Pattern Analysis and Applications (2023) 26:797–808
https://doi.org/10.1007/s10044-022-01109-9

INDUSTRIAL AND COMMERCIAL APPLICATION

Dual autoencoder based zero shot learning in special domain

Qiong Li[1,2] · Eric Rigall[2] · Xin Sun[2,4] · Kin Man Lam[3] · Junyu Dong[2]

Received: 27 August 2021 / Accepted: 27 August 2022 / Published online: 6 October 2022

Abstract
Zero-shot learning aims to learn a visual classifier for a category which has no training samples leveraging its semantic information and its relationship to other categories. It is common, yet vital, in practical visual scenarios, and particularly prominent in the uncharted ocean field. Phytoplankton plays an important part in the marine ecological environment. It is common to encounter the zero-shot recognition problem during the in situ observation. Therefore, we propose a dual autoencoder model, which contains two similar encoder–decoder structures, to tackle the zero-shot recognition problem. The first one is used for the projection from the visual feature space to a latent space, then to the semantic space. Inversely, the second one projects from the semantic space to another latent space, then back to the visual feature space. This structure guarantees the projection from the visual feature space to the semantic space to be more effective, through the stable mutual mapping. Experimental results on four benchmarks demonstrate that the proposed dual autoencoder model achieves competitive performance compared with six recent state-of-the-art methods. Furthermore, we apply our algorithm to phytoplankton classification. We manually annotated phytoplankton attributes to develop a practical dataset for this real and special domain application, i.e., Zero-shot learning dataset for PHYtoplankton (ZeroPHY). Experiment results show that our method achieves the best performance on this real-world application.

Keywords Zero shot learning · Autoencoder · Phytoplankton · Classification

1 Introduction

The research on visual object recognition has made tremendous achievements in the last few years with the availability of large-scale training data. Deep neural models, trained with a sufficiently large amount of training datasets, can achieve better accuracy and generalization ability, which promote the rapid research progress on various domains, such as the study of phytoplankton in marine science. However, their superior performances are obtained at the cost of much more expensive human labor to collect and annotate training data, which is time-consuming, labor-intensive, and sometimes very difficult. Furthermore, the collection of a sufficient amount of annotated training data for rare categories is a real challenge.

Humans can distinguish at least 30,000 relevant object classes. To reach this level, conventional object detectors require millions of well-annotated training samples for training. More importantly, with the help of semantic descriptions, humans have the ability to recognize objects, which they have never seen before, by attribute combinations [1]. For example, a child, who has ever seen horses and Siberian

✉ Xin Sun
sunxin1984@ieee.org

✉ Junyu Dong
dongjunyu@ouc.edu.cn

Qiong Li
liqiong@stu.ouc.edu.cn

Eric Rigall
rigallericcharlesalaric@stu.ouc.edu.cn

Kin Man Lam
kin.man.lam@polyu.edu.hk

1 Department of Science and Information Science, Qingdao Agricultural University, No.700 Changcheng Road, Qingdao 266109, Shandong, China

2 Department of Computer Science and Technology, Ocean University of China, No.238 Songling Road, Qingdao 266100, Shandong, China

3 Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon 999077, Hong Kong, China

4 Department of Aerospace and Geodesy, Technical University of Munich, 80333 Munich, Germany

tigers, could recognize a "zebra," even though the child has never seen a zebra, because the child knows that zebras have the same shape as horses and the same fur-pattern color as Siberian tigers. Therefore, in the past few years, techniques based on semantic descriptions have been proposed, with the aim of reducing the required number of training samples for recognition [2–6]. Zero-shot learning [7–9] is undoubtedly the most forward-looking task, where the training and testing classes are disjointed. The source domain contains seen categories with labeled training images, whereas the target domain gathers unseen categories without any associated annotated samples. These two domains are connected with semantic information, e.g., attributes [10, 11]. To recognize the unseen categories in the target domain, the information learned from the source domain can be utilized.

Zero-shot learning is necessary for identifying tasks in many practical domains. In the area of marine science, phytoplankton is the foundation of marine ecosystems, which activates the marine food chain as the basic producer in the ecological cycle. Therefore, some poisonous phytoplankton, which accumulates in marine creatures, will ultimately affect human health. It also plays an important role in aquaculture. Red tides and canola caused by marine phytoplankton in coastal areas are typically harmful ecological phenomena, which can directly lead to the deaths of numerous aquatic organisms, due to the lack of oxygen. In order to better predict the occurrence of marine ecological disasters, the biomass and abundance of phytoplankton in sea water should be observed in situ. However, during the in situ monitoring of marine phytoplankton, new species may emerge inevitably, which cannot be recognized by the existing detection models. The detectors trained based on the existing images do not have the ability to detect the new species, which reduces the detection accuracy and hinders the development of marine ecological environment monitoring [12]. Thus, in this paper, we propose an accurate zero-shot detector for phytoplankton. We also propose and create a new dataset, namely ZeroPHY, which combines with the semantic description of those unseen classes. The dataset includes phytoplankton images and human-specified high-level descriptions of the categories.

In this paper, we firstly propose a dual encoder–decoder model for zero-shot learning. The first encoder–decoder projects a visual feature representation of an image into a latent representation space, which is then projected into a semantic representation space. The second encoder–decoder firstly projects the semantic representation back into the latent representation space, then to the original visual feature representation. This dual autoencoder imposes a new constraint in the construction of the visual-to-semantic projection function, so that the projection will preserve the information from the original visual feature representation. Finally, we evaluate our proposed approach on both existing benchmark datasets for zero-shot learning and the practical ZeroPHY dataset. Compared with other zero-shot learning methods, our method achieves the best performance. The contributions of this paper are summarized as follows:

- We propose a novel dual autoencoder model, which can learn a robust projection from the input visual-feature space into a low-dimensional semantic representation that can be utilized for zero-shot learning.
- For object detection in real and special domains, we develop a novel dataset for zero-shot learning of phytoplankton, named ZeroPHY, which contains images of 20 categories, with 56 high-level description attributes.
- We evaluate our proposed method on four benchmark datasets. Compared with state-of-the-art zero-shot learning methods, our model demonstrates outstanding performance. Furthermore, we conducted extensive experiments on the ZeroPHY dataset for practical application in a special domain. Our method outperforms all the compared methods.

2 Related work

In this work, we apply zero-shot learning to the identification of phytoplankton, which may have unseen species during in situ identification. Therefore, in this paper, we will first introduce the in situ identification followed by zero-shot learning.

2.1 Phytoplankton in situ identification

As phytoplankton is sensitive to the environmental changes, the study and monitoring of their development in situ are critical in evaluating the evolution of marine ecosystems and ocean dynamics, as well as protecting the global climate and environment. The research community of marine science has explored automated methods for performing plankton identification in situ. Acoustic systems, such as Acoustic Doppler Current Profiler [13], Multifrequency Hydroacoustic Probing Systems [14], and wideband sonar [15], have played an essential role in plankton identification. Compared to visible light and electromagnetic waves, sound waves are hardly affected by the underwater environment, thus, can be utilized for long-distance monitoring. However, the low discriminant ability, unapparent fine-grained features and inaccurate positioning of underwater observation data do affect their reliability and accuracy [16]. The chlorophyll fluorescence instrument for phytoplankton monitoring [17] is the most well-developed and widely used in situ monitoring device for underwater organisms. The main drawback of this device is that it can only detect the organisms with fluorescence [18]. Profiting from the emergence of optical microscopes,

microscopic images have gradually gained a foothold in the basic micro-organism research. In 1992, the Video Plankton Recorder was the first in situ automated monitoring device and the pioneer of modern in situ plankton image devices [19]. In 2004, Yu [20] proposed an "automated underwater digital microscope image," based on optical microscopes, which can automatically capture plankton images in situ, but this instrument is only suitable for high-concentration conditions. In terms of small and micro plankton, Olson et al. [21] utilized the Imaging FlowCytobot to monitor microzooplankton and phytoplankton, with the size ranging from 10 to 100 μm. Systems, such as FlowCAM, CytoSub, and CytoSense, are also available for the acquisition of images from micro to small phytoplankton, and provide effective methods for in situ observation of full-grained phytoplankton.

The above mentioned methods have laid a solid foundation of in situ observation of phytoplankton. In particular, the acquisition of images fosters the possible use of artificial intelligence in the field of phytoplankton monitoring. Biological oceanographers increasingly use imaging-based technologies to study marine creatures. In 2015, Orenstein et al. [22] introduced WHOI-Plankton as a large scale, fine-grained visual recognition benchmark dataset for plankton classification. Convolutional neural networks and residual networks have been widely used for plankton classification in the last 5 years [23–25]. All these research works are based on a large number of training samples, and do not consider the probable occurrence of unknown samples. Thus, in this work, we particularly pay attention to the unseen categories that can be encountered during in situ observation.

In this paper, in addition to proposing a dual autoencoder model for zero-shot learning, we also apply it to a special domain for the classification of in situ phytoplankton observation scenes.

2.2 Zero-shot learning

This paper focuses on the zero-shot learning problem, where the test classes are disjointed from the training classes [7, 26]. As the visual information from unseen classes is not available during training, additional information is required to make up the missing information. The additional information, which belongs to the semantic level, can be acquired in different forms: text corpora, e.g., glove [27] and word2vec [28]; structured semantic sources, e.g., wordnet hierarchies [29]; and human annotations, e.g., manually specified attributes, which were the most widely used in previous works [10]. How to establish the projection function from the visual space to the semantic space is the main task in existing zero-shot learning models.

With the rapid development of deep learning, most of the recent zero-shot learning methods [30] can be divided into two main categories. The first category is based on the use of semantic embeddings to represent each category with learned vector representations, which can then be projected to visual classifiers [31]. Frome et al. [2] proposed a system, named DeViSE, to train a projection function, from image to word embedding, using a convolutional neural network with a transformation layer. Norouzi et al. [32], instead of predicting the word embedding directly through convolutional neural networks, proposed another method, named ConSE, to predict the image embedding by combining a convolutional neural network for image classification with a word embedding model. Kodirov et al. [33] proposed a novel solution, by using the encoder–decoder paradigm to learn the projection function. Different from DeViSE and ConSE, which are based on regression with deep neural networks, they applied linear and symmetric projection in both the encoder and decoder.

The second category is based on distilling knowledge from a knowledge graph for object recognition [34, 35]. Salakhutdinov et al. [36] proposed a method to share the representations among different object classifiers using WordNet. This representation shares the outputs' statistical strength from related objects for those objects with a few training examples. Wang et al. [37] proposed to distill information via both semantic embeddings and knowledge graphs. They used the word embedding of every category and their relationship to each other to learn a visual classifier for those categories without any training examples.

3 Dual autoencoder

In this section, we will describe the proposed method in detail, including a basic introduction to deep autoencoder, the model formulation, and the loss function of our model.

3.1 Deep autoencoder

In this section, we briefly describe the structure of a standard autoencoder with several hidden layers. The encoder projects the input data into a lower-dimensional space representation in the latent space, while the decoder projects the lower-dimensional data from the latent space back to the original feature space to reconstruct the input data. Denote an input data matrix as $\mathbf{X} \in \mathbb{R}(d \times N)$, which is composed of N d-dimensional feature vectors. The input is projected across L hidden layers from $\mathbf{H}_1$ to $\mathbf{H}_L$, with the corresponding projection matrixes $\mathbf{W}_i$, where $\mathbf{i} = 1, ..., L$. The resulting matrix in the k-dimensional latent space representation is denoted as $\mathbf{S} \in \mathbb{R}^{k \times N}$, composed of N feature vectors of dimension k, where $k < d$. This part forms the encoder, whose function is denoted as $s = f(x)$. The generated latent space representation can be projected back to the feature

space representation using the corresponding hidden layers from H_1^* to H_i^*, resulting in $\hat{\mathbf{X}} \in \mathbb{R}^{d \times N}$. This part forms the decoder, whose function is to reconstruct the input data, and is denoted as $r = g(h)$. The reconstruction error, i.e., the difference between $\mathbf{X}$ to $\hat{\mathbf{X}}$, is minimized by optimizing the following objective function:

$$\min \parallel g(f(\mathbf{X})) - \mathbf{X} \parallel_F^2. \tag{1}$$

3.2 Model formulation

To make the reconstruction of the latent space robust and meaningful, we propose a dual autoencoder model, to be utilized for zero-shot learning. The model is essentially a two-stage autoencoder, composed of two autoencoders. Different from the semantic encoder–decoder model [33], our autoencoder has hidden layers. Figure 1 shows the overall structure of our model. During the training process, indicated by the dark blue arrows in Fig. 1, the feature extractor f_φ, based on a convolutional neural network, maps an input image of a seen category from the source domain $\mathbf{x} \in \mathbb{R}^N$ to a d-dimensional feature vector $f_\varphi(x) \in \mathbb{R}^d$. For zero-shot learning, it is pivotal to build a solid projection from the visual feature space to the semantic space. Therefore, we propose a dual autoencoder to realize the projection. As shown in Fig. 1, we firstly apply the input feature vector to the Feature Encoder, which has two hidden layers, and then project the input to the latent feature space. The Feature Encoder is denoted as $\mathcal{G}_f()$, so the latent feature can be expressed as follows:

$$\mathcal{L}_f = \mathcal{G}_f(f_\varphi(x)). \tag{2}$$

The representation in the latent feature space is in a lower dimension, and is fed to the Semantic Decoder, which has one hidden layer. By denoting the Semantic Decoder as $\mathcal{G}_s()$, the latent feature is mapped to the semantic space, as follows:

$$\hat{S} = \mathcal{G}_s(\mathcal{L}_f). \tag{3}$$

The semantic representation $\hat{S}$, represented in Fig. 1 as the dark blue vertical block, has the projection domain shift problem. To alleviate the domain shift, we apply a reconstruction constraint to map the semantic representation back to the visual feature representation. The Semantic Encoder projects $\hat{S}$ to the latent semantic space, i.e., the green block on the right in Fig. 1, is given as follows:

$$\mathcal{L}_s = \mathcal{H}_s(\hat{S}) \tag{4}$$

Finally, the Feature Decoder maps the latent representation back to the visual feature space, as follows:

$$\mathcal{F} = \mathcal{H}_f(\mathcal{L}_s). \tag{5}$$

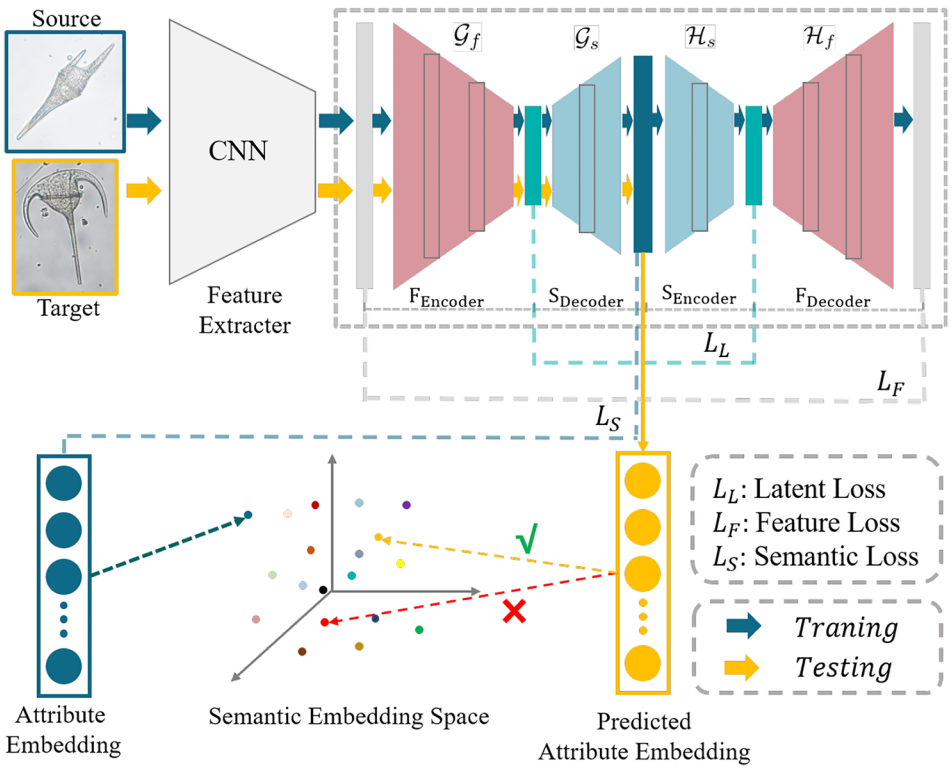

Fig. 1 The architecture of the dual autoencoder. The input with a blue border is a seen image from the source domain, and the input with a yellow border is an unseen image from the target domain. The path with blue arrows represents the training process, while the path with yellow arrows is the testing process. After extracting the features from the convolutional neural network (CNN), the feature vectors are input to the dual autoencoder, which is enclosed in the gray dotted box. The dual autoencoder includes a Feature Encoder, a Semantic Decoder, a Semantic Encoder, and a Feature Decoder. During testing, the semantic representation in the dark blue vertical box is the output. Three loss functions are utilized to train the model (color figure online)

In the testing process, as indicated by the yellow arrows in Fig. 1, an image of an unseen category from the target domain is input to the model. The model outputs the predicted attribute embedding for classification in the semantic embedding space, shown as the yellow feature vector in Fig. 1. The algorithm is shown in Algorithm 1.

Algorithm 1 Dual Autoencoder.

Input:
Test images from unseen categories, $x_i^u,\ i \in \{1,\ ..., n_u\}$;
The semantic vectors of the unseen categories, $s_i^u,\ i \in \{1,\ ..., N_u\}$;
The weights of the dual aotoencoder: $\mathcal{W}_{FE}$, $\mathcal{W}_{SD}$;
The max iteration.
Output:
The category y_i^u of the unseen image x_i^u.

1: **for** $j \leq$ max iteration **do**
2: Extract the unseen image features, $f(x_i^u)$;
3: Use the Feature Encoder to get the latent representation $\mathcal{L}_f = \mathcal{G}_f(f_\varphi(x_i^u))$;
4: Use the Semantic Decoder to obtain the semantic representation $\hat{S} = \mathcal{G}_s(\mathcal{L}_f)$;
5: Calculate the cosine similarity of the semantic representation $\hat{S}$ of x_i^u with the semantic vectors of the unseen categories to find the category y_i^u it belongs to;
6: **end for**

4 Loss function

In order to make the constraints of the semantic space representation as robust as possible, we design three loss functions, which provide reconstruction constraints, for training our proposed model. The first loss function is the visual feature loss, which constrains the reconstructed visual feature to be as close to the original visual feature as possible. The visual feature loss is defined as follow:

$$\mathcal{L}_F = \| \mathcal{F} - f_\varphi(x) \|_F^2 \tag{6}$$

The second loss is the semantic loss function, which makes sure that the reconstructed semantic representation is effective. The semantic loss function is defined as follows:

$$\mathcal{L}_S = \| S - \hat{S} \|_F^2 \tag{7}$$

where S is the original semantic embedding represented by the attribute vector.

The third loss is the latent space loss function, which further constrains the reconstruction accuracy in the latent space. It is utilized to align the latent spaces to ensure that the projection between feature space and the semantic space is smoother.The latent loss is defined as follows:

$$\mathcal{L}_L = \| \mathcal{L}_s - \mathcal{L}_f \|_F^2 \tag{8}$$

where $\mathcal{L}_s$ and $\mathcal{L}_f$ represent the latent space embedding of latent semantic space and latent feature space, respectively.

Ultimately, the overall loss function $\mathcal{L}$ is the combination of the three loss functions, as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_F + \lambda_2 \mathcal{L}_S + \lambda_3 \mathcal{L}_L \tag{9}$$

where λ_1, λ_2 and λ_3 are the hyperparameters that balance the weights of $\mathcal{L}_F$, $\mathcal{L}_S$ and $\mathcal{L}_L$, respectively.

By training with the three loss functions, the proposed model can finally construct an effective projection from the visual feature space to the semantic space.

5 The ZeroPHY dataset

In order to evaluate our proposed method in a special domain, we conducted experiments on recognition of unseen phytoplankton classes. Therefore, we developed a phytoplankton dataset, namely ZeroPHY, with attributes. The dataset has two parts: the images and their corresponding attributes. The images of our dataset are cropped according to our previous work in [38]. The semantic attributes of the seen and unseen classes are suggested by experts in this domain. In this section, we will introduce the ZeroPHY dataset in two aspects. The first one is the images of the

different categories in the dataset, and the second one is the definitions of the attributes.

5.1 The images of different categories in ZeroPHY

Phytoplankton plays an important role in aquaculture, which causes red tides and Enteromorpha prolifera in coastal areas. These harmful phytoplankton ecological phenomena led to the death of numerous aquatic organisms [39]. In our previous work [38], we developed a phytoplankton microscopic image dataset (PMID2019) for in situ detection, which contains 10,819 phytoplankton microscopic images of 24 different categories. In this paper, we cropped objects from the images of PMID2019, to form separate images for the recognition task.

According to the number and quality of the images in the different categories in PMID2019, we finally selected 20 categories from its 24 classes to develop the novel dataset, including Chaetoceros, Eucampia, Cosinodiscus, Thanlassionema nitzshioides, Guinardia flaccida, Ceratium furca, Ceratium fusus, Bacteriastrum, Skeletonema, Ceratium trichoceros, Dinophysis caudata, Biddulphia, Helicotheca, Thalassionema frauenfeldii, Pleurosigma pelagicum, Ceratium tripos, Ditylum, Navicular, Protoperidinium, and Rhizosolenia. The distribution of the number of samples for the selected categories is shown in Fig. 2. The categories, represented in green, are the seen classes used for training, while the categories in blue are the unseen classes used for testing. This dataset follows a long tail distribution, which is a big challenge for the classification task.

5.2 Attribute definition of the dataset

To develop a zero-shot learning dataset for phytopalnkton, we need to define reasonable semantic descriptions for each category. In order to define effective semantic attributes, the attributes (1) should reflect the characteristics of the different object classes; (2) should be nameable; and (3) should have semantics, i.e., they must also have the following three basic characteristics: (i) they should be shared between classes, not unique to a certain class, so that the model can learn attributes from mutually exclusive training data; (ii) they must be open to public criticism, which requires us not only to clearly know the shared visual information between the classes when defining the attributes, but also to understand the visual differences between them, so as to better recognize the different classes; and (iii) they do not necessarily have direct visual meaning, but must be visually

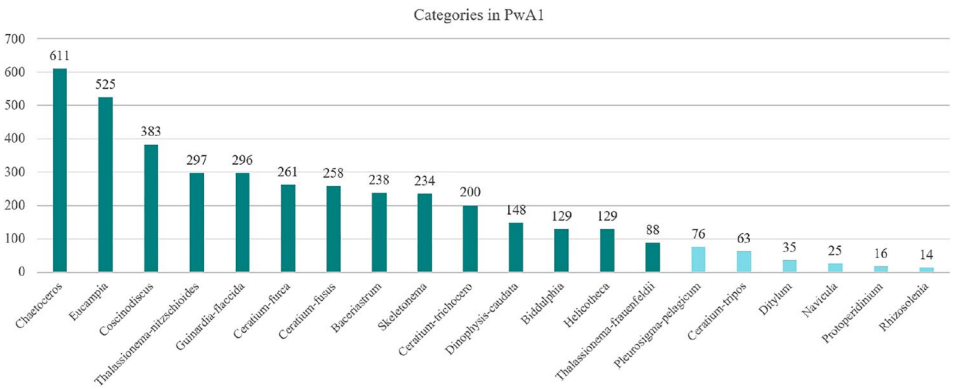

Fig. 2 The number of samples in each of the categories of the ZeroPHY dataset

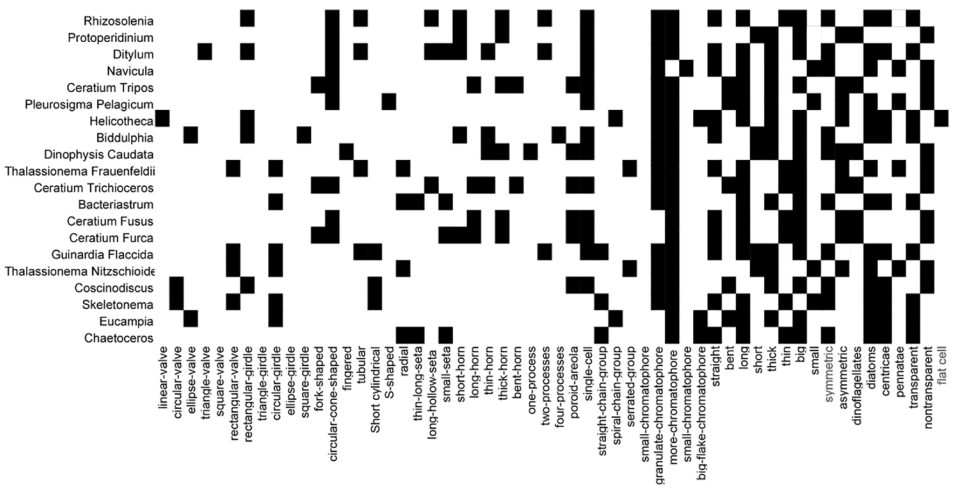

Fig. 3 The 56 shared attributes among the 20 categories in the ZeroPHY dataset

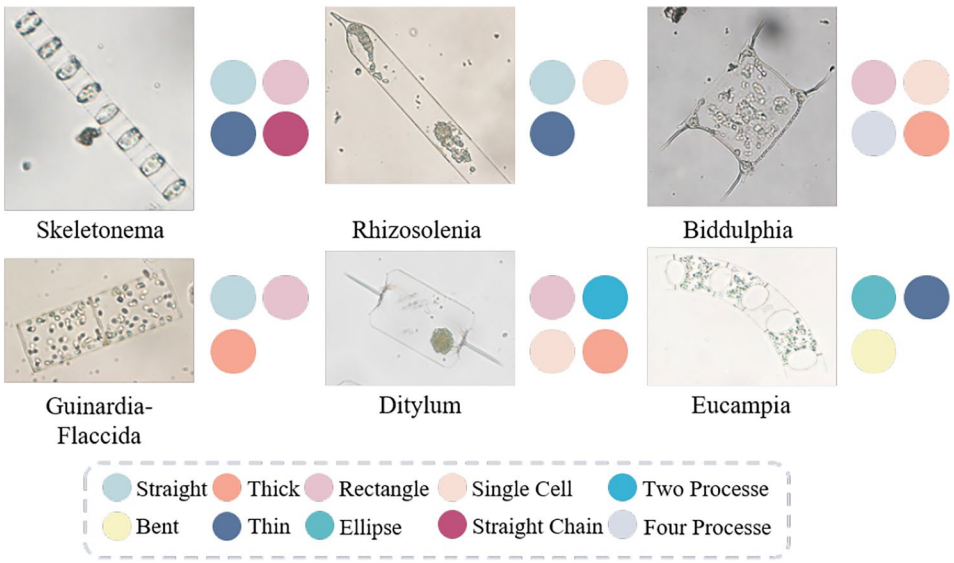

Fig. 4 This figure shows six images from different categories with some of their attributes represented by various colors

detectable, so that we can learn reliable attribute classifiers from visual images.

According to the above criteria, we define 56 attributes for the 20 categories in our dataset, as shown in Fig. 3. The x-axis represents the 56 attributes designed by experts in the area of marine science. The y-axis represents the 20 categories of our dataset. The description consists of arbitrary semantic attributes. In Fig. 4, we present six images from the different categories, with some attributes shared by each other. We can see that the Skeletonema category has four attributes present in the figure, including "Straight," "Rectangle," "Thin," and "Straight Chain." The attribute "Thin," shown as dark blue bubbles in Fig. 4, is shared by Skeletonema, Rhizosolenia, and Eucampia. However, in the six categories, the attribute 'Straight Chain' is used to describe Skeletonema only. This shows that this attribute is a unique characteristic of Skeletonema, i.e., other categories do not have this attribute.

The semantic space is shared by both the seen and unseen classes. Different categories provide knowledge about attributes, and vice versa. Each attribute can be utilized for the detection of numerous object classes. This reciprocity between classes and attributes makes our proposed learning method statistically efficient. Each category from our dataset can be described by their associated attributes, with the relationships shown in Fig. 3. For example, we can use images of the Skeletonema and Biddulphia categories to learn the shared attributes, e.g., rectangle. For the attribute single-cell, Skeletonema is shown not to be useful to learn this attribute, but Biddulphia and Rhizosolenia are useful, possibly together with Ditylum.

Table 1 A summary of the datasets used in experiments, where "SS" and "SS-D" mean semantic space and the dimension of the semantic space, respectively

D	Instances	SS	SS-D	# seen/unseen
AWA	30,475	A	85	40/10
ap & Y	15,339	A	64	20/12
CUB	11,788	A	312	150/50
SUN	14,340	A	102	645/72

'A' means attributes

6 Experiments

In this section, experiments were carried out to measure the performance of the proposed method and compare it with other state-of-the-art models, on both benchmark datasets and our ZeroPHY dataset.

6.1 Datasets

Four small-scale benchmark datasets for zero-shot learning were used in the experiments. These include Animals with Attributes (AwA [26]), aPascal &Yahoo (aP &Y [40]), Caltech-UCSD-Birds 200-2011 (CUB [41]), and SUN Attribute (SUN [42]). A summary of these datasets, as well as our new ZeroPHY dataset, is shown in Table 1, including the number of instances, the semantic-space dimension, and the number of seen and unseen categories. AwA contains 30,475 images from 50 kinds of animals with 85 attributes. Ap &Y contains 15,339 images, including 32 classes described by 64 attributes. CUB contains 11,788 images from 200 different types of birds, annotated with 312 attributes. SUN comprises 14,340 images from 817 categories with 102 attributes. Our ZeroPHY dataset for the special

domain contains 4,026 images from 20 classes, described by 56 attributes.

6.2 Semantic spaces

Attributes are the distinguishable object properties obtained through human annotation. All the semantic spaces are composed of the annotated attributes. Our attribute-class embedding is a per-class vector, which measures the strength of each attribute based on human evaluation.

6.3 Features

The deep visual features are extracted by using ResNet [43, 44], with a dimension of 2048.

6.4 Parameter settings

We first verify the coefficients of each part in the loss function through experiments. The results in Table 3 turns out that the accuracy is the best when all three coefficients equal one. As shown in Fig. 6, the loss converges during training.

Table 2 The top-5 accuracy of different settings of the layers

Fh_Dim1	Fh_Dim2	L_Dim	Sh_Dim	Accuracy
512	**128**	**32**	**48**	**73.3**
512	128	16	32	58.9
1024	128	32	48	63.3
1024	128	16	32	60.0
512	256	32	48	71.1
512	256	16	32	55.6
1024	256	32	48	72.2
1024	256	16	32	62.2

Notation: Fh_Dim1 represents the dimension of the first Feature encoder hidden layer, and the second Feature decoder hidden layer. Fh_Dim2 means the dimension of the second Feature encoder hidden layer, and the first Feature decoder hidden layer. L_Dim means the dimension of the latent representation, and Sh_Dim means the dimension of the Semantic autoencoder hidden layer

Bold indicates the best dimension of each layer

Adam [45] was adopted to optimize our model. The Feature Autoencoder has two hidden layers, while the Semantic Autoencoder has one hidden layer only, as illustrated in

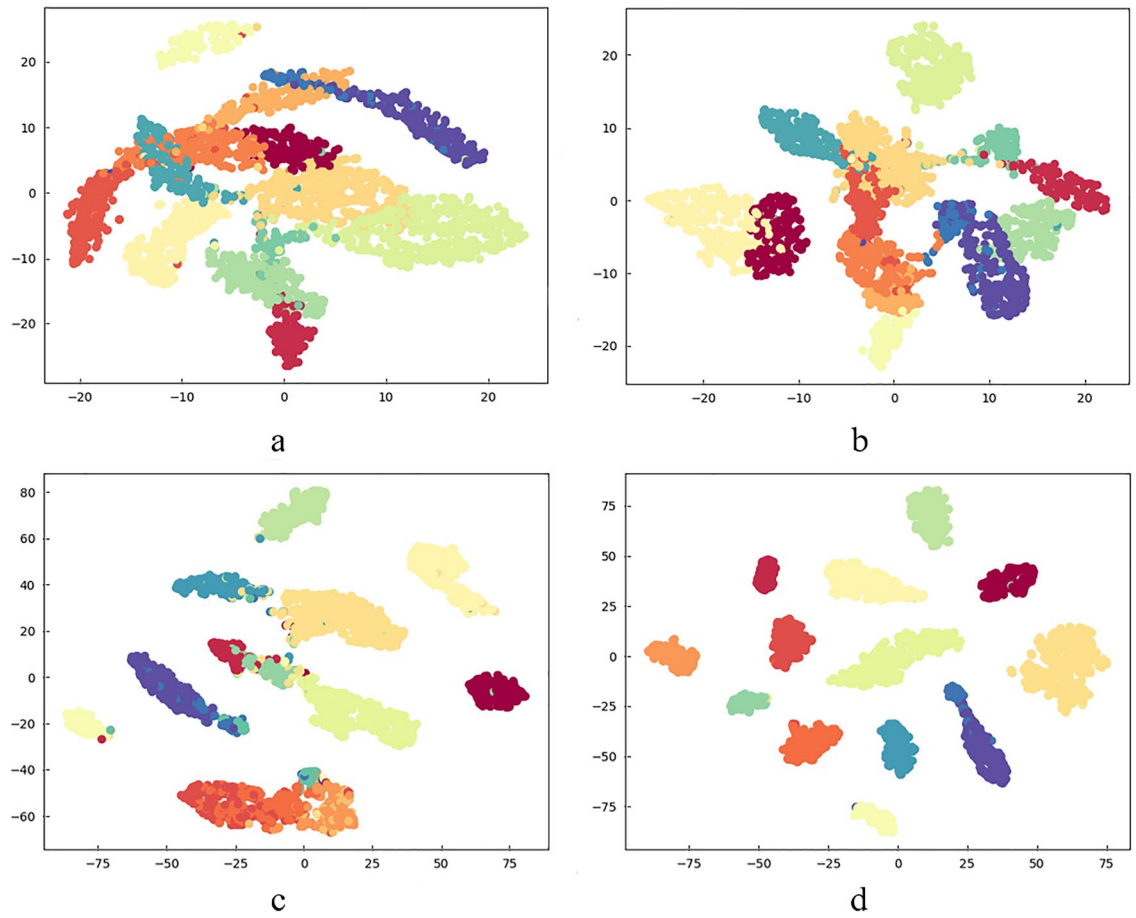

Fig. 5 The visualization of t-SNE to show the classification effect with different models (**a**: DAP; **b**: IAP; **c**: SAE; **d**: Our model)

Table 3 The top-5 accuracy of different settings of the coefficients in the loss function

λ_1	1	0	0	0.5	0.5	1	1	1	1
λ_2	0	1	0	1	1	0.5	0.5	1	1
λ_3	0	0	1	0	1	0	1	0	1
Accuracy	25.1	26.9	24.3	40.7	42.6	38.2	41.9	43.8	**49.2**

Bold indicates the best accuracies

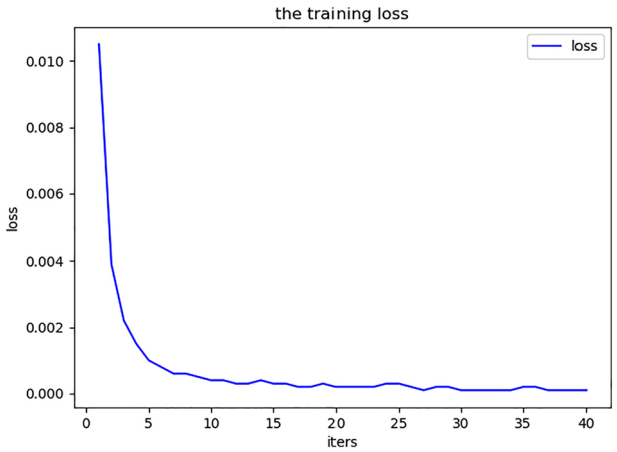

Fig. 6 The loss evolution in training

Fig. 1. A number of ablation experiments on our dataset were performed to evaluate the performance of our model with different hidden-layer widths. As shown in Table 2, the best performance corresponds to the settings in the first row. The dimension of the first Feature hidden layer, the second Feature hidden layer, the Latent hidden layer, and the Semantic hidden layer are set to 512, 128, 32, and 48, respectively.

6.5 Competitors

Six zero-shot learning models, including semantic autoencoder (SAE) [33], direct attribute prediction (DAP) [46], indirect attribute prediction (IAP) [46], structured joint embedding (SJE) [47], structured similarity embedding (SSE) [48], and deep visual-semantic embedding (DeViSE) [2], are compared with our model.

6.6 Evaluation metric

In the experiments, similar to the previous works, the multi-way classification accuracy is used to measure the

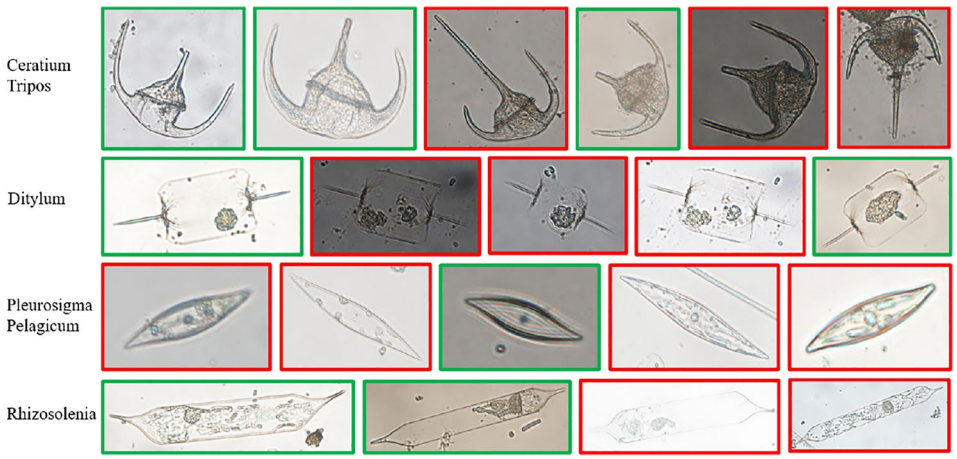

Fig. 7 The top-1 recognition results of the four unseen classes using our approach. The images with a green border are correctly classified (color figure online)

Table 4 The top-1 accuracy compared with the state-of-the-art models on benchmark datasets

Dataset	Dual AE	SAE	DAP	IAP	SJE	SSE	DeViSE
AWA	57.6	49.8	44.1	35.9	**60.5**	60.1	52.8
ap & Y	**35.2**	12.0	33.8	34.9	34.4	34.0	35.1
CUB	34.7	32.4	34.3	24.0	**40.9**	36.2	37.6
SUN	**36.3**	31.8	35.7	19.4	35.1	32.9	33.7
ZeroPHY	**49.2**	38.1	37.6	34.5	47.2	46.5	47.9

Bold indicates the best accuracies

Table 5 The f1 score compared with the state-of-the-art models on benchmark datasets

Dataset	Dual AE	SAE	DAP	IAP	SJE	SSE	DeViSE
AWA	**0.586**	0.482	0.512	0.375	0.582	0.579	0.496
ap & Y	**0.297**	0.138	0.261	0.291	0.294	0.284	0.273
CUB	0.496	0.396	0.416	0.263	**0.516**	0.343	0.365
SUN	**0.385**	0.292	0.304	0.233	0.369	0.299	0.346
ZeroPHY	**0.534**	0.358	0.397	0.296	0.511	0.492	0.503

Bold indicates the best accuracies

performance of the different methods. We also use t-SNE to visualize the classification ability. As shown in Fig. 5, compared with DAP, IAP, and SAE, our method can achieve a better classification performance than the other methods.

We firstly compare the performance of those compared models with our proposed model on the four benchmark datasets. Table 4 illustrates the top-1 accuracy of all these methods on all the datasets. Table 5 shows the f1 score of all method on all datasets. As can be seen in both Tables 4 and 5, on most of the datasets, our proposed model, based on dual autoencoder, yields better classification performance compared to the other models. SJE uses structured joint embedding and combine unsupervised multiple output embeddings which provides complementary information. Therefore, SJE performs a little better in two benchmark datasets. However, our proposed method can achieve comparable results using attributes alone.

Furthermore, for the special domain application, we compare the performance of the different models on our developed ZeroPHY dataset. As shown in Table 4, the identification results on our dataset are shown in the last row. Obviously, our proposed model outperforms the other methods, which indicates that the dual autoencoder structure can establish a more effective projection from the feature space to the semantic space for the identification of phytoplankton. In the meanwhile, Fig. 7 shows the top-1 recognition results of our method on the four unseen classes, including Ceratium Tripos, Ditylum, Pleurosigma Pelagicum, and Rhizosolenia. Those images with a green border mean that they are classified correctly, while those with a red border are incorrectly classified. As shown in Fig. 7, the third and the sixth images in the first row are different from the other images in Ceratium Tripos, because they have a longer horn. In the third row, our method achieves unsatisfactory performance on Pleurosigma Pelagicum, mainly because the dataset lacks the attributes that can be shared by this unseen category and the other seen categories. Therefore, a more complete semantic space, shared by both the seen and the unseen categories, is the key to constructing a more effective projection from visual features to semantic representations.

7 Conclusion

In this paper, we proposed a novel zero-shot learning model, based on a dual autoencoder, which is composed of two encoder–decoders, connected in cascade. This structure can guarantee a more effective and accurate projection from the visual feature space to the semantic space. To evaluate the performance of our proposed method for a special domain in marine applications, we manually annotated phytoplankton attributes and developed the ZeroPHY dataset, which includes 20 categories with 56 attributes. To the best of our knowledge, this work is the first to apply zero-shot learning to the marine phytoplankton field. Extensive experiments were carried out on four existing benchmarks and the new ZeroPHY dataset. Experiment results show that our proposed dual autoencoder model achieves superior performance, compared with other state-of-the-art methods. In our future work, we will study the use of a knowledge graph for phytoplankton detection, based on zero-shot learning.

The dataset generated during the current study are available from the corresponding author on reasonable request.

Acknowledgments This work was supported by the National Natural Science Foundation of China (61971388, U1706218), and a research grant from the Key-Area Research and Development Program of Guangdong Province 2020 under Project 76.

References

1. Huang S, Elhoseiny M, Elgammal A, Yang D (2015) Learning hypergraph-regularized attribute predictors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 409–417
2. Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Ranzato M, Mikolov T (2013) Devise: a deep visual-semantic embedding model. Adv Neural Inf Process Syst 26:2121–2129
3. Liu Z, Zhang X, Zhu Z, Zheng S, Zhao Y, Cheng J (2020) Convolutional prototype learning for zero-shot recognition. Image Vis Comput 98:103924
4. Sun X, Xv H, Dong J, Zhou H, Chen C, Li Q (2020) Few-shot learning for domain-specific fine-grained image classification. IEEE Trans Ind Electron 68(4):3588–3598
5. Fu Y, Sigal L (2016) Semi-supervised vocabulary-informed learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 5337–5346

6. Ji Z, Sun Y, Yu Y, Guo J, Pang Y (2018) Semantic softmax loss for zero-shot learning. Neurocomputing 316:369–375
7. Larochelle H, Erhan D, Bengio Y (2008) Zero-data learning of new tasks. AAAI 1:3
8. Qin J, Wang Y, Liu L, Chen J, Shao L (2016) Beyond semantic attributes: Discrete latent attributes learning for zero-shot recognition. IEEE Signal Process Lett 23(11):1667–1671
9. Ma Y, Xu X, Shen F, Shen HT (2020) Similarity preserving feature generating networks for zero-shot learning. Neurocomputing 406:333–342
10. Duan K, Parikh D, Crandall D, Grauman K (2012) Discovering localized attributes for fine-grained recognition. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp 3474–3481
11. Lv H, Chen J, Pan T, Zhou Z (2020) Hybrid attribute conditional adversarial denoising autoencoder for zero-shot classification of mechanical intelligent fault diagnosis. Appl Soft Comput 95:106577
12. Lumini A, Nanni L (2019) Deep learning and transfer learning features for plankton classification. Ecol Inf 51:33–43
13. Potiris E, Frangoulis C, Kalampokis A, Ntoumas M, Pettas M, Petihakis G, Zervakis V (2018) Acoustic doppler current profiler observations of migration patternsof zooplankton in the cretan sea. Ocean Sci 14(4):783–800
14. Sknarya AV, Razin AA, Toshchov SA, Demidov AI (2018) Ultra wideband sounding signals in hydroacoustic systems. Rensit Radioelectron Nanosyst Inf Technol 10(2):209–212
15. Dmitrieva Md et al. (2018) Object characterisation using wideband sonar pulses. PhD thesis, Heriot-Watt University
16. Warren JD, Stanton TK, Benfield MC, Wiebe PH, Chu D, Sutor M (2001) In situ measurements of acoustic target strengths of gas-bearing siphonophores. ICES J Mar Sci 58(4):740–749
17. Proctor CW, Roesler CS (2010) New insights on obtaining phytoplankton concentration and composition from in situ multispectral chlorophyll fluorescence. Limnol Oceanogr Methods 8(12):695–708
18. Kolber Z, Falkowski PG (1993) Use of active fluorescence to estimate phytoplankton photosynthesis in situ. Limnol Oceanog 38(8):1646–1665
19. Sullivan-Silva KB, Forbes MJ (1992) Behavioral study of zooplankton response to high-frequency acoustics. J Acoust Soc Am 92(4):2423–2423
20. Yu L, Xu L (2004) Calibration method of the red tide species coun precision based on digital microscope. Ocean Tech China 23(1):31–34 (Chinese)
21. Olson RJ, Sosik HM (2007) A submersible imaging-in-flow instrument to analyze nano- and microplankton: imaging flowcytobot. Limnol Oceanogr Methods 5(6):195–203
22. Orenstein EC, Beijbom O, Peacock EE, Sosik HM (2015) Whoi-plankton-a large scale fine grained visual recognition benchmark dataset for plankton classification. arXiv preprint arXiv:1510.00745
23. Li X, Cui Z (2016) Deep residual networks for plankton classification. In: OCEANS 2016 MTS/IEEE Monterey. IEEE pp 1–4
24. Py O, Hong H, Zhongzhi S (2016) Plankton classification with deep convolutional neural networks. In: 2016 IEEE information technology, networking, electronic and automation control conference. IEEE, pp 132–136
25. Rawat S.S, Bisht A, Nijhawan R (2019) A deep learning based cnn framework approach for plankton classification. In: 2019 fifth international conference on image information processing (ICIIP). IEEE, pp 268–273
26. Lampert CH, Nickisch H, Harmeling S (2014) Attribute-based classification for zero-shot visual object categorization. IEEE Trans Pattern Anal Mach Intell 36(3):453–465
27. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP). pp 1532–1543
28. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th international conference on neural information processing systems. Vol. 2, pp 3111–3119
29. Miller GA (1995) Wordnet: a lexical database for English. Commun ACM 38(11):39–41
30. Wang W, Zheng VW, Yu H, Miao C (2019) A survey of zero-shot learning: settings, methods, and applications. ACM Trans Intell Syst Technol (TIST) 10(2):1–37
31. Huang C, Loy C.C, Tang X (2016) Local similarity-aware deep feature embedding. In: Proceedings of the 30th international conference on neural information processing systems. pp 1270–1278
32. Norouzi M, Mikolov T, Bengio S, Singer Y, Shlens J, Frome A, Corrado GS, Dean J (2014) Zero-shot learning by convex combination of semantic embeddings. The 2nd International Conference on Learning Representations, ICLR 2014
33. Kodirov E, Xiang T, Gong S (2017) Semantic autoencoder for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 3174–3183
34. Deng J, Ding N, Jia Y, Frome A, Murphy K, Bengio S, Li Y, Neven H, Adam H (2014) Large-scale object classification using label relation graphs. In: European conference on computer vision. Springer, pp 48–64
35. Leksut JT, Zhao J, Itti L (2020) Learning visual variation for object recognition. Image Vis Comput 98:103912
36. Salakhutdinov R, Hinton G (2009) Semantic hashing. Int J Approx Reason 50(7):969–978
37. Wang X, Ye Y, Gupta A (2018) Zero-shot recognition via semantic embeddings and knowledge graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 6857–6866
38. Li Q, Sun X, Dong J, Song S, Zhang T, Liu D, Zhang H, Han S (2019) Developing a microscopic image dataset in support of intelligent phytoplankton detection using deep learning. ICES J Mar Sci 77(4):1427–1439. https://doi.org/10.1093/icesjms/fsz171
39. Charlson RJ, Lovelock JE, Andreae MO, Warren SG (1987) Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. Nature 326(6114):655–661
40. Farhadi A, Endres I, Hoiem D, Forsyth D (2009) Describing objects by their attributes. In: IEEE conference on computer vision and pattern recognition. IEEE, pp 1778–1785
41. Wah C, Branson S, Welinder P, Perona P, Belongie S (2011) The caltech-ucsd birds-200-2011 dataset. Adv Water Resour
42. Patterson G, Xu C, Su H, Hays J (2014) The sun attribute database: beyond categories for deeper scene understanding. Int J Comput Vis 108(1–2):59–81
43. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778
44. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
45. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
46. Lampert CH, Nickisch H, Harmeling S (2008) Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 951–958
47. Akata Z, Reed S, Walter D, Lee H, Schiele B (2015) Evaluation of output embeddings for fine-grained image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 2927–2936

48. Zhang Z, Saligrama V (2015) Zero-shot learning via semantic similarity embedding. In: Proceedings of the IEEE international conference on computer vision. pp 4166–4174