Anomaly Detection for *In situ* Marine Plankton Images

Yuchun Pu[1] Zhenghui Feng[1] Zhonglei Wang[1,†] Zhenyu Yang[2,3] Jianping Li[2,3,‡]

[1]Xiamen University, China
[2]CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
[3]University of Chinese Academy of Sciences, China
[†]wangzl@xmu.edu.cn, [‡]jp.li@siat.ac.cn

Abstract

Machine learning and deep learning algorithms have achieved great success in plankton image recognition, but most of them are proposed to deal with closed-set tasks, where the distribution of the test data is the same as the training one. In reality, however, we face the challenges of open-set tasks, which are also recognized as the anomaly detection problems. In these tasks, there often exist abnormal classes, which are not in the training set, and the final goal of anomaly detection is to detect the anomalies correctly so that the misclassification of them can be reduced. However, little attention has been paid to anomaly detection in marine related fields. In this paper, to help marine plankton observers to detect anomalies conveniently and efficiently, we propose an anomaly detection pipeline including both the training and the testing phases. The training phase includes two parts, the pre-training and the post-training. In the pre-training phase, we propose a new loss function to better detect the abnormal classes and classify the normal classes simultaneously, which incorporates the expected cross-entropy loss, the expected Kullback-Leibler divergence, and the Anchor loss. We conduct several experiments to show the efficacy of the proposed method and compare its performance with other competitors based on a newly released dataset of in situ marine plankton images. Numerical results show that the proposed method outperforms its competitors in terms of classification accuracy and other commonly used criteria.

1. Introduction

Marine plankton play important roles in marine ecology and have a great impact on aquaculture and global climate change. To understand the biological and ecological processes that regulate plankton populations, a fundamental task is to observe the abundance and taxonomy of plankton over time and space [51]. With the development of underwater optical imaging technology and machine vision, many *in situ* plankton imaging instruments have emerged [34]. Compared with the traditional method that collects water samples by nets and classifies them by light microscope manually, the machine vision technology that based on *in situ* plankton image capturing greatly improves the efficiency for plankton observation.

Although some recent research shows that the machine learning methods have achieved success in other objects image classification [5, 13, 14, 25, 29, 41, 58], the classification of natural seawater samples still encounters a lot of challenges in practice. Schulze *et al*. [57] developed an automated analysis system for the identification of phytoplankton. Dai *et al*. [8] proposed a deep learning architecture for automatic classification of zooplankton images. Sosik *et al*. [59] used a combination of feature selection and a support vector machine to classify phytoplankton images. Blaschko *et al*. [3] utilized a variety of features and classification methods for automatic identification of plankton. Zheng *et al*. [62] proposed an automatic image classification system incoporating multiple view features by multiple kernel learning. All the above efforts belong to closed-set classification tasks, where the distribution of the test data is the same as the training one. As a matter of fact, however, we encounter lots of open-set tasks, which are also recognized as anomaly detection. In these tasks, there often exist abnormal classes, which are not in the training set, and the final goal of anomaly detection is to detect the anomalies correctly so that the misclassification [1, 12, 24, 53, 54] of them can be reduced. However, little attention has been paid to anomaly detection in marine plankton image classification tasks. It is known that in addition to marine plankton, abnormal classes, such as bubbles and various suspending particles exist in the natural seawater. These non-plankton

particles were reported to even occupy more than 97% in the coastal waters [34], which make the classification of the plankton images even harder in the real world. Thus, additional manpower is often needed to identify those images, which is impractical if the data volume is large.

In the field of anomaly detection, a number of traditional machine learning algorithms have been put forward [12, 24, 43, 53, 55]. To make the model acquire prior knowledge of the abnormal classes by utilizing auxiliary datasets, the idea of outlier exposure (OE) [31] has been used widely. [21, 31, 44] proposed supervised deep learning methods using GAN [16, 27] or real-world dataset to generate an artificially abnormal class, which is used to train the network along with the normal ones. However, these methods may result in inaccurate classification of the normal classes; see [2, 26, 32, 52] for other supervised learning methods for anomaly detection. Besides, to achieve desired detection rate of the anomalous classes, Liu *et al*. [37] proposed a semi-supervised learning method based on an additional unlabeled "training" set containing a fraction of anomalies; also see [50]. However, semi-supervised learning methods [37, 50, 60] perform worse than the supervised ones if the model and parameters are not chosen correctly [11]. In addition, unsupervised learning methods have also been developed for anomaly detection [6, 10, 20, 30, 36, 42, 56], but the prior knowledge about anomalous classes is generally unavailable. Thus, the anomaly detection by unsupervised learning methods may be highly uncertain, and their performance fail to outperform the supervised learning methods.

In this paper, we design an anomaly detection pipeline that consists of both a training phase and a testing phase for marine plankton images classification in a supervised way. The training phase includes two parts of pre-training and post-training. In the pre-training phase, a convolutional neural network (CNN) is trained to classify the normal input as well as the auxiliary input as abnormal in advance. The normal input possesses high classification confidence, while the confidence of the auxiliary one is low. In the post-training phase, a detector is trained by transforming the features of an input image extracted by the pre-trained model into a corresponding score. If the score is smaller than a specific threshold ε, the detector will discriminate the corresponding input as abnormal, otherwise it recognizes the input as normal, and the pre-trained CNN will further classify it into one normal class. Based on a specific true positive rate (TPR), the whole post-training process is able to determine the threshold ε. For testing, the feature information of the input is extracted by the pre-trained CNN as a feature extractor. Then, the detector transforms the features into a score. Based on the threshold determined in the post-training phase, the detector is capable of discriminating the input. This discriminating process is similar to the one in the post-training phase. In addition, we provide concrete execution methods for the pipeline. In the pre-training phase, not only do we propose a general loss function based on the OE technique, but also a specific one named CKA loss by incorporating the expected cross-entropy loss, the expected Kullback-Leibler (KL) divergence, and the Anchor loss. The proposed loss function achieves the goal of detecting abnormal classes and accurately classifying the ones belonging to the normal classes simultaneously. To make use of the prior knowledge of the anomalies, one possible approach is to utilize the real-world images as an auxiliary dataset, which is trained together with the original one. However, this idea fails in specific tasks, such as the marine related one in this paper because we lack enough additional plankton datasets. Thus, we propose a data augmentation technique, including image rotation, flipping, blurring [7, 33, 49], and noise addition, to generate abnormal images artificially only based on the training dataset. In the post-training phase, we provide several existing post-training models as references.

The main contributions of this paper are as the following. Firstly, an anomaly detection pipeline consisting of both a training and a testing phase is designed. Secondly, a data augmentation technique is used to generate auxiliary datasets so that the model is able to acquire prior knowledge of the abnormal classes. Thirdly, in the training phase, we propose a CKA loss function to detect abnormal classes and classify the ones belonging to the normal classes simultaneously.

We conduct experiments on a newly released dataset of *in situ* marine plankton images named DYB-PlanktonNet [35] and compare the proposed method with the state-of-the-art anomaly detection methods [21, 42, 44]. Numerical results show that the proposed method outperforms its competitors in terms of classification accuracy, TNR95, AUROC, AUPR and DTACC. We also consider the case in which sufficient *in situ* images are available in advance and can serve as the auxiliary dataset directly. Compared with the augmented auxiliary dataset, we find that the real-world one improves classification accuracy but undermines the anomaly detection performance. Additionally, we conduct the case study to show the comparable performance of our method in the testing phase.

2. Related Work

Anomaly Detection for Plankton. Gonzalez *et al*. [15] utilized the Hellinger distance to study the difference of the distributions between training and test datasets and showed how this distance influences the accuracy of the classifier and the corresponding validation methods. In addition, they emphasized the necessity to focus on designing new learning algorithms which are more robust to anomaly detection. Based on a specific threshold, Zimmerman *et al*. [63] proposed an image quantization method for anomaly detection

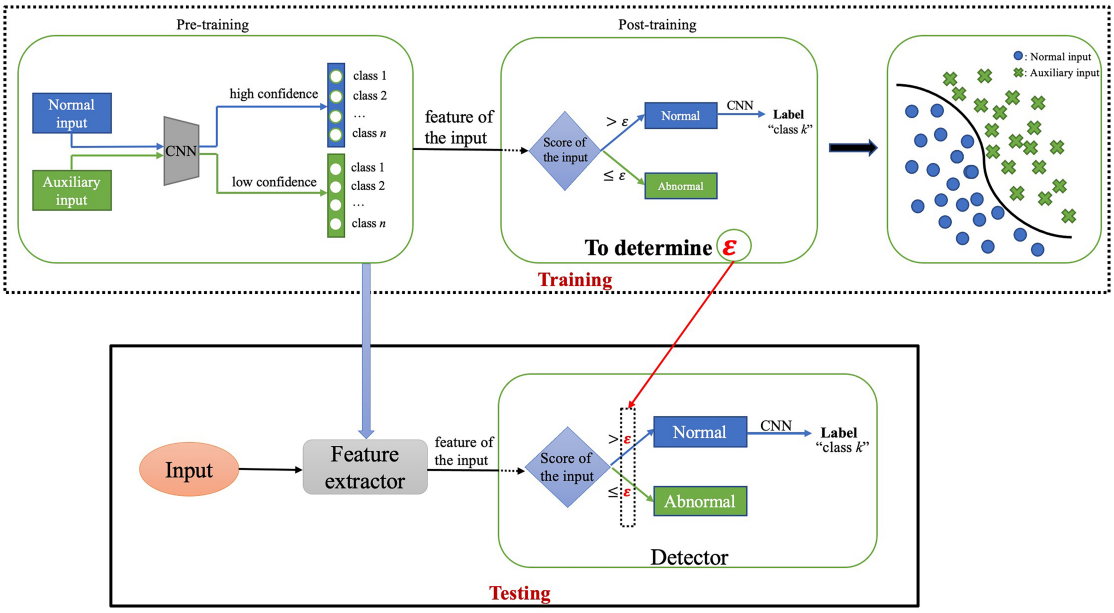

Figure 1. The proposed anomaly detection pipeline, which consists of both the training and the testing phases.

based on the edge strength. Pastore *et al.* [45] adopted one-class SVM to detect anomalies, which is defined as a significant deviation from the established classification.

Outlier Exposure. Lee *et al.* [31] proposed a training method by a deep neural network (DNN) to detect anomalies while maintaining the classification accuracy of the normal classes. Specifically, Lee *et al.* [31] considered an additional cross-entropy loss to guarantee that the predicted distribution of anomalies is uniform, and two models are jointly trained to detect anomalies by alternately minimizing their loss functions. Thus, the trained DNN is extremely unconfident about the anomalies. Hendrycks *et al.* [21] generalized the method of Lee *et al.* [31] by OE, which enables the DNN to learn anomalies in advance with the help of an auxiliary dataset. Based on the OE technique, Papadopoulos *et al.* [44] proposed a new loss function consisting of two constraints. The first constraint minimizes the Euclidean distance between the training accuracy and its average confidence for the training dataset. Thus, the DNN can accurately classify the normal classes. The second one guarantees that the softmax probability of the anomalies is approximately uniform, making the DNN extremely uncertain about them.

Auxiliary Datasets. OE used an auxiliary dataset, which is disjoint from the test one, to train a DNN with better representation for anomaly detection [21]. Goodfellow *et al.* [17] proposed to use adversarial samples to achieve better robustness. Mahajan *et al.* [40] showed that the performance of object detection could be improved by representations, abstracted from sources including search engines and photo-sharing websites. Radford *et al.* [48] trained an unsupervised network by a corpus of Amazon reviews to obtain high-quality sentiment representations. Adhikari *et al.* [23] built six random forest classification models with different sets of objective features, and the inclusion of auxiliary features significantly improved the classification accuracy. Qu *et al.* [47] comprehensively analyzed the performance of different auxiliary features in improving the accuracy of pixel- and object-based land use and land cover classification models, and they showed that the overall classification accuracy can be improved regardless of the types of auxiliary features. Lee *et al.* [31] proposed to use GAN to generate auxiliary datasets for anomaly detection, and Hendrycks *et al.* [21] and Papadopoulos *et al.* [44] used real-world datasets instead.

3. Method

In this section, we design an anomaly detection pipeline to help marine plankton observers to deal with plankton image recognition tasks conveniently and efficiently. The pipeline is schematically illustrated in Figure 1. In addition, we provide concrete execution approaches for the pipeline. It's worth mentioning that a data augmentation technique is proposed to generate auxiliary datasets so that the model is able to acquire prior knowledge of the abnormal classes. Furthermore, in the pre-training phase, we propose a new CKA loss function to help the model to achieve the goal of correctly identifying the normal classes and detecting the anomalies simultaneously.

3.1. Anomaly Detection Pipeline

As shown in Figure 1, we have designed an anomaly detection pipeline, which consists of a training phase and a testing phase. The training phase includes pre-training and

post-training. To classify the normal input as well as the auxiliaries, a CNN is pre-trained in advance. The normal input possesses high classification confidence [44], while the confidence of the auxiliary one is low. In the post-training phase, a detector is trained by transforming the feature information extracted by the pre-trained CNN into the corresponding score. For discrimination, the input will be judged as abnormal if the score is smaller than a specific threshold ε, otherwise it is recognized as normal, and the pre-trained CNN will further classify it into one of the normal classes. Based on a specific TPR, the whole post-training process is able to determine a threshold ε, which is used directly in the testing phase. For testing, the feature information of the input is extracted by the pre-trained CNN feature extractor. Then, the detector transforms the feature into the corresponding score. Based on the threshold determined in the post-training phase, the detector is capable of discriminating the input, and the discriminating process is similar as the one in the post-training phase.

For pre-training, we choose Wide ResNet (WRN) [61] as the CNN model, but other CNNs can also be used such as the ResNet [19] and the DenseNet [22]. We train the WRN with our CKA loss in the pre-training phase; see the CKA loss in Section 3.3 for details. Of course, existing losses such as OE [21], OECC [44] and CAC [42] can also be chosen. Since ensemble models tend to perform better than a single one, we propose to combine an existing post-training detector with the pre-trained CNN. For the choice of the post-training detector, it should have good compatibility with the pre-trained CNN, and the model should have good capability in fully using the previous information. For post-training, the Mahalanobis Distance (MD) classifier [9, 32] is used by default. The existing Maximum Softmax Probability (MSP) [20] classifier and the energy-based one [38] are also good alternatives.

3.2. Data Augmentation

For open-set learning, OE techniques are commonly used for anomaly detection, and auxiliary datasets, such as other real-world or artificially generated ones, should be utilized in advance. The auxiliary datasets can be viewed as train-time anomalies. Existing methods mainly assume the availability of external open-source images [28, 43], but it is not the case for a number of tasks like the one presented in this paper, where we consider the anomaly detection based on the plankton images taken in the coastal waters. However, seldom can external open-source images serve as the auxiliary data for this specific task.

In this paper, we propose to use data augmentation techniques, including image rotation, flipping, blurring, and noise addition, to obtain train-time anomalies from the normal classes for improving the generalization of the proposed method. Specifically, the probability of clockwise rotation is set to 0.8 for each image in the normal classes, and the corresponding rotation angle is uniformly generated from $[-30°, 30°]$. The probability of both horizontal and vertical flip is set to 0.7, and 0.8 for blurring. The random noise [4] is added to each of the normal images, and the noise is generated randomly by one of the following mechanisms, including Gaussian noise, local variance noise, Poisson noise, salt noise, pepper noise, salt and pepper noise, and speckle noise; see [3, 20, 21, 44] for details. Figure 2 shows the generated images by each of the above data augmentation techniques using an original plankton image. In this paper, we consider a combination of these techniques to generate train-time anomalies, and the generated anomalies not only maintain the basic structure of the original images but also provide the "boundary" information; see Figure 3 for details.

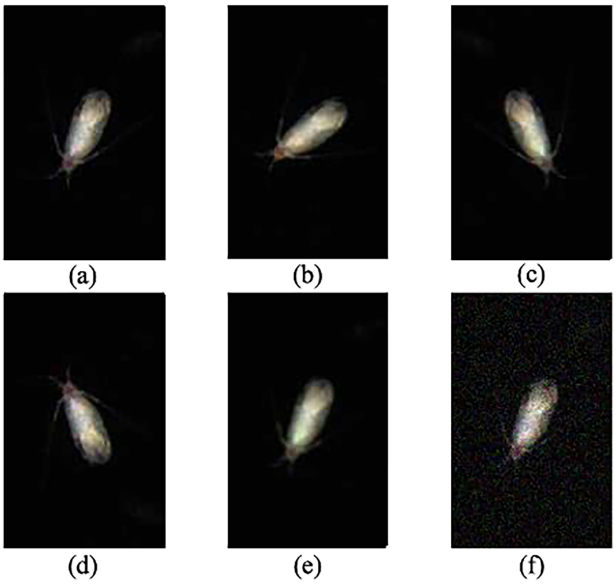

Figure 2. The output of a marine plankton image after each of the data augmentation methods. (a): original image; (b): rotation with $30°$; (c): horizontal flip; (d): vertical flip; (e): blurring; (f): noise addition.

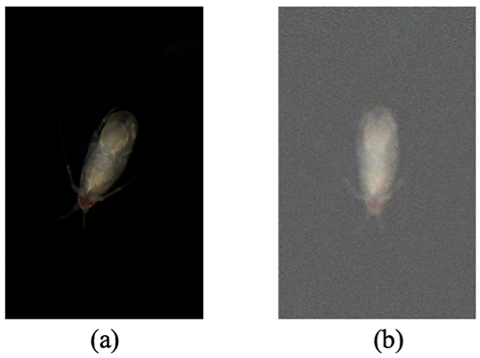

Figure 3. (a) is an original plankton image, and (b) is a generated anomaly from (a) by a combination of data augmentation techniques, including rotation, flipping, noise addition, and blurring.

3.3. Loss Function

Conventional image classification tasks train the DNN by minimizing the cross-entropy loss, which aims to make the predicted distribution as close to the training one as possible, so they are only applicable for closed-set learning. For open-set problems, however, they may lead to overconfident classification, resulting in misclassifying abnormal images to one of the normal classes. In this part, based on the OE technique, we propose a CKA loss function containing three loss terms to detect the anomalies and maintain the classification accuracy for the images belonging to the normal classes.

Equation (1) shows the proposed loss function:

$$L(\boldsymbol{\theta}) = l_1(\boldsymbol{\theta}; \boldsymbol{X_1}, \boldsymbol{Y_1}) + \lambda_1 l_2(\boldsymbol{\theta}; \boldsymbol{X_2}, \boldsymbol{Y_2}) + \lambda_2 l_3(\boldsymbol{\theta}; \boldsymbol{X_1}, \boldsymbol{Y_1}). \quad (1)$$

In the proposed loss function, l_1 is used for classification of the normal images, and various loss functions can be used, such as the expected cross-entropy loss and the mean squared error loss. Based on the auxiliary dataset, l_2 should make the DNN extremely unconfident for detecting the anomalies, that is, the loss term should guarantee the estimated distribution of the auxiliary data is approximately uniform. Since l_2 makes use of the OE technique, it inevitably reduces the classification accuracy for the normal classes. Thus, we add an additional term l_3 to improve the classification accuracy. For the parameters in the loss function, $\boldsymbol{\theta}$ represents the model parameters of the DNN; $(\boldsymbol{X_1}, \boldsymbol{Y_1}) = \left\{\left(\boldsymbol{x_i^1}, \boldsymbol{y_i^1}\right)\right\}_{i=1}^{N_1}$ is the training set of N_1 normal images of size $s \times s$, $\boldsymbol{y}_i^1 = \{y_{i1}, y_{i2}, \ldots, y_{iK}\}$ is the one-hot encoder for $\boldsymbol{x}_i^1$ and K is the number of the normal classes; $(\boldsymbol{X_2}, \boldsymbol{Y_2}) = \left\{\left(\boldsymbol{x_j^2}, \boldsymbol{y_j^2}\right)\right\}_{j=N_1+1}^{N_1+N_2}$ is the set of N_2 auxiliary images of the same size as those in the normal ones, $\boldsymbol{y}_j^2 = \{y_{j1}, y_{j2}, \cdots, y_{jK}\}$ and $y_{j1} = \cdots = y_{jK} = K^{-1}$; $\lambda_1 > 0$ and $\lambda_2 > 0$ are two tuning parameters.

Based on (1), by incorporating the expected cross-entropy loss, the expected KL divergence, and the Anchor loss, we propose the CKA loss. In the CKA loss, l_1 is the following expected cross-entropy:

$$l_1(\boldsymbol{\theta}; \boldsymbol{X_1}, \boldsymbol{Y_1}) = -\frac{1}{N_1} \sum_{i=1}^{N_1} \sum_{k=1}^{K} y_{ik} \log P_{ik}(\boldsymbol{\theta}), \quad (2)$$

where $P_{ik}(\boldsymbol{\theta})$ is the softmax probability with respect to the i-th image and the k-th class. The expected cross-entropy loss is only applied to the images of the normal classes to guarantee the consistency of the predicted distribution associated with the normal classes as close as the training one.

For open-set problems, if the DNN is trained only by minimizing the cross-entropy loss (2), it leads to erroneously misclassifying anomalies into the normal classes. To overcome this difficulty, based on the generated anomalies by the proposed data augmentation techniques, we use KL divergence [46] as l_2 in the CKA loss so that the DNN can be extremely unconfident for classifying the anomalies. The expected KL divergence is as the following:

$$l_2(\boldsymbol{\theta}; \boldsymbol{X_2}, \boldsymbol{Y_2}) = \frac{1}{N_2} \left[\sum_{i=N_1+1}^{N_1+N_2} \sum_{k=1}^{K} \frac{1}{K} \log \frac{K^{-1}}{P_{ik}(\boldsymbol{\theta})} \right]. \quad (3)$$

Except for the expected cross-entropy (2) and the expected KL divergence (3), we consider an additional Anchor loss as l_3 to ensure clustering of the normal images, so that classification accuracy of the normal images can be maintained. The Anchor loss is described as the following:

$$l_3(\boldsymbol{\theta}; \boldsymbol{X_1}, \boldsymbol{Y_1}) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (z_{ik} - \mu_k)^2, \quad (4)$$

where $\mu_k = (n_k)^{-1} \sum_{i=1}^{n_k} z_{ik}$, n_k is number of the original images in the k-th class, and $\{z_{i1}, ..., z_{iK}\}$ are the output features of the i-th original image in the k-th class.

4. Experiments

In this section, we evaluate the anomaly detection performance of our method and compare it with other competitors based on a subset of the DYB-PlanktonNet dataset [35], which is collected by a buoy-borne underwater plankton imaging system [34] deployed in Daya Bay, Shenzhen, China. We conduct our experiments on a subset of the dataset, and it contains 43 classes with 24,880 images in total, including 41 classes of normal plankton, a class of suspending particles and a class of bubbles. The 41 classes of plankton are treated as normal, and the particles and the bubbles are treated as abnormal; see Figure 4 for their representatives. In the experiments, the training dataset consists of 5/6 of the 41 classes of normal plankton images, and they are randomly selected. The remaining 1/6 normal plankton images as well as the abnormal ones are treated as the testing dataset.

In the following experiments, the WRN is chosen as the basic CNN model in the pre-training phase, and the training loss is the CKA loss by default; see Table 1 for the model parameters. As mentioned in the previous part, the MD classifier is used for detecting the anomalies by default.

Parameter	Value	Parameter	Value
Layers	40	**Momentum**	0.9
Widen Factor	2	**Weight Decay**	0.0005
Optimizer	SGD	**Batch Size**	128
Learning Rate	0.01	**Epoch**	100
Dropout Rate	0.3		

Table 1. Model parameters for training the WRN.

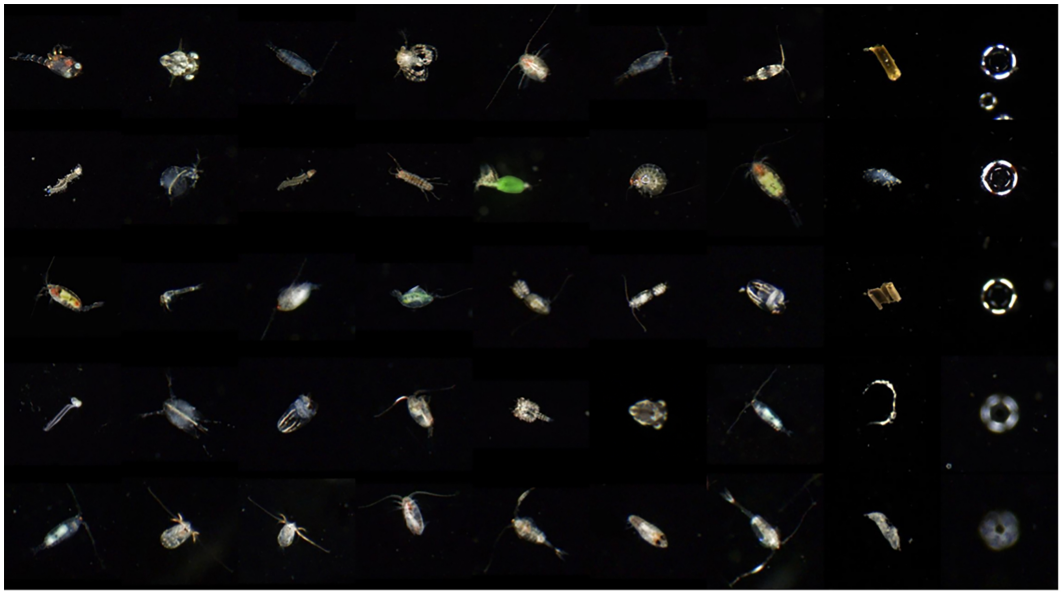

Figure 4. Representative images of plankton, bubbles, and invalid suspending particles of the DYB-PlanktonNet. The first seven columns are plankton images, and the last two columns are suspended particles and bubbles.

4.1. Evaluation Metrics

We consider the following criteria to evaluate the performance of the proposed method and compare it with its competitors by treating the anomalies as positive and the normal as negative. The Accuracy is the classification criterion for normal classes, and the other four are used for the anomaly detection performance.

- **Accuracy (ACC):** (TN+TP) /(FP+TN+TP+FN), where TP, TN, FP, and FN are true positive, true negative, false positive and false negative, respectively.
- **True negative rate at 95% true positive rate (TNR95):** TNR = TN / (FP+TN) when TPR = TP / (TP+FN) is 95%.
- **Area under the receiver operating characteristic curve (AUROC)**, where the ROC curve is a graph plotting TPR against the false positive rate, FP / (FP+TN).
- **Area under the precision-recall curve (AUPR)**, where the PR curve is a graph plotting the precision, TP / (TP+FP), against the recall, TP / (TP+FN), by various thresholds.
- **Detection accuracy (DTACC)**. This criterion corresponds to the maximum classification probability over all possible thresholds ε:

$$1-\min_{\varepsilon}\left\{P_{\text{normal}}\left(q\left(\boldsymbol{x}\right)\leq\varepsilon\right)P\left(\boldsymbol{x}\text{ is from }D_{\text{normal}}\right)\right.$$
$$\left.+P_{\text{abnormal}}\left(q\left(\boldsymbol{x}\right)>\varepsilon\right)P\left(\boldsymbol{x}\text{ is from }D_{\text{abnormal}}\right)\right\}, \tag{5}$$

where $q\left(\boldsymbol{x}\right)$ is the confidence score, $P_{\text{normal}}\left(q\left(\boldsymbol{x}\right)\leq\varepsilon\right)$ is the probability of discriminating a normal image $\boldsymbol{x}$ as abnormal, $P\left(\boldsymbol{x}\text{ is from }D_{\text{normal}}\right)$ is the probability that the image $\boldsymbol{x}$ belongs to a normal class, and $P_{\text{abnormal}}\left(q\left(\boldsymbol{x}\right)>\varepsilon\right)$ and $P\left(\boldsymbol{x}\text{ is from }D_{\text{abnormal}}\right)$ are defined in a similar manner; see [39] for details.

4.2. Data Augmentation

Based on the 41 classes of the normal plankton, the goal of this part is to look for the best combination of the data augmentation techniques for generating an auxiliary dataset. Since rotation and flipping only change the orientation of the object, they do not generate images on the "boundary" of the distribution associated with the normal classes. That is, the combination of these two techniques improves the generalization but fails to generate anomalies. However, noise addition and blurring can generate abnormal images, which have a different distribution as the original one. To generate anomalies with wider generalization, we consider the following combinations: rotation + flipping + noise (RFN), rotation + flipping + blurring (RFB) and rotation + flipping + noise + blurring (RFNB). We also compare the above three combinations with two artificially synthetic methods [21] by arithmetic and geometric averaging of the pixel values of the RGB channels, as depicted in Figure 5.

The classification accuracy of the normal classes is 93.13% by the RFN, 93.16% by the RFB, and 93.21% by the RFNB, respectively. Compared with the results by the two artificial synthetic methods of arithmetic mean (AM) and geometric mean (GM) in [21], which are 85.39% and 81.77%, the three data augmentation techniques achieve more than 8% improvement in terms of accuracy. However, the difference among the three data augmentation techniques is not significant, and it is necessary to compare them on the performance of anomaly detection. In the test-

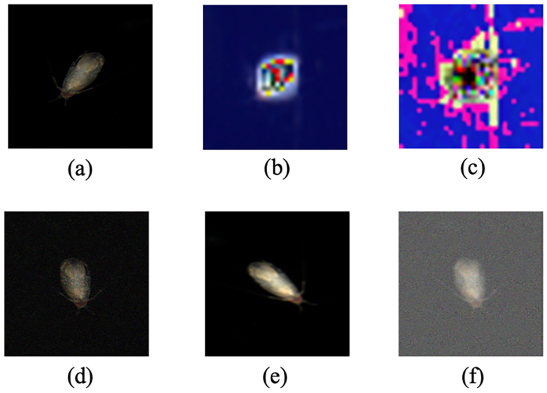

Figure 5. The generated anomalies by different data augmentation methods. (a): original image; (b): arithmetic mean; (c): geometric mean; (d): RFN; (e): RFB; (f): RFNB.

ing phase, we treat the suspended particles and the bubbles as the abnormal classes. The experimental results are shown in Table 2, and we conclude that the third data augmentation technique, *e.g.*, RFNB, performs better than the other two for detecting anomalies.

4.3. Ablation Studies

In the following, based on the MD classifier, we validate the efficacy of the proposed loss function, and we take the CKA loss as the example of validation.

Based on the expected cross-entropy loss with and without the expected KL divergence, we first compare the classification accuracy and the anomaly detection performance. The experimental results are shown in Table 3. Although the classification accuracy of the normal classes reduces a little after adding the expected KL divergence, its anomaly detection performance is significantly improved. Thus, it's necessary to add the expected KL divergence in the tasks of anomaly detection. Furthermore, we consider the following two different loss functions. One is the expected cross-entropy loss incorporated with the expected KL divergence as mentioned above, and the other is the expected cross-entropy loss incorporated with both the expected KL divergence and the Anchor loss. The experimental results are also shown in Table 3. The loss function based on all the three loss terms performs better than the one with the expected cross-entropy loss and the expected KL divergence only, since the Anchor loss term clusters the normal classes and mitigates the interference of the abnormal one. As shown in Table 3, although l_1 performs better than $l_1+l_2+l_3$ in terms of the classification accuracy, its anomaly detection performance is not so good. All in all, the proposed loss function shows its superiority on both classification and anomaly detection.

4.4. Comparison of experimental performance

In this part, based on the MD classifier, we compare the proposed CKA loss with the existing ones on the classification accuracy and the anomaly detection performance, and the auxiliary dataset is the RFNB generated one in Section 4.3. The experimental results are shown in the first four rows of Table 4. The classification accuracy, TNR95 and DTACC by OECC are 93.80%, 80.52%, and 90.22%, respectively, which are higher than the corresponding results by other methods. However, the CKA loss performs better on AUROC and AUPR, which are 96.30% and 98.23%, respectively. Thus, we conclude that the proposed CKA loss outperforms the OE and the CAC since we combine the advantage of these two methods. Although OECC performs slightly better than the proposed method in terms of classification accuracy, TNR95 and DTACC, it trains the network twice. Thus, the training process of the proposed method is much simpler compared with the OECC. Besides, the proposed method does perform better than OECC in terms of AUROC and AUPR. Overall, the proposed CKA loss is better than the existing state-of-art ones.

Additionally, we want to compare the performance of different choices for the post-training model. As an example, based on the CKA loss, we replace the MD classifier by the MSP mentioned in Section 3.1. Since the MSP classifier may lead to overconfident posterior distributions [18], it should perform no better than the MD, and numerical results in the last two rows of Table 4 have also proved this conjecture. Compared with the MD classifier, the MSP one leads to more than 12% reduction in terms of TNR95, AUROC and DTACC. Besides, the MSP classifier reduces AUPR from 98.23% to 95.71%. Thus, the MD classifier performs much better than the MSP one.

	TNR95	AUROC	AUPR	DTACC
RFN	77.13	95.28	98.16	88.36
RFB	76.22	94.17	98.01	87.29
RFNB	**79.94**	**96.30**	**98.23**	**89.34**

Table 2. Anomaly detection performance (%) comparison based on three different data augmentation techniques.

	ACC	TNR95	AUROC	AUPR	DTACC
l_1	**95.51**	73.98	91.02	92.76	82.33
$l_1 + l_2$	92.89	79.13	95.98	98.21	88.95
$l_1 + l_2 + l_3$	93.21	**79.94**	**96.30**	**98.23**	**89.34**

Table 3. The classification accuracy (%) and the anomaly detection performance (%) based on the combination of the different loss terms.

4.5. Comparison of auxiliary datasets

In this part, we consider the case in which sufficient *in situ* images are available in advance so that these images can be used as the auxiliary ones directly. We compare the performance of the proposed method by using different aux-

	ACC	TNR95	AUROC	AUPR	DTACC
OE	92.81	78.12	94.28	96.84	87.94
CAC	93.15	79.01	94.88	97.73	89.07
OECC	**93.80**	**80.52**	95.65	97.86	**90.22**
CKA	93.21	79.94	**96.30**	**98.23**	89.34
CKA (+MSP)	93.21	65.59	80.66	95.71	76.47

Table 4. Comparison of experimental performance (%).

iliary datasets, *e.g*., obtained by the proposed data augmentation technique, RFNB, and a real-world source.

The settings are the same as the previous parts. We treat the 41 classes of the plankton as normal, and a class of suspending particles and a class of bubbles as abnormal. From the DYB-PlanktonNet, we randomly choose 30 classes of the plankton as the real-world auxiliary data, which are not in the normal and abnormal set.

The experimental results are shown in Table 5. Compared with the RFNB, the real-world auxiliary dataset leads to better classification accuracy of the normal classes, since it possesses real semantic features of the *in situ* plankton images, and it has less influence on the model prediction. However, the auxiliary dataset generated by the proposed RFNB performs better for anomaly detection. One possible reason is that the real-world auxiliary dataset only contains the feature information of some plankton, thus it lacks sufficient prior knowledge of the other species.

	ACC	TNR95	AUROC	AUPR	DTACC
RFNB	93.21	**79.94**	**96.30**	**98.23**	**89.34**
Real-world	**94.18**	76.86	94.84	97.12	87.65

Table 5. Evaluation performance (%) using the RFNB generated data and using the real-world auxiliary data, respectively.

4.6. Case Study

To demonstrate the application of the anomaly detection model for practical problems, we carry out a case study of using the method for discriminating suspended particles from plankton in coastal water samples, in which the number and heterogeneity of the particles are known to be overwhelmingly larger and more complex than that of plankton [34]. If the observation interest is plankton, it will be very difficult to use just one CNN model to well achieve the classification task, as the data is so imbalanced and many particles also look very alike with certain plankton species. Li *et al*. considered a relayed usage of two classification CNNs for excluding the interference from the suspended particles and bubbles, such that they could better classify the plankton (normal) species in the end [34]. Here we design a toy dataset to compare the performance of our method with the VGG-11 model used in [34] for the same purpose. The dataset consists of 1015 suspended particles and 99 plankton ROIs that are never seen by both methods during their training phases. Specifically, the plankton ROIs are extracted from the DYB-PlanktonNet, and the particles ROIs are collected differently in space and time. So, the particle data can be regarded as shifts from those used in the training phase.

	ACC	Precision	Recall	F1 Score
VGG-11 [34]	95.42	68.15	**92.93**	78.63
Our Method	**96.08**	**72.73**	88.89	**80.00**

Table 6. Classification performance (%) of the case study.

The result shown in Table 6 proves that both methods achieve comparable performance in the task of excluding particles from the plankton, and our approach is a little bit superior. Compared to using two cascaded CNNs, our model is more integrated when applied for finer plankton taxa classification. It can achieve similar performance to a conventional classifier with a relatively small augmented dataset, so the requirements for number and diversity of the training data are much lower. Limited by the real-world data availability, the dataset shift in this case is still not obvious enough. We believe that our method will have better performance on data with more serious shift in practice, *e.g*., better generalizability.

5. Conclusion

In this paper, we propose an anomaly detection pipeline as well as concrete methods for the execution of this pipeline for the open-set problem in marine plankton image classification task. Especially, we discuss different data augmentation techniques to generate auxiliary datasets, and propose a CKA loss function in the pre-training phase, which achieves satisfactory performance for detecting anomalies and preserves high classification accuracy of the normal in the experiments. We conduct experiments on the DYB-PlanktonNet dataset and show the usefulness of our approach achieving state-of-the-art performance, and we also emphasize the importance of the choices of the post-training models and the auxiliary dataset. Additionally, we conduct the case study to show the comparable performance of our method in the testing phase. The proposed open-set classification methods are expected to help marine biologists to better identify their observation targets of interest, so that the *in situ* monitoring of marine plankton could become more convenient and efficient.

In the future, we will investigate to improve the robustness of our approach for both classification and anomaly detection, and apply it on more challenging applications.

6. Acknowledgements

This work was supported in part by the Scientific Instrument Development Project of the Chinese Academy of Sciences (Grant No. YJKYYQ20190028) and the Shenzhen Science and Technology Innovation Program (JCYJ20200109105823170).

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv:1606.06565*, 2016. 1

[2] Yuval Bahat and Gregory Shakhnarovich. Confidence from invariance to image transformations. *arXiv:1804.00657*, 2018. 2

[3] Matthew B. Blaschko, Gary Holness, Marwan A. Mattar, D. Lisin, P. Utgoff, A. Hanson, H. Schultz, E. Riseman, M. Sieracki, W. Balch, and B. Tupper. Automatic in situ identification of plankton. In *2005 Seventh IEEE Workshops on Applications of Computer Vision*, volume 1, pages 79–86, 2005. 1, 4

[4] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2):125, 2020. 4

[5] Ken Chatfield, Victor S Lempitsky, Andrea Vedaldi, and Andrew Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, volume 2, page 8, 2011. 1

[6] Wenhu Chen, Yilin Shen, Xin Eric Wang, and William Wang. Enhancing the robustness of prior network in out-of-distribution detection. *arXiv:1811.07308*, 2018. 2

[7] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. In *Computer Vision and Pattern Recognition*, 2019. 2

[8] Jialun Dai, Ruchen Wang, Haiyong Zheng, Guangrong Ji, and Xiaoyan Qiao. Zooplanktonet: Deep convolutional network for zooplankton classification. In *OCEANS 2016 - Shanghai*, pages 1–6, 2016. 1

[9] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000. 4

[10] Taylor Denouden, Rick Salay, Krzysztof Czarnecki, Vahdat Abdelzad, Buu Phan, and Sachin Vernekar. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *arXiv:1812.02765*, 2018. 2

[11] Jie Ding, Vahid Tarokh, and Yuhong Yang. Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35:16–34, 2018. 2

[12] Cassio Elias Dos Santos and William Robson Schwartz. Extending face identification to open-set face recognition. In *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 188–195, 2014. 1, 2

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 1

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 1

[15] Pablo González, E. Álvarez, J. Díez, A. Lopez-Urrutia, and J. J. Coz. Validation methods for plankton image classification systems. *Limnology and Oceanography: Methods*, 15:221–237, 2017. 2

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014. 2

[17] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 3

[18] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, volume 70, pages 1321–1330, 2017. 7

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[20] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 2, 4

[21] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. 2, 3, 4, 6

[22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2261–2269, 2017. 4

[23] Pekka Hurskainen, Hari Adhikari, Mika Siljander, PKE Pellikka, and Andreas Hemp. Auxiliary datasets improve accuracy of object-based land use/land cover classification in heterogeneous savanna landscapes. *Remote Sensing of Environment*, 233:111354, 2019. 3

[24] Lalit P. Jain, Walter J. Scheirer, and Terrance E. Boult. Multi-class open set recognition using probability of inclusion. In *European Conference on Computer Vision*, pages 393–409. Springer, 2014. 1, 2

[25] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2011. 1

[26] Heinrich Jiang, Been Kim, Melody Y Guan, and Maya Gupta. To trust or not to trust a classifier. In *32nd Conference on Neural Information Processing Systems*, 2018. 2

[27] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894, 2017. 2

[28] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. In *Techncal Report. University of Toronto*, 2009. 4

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *The 25th International Conference on Neural*

Information Processing Systems, volume 25, pages 1097–1105, 2012. 1

[30] Wallace Lawson, Esube Bekele, and Keith Sullivan. Finding anomalies with generative adversarial networks for a patrolbot. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 12–13, 2017. 2

[31] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018. 2, 3

[32] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *International Conference on Neural Information Processing Systems*, pages 7167–7177, 2018. 2, 4

[33] Joseph Lemley, Shabab Bazrafkan, and Peter Corcoran. Smart augmentation learning an optimal data augmentation strategy. *IEEE Access*, 5:5858–5869, 2017. 2

[34] Jiangping Li, Tao Chen, Yangzhen Yu, Liangpei Chen, Peng Liu, Yizhou Zhang, Guangwen Yu, Jixin Chen, Haitao Li, , and Xiaohong Sun. Development of a buoy-borne underwater imaging system for in situ mesoplankton monitoring of coastal waters (accepted). *IEEE Journal of Oceanic Engineering*, 2021. 1, 2, 5, 8

[35] Jianping Li, Zhenyu Yang, and Tao Chen. Dybplanktonnet. *IEEE Dataport*, 2021. doi: https://dx.doi.org/10.21227/875n-f104. 2, 5

[36] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 2

[37] Si Liu, Risheek Garrepalli, Thomas Dietterich, Alan Fern, and Dan Hendrycks. Open category detection with pac guarantees. In *International Conference on Machine Learning*, pages 3169–3178, 2018. 2

[38] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. *arXiv:2010.03759*, 2020. 4

[39] Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E. Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*, 2018. 6

[40] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision*, pages 181–196, 2018. 3

[41] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2624–2637, 2013. 1

[42] Dimity Miller, Niko Sunderhauf, Michael Milford, and Feras Dayoub. Class anchor clustering: A loss for distance-based open set recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3570–3578, 2021. 2, 4

[43] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *International Conference on Neural Information Processing Systems Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. 2, 4

[44] Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing*, 441:138–150, 2021. 2, 3, 4

[45] Vito P. Pastore, Thomas G. Zimmerman, Sujoy K. Biswas, and Simone Bianco. Annotation-free learning of plankton for classification and anomaly detection. *Scientific Reports*, 10(1):1–15, 2020. 3

[46] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *International Conference on Learning Representations*, 2017. 5

[47] Le'an Qu, Zhenjie Chen, Manchun Li, Junjun Zhi, and Huiming Wang. Accuracy improvements to pixel-based and object-based lulc classification with auxiliary datasets from google earth engine. *Remote Sensing*, 13(3):453, 2021. 3

[48] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. *arXiv:1704.01444*, 2017. 3

[49] Alexander J. Ratner, Henry R. Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. Learning to compose domain-specific transformations for data augmentation. In *31st Conference on Neural Information Processing Systems*, volume 30, page 3239. NIH Public Access, 2017. 2

[50] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A DePristo, Joshua V Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Conference on Neural Information Processing Systems*, 2019. 2

[51] Daniel L Roelke and Sofie Spatharis. Phytoplankton succession in recurrently fluctuating environments. *PLoS One*, 10(3):e0121392, 2015. 1

[52] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with in-distribution examples and gram matrices. In *International Conference on Machine Learning*, pages 8491–8501, 2019. 2

[53] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2012. 1, 2

[54] Walter J. Scheirer, Lalit P. Jain, and Terrance E. Boult. Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2317–2324, 2014. 1

[55] Matthew D. Scherreik and Brian D. Rigling. Open set recognition for automatic target classification with rejection. *IEEE Transactions on Aerospace and Electronic Systems*, 52(2):632–642, 2016. 2

[56] Peter Schulam and Suchi Saria. Can you trust this prediction? auditing pointwise reliability after learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1022–1031, 2019. 2

[57] Katja Schulze, Ulrich M. Tillich, Thomas Dandekar, and Marcus Frohme. Planktovision-an automated analysis system for the identification of phytoplankton. *BMC Bioinformatics*, 14(1):1–10, 2013. 1

[58] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations*, 2014. 1

[59] Heidi M. Sosik and Robert J. Olson. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnology and Oceanography: Methods*, 5(6):204–216, 2007. 1

[60] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9518–9526, 2019. 2

[61] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv:1605.07146*, 2016. 4

[62] Haiyong Zheng, Ruchen Wang, Zhibin Yu, Nan Wang, Zhaorui Gu, and Bing Zheng. Automatic plankton image classification combining multiple view features via multiple kernel learning. *BMC Bioinformatics*, 18(16):1–18, 2017. 1

[63] Thomas G. Zimmerman, Vito P. Pastore, Sujoy K. Biswas, and Simone Bianco. Embedded system to detect, track and classify plankton using a lensless video microscope. *arXiv:2005.13064*, 2020. 2