Please cite this paper as X. Sun, H. Xv, J. Dong, H. Zhou, C. Chen, and Q. Li, "Few-shot Learning for Domain-specific Fine-grained Image Classification," IEEE Transactions on Industrial Electronics, pp. 1-1, 2020.

Few-shot Learning for Domain-specific Fine-grained Image Classification

Xin Sun, *Member, IEEE*, Hongwei Xv, Junyu Dong, *Member, IEEE*,
Huiyu Zhou, *Member, IEEE*, Changrui Chen, and Qiong Li

***Abstract*—Learning to recognize novel visual categories from a few examples is a challenging task for machines in real-world industrial applications. In contrast, humans have the ability to discriminate even similar objects with little supervision. This paper attempts to address the few-shot fine-grained image classification problem. We propose a feature fusion model to explore discriminative features by focusing on key regions. The model utilizes the focus-area location mechanism to discover the perceptually similar regions among objects. High-order integration is employed to capture the interaction information among intra-parts. We also design a Center Neighbor Loss to form robust embedding space distributions. Furthermore, we build a typical fine-grained and few-shot learning dataset *mini*PPlankton from the real-world application in the area of marine ecological environments. Extensive experiments are carried out to validate the performance of our method. The results demonstrate that our model achieves competitive performance compared with state-of-the-art models. Our work is a valuable complement to the model domain-specific industrial applications.**

***Index Terms*—Computer vision; Few-shot learning; Representation learning**

I. Introduction

IN RECENT YEARS, we have witnessed significant progress in computer vision [1], [2]. Thanks to large-scale of labeled training data, e.g., ImageNet, deep convolutional neural networks (ConvNets) are able to successfully learn robust feature representations and achieve excellent performance in recognition tasks. Although it has high accuracy in various labeled datasets, the generalization ability of the ConvNet model is still weak. In particular, the ConvNet model is difficult to quickly identify a novel category using only one or a few labeled samples. However, humans are able to recognize new objects easily with very little supervision [3]. For example, kids have no problem to generalize the concept of "panda" from only one picture. Furthermore, experts will be faster to understand novel concepts with prior professional knowledge. This work focuses on the task that recognizing novel visual categories after seeing just a few labeled examples. Research on this subject is often termed *few-shot learning*.

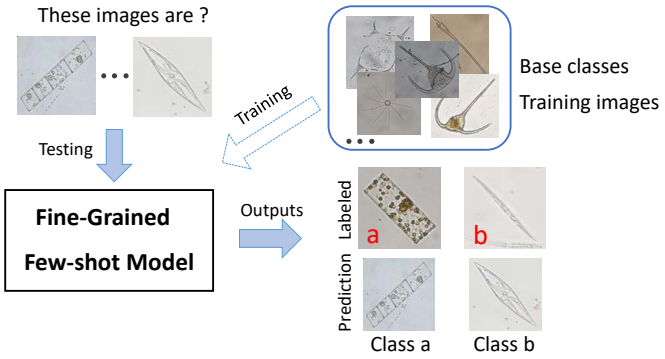


Fig. 1: A brief illustration of fine-grained few-shot recognition.

In contrast to the common image classification problem in daily life, most of the real-world scenarios face few-shot problems. For example, marine biologist pays great attention to the phytoplankton recognition problem which is a typical fine-grained and few-shot learning issue. The change of their abundance, e.g. eutrophication, is a significant indicator of the oceanic ecosystem's health. It is therefore very important to automatically identify phytoplankton in a certain area of the ocean. However, collections of phytoplankton images are very difficult. It is commonly accomplished by professional instruments such as electron microscope. Only a few samples of valuable categories can be discovered in one expensive sampling task. Therefore, the fine-grained and few-shot model is critical for domain-specific issues and has become one of the important topics in computer vision.

Most of the few-shot learning methods fall under the umbrella of metric-learning. The metric-learning approaches try to solve these problems by placing new classes in a metric space (e.g., Euclidean or cosine distances) that can easily separate classes. For instance, Matching Networks [4] can be interpreted as a nearest-neighbor classifier which can be trained end-to-end over the cosine distance. Notably, the

Manuscript received Month xx, 2xxx; revised Month xx, xxxx; accepted Month x, xxxx. This work was supported in part by the National Natural Science Foundation of China (No. U1706218, 61971388, L1824025), Key Research and Development Program of Shandong Province (No. GG201703140154), and Major Program of Natural Science Foundation of Shandong Province (No. ZR2018ZB0852). H. Zhou was supported by Royal Society-Newton Advanced Fellowship under Grant NA160342, and European Union's Horizon 2020 research and innovation program under the Marie-Sklodowska-Curie grant agreement No 720325.

X. Sun, H. Xv, J. Dong, C. Chen, Q. Li are with the Department of Computer Science and Technology, Ocean University of China, Qingdao, PR.China. (e-mail: sunxin1984@ieee.org, dongjunyu@ouc.edu.cn).

H. Zhou is with the School of Informatics, University of Leicester, UK. (e-mail: hz143@leicester.ac.uk)

training procedure has to be chosen carefully so as to match inference at the test stage. Each episode is designed to mimic the few-shot tasks by subsampling classes as well as data points (e.g., every episode sampling 5 classes and each class has 5 labeled samples). Prototypical Networks [5] handles the few-shot tasks by calculating the Euclidean distance between the embedding points of query set and prototype representation of support set. Meanwhile, the pre-defined metric is no longer used in Relation Networks [6]. It uses concatenated feature maps from the query and support images to distinguish similar and dissimilar samples.

It is very important to explore the relationship between feature representation of template images and that of the query image. Thus, to succeed in few-shot metric tasks, we shall make sure two aspects. First, we shall have a well-trained feature extractor. The other is an effective classifier including good metrics. However, the above-mentioned methods are not conducive to ConvNets for extracting robust features and can sacrifice the accuracy of initial categories [7]. Most of the few-shot methods pay attention to learning a deep distance metric to compare query images with the labeled images, while ignoring the importance of mining the better features from the existing few categories. That means it is critical to mining rich information from the labeled samples of few categories. Motivated by the above observation, we propose a Feature Fusion Model for obtaining more discriminative information from focus areas. We also design a loss function (Center Neighbor Loss) to help the whole architecture to learn better feature space distributions.

For the special fine-grained few-shot visual problem, we further build a microimage dataset of phytoplankton, i.e., ***mini*PPlankton**. Unlike toy datasets for few-shot learning in literature, the ***mini*PPlankton** dataset comes from the real-world tasks and can be used to evaluate fine-grained and few-shot methods. It illustrates a typical fine-grained and few-shot problem in marine biological science.

The main contributions of this paper are as follows:

1) We propose a feature fusion model to explore the features by focusing on the key regions. It utilizes the focus-area location mechanism to discover the similarity regions between objects. Meanwhile, high-order integration is used to capture the intra-parts discriminative information.
2) We design a Center Neighbor Loss function to form robust feature space distributions for generating discriminative features, to accomplish the fine-grained few-shot visual categorization task.
3) We build a domain-specific fine-grained and few-shot dataset ***mini*PPlankton** for the real-world phytoplankton recognition problem. Experiments on the ***mini*PPlankton** show the superiority of the proposed model compared with other models.

The rest of this paper is organized as follows. Section II summarizes the related works. Section III formally describes our model. Section IV presents the experimental results. Finally, we conclude in Section V.

II. Related work

Deep convolutional neural networks have made significant achievements for a wide range of visual tasks [8]–[10]. Nevertheless, for fine-grained image categorization [11], it remains quite challenging to obtain the discriminative representations. In particular, it is a novel challenge to classify fine-grained images using only a few labeled sample images . The convolutional neural networks usually require thousands of labeled examples of each class to saturate performance. However, it is impractical to collect large amounts of annotated data, especially the domain-specific industrial applications that requires expert knowledge, such as oceanography [12], [13]. Recently, there is a resurgence of interest on few-shot learning [4]–[6]. And a few research works are already pay attention to the fine-grained few-shot visual problem [14]–[17].

Among the recent literature of few-shot learning, the metric learning and attention mechanism are most relevant proposed method. Metric learning has been successfully applied to face recognition [18] and fine-grained image classification [11]. The core idea is to learn an embedding function that the samples of the same category are closer than those of different classes. Once the embedding function is learned, the query images will be classified. Siamese network [19] consists of two identical sub-ConvNets that minimize the distances between paired data with the same labels while keeping the distances with different labels far apart. Triplet loss [20] attempts to focus on relative distances rather than absolute pair-wise distances. It has been widely implemented in fine-grained tasks [21]. However, the problem of triplet loss is dramatic data expansion when selecting triplets. Furthermore, center loss [22] can obtain highly discriminative features for robust face recognition. And it is unnecessary to design the sampling strategy carefully as contrastive loss and triplet loss do. The center loss has shown benefits in face identification. However, its performance is unknown for the fine-grained few-shot tasks. Then we further design a Center Neighbor Loss for achieving a robust embedding space.

It is critical to know which part of the images worth paying attention to. To acquire the attention feature representation, Li et al. [23] proposed a zoom network which utilized the candidate region to crop the original images. Wei et al. [24] adopted the unsupervised object discovery and co-localization mechanism by deep descriptor transformation to discover the attention area. The attention mechanism is a possible way for learning robust representation. In this work, we introduce the focus-area location mechanism Grad-CAM [25] to find regions with discriminative features, which are critical for fine-grained classification.

Few-shot learning is critical in model industrial applications, such as novel species discovering. In this work, we take one typical real-world industrial problem to verify our method, i.e., phytoplankton classification. Marine phytoplankton is the foundation of the marine ecosystem [26]. It is an ecological concept that refers to tiny plants that float in the water. Plankton image classification [1] is becoming critically important for

[1]We no longer distinguish the image classification of phytoplankton and zooplankton separately.

marine observations and aquaculture.

The research of phytoplankton detection mainly relies on people to manually identify and count through the microscope. Current monitoring systems (e.g. ZooScan and FlowCAM [12], [13]) yield large amounts of images every day. They are usually time-consuming, labor-intensive and needs strong professional knowledge. Schroder et al. [27] also notice the importance of classifying plankton only using a few labeled samples. They directly use weight Imprinting [28] to enable a neural network to recognize small classes immediately without re-training.

III. METHODOLOGY

A. Notation

For few-shot classification, there is a base train dataset $\mathcal{D}_{base} = \{(x_i, y_i)\}_{i=1}^{N}$ consisting of N labeled images, where y_i is the label of image x_i. Crucially, the model must distinguish a set of novel categories $\mathcal{Q} = \{(x_j, y_j)\}_{j=1}^{N_q}$ with a few training examples per category. These training examples are called support set, i.e., $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{N_s} (N_s = K * C)$ which contains K labeled examples for each of C unique novel classes. $\mathcal{Q}$ acts as the unlabeled query set. Here $\mathcal{S} \bigcup \mathcal{Q} = \mathcal{D}_{novel}$ and $\mathcal{D}_{base} \bigcap \mathcal{D}_{novel} = \emptyset$. This target few-shot task is named C-way K-shot.

B. Model

An overview of our method is illustrated in Fig. 2, which mainly consists of three parts.

1) ConvNet-based feature extractor: A feature extractor f_φ, which parameterized by a ConvNet (e.g., ResNet [29]), maps an input image $x \in \mathbb{R}^N$ to a d-dimensional feature vector $f_\varphi(x) \in \mathbb{R}^d$. As a classification model, f_φ has a dot-product based classifier $C(.|W)$ (i.e., Last linear layer), where $W = \{w_i \in \mathbb{R}^d\}_{i=1}^{K}$ is the set of weight vectors of the K base classes. We can get the probability scores of the base training categories by calculating $C(f_\varphi(x)|W)$ and optimize the feature extractor by back-propagation.

2) Feature fusion module: For few-shot learning, it is pivotal to mine the largest support information from the support set $\mathcal{S}$. We propose a feature fusion model which utilizes the focus-area location and high-order integration to generate feature representation for the few-shot tasks. As shown in Fig. 2, it consists of two components: (1) high-order integration, and (2) focus-area location.

High-order integration. The recent progress of fine-grained classification demonstrates that the high-order representations with ConvNets can greatly improve its performance [30], [31]. Intuitively, the key for fine-grained few-shot tasks is to represent the regions within same category that have a closer appearance and to exhibit discriminative areas between the different categories.

We assume that $\mathcal{X} \in \mathbb{R}^{K \times M \times N}$ is a $3D$ feature map from the convolutional layers, where $x \in \mathcal{X}$ is a K-dimensional descriptor of one particular location region $p \in M \times N$. The linear predictor $\mathcal{W}$ on the high-order statistics of $\mathcal{X}$ could be formulated as follow.

$$f(\mathcal{X}) =< \mathcal{W}, \sum_{x \in \mathcal{X}} \phi(x) > \tag{1}$$

where $\sum_{x \in \mathcal{X}} \phi(x)$ denotes the high-order statistics characterized by a homogenous polynomial kernel [32]. The $\mathcal{W}$ can be approximated by rank-one decomposition. The tensor rank decomposition expresses a tensor as a minimum-length linear combination of rank-1 tensors. The outer product of vectors $\mathbf{u}_1 \in \mathbb{R}^{K_1}, \ldots, \mathbf{u}_r \in \mathbb{R}^{K_r}$ is the $K_1 \times \ldots \times K_r$ rank-1 tensor that satisfies $(\mathbf{u}_1 \otimes \ldots \otimes \mathbf{u}_r)_{k_1 \ldots, k_r} = (\mathbf{u}_1)_{k_1} \ldots (\mathbf{u}_r)_{k_r}$. The $\mathcal{W}$ can be rewritten as $\mathcal{W} = \sum_{d=1}^{D} a^d \mathbf{u}_1^d \otimes \ldots \otimes \mathbf{u}_r^d$, where a^d is the weight for d-th rank-one tensor and D is the rank of the tensor if D is minimal. Thus, Equation 1 can be reformulated as follow.

$$\begin{aligned} f(\mathcal{X}) &= \sum_{\mathrm{x} \in \mathcal{X}} \left\{ \langle \mathbf{w}^1, \mathbf{x} \rangle + \sum_{r=2}^{R} \sum_{d=1}^{D^r} a^{r,d} \prod_{s=1}^{r} \langle \mathbf{u}_s^{r,d}, \mathrm{x} \rangle \right\}, \\ &= \left\langle \mathbf{w}^1, \sum_{\mathrm{x} \in \mathcal{X}} \mathbf{x} \right\rangle + \sum_{r=2}^{R} \left\langle \boldsymbol{a}^r, \sum_{z^r \in \mathcal{Z}^r} \boldsymbol{z}^r \right\rangle \end{aligned} \tag{2}$$

where the $\boldsymbol{z}^r = [z^{r,1}, \ldots, z^{r,D^r}]^\top$ with $z^{r,d} = \prod_{s=1}^{r} \langle \mathbf{u}_s^{r,d}, \mathbf{x} \rangle$ characterizes the degree-r variable interactions under a single rank-1 tensors, and $\boldsymbol{a}^r$ is the weight vector. The $\boldsymbol{z}^r$ can be calculated by performing r-th 1×1 convolutions with D^r channel [33], i.e., $\mathcal{Z}^r = \{\boldsymbol{z}^r\} = \prod_{i=1}^{r} conv_{1 \times 1 \times D^r}^i(\mathcal{X})$. In our feature fusion operation as shown in Fig. 2, we integrate 2nd-order representations to capture more complex and high-order relationships among parts. After that, we perform global average pooling (GAP) [34] to further aggregate features.

Focus-area location. Existing studies show that learning from object regions could benefit object recognition at image-level [23]. Such focus-area in an image which benefit few-shot learning. During the training procedure, f_φ can generate focus-areas of images by Grad-CAM [25], as formulated below.

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k) \tag{3}$$

where α_k^c denotes the weight of the k-th feature map for category c. α_k^c can be calculated by the following formula.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{4}$$

where Z is the number of pixels in feature map, y^c is the classification score corresponding to the category c, and A_{ij}^k denotes the pixel value at the location of (i, j) of the k-th feature map.

Grad-CAM has the ability to locate the focus areas that belong to the corresponding category. As shown in Fig. 2, the dot line is a diagram of Grad-CAM, which represents the focus-area is obtained by weighted summing the feature maps. In this work, we utilize Grad-CAM to generate base categories' focus regions $\mathcal{H}_{base} = \{(x_i^h, y_i)\}_{i=1}^{N}$. However, the ConvNet extractor can not give a correct response of c in formula (4), when a novel category appears. To our delight, we find that the model has accumulated lots of meta-knowledge in the domain field (e.g., Ornithology) during the training process of $\mathcal{D}_{base} = \{(x_i, y_i)\}_{i=1}^{N}$. The concepts of novel categories can be made up of various meta-knowledge, which are already embedded in the neural networks. For example, if someone

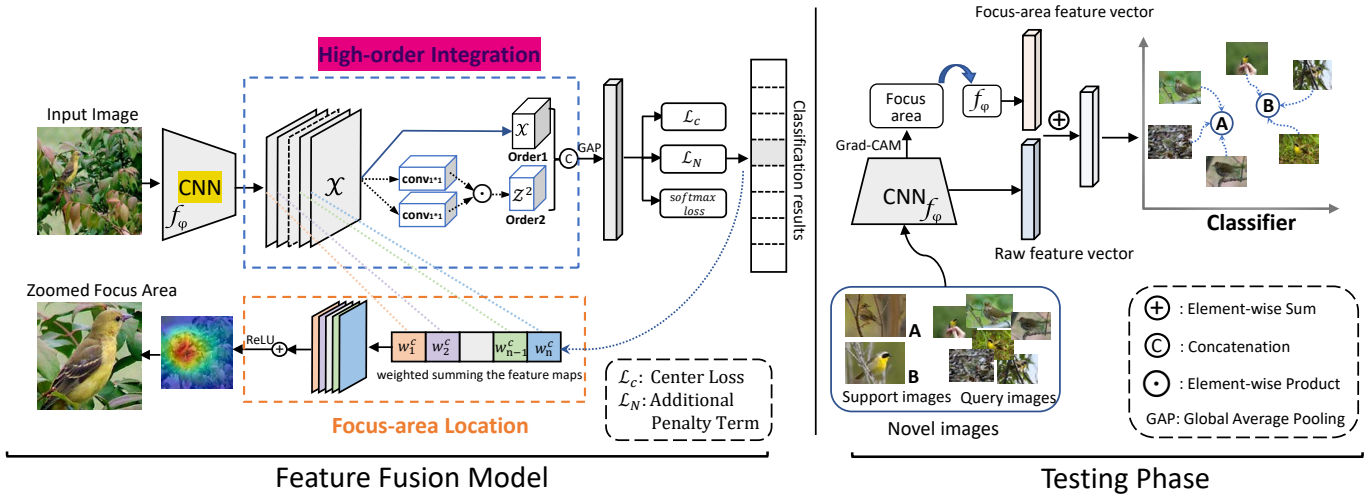


Fig. 2: The overview framework of our method. It consists of a *ConvNet-based feature extractor* f_φ , a *feature fusion model* which is formed by focus-area location mechanism and high-order integration , and a *cosine-similarity based classifier*. During the testing process, we classify unlabeled samples by comparing the cosine similarities of support set $\mathcal{S}$ and query set $\mathcal{Q}$.

has never seen the tiger, she/he might think it has many close parallels to a cat (learned before). The reason is that the attention locations of human on the new category tiger and the known category cat are similar to each other. Although we don't know the ground truth of the novel samples for the fine-grained few-shot tasks, the unseen class always has similar regions to the $\mathcal{D}_{base}$, such as bird's mouths and wings. And the base classifier will classify the new sample into the most similar class in $\mathcal{D}_{base}$. Therefore, it is possible to utilize Grad-CAM tp generate good focus-area location $\mathcal{H}_{novel}$ on the unseen categories for enhancing feature representation.

Telling the neural network the regions of rich discriminative information will form a more robust representation. This step is similar to the data augmentation of input space. However we only mine the available information on the input data itself without using the extra data augmentation.

3) Classifiers: Generally, the ConvNet's classifier uses the dot-product operator to compute classification scores: $s = z^T * w_k^b$, where z is the feature vector extracted by ConvNets and w_k^b is the k-th classification weight vector in W_{base}. It is trained from scratch by thousands of optimization steps (e.g., SGD). In contrast, the W_{base} is not adapted to the new categories and it is difficult to find the proper classification weights W_{novel} with only a few samples and optimization steps. To address this critical problem, a classifier should be implemented to distinguish the new categories. To the best of our knowledge, current researches commonly choose one of the following classifiers to gain their best performance, i.e., SVM [35], cosine-similarity [28] and nearest neighbor.

SVM. SVM classifier has achieved excellent performance for small training data in few-shot learning [35]. Essentially, unlike deep learning methods which need large-scale training data to learn generalization ability within classes, SVM is a classical transductive inference method aiming to build a model that is applicable to the problem domain.

Nearest neighbor. The Euclidean-based nearest neighbor method uses feature vector z_s to build a prototype representation of each novel class for the few-shot learning scenario. Then it classifies the unlabeled data by calculating the distance from each query embedding point to the prototype.

Cosine classifier. The cosine classifier has been well established as an effective similarity function for few-shot tasks [4], which classifies samples by comparing the cosine similarity between z_s and z_q.

C. Objective function

The loss function is important to let the neural network generate separable representations for the unseen classes. For example, Siamese Nets [36] applies contractive loss to few-shot tasks, so that neural networks can learn to distinguish similarities from dissimilarities. For fine-grained few-shot tasks, it is critical to develop an effective similarity constraint function to improve the discriminative power of the feature representations. Center loss [22], which was first proposed for the face recognition problem, simultaneously learns a center for deep features of each class and penalizes the distances between the deep features and their corresponding class centers. Suppose there are K classes for samples, k^i is the category of the image x_i and $z_i = f_\varphi(x_i)$ denotes the deep features extracted from x_i. Here is the formulation for center loss:

$$\mathcal{L}_c = \frac{1}{2}\sum_{i}^{n} \|z_i - c_{k^i}\|_2^2. \tag{5}$$

The $c_k \in \mathbb{R}^K$ denotes the k-th class center of deep features. The formulation effectively characterizes the intra-class variations. However, all training samples are treated equally when a center loss function minimizes the intra-class variations, regardless of whether the sample is easy or hard to pull into the center point. Intuitively, for the fine-grained tasks, the difference among classes is extremely small. It is not enough to form a good distribution by simply pulling the feature vector

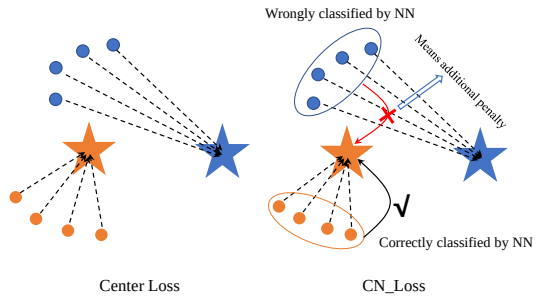


Fig. 3: The center loss is simply pulling the samples into the class-center (entagram). While CN_loss adds additional penalties to the sample of the wrong classification (red fork symbol) using nearest-neighbor.

into the class center. It is critical to impose special penalties on the samples which are difficult to approach the center during the training process. To this end, we further propose the Center Neighbor Loss (CN_Loss) function $\mathcal{L}_s$ to form robust embedding space distribution as following.

$$\mathcal{L}_s = \mathcal{L}_c + \beta \cdot \mathcal{L}_N \tag{6}$$

β is the balance parameter for penalty term $\mathcal{L}_N$. $\mathcal{L}_N$ is a negative log-probability for samples that are not classified to the correct class center. The $\mathcal{L}_N$ can be formulated as following.

$$\mathcal{L}_N = -\log \frac{\exp(-E(\bar{z^k}, \mathbf{c}_k))}{\sum_{k' \in K} \exp(-E(\bar{z^{k'}}, \mathbf{c}_{k'}))} \tag{7}$$

$\bar{z^k} = Avg(\sum_{x_i^k \in \mathcal{D}_{base}} f_\varphi(x_i^k))$ is the k-th class average feature vector contained in every batch, and $E(\cdot,\cdot)$ denotes the $Euclidean$ distance.

The schematic is shown in Fig. 3. We take the center points learned from the last iteration as support points and use Euclidean-based nearest neighbors to classify the current batch of samples. With the penalty $\mathcal{L}_N$, each cluster will gather faster and perform robustly.

Ideally, the class center c_k should be updated as feature vectors change. That means we should take the entire training set into account and average the deep features of each class in each iteration, which is not feasible in practice. To solve this problem, we implement the solution suggested for center loss [22]. First of all, we perform the update procedure based on mini-batch. The centers are computed by averaging the features of every category in each iteration. Secondly, we use the centers learned from the last iteration to classify the current batch samples by Nearest Neighbor Algorithm and punish the mislabeled samples. At last, we fix the learning rate of the centers as 0.5 to avoid large perturbations caused by mislabeled samples [22].

IV. Experiments

A. Experimental design

For rare categories, it's extremely difficult to collect sufficient and diverse training images. Currently, most of the previous few-shot learning methods take the *mini*ImageNet dataset [4] to test their performance with 5-way 1-shot or 5-way 5-shot assumptions. However, the *mini*ImageNet consists of 60,000 color images with 100 classes of which 64 classes for training. The training data is enough to learn a good feature extractor for a common few-shot classification task, and nearly 80% accuracy has been already achieved recently [35]. In this paper, we focus on the fine-grained few-shot classification tasks. To this end, we design three different experiments on Caltech-UCSD Birds [37] datasets, *mini*DogsNet [38] and miniPPlankton.

For the *mini*DogsNet dataset [38], we only use **10 classes** for training, and conduct 5-way experiments with both 1-shot and 5-shot settings. We will compare our method with other well known techniques [4], [5], [6], [39], [40]. All methods are also training on these 10 classes. In order to ensure the fairness of comparison, we unify the MatchingNets [4], PrototypicalNets [5] and Imprint [28]' feature extractor to ResNet. As the meta-learning training strategy of the Relation Networks [6] and MAML [39] is difficult to be trained via deep ConvNets, we keep their original network architecture.

In real-world scenarios, humans face a large number of novel categories to be recognized. 5-way experiments only for toy examples in papers. Currently, one state-of-the-art research work Imprint [28] implemented the Caltech-UCSD Birds dataset [37] for 100-way few-shot learning problem, which is much practical. Here we will carry out experiments with the same setting of Imprint [28]. That means we investigate the accuracy on all the novel classes. As the above-mentioned methods including RelationNets [6] are designed for only 5-way experiments, it is difficult to accomplish the 100-way procedure. For example, RelationNets [6] requires huge GPU memory spaces for the 100-way training. Therefore, we only set the recent work Imprinted Weights [28] as the comparison. For few-shot tasks, the Imprinted Weights [28] described how to add a similar capability to ConvNet classifier by directly setting the weights of the final layer from novel labeled samples. Essentially, the core of Imprinted Weights method is cosine similar function. Therefore, in the following experiments, the baseline (ResNet + cosine classifier) here is the same as the Imprint.

In a real-world scenario application, for *mini*PPlankton, we will compare our method with MatchingNets, Prototypical-Nets, Relation Net, MAML and Imprint. All of above the methods are re-implemented with ResNet as the backbone feature extractor.

B. Implementation details

ResNet18 [29] is employed as the feature extractor f_φ. Following the similar strategy of Wen [22], we train the feature extractor with the joint supervision of softmax loss and CN_loss. We initialize the learning rate of the softmax loss as 0.001 and half it every 20 epochs. And we only use the last feature map as the input of high-order integration. During the testing phase, for *mini*DogsNet and Caltech-UCSD Birds, raw support image and zoomed focus-area are uniformly resizing into 224*224 and be sent to f_φ to form a robust feature

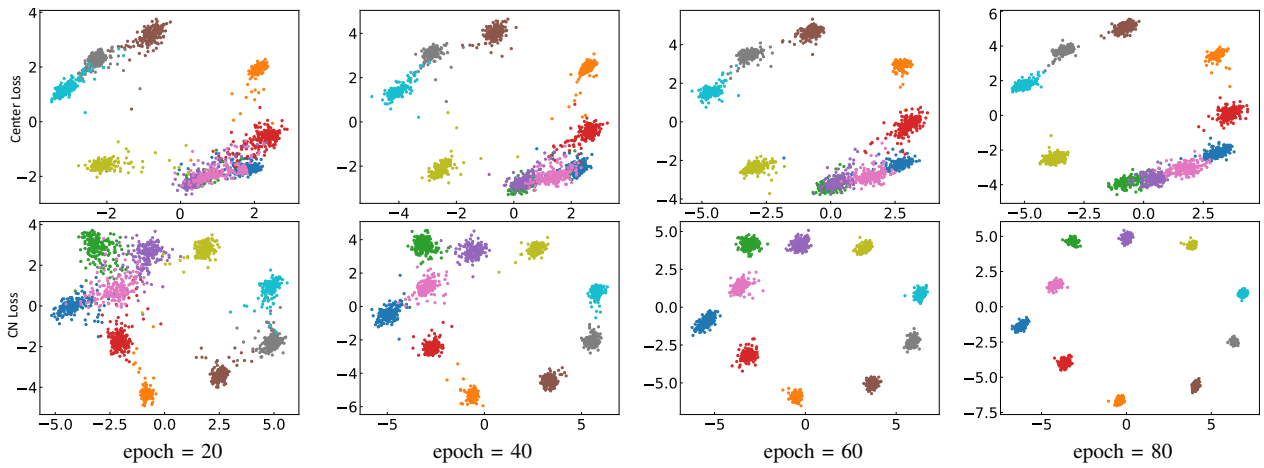


Fig. 4: The distribution of deeply learned features under Center loss and CN_loss. Different colors denote different classes.

vector using element-sum operation. For *mini*PPlankton, due to the specificity of the phytoplankton image, e.g., the target is scattered shape. Therefore, slightly differing from the structure Fig. 2, we do not use the backbone network to extract the focus-areas' feature. Here we resize focus-area into 84*84 to train a shallow CNN (four convolution blocks). Through the shallow CNN, the testing focus-area's feature will be concatenated with the original image's feature.

C. Configuration variants

CN_Loss. Fashion-MNIST is commonly used to evaluate the the loss function [22], [41]. We conduct similar experiments as suggested [22], [41] to visualize the performance of Center loss and CN_loss on Fashion-MNIST. Fashion-MNIST consists of 60,000 training examples and 10,000 for testing. Each example is a 28x28 gray-scale image, associated with a label from 10 classes. The space distribution results are shown in Fig. 4. We can see that, CN_loss can quickly form the cluster of each class. A more robust feature space distribution usually means a better feature extractor. And from the Table I, the CN_loss shows better performance on classification tasks.

Loss Function	Accuracy(%)
Softmax Loss	89.5 ± 0.2
Center Loss	90.0 ± 0.2
CN Loss	**91.42 ± 0.3**

TABLE I: The general classification performance of the three loss functions on the Fashion-MNIST dataset.

In addition, the hyperparameter β in (6) is the balance for penalty term $\mathcal{L}_N$. We investigate the performance of our model with different hyperparameter β on *mini*DogsNet's validation set. As shown in Fig. 5, it is very clear that the center loss (i.e., $\beta = 0$) is not a good choice for few-shot classification problem. The best performance can be achieved in the case of $\beta \in [0.4, 0.6]$.

High-order integration. The high-order integration could help us capture more complex and high-order relationships

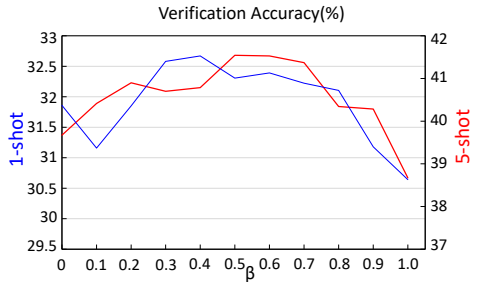


Fig. 5: The verification accuracy with different β.

among different intra-parts to get better attention maps. As shown in Fig. 6, it helps to focus on the discriminative regions of the image. We have conducted experiments with different orders on the performance of our method. And we found that 2-order performs stable on the novel classes classification. For instance, the accuracy of 2-order (49.52%) is higher than 1-order (48.56%) and 3-order (47.00%) on the CUB-200-2011 dataset for 5-way setting.

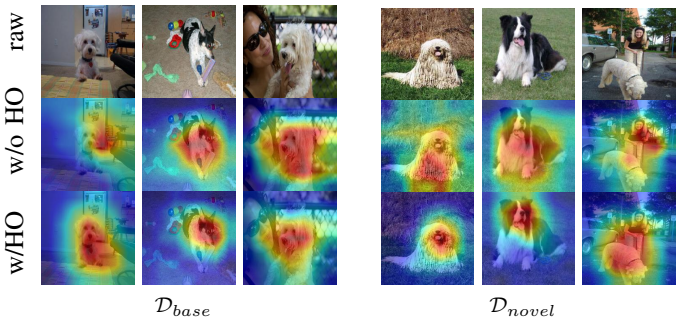


Fig. 6: Visualization results with Higher-order Integration and without it. The left three columns show the focus regions on the $\mathcal{D}_{base}$, while the right three denote focus regions of novel samples from $\mathcal{D}_{novel}$.

Focus-area location. We investigate the role of Focus-

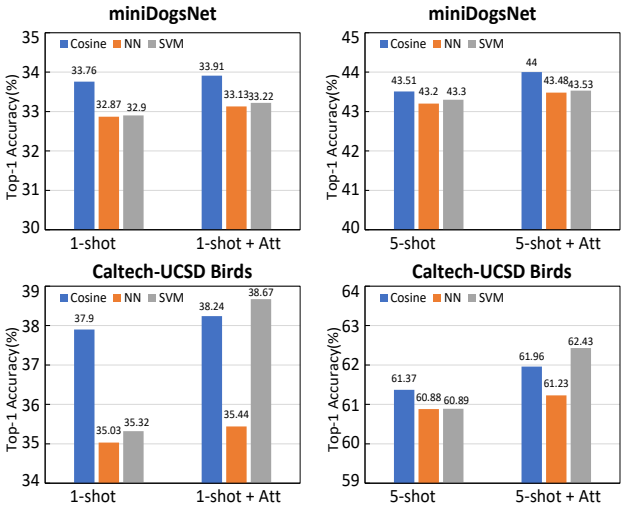


Fig. 7: The accuracy of three classifiers with and without Focus-area Location under 1 shot and 5 shot assumptions.

area Location on our fine-grained tasks. Section III-B3 briefly described the cosine classifier used in our task. We will also illustrate the performance of different classifiers on the fine-grained few-shot classification. Figure 7 shows the accuracy on different tasks with the same feature extractor setting. For the *mini*DogsNet dataset, the cosine classifier could achieve the highest accuracy on the validation set. And focus-area location achieves positive improvement. For the Caltech-UCSD Birds, both cosine classifier and SVM can achieve nice performance. Especially, we also find that the focus-area location greatly improves the accuracy of SVM classifier.

D. Caltech-UCSD Birds

The Caltech-UCSD Birds dataset [37] includes 200 fine-grained categories of birds with 11,788 images. We take the pre-trained ResNet18 [29] with ImageNet as the feature extractor f_φ. Train/test split setting is followed the suggestion of Imprinted Weights [28]. Here, 100 novel classes are required to be distinguished, which is very challenging and similar to the real-world scenario. The cosine classifier is employed to recognize the novel categories.

As shown in Table II, for all novel categories classification, we observe that the high-order module and CN_Loss function are beneficial to our tasks. In particular, the information of focus-areas brings considerable improvement in accuracy on the 5-shot setting. It is also important to illustrate the capability of recognition performance on all the categories [7]. We further evaluate the performance on dataset of $\mathcal{D}_{base} \cup \mathcal{D}_{novel}$ [28]. Table II and table III show that our model achieves promising accuracies on the novel categories while at the same time it does not sacrifice the recognition performance of the base categories $\mathcal{D}_{base}$.

N-shot	1	2	5
Imprint (ori) [28]	21.26%	28.69%	39.52%
Imprint + Aug (ori)	21.40%	30.03%	39.35%
Imprint (re)	28.77%	39.25%	49.33%
ResNet + H	30.14%	38.46%	49.83%
ResNet + CNloss	29.86%	39.45%	50.68%
ResNet + CNloss + H	30.17%	40.10%	50.78%
ResNet + CNlos s + H + Att	**30.82%**	**40.85%**	**51.95%**

TABLE II: The top-1 accuracy measured across all 100 novel classes of Caltech-UCSD Birds. 'H' denotes the High-Order Integration and 'Att' means Focus-area Location. '(ori)' means the original data in the paper and '(re)' represents the data we re-implement with ResNet as backbone.

N-shot	1	2	5
Imprint(ori) [28]	44.75%	48.21%	52.95%
Imprint + Aug (ori)	44.60%	48.48%	52.78%
Imprint(re)	44.68%	52.19%	59.27%
ResNet + H	45.72%	52.64%	59.96%
ResNet + CNloss	45.06%	51.69%	58.73%
ResNet + CNloss + H	47.23%	54.38%	60.27%
ResNet + CNloss + H + Att	**47.89%**	**54.83%**	**61.30%**

TABLE III: Top-1 accuracy measured across base plus novel categories of Caltech-UCSD Birds.

E. *mini* DogsNet

Hilliard et al. [38] created a *mini*DogsNet which consists images of dog categories from the ImageNet to test the model's fine-grained ability. They selected 100 of those classes and use the 64/16/20 random classes split for training, validation, and testing. In our work, we further increase the difficulty by random selecting 10 of 64 classes to form our training set. That means only 10 classes are used for training the feature extractor and 20 novel classes should be distinguished. And we train the ResNet18 [29] from scratch.

We conduct 5-way experiments with both 1-shot and 5-shot trials. Table IV shows that our model could achieve promising performance both on 1-shot and 5-shot tasks. To verify the effectiveness of our different modules, we use the ResNet18 and cosine classifier as the baseline. To our surprise, the baseline can also achieve nice performance. Relation Nets uses the deep non-linear metric to capture the similarity between samples and is well performed even using 10 classes training data. For our method, we can see that CN_Loss and high-order integration can bring promising improvements. And the focus-area location mechanism is still beneficial to the task.

F. *mini* PPlankton

For a real-world task in specific domain such as phytoplankton classification, it is infeasible to collect large-scale samples and it always requires experts to label the data. Meanwhile, it is also quite difficult to search for the relevant open-source web-data. Current monitoring systems (e.g. ZooScan

5way N-shot	Dist.	1	5
Matching Net [4]	Cosine	30.39%	37.97%
Prototypical Net [5]	Euclid.	31.37%	39.33%
Relation Net [6]	Deep metric	32.42%	38.53%
MAML [39]	-	26.66%	35.60%
Imprint [28]	Cosine	30.14%	38.31%
Resnet + H	Cosine	30.17%	38.77%
ResNet + CNloss	Cosine	31.25%	40.63%
ResNet + CNloss + H	Cosine	31.95%	41.40%
ResNet+CNloss+H+Att	Cosine	**33.13%**	**42.53%**

TABLE IV: The top-1 accuracy on the test set of *mini*DogsNet, all accuracy results are averaged over 100 test episodes and each episode contains 100 query samples from 5 classes. All results are reported with 95% confidence intervals.

and FlowCAM [12], [13]) yield large amounts of images every day. It requires many marine biologists to manual classify the sample images. Nevertheless, new and scarce categories are valuable for marine science. **PPlankton** is a large-scale public dataset for machine learning with the help of marine biologists [42]. And for few-shot tasks, we further construct a phytoplankton dataset ***mini*PPlankton**. It is a particular image dataset for few-shot fine-grained classification problem.

Some examples of the dataset are shown in Fig. 8. To construct the dataset, we collect seawater samples from the Bohai Sea, and we photograph phytoplankton images contained in the sampled seawater by optical microscopes. With the help of marine biologists, we label each object with its confident catergory. The *mini*PPlankton includes 20 classes each of which contains about 70 samples. From Fig. 8, we can observe that our dataset faces the challenge problem of fine-grained classification. For example, their shapes between different categories are similar, such as tripos and trichocero.

N-shot	Dist.	1	5
Matching Net [4]	Cosine.	48.76%	60.78%
Prototypical Net [5]	Euclid.	50.84%	66.67%
Relation Net [6]	Deep-metric	46.79%	58.48%
MAML [39]	-	46.0%	60.63%
Imprint [28]	Cosine	57.72%	72.99%
ResNet + CNloss	Cosine	59.0%	74.84%
ResNet + CNloss + H	Cosine	56.29%	70.8%
ResNet + CNloss + Att	Cosine	**60.03%**	**75.56%**

TABLE V: The top-1 accuracy on the test set of *mini*PPlankton.

For this dataset, we conduct 5-way experiments with both 1-shot and 5-shot trials on the $\mathcal{D}_{novel}$ and we use the ResNet18 with cosine-classifier as the baseline (the same as Imprint). We randomly selected 10 classes as the basic training classes, and the remaining classes as the novel classes to evaluate few-shot tasks. As shown in table V, we can see that the proposed model with CN_loss outperforms the baselines by a significant margin, from 72.99% to 74.84% in the 5-shot trial. However, to our surprise, the high-order module does not work for this dataset, and even leads to decline of test accuracy. The reason is that phytoplankton images are not "closed-shape" (target and background are separate) like normal images. For example, as shown in Fig. 8, the object of Biddulphia is interspersed with the background.

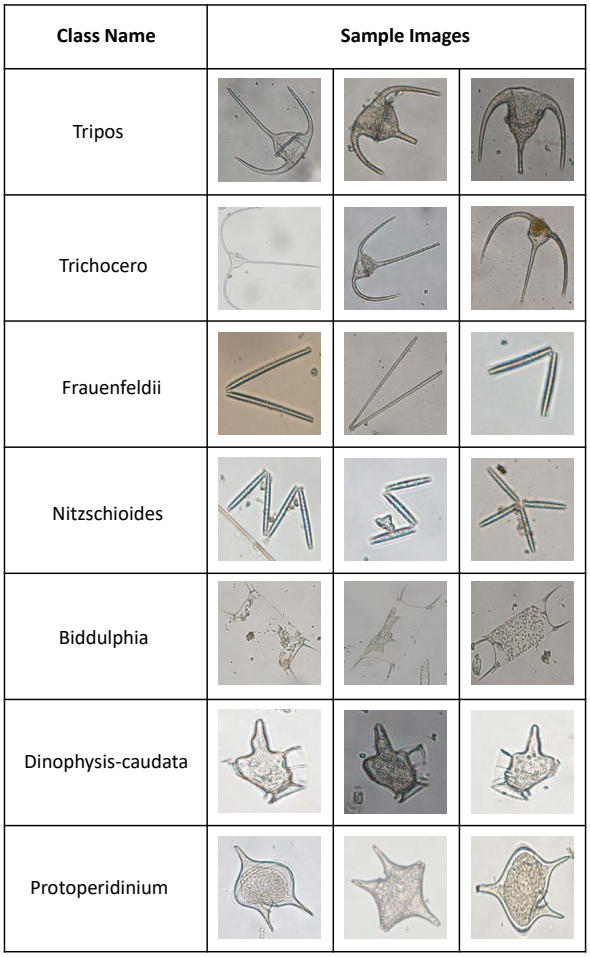
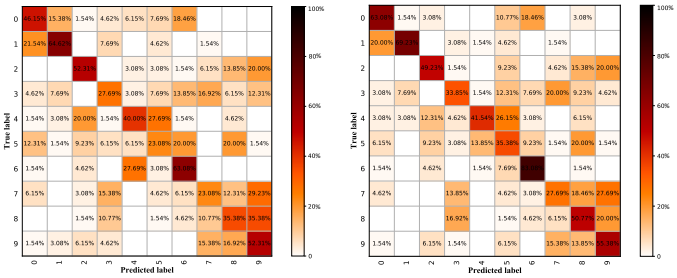


Fig. 8: Random samples of nine categories from our Phytoplankton dataset. The morphological differences among different categories are very small (such as the first two categories). It is a typical dataset for fine-grained challenge problem.

We further illustrate the improvement of classification performance for each category. Fig. 9 shows the confusion matrices of the baseline and our method on $\mathcal{D}_{novel}$ of *mini*PPlankton. We can see that our model greatly improves the accuracy of category 1 (pleurosigma-pelagicum) and category 6 (nitzschioides). At the same time, we reduce the possibility of misclassification of category 5 into category 6. However, it is still very challenge for some categories. Moreover, we visualize the focus area of some examples in Fig. 10. We can see that our method can capture the key area of the object. It helps the model to extract discriminative features for classification. Fig. 11 shows the most difficult category pairs. For instance, samples of category 8 are usually



(a) Imprint (b) Ours

Fig. 9: The confusion matrix of the baseline (ResNet with cosine classifier, also equivalent to Imprint [28]) and our methods.

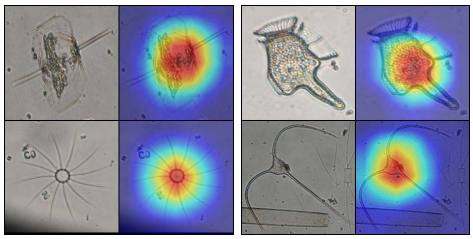


Fig. 10: The focus-area on some examples of the phytoplankton dataset.

classified into category 9. It can be seen from Fig. 11 that the difference between these categories are very small. Such similarity even confuses marine biologists to distinguish them from each other.

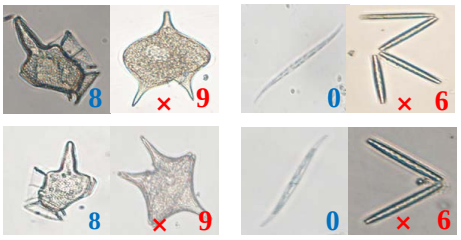


Fig. 11: The blue number in the lower right corner of the image represents the ground-truth category, and the red number represents the wrongly predicted category.

V. Conclusion

In this paper, we focus on the challenge of the domain-specific few-shot fine-grained classification problem via exploring the attention features from a few labeled examples. The Feature Fusion Model and CN_Loss are our two contributions on mining features for such a challenge task. The fusion model utilizes the focus-area location and high-order integration to generate features from discriminative regions. High-order integration has the ability to capture the intra-parts discriminative information. And Grad-CAM can generate focus-area locations for the novel labeled samples. For few-shot learning, we want to learn a more robust feature extractor through basic training classes. As the fine-grained visual categories are quite similar to each other, we design CN_Loss to penalize the special samples which are difficult to approach class centers in each iteration. Furthermore, we build a typical fine-grained and few-shot learning dataset ***mini*PPlankton** from the real-world application in the area of marine ecological environment. We not only build a few-shot phytoplankton dataset but also design an universal model to accomplish the few-shot classification task of natural images and phytoplankton images in the real-world industrial applications. Extensive experiments are carried out to investigate the effects of these proposed modules. We believe that our method is a valuable complement to few-shot classification problem and the new ***mini*PPlankton** is attractive for the marine industrial applications.

Acknowledgment

The authors would like to thank anonymous referees for their useful comments and editors for their work. The authors gratefully thank the GPU computation support from Center for High Performance Computing and System Simulation, Pilot National Laboratory for Marine Science and Technology (Qingdao).

References

[1] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017.

[2] H. Zhuang, K.-S. Low, and W.-Y. Yau, "Multichannel pulse-coupled-neural-network-based color image segmentation for object detection," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 8, pp. 3299–3308, 2012.

[3] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, pp. 1332–1338, 2015.

[4] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching Networks for One Shot Learning," *NIPS*, 2016.

[5] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical Networks for Few-shot Learning," *NIPS*, 2017.

[6] F. Sung and Y. Yang, "Learning to Compare: Relation Network for Few-Shot Learning," *CVPR*, 2018.

[7] S. Gidaris and N. Komodakis, "Dynamic Few-Shot Visual Learning without Forgetting," *CVPR*, 2018.

[8] J. Jiao, M. Zhao, J. Lin, and C. Ding, "Deep coupled dense convolutional network with complementary data for intelligent fault diagnosis," *IEEE Transactions on Industrial Electronics*, pp. 1–1, 2019.

[9] X. Qi, B. Fang, L. Yi, J. Wang, Z. Jian, Y. Zheng, and G. Bao, "Automatic pearl classification machine based on multi-stream convolutional neural network," *IEEE Transactions on Industrial Electronics*, vol. PP, no. 99, pp. 1–1, 2017.

[10] C. Chen, X. Sun, Y. Hua, J. Dong, and H. Xv, "Learning deep relations to promote saliency detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[11] L. Zhang, Y. Gao, Y. Xia, Q. Dai, and X. Li, "A fine-grained image categorization system by cellet-encoded spatial pyramid modeling," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 1, pp. 564–571, 2015.

[12] G. Gorsky and Ohman, "Digital zooplankton image analysis using the ZooScan integrated system," 2010.

[13] H. Jakobsen and J. Carstensen, "FlowCAM: Sizing cells and understanding the impact of size distributions on biovolume of Âplanktonic community structure," *Aquatic Microbial Ecology*, 2011.

[14] H. Huang, J. Zhang, J. Zhang, J. Xu, and Q. Wu, "Low-Rank Pairwise Alignment Bilinear Network For Few-Shot Fine-Grained Image Classification," 2019. [Online]. Available: http://arxiv.org/abs/1908.01313

[15] H. Huang, J. Zheng, J. Zhang, Q. Wu, and J. Xu, "Compare more nuanced: Pairwise alignment bilinear network for few-shot fine-grained learning," *Proceedings - IEEE International Conference on Multimedia and Expo*, vol. 2019-July, DOI 10.1109/ICME.2019.00024, pp. 91–96, 2019.

[16] F. Pahde, P. Jähnichen, T. Klein, and M. Nabi, "Cross-modal Hallucination for Few-shot Fine-grained Recognition," 2018. [Online]. Available: http://arxiv.org/abs/1806.05147

[17] D. Das and C. S. G. Lee, "A two-stage approach to few-shot learning for image recognition," *IEEE Transactions on Image Processing*, pp. 1–1, 2019.

[18] Y. Sun, X. Wang, and X. Tang, "DeepID2: Deep Learning Face Representation by Joint Identification-Verification," *NIPS*, 2014.

[19] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR*, 2005.

[20] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Web-scale training for face identification," in *CVPR*, 2015.

[21] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *CVPR*, 2014.

[22] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Lecture Notes in Computer Science)*, 2016.

[23] Y. Li, J. Zhang, J. Zhang, and K. Huang, "Discriminative Learning of Latent Features for Zero-Shot Recognition," *CVPR*, 2018.

[24] X. S. Wei, C. L. Zhang, J. Wu, C. Shen, and Z. H. Zhou, "Unsupervised object discovery and co-localization by deep descriptor transformation," *Pattern Recognition*, vol. 88, pp. 113–126, 2019.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *ICCV*, 2017.

[26] R. J. Charlson, J. E. Lovelock, M. O. Andreae, and S. G. Warren, "Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate," *Nature*, 1987.

[27] R. K. Simon-Martin Schröder, Rainer Kiko, Jean-Olivier Irisson, "Low-Shot Learning of Plankton Categories," *GCPR*, 2019.

[28] H. Qi, M. Brown, and D. G. Lowe, "Low-Shot Learning with Imprinted Weights," *CVPR*, 2018.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.

[30] S. Cai, W. Zuo, and L. Zhang, "Higher-Order Integration of Hierarchical Convolutional Activations for Fine-Grained Visual Categorization," *ICCV*, pp. 511–520, 2017.

[31] P. Koniusz, F. Yan, P.-H. Gosselin, and K. Mikolajczyk, "Higher-order Occurrence Pooling on Mid- and Low-level Features: Visual Concept Detection," Technical Report, Sep. 2013. [Online]. Available: https://hal.inria.fr/hal-00922524

[32] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," *Proceedings of the 19th ACM SIGKDD*, DOI 10.1145/2487575.2487591, 2013.

[33] H. Wang, Q. Wang, and M. Gao, "Multi-scale location-aware kernel representation for object detection," *CVPR*, 2018.

[34] M. Lin, Q. Chen, and S. Yan, "Network In Network," *ICLR*, 2014.

[35] Z. Chen, Y. Fu, Y. Zhang, Y.-G. Jiang, X. Xue, and L. Sigal, "Multi-level Semantic Feature Augmentation for One-shot Learning," *IEEE Transactions on Image Processing*, 2018.

[36] G. Koch and Zemel, "Siamese Neural Networks for One-shot Image Recognition," *ICML*, 2015.

[37] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," Tech. Rep., 2011.

[38] N. Hilliard, L. Phillips, S. Howland, A. Yankov, C. D. Corley, and N. O. Hodas, "Few-Shot Learning with Metric-Agnostic Conditional Embeddings," *arXiv preprint arXiv:1802.04376*, 2018.

[39] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," *International Conference on Machine Learning(ICML)*, 2017.

[40] V. Garcia and J. Bruna, "Few-Shot Learning with Graph Neural Networks," *ICLR*, 2018.

[41] S. A. Calefati, M. Janjua, "Git Loss for Deep Face Recognition," *BMVC*, 2018.

[42] Q. Li, X. Sun, J. Dong, S. Song, T. Zhang, D. Liu, H. Zhang, and S. Han, "Developing a microscopic image dataset in support of intelligent phytoplankton detection using deep learning," *ICES Journal of Marine Science*, 09 2019, fsz171. [Online]. Available: https://doi.org/10.1093/icesjms/fsz171