPR*, 2019. 2

[19] Erhan Gundogdu, Berkan Solmaz, Veysel Yücesoy, and Aykut Koç. MARVEL: A Large-Scale Image Dataset for Maritime Vessels. In *ACCV*, 2016. 1, 2

[20] Ju He, Jieneng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, Changhu Wang, and Alan Loddon Yuille. TransFG: A Transformer Architecture for Fine-grained Recognition. In *AAAI Conference on Artificial Intelligence*, 2021. 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE CVPR*, 2016. 4, 5

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:386–397, 2017. 2

[23] Brian Hu, Bhavan Kumar Vasu, and Anthony J. Hoogs. X-MIR: EXplainable Medical Image Retrieval. *IEEE WACV*, 2022. 1, 2

[24] Liang Huang, Fengxiang Wang, Yalun Zhang, and Qingxia Xu. Fine-grained ship classification by combining cnn and swin transformer. *Remote Sensing*, 14(13):3087, 2022. 1

[25] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. A Survey of Deep Learning-Based Object Detection. *IEEE Access*, 7:128837–128868, 2019.

[26] Urška Kanjir, Harm Greidanus, and Krištof Oštir. Vessel detection and classification from spaceborne optical images: A literature survey. *Remote Sensing of Environment*, 207: 1–26, 2018. 1

[27] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy Anchor Loss for Deep Metric Learning. In *IEEE CVPR*, 2020. 1, 2, 4, 5

[28] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In *IEEE ICCV Workshops*, 2013. 4, 5

[29] Matej Kristan, Vildana Sulić Kenk, Stanislav Kovačič, and Janez Perš. Fast Image-Based Obstacle Detection From Unmanned Surface Vehicles. *IEEE Transactions on Cybernetics*, 46(3):641–654, 2016. 1

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NeuRIPS*, 2012. 2

[31] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *IEEE ICCV*, 2015. 2

[32] Xudong Lv and ZJane Wang. An Extended Image Hashing Concept: Content-Based Fingerprinting Using FJLT. *EURASIP Journal on Information Security*, 2009:1–16, 2009. 2

[33] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-Grained Visual Classification of Aircraft. *ArXiv*, abs/1306.5151, 2013. 3, 4

[34] Sebastian Moosbauer, Daniel König, Jens Jäkel, and Michael Teutsch. A Benchmark for Deep Learning Based Object Detection in Maritime Environments. In *IEEE CVPR Workshops*, 2019. 1

[35] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No Fuss Distance Metric Learning Using Proxies. In *IEEE ICCV*, 2017. 2

[36] Dat Tien Nguyen, Firoj Alam, Ferda Ofli, and Muhammad Imran. Automatic image filtering on social networks using deep learning and perceptual hashing during crises. In *International Conference on Information Systems For Crisis Response and Management (ISCRAM)*, 2017. 2

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeuRIPS*, 2019. 5

[38] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *BMVC*, 2018. 1, 2, 6, 7

[39] Dilip K. Prasad, Deepu Rajan, Lily Rachmawati, Eshan Rajabally, and Chai Quek. Video Processing From Electro-Optical Sensors for Object Detection and Tracking in a Maritime Environment: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 18(8):1993–2016, 2017. 1

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016. 2

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115:211–252, 2014. 2, 4

[42] Rabia Saleem, Bo Yuan, Fatih Kurugollu, Ashiq Anjum, and Lu Liu. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing*, 513:165–180, 2022. 6

[43] Mostafa Salem, Yujian Li, Zhaoying Liu, and Ahmed AbdelTawab. A Transfer Learning and Optimized CNN Based Maritime Vessel Classification System. *Applied Sciences*, 13, 2023. 1

[44] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *NeurIPS*, 2018. 2

[45] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *IEEE ICCV*, 2017. 1, 2, 6, 7

[46] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, 2016. 2

[47] Raphael Spraul, Lars Sommer, and Arne Schumann. A comprehensive analysis of modern object detection methods for maritime vessel detection. In *Proceedings of SPIE Vol. 11543*, 2020. 1

[48] Divya Srivastava, Shashank Sheshar Singh, B. Rajitha, Madhushi Verma, Manjit Kaur, and Heung-No Lee. Content-Based Image Retrieval: A Survey on Local and Global Features Selection, Extraction, Representation, and Evaluation Parameters. *IEEE Access*, 11:95410–95431, 2023. 1

[49] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. *Transactions on Machine Learning Research*, 2022. 5

[50] Abby Stylianou, Richard Souvenir, and Robert Pless. Visualizing deep similarity networks. In *IEEE WACV*, 2019. 2, 6, 7

[51] Laurens Van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008. 5

[52] Ramarathnam Venkatesan, S.-M. Koon, Mariusz H. Jakubowski, and Pierre Moulin. Robust image hashing. In *IEEE ICIP*, 2000. 2

[53] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011. 3, 4, 5

[54] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *IEEE CVPR Workshops*, 2020. 2, 6, 7

[55] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 2

[56] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-Grained Image Analysis With Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8927–8948, 2022. 1, 2

[57] Yihan Wu, Hongyang Zhang, and Heng Huang. RetrievalGuard: Provably Robust 1-Nearest Neighbor Image Retrieval. In *International Conference on Machine Learning (ICML)*, 2022. 5

[58] Lingxi Xie, Richang Hong, Bo Zhang, and Qi Tian. Image Classification and Retrieval are ONE. In *ACM on International Conference on Multimedia Retrieval*, 2015. 1

[59] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. PCL: Proxy-based Contrastive Learning for Domain Generalization. In *IEEE CVPR*, 2022. 2

[60] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *ECCV*, 2014. 2, 6, 7

[61] Andrew Zhai and Hao-Yu Wu. Classification is a Strong Baseline for Deep Metric Learning. In *BMVC*, 2018. 1

[62] Mabel M. Zhang, Jean Choi, Kostas Daniilidis, Michael T. Wolf, and Christopher Kanan. VAIS: A dataset for recogniz-

ing maritime imagery in the visible and infrared spectrums. In *IEEE CVPR Workshops*, 2015. 1

[63] Ruolan Zhang, Shaoxi Li, Guanfeng Ji, Xiuping Zhao, Jing Li, and Mingyang Pan. Survey on Deep Learning-Based Marine Object Detection. *Journal of Advanced Transportation*, 2021:1–18, 2021. 1

[64] Liang Zheng, Yi Yang, and Qi Tian. SIFT Meets CNN: A Decade Survey of Instance Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1224–1244, 2018. 1, 2

[65] Sijie Zhu, Taojiannan Yang, and Chen Chen. Visual Explanation for Deep Metric Learning. *IEEE Transactions on Image Processing*, 30(9):7593–7607, 2021. 2

[66] Xiangxin Zhu, Carl Vondrick, Deva Ramanan, and Charless C. Fowlkes. Do We Need More Training Data or Better Models for Object Detection? In *BMVC*, 2012. 2