Relation-based Discriminative Cooperation Network for Zero-Shot Classification

Liu, Yang; Gao, Xinbo; Gao, Quanxue; Han, Jungong; Shao, Ling

Published in:
Pattern Recognition

DOI:
10.1016/j.patcog.2021.108024

Publication date:
2021

Citation for published version (APA):
Liu, Y., Gao, X., Gao, Q., Han, J., & Shao, L. (2021). Relation-based Discriminative Cooperation Network for Zero-Shot Classification. *Pattern Recognition*, *118*, Article 108024. https://doi.org/10.1016/j.patcog.2021.108024

Pattern Recognition

Relation-based Discriminative Cooperation Network for Zero-Shot Classification

--Manuscript Draft--

Manuscript Number:	PR-D-20-01091R3
Article Type:	VSI:Deep Relational Learning
Section/Category:	Objects and image analysis
Keywords:	zero-shot learning; Bias; Discriminative; Relation
Corresponding Author:	Yang Liu Xi‘an, CHINA
First Author:	Yang Liu
Order of Authors:	Yang Liu
	Xinbo Gao
	Quanxue Gao
	Jungong Han
	Ling Shao
Abstract:	Zero-shot learning (ZSL) aims to assign the category corresponding to the relevant semantic as the label of the unseen sample based on the relationship between the learned visual and semantic features. However, most typical ZSL models faced with the domain bias problem, which leads to unseen samples being easily misclassified into seen classes. To handle this problem, we propose a relation-based discriminative cooperation network (RDCN) model for ZSL in this work. The proposed model effectively utilize the structure of the space spanned by the cooperated semantics with the help of a set of relations. On the other hand, we devise the relation network to measure the relationship between the embedded semantic and visual features, and the validation information will guide the embedding module to learn more discriminative information. At last, the proposed RDCN model is validated on six benchmarks, and extensive experiments demonstrate the superiority of proposed method over most existing ZSL models on the standard zero-shot setting as well as the more realistic generalized zero-shot setting.

Title page

Relation-based Discriminative Cooperation Network for Zero-Shot Classification

Yang Liu. State Key Laboratory of Integrated Services Networks, Xidian University, Shaanxi 710071, China.

Email: yangl@xidian.edu.cn

Xinbo Gao. State Key Laboratory of Integrated Services Networks, Xidian University, Shaanxi 710071, China.

Email: xbgao@mail.xidian.edu.cn

Quanxue Gao. State Key Laboratory of Integrated Services Networks, Xidian University, Shaanxi 710071, China.

Email: qxgao@xidian.edu.cn

Jungong Han. WMG Data Science, University of Warwick, CV4 7AL Coventry, United Kingdom.

Email: jungonghan77@gmail.com

Ling Shao. Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates.

Email: ling.shao@ieee.org

Highlights

- The discriminative visual embedding preserves the discriminative information of the image features by separating the inter-classes and clustering the intra-classes with a margin.

- The discriminative semantic embedding acts as a pivot regularization to ensure the cooperated structures representative by utilizing semantic relations between classes.

- Extensive experimental evaluation on multiple datasets, including the large scale ImageNet shows that the proposed model performs favorably against state-of-the-art ZSL methods.

Relation-based Discriminative Cooperation Network for Zero-Shot Classification

Yang Liu[a], Xinbo Gao[a,b], Quanxue Gao[a], Jungong Han[c], Ling Shao[d]

[a]*State Key Laboratory of Integrated Services Networks, Xidian University, Shaanxi 710071, China.*
[b]*Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing, China*
[c]*Computer Science Department, Aberystwyth University, Aberystwyth SY23 3FL, U.K.*
[d]*Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates.*

Abstract

Zero-shot learning (ZSL) aims to assign the category corresponding to the relevant semantic as the label of the unseen sample based on the relationship between the learned visual and semantic features. However, most typical ZSL models faced with the domain bias problem, which leads to unseen or test samples being easily misclassified into seen or training categories. To handle this problem, we propose a relation-based discriminative cooperation network (RDCN) model for ZSL in this work. The proposed model effectively utilize the robust metric space spanned by the cooperated semantics with the help of a set of relations. On the other hand, we devise the relation network to measure the relationship between the visual features and embedded semantics, and the validation information will guide the embedding module to learn more discriminative information. At last, the proposed RDCN model is validated on six benchmarks, and extensive experiments demonstrate the superiority of proposed method over most existing ZSL models on the traditional zero-shot setting and the more realistic generalized zero-shot setting.

Keywords: Zero-shot learning, Bias, Discriminative, Relation

[*]Corresponding author
Email address: yangl@xidian.edu.cn (Yang Liu)

1. Introduction

Humans can recognize many categories, including about 30,000 basic categories and even more subcategories. On the other hand, humans are also very good at recognizing objects even if they have never seen any of their examples. For example, if a child has seen cattle before, he/she can easily recognize a cow and learn that a cow looks like cattle with black-and-white color. Inspired by the ability of humans to identify unseen categories, that the research area of Zero-Shot Learning (ZSL) [1, 2, 3, 4] aims to recognize classes whose samples haven't been available during training time has received increasing interests.

Different from traditional supervised learning, ZSL considers an extreme case where testing classes is unavailable during training, *i.e.,* the training (seen) classes and testing (unseen) classes are disjoint [5]. ZSL links the seen classes and unseen classes through the semantic information to complete the recognition task. The semantic is defined as a high dimensional vector space where unseen and seen classes are connected together, which can be a semantic attribute space [6, 7, 8] or a semantic word vector space [9, 10]. In adddition, test images may come from both unseen and seen classes, which is named Generalized Zero-Shot Learning (GZSL). In real-world applications, since we cannot predict whether a new sample comes from an unseen class or a seen class, GZSL is more practical and challenging than ZSL.

Regarding the bridge between visual space and semantic space, most traditional ZSL methods tend to learn a mapping that project samples from the visual space to the semantic space with the labelled training set including seen classes only. When classifying unseen images, the learned embedding is used to project the visual representation of unseen samples into the semantic space including unseen and seen classes. Then the Nearest Neighbor (NN) search method is used to recognize the sample of unseen class, which is the testing process. However, the NN search method is easy to cause hubness problem [11]. To solve this problem, Sung *et al.* [12] recently proposed a model named Relation Network (RN) to compare the test samples with the embedded semantics in a self-adaptive way. Different from NN search, RN tries to measure the relation score between unseen samples and semantics by learning a distance metric.

Unfortunately, there is a strong domain bias problem [13] when applying almost all standard ZSL models to deal with a GZSL task, which leads to unseen images being misclassified into seen classes. To alleviate this problem, Zhang and Shi [14] proposed a Co-Representative network (CRnet) based on RN. CRnet tries to learn a uniform embedding space by a single-layer module with parallel structure and high local linearity.

However, CRnet still relies heavily on obtaining human-defined semantics for knowledge transfer. Similar to other embedding models, CRnet focuses on the original visual features and semantics, but completely ignores the discriminative information among them. Although a few works [15, 16, 17] have been proposed to maintain the relationship between semantics by using complementary features or extracting deep local features, they are rarely used in ZSL task. To address this point, we formulate a novel framework named Relation-based Discriminative Cooperation Network (RDCN) for ZSL task in this paper.

The RDCN model aims to preserve the discriminative information of the visual features and semantics. At first, RDCN adopts the encoder-decoder paradigm to obtain discriminative visual features. Specifically, the encoder aims to learn a mapping from the visual space to the embedding space where the distance between classes is adjusted by the triplet loss [18], while decoder reconstructs the original input features. On the other hand, RDCN uses a decomposition structure to alleviate the bias problem in the semantic space, and adopts a novel semantic pivot regularization to obtain discriminative semantic features. At last, RDCN adopts a relation network as the similarity function to measure the relationship between the discriminative visual features and semantics. In summary, our contributions are concluded into the following three-fold:

- The discriminative visual embedding preserves the discriminative information of the input image features by separating inter-classes and clustering intra-classes with a margin.

- The discriminative semantic embedding acts as a pivot regularization to ensure the cooperated structures representative by utilizing semantic relations between classes.

- The experimental evaluation on several popular datasets, including the ImageNet demonstrates that the proposed RDCN performs favorably against state-of-the-art ZSL models.

The remainder of this paper is organized as follows: Section 2 reviews the related ZSL work followed by Section 3 which describes the proposed model in detail. Experimental results with some detailed analysis are given in Section 4. At last, the conclusion of the work is given in Section 5.

2. Related Work

In this section, we introduce the related works of zero-shot learning from two aspects: relation-based models and synthesis-based models.

2.1. Relation-based Models

Relation-based models aims to learn the relationship between images and the semantics. In early works of ZSL, most algorithms focus on building this relationship by Visual-Semantic Embedding (VSE) framework. According to different directions of mappings, the VSE framework is divided into three types as follows.

(1) *Visual→Semantic Embedding* tries to learn a mapping from the visual space to the semantic space either using linear function [19, 20, 21, 22] or by deep neural network regression [23]. For example, Deep Visual-Semantic Embedding (DeViSE) model [24] is one of the earlier attempts. It inputs Convolutional Neural Networks (C-NN) [25] and Word2Vec [26] features to learn an end-to-end deep classification model. Socher *et al.* [27] mapped the improved visual features to the semantic space by the two-layer or three-layer neural network, and used the least squares loss to train the network.

(2) *Semantic→Visual Embedding* tries to learn a mapping from the semantic space to the visual space, such as [28, 29, 30]. The training and test process are similar with the first mapping manner. For example, Deep Transductive Network (DTN) [31] exploits the high confidence assignments with the assistance of an auxiliary target distribution to reduce the impact of the hubness problem [11]. Shojaee *et al.* [32] proposed a

semi-supervised ZSL method based on the visual features of similar samples clustered together in the visual space.

(3) *Visual→Common Space←Semantic* learns a common space where both the visual features and the semantics are embedded to, such as [33, 34, 35, 36]. For the test phase, the visual features and attributes are embedded into the common space for final classification task. Semantic Similarity Embedding (SSE) [37] is a typical method which learns an embedding for visual features and semantics to find mixture features that used to measure the similarity.

Different from traditional VSE framework, Sung *et al.* [12] recently proposed a Relation Network (RN) that aims to compare the query visual features with the target embedded semantics through a self-adaptive way. Inspired by RN, Zhang and Shi [14] proposed a Co-Representative network (CRnet) with the help of high local linearity. In this paper, the proposed model extract discriminative visual and semantic features by VSE framework and measure their relationship with RN framework.

2.2. Synthesis-based Models

Synthesis-based models tries to learn a generator that generates samples from the semantics, and to then train a classifier to predict classes with the generated synthesis samples. For example, f-CLSWGAN [38] generates sufficiently discriminative CNN features to train softmax classifiers with the help of a Wasserstein Generative Adversarial Networks (WGAN) [39]. CVAE-ZSL [40] implements the generation by a Conditional Variational AutoEncoder (CVAE) [41]. SE-GZSL [42] also designs a generator based on a Variational AutoEncoder (VAE) [43] but generates synthesis samples in a feedback-driven way. Recently, Zero-VAE-GAN [5] is proposed to convert ZSL problems into supervised tasks by a combination of VAE and GAN.

The advantage of synthesis-based models is that both seen and unseen samples may be obtained by generator, which contributes to significantly alleviate the domain bias problem. However, synthesis-based methods require the generator to generate a large number of high-quality and diverse samples (including unseen ones) for each class, which is costly and requires additional classifier learning. In this work, we aim to

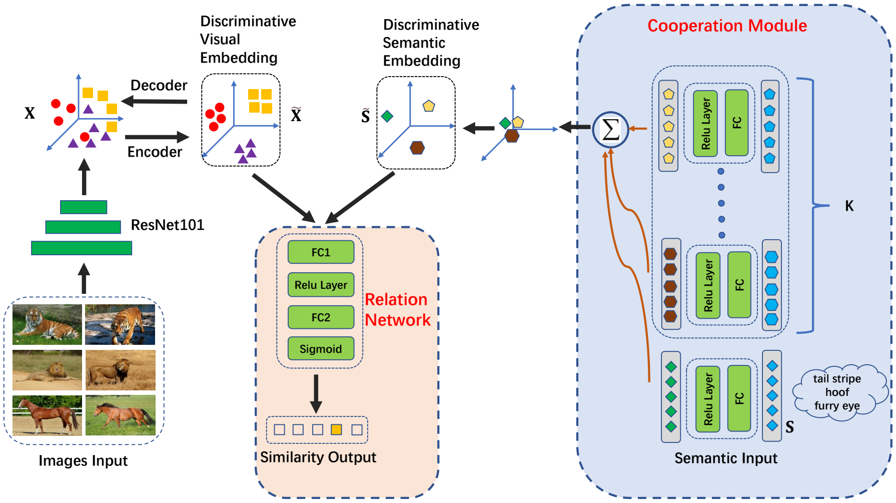


Figure 1: The framework of RDCN.

achieve high performance in both ZSL and GZSL tasks by the end-to-end relation-based model.

3. Proposed Approach

In this section, firstly, we provide the problem definition of ZSL by mathematical notation. Secondly, we give the model structure and detailed description of different modules. At last, we conclude the overall objective function.

3.1. Problem Definition

Suppose there are n labeled samples with c seen classes $\{\mathbf{X}, \mathbf{S}, \mathbf{Y}\}$ and n_u unlabeled samples with c_u unseen classes $\{\mathbf{X}^u,\mathbf{S}^u,\mathbf{Y}^u\}$, where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbf{R}^{d\times n}$ and $\mathbf{X}^u = [\mathbf{x}^u_1, \mathbf{x}^u_2, \cdots, \mathbf{x}^u_{n_u}] \in \mathbf{R}^{d\times n_u}$ are d-dimensional visual features, while the corresponding labels are $\mathbf{Y}$ and $\mathbf{Y}^u$, respectively. It is noteworthy that the labels of seen and unseen samples are disjoint, *i.e.*, $\mathbf{Y} \cap \mathbf{Y}^u = \emptyset$. $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_c] \in \mathbf{R}^{k\times c}$ and $\mathbf{S}^u = [\mathbf{s}^u_1, \mathbf{s}^u_2, \cdots, \mathbf{s}^u_{c_u}] \in \mathbf{R}^{k\times c_u}$ are k-dimensional semantic features of seen and unseen samples. The ZSL task aims to learn a classifier $f : \mathbf{X}^u \to \mathbf{Y}^u$, where classes of testing data $\mathbf{X}^u$ are unavailable in training phrase.

3.2. Model Architecture

Figure 1 shows the framework of proposed model RDCN. Specifically, the RDCN consists of four parts: (1) The visual features $\mathbf{X}$ extracted by ResNet101 are encoded with the help of the discriminative visual embeddings $\tilde{\mathbf{X}}$. (2) The discriminative visual embeddings are decoded to reconstruct the input visual features $\mathbf{X}$. (3) The cooperated semantic features obtained by cooperation network are send into a discriminative semantic embedding space, where $\tilde{\mathbf{S}}$ is generated. (4) A relation network is adopted as the similarity function to measure the relationship between the discriminative visual features $\tilde{\mathbf{X}}$ and discriminative semantics $\tilde{\mathbf{S}}$.

3.3. Encoder

The deep image features $\mathbf{X}$ are trained by following function to obtain the discriminative visual embeddings:

$$\tilde{\mathbf{X}} = f_e\left(\mathbf{X}; \theta_e\right), \tag{1}$$

where f_e indicates the operation of the encoder whose parameters are denoted by θ_e. In detail, the deep visual features $\mathbf{X}$ extracted by RessNet101 pass through multilayer perceptron (MLP) with two hidden layers (h1 = 1024-D and h2 = 512-D), followed by a dense layer with the LeakyReLU [44] activation. The output discriminative visual embeddings $\tilde{\mathbf{X}}$ have the same dimension with the semantic embeddings.

3.4. Discriminative Visual Embedding

Most embedding models that solve the ZSL problem focus on calculating a typical description of images in all classes, which makes the encoder non-discriminatory. Motivated by [45], adding discriminative embedding operations in the encoding process can make the learned low-dimensional features more discriminative, which is helpful for classification.

In order to ensure that an embedding visual feature $\tilde{\mathbf{x}}_i$ $(\tilde{\mathbf{x}}_i \in \mathbf{X})$ is closer to each image feature $\tilde{\mathbf{x}}_j$ from the same class than any image feature $\tilde{\mathbf{x}}_k$ from different classes. We use the triplet loss [18] to learn a discriminative embedding by adjusting the intra-class and inter-class distance between the learned features:

$$\ell_{tri} = \frac{1}{n}\sum_{i=1}^{n} \max\left(0, m + d_{intra} - d_{inter}\right), \tag{2}$$

Where $m > 0$ is a margin that is enforced between positive (same class) and negative (different classes) pairs. d_{intra} denotes the squared Euclidean distance between $\tilde{\mathbf{x}}_i$ and visual features from the same class, meanwhile, d_{inter} denotes the squared Euclidean distance between $\tilde{\mathbf{x}}_i$ and visual features from the different class, *i.e.,*

$$d_{intra} = \sum_{\tilde{\mathbf{x}}_j \in c_i} \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\|_2^2, \tag{3}$$

$$d_{inter} = \sum_{\tilde{\mathbf{x}}_k \notin c_i} \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_k\|_2^2. \tag{4}$$

3.5. Decoder

The encoder and the decoder are connected by the discriminative visual embeddings. The discriminative visual feature $\tilde{\mathbf{X}}$ is the input of the decoder, then the reconstructed visual feature is denoted by following equation:

$$\hat{\mathbf{X}} = f_d\left(\tilde{\mathbf{X}}; \theta_d\right), \tag{5}$$

where f_d indicates the operation of the decoder whose parameters are denoted by θ_d. Similar with the encoder, the decoder is a multilayer perceptron (MLP) includes two hidden layers (h1 = 512-D and h2 = 1024-D) and a dense layer with the LeakyReLU activation.

Since the proposed framework involves a decoder which reconstructs the original visual features, there is an accompanying reconstruction loss:

$$\ell_{rec} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2, \;\; \hat{\mathbf{x}}_i \in \hat{\mathbf{X}}. \tag{6}$$

3.6. Cooperation Module

Motivated by the CRnet [14], we use a decomposition structure to alleviate the bias problem in cooperation module. Specifically, We first adopt the unsupervised K-means clustering method to divide the semantic features $\mathbf{S}$ into $K(K < c)$ subsets. The clustering center of k-th subset is denoted by $\bar{\mathbf{s}}_k(k = 1, 2, \cdots, K)$. Then, the semantic feature $\mathbf{s}_i(i = 1, 2, \cdots, c)$ with a combination of K clustering center are trained by following cooperation module to obtain the discriminative semantic embedding:

$$\tilde{\mathbf{s}}_i = \sum_{k=0}^{K} f_c\left(\left[\mathbf{s}_i - \bar{\mathbf{s}}_k\right]; \theta_c\right), \tag{7}$$

where f_c indicates the operation of the cooperation module whose parameters are denoted by θ_c. In detail, the vector calculated by $\mathbf{s}_i - \bar{\mathbf{s}}_k$ is fed into a single FC layer with a ReLU activation. Moreover, different from the CRnet, we also add the original semantic features as the input to retain its own sparsity feature, *i.e.,* $\bar{\mathbf{s}}_0 = \mathbf{0}$.

3.7. Discriminative Semantic Embedding

Intuitively, maximizing the distance of the semantic embeddings by following function can maintain the discriminative information between different semantic features,

$$\max \sum_{i=1}^{c} \sum_{j=1}^{c} \|\tilde{\mathbf{s}}_i - \tilde{\mathbf{s}}_j\|_2^2, \tag{8}$$

where $\tilde{\mathbf{s}}_i$ and $\tilde{\mathbf{s}}_j$ are different semantic embeddings. In order to reduce the computational complexity of the above calculation methods, a semantic pivot is used to simplify the calculation. The semantic pivot of semantics is defined as the center of semantic embedding. It can be solved by the average embedded feature, or it can be calculated by the mean shift technique. In fact, there is almost no difference in performance between these two calculations [17]. For simplicity, in this paper, the semantic pivot $\bar{\mathbf{s}}$ is calculated by the center of the semantic embeddings, *i.e.,* $\bar{\mathbf{s}} = \frac{1}{c} \sum_{i=1}^{c} \tilde{\mathbf{s}}_i$. Then, we get the following loss function:

$$\ell_{piv} = - \sum_{i=1}^{c} \|\tilde{\mathbf{s}}_i - \bar{\mathbf{s}}\|_2^2. \tag{9}$$

3.8. Relation Module

After obtaining discriminative visual feature $\tilde{\mathbf{x}}_i$ and discriminative semantic feature $\tilde{\mathbf{s}}_j$, we adopt the RN [12] to measure their relationship. Specifically, the relation module is a two-layer neural network and the input is the concatenation of $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{s}}_j$; the hidden layer increase nonlinearity and the output of the network is a scalar in range of 0 to 1 representing the similarity between discriminative visual and semantic features, which is called relation score.

In this module, we adopt RN as the similarity function $g(\cdot)$ and follow the original settings. Thus the output relation score of training pairs $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{s}}_j$ be denoted as

$g(\tilde{\mathbf{x}}_i, \tilde{\mathbf{s}}_j)$. In the training process, we randomly sample the entire training set to generate training pairs of $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{s}}_j$, and control the ratio of matched pairs(discriminative visual and semantic features belong to the same class) to mismatched pairs(discriminative visual and semantic features come form different classes) at about $1 : 30$. The similarities of matched pairs and mismatched pairs are set to 1 and 0, respectively. The relation module is trained by mean square error loss:

$$\ell_{rel} = \sum_{j=1}^{c} \sum_{i=1}^{n} \left[g(\tilde{\mathbf{x}}_i, \tilde{\mathbf{s}}_j) - l(\tilde{\mathbf{x}}_i, \tilde{\mathbf{s}}_j)\right]^2 \tag{10}$$

where $l(\cdot)$ is the similarity ground-truth, $l(\tilde{\mathbf{x}}_i, \tilde{\mathbf{s}}_j) = 1$ when $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{s}}_j$ belong to the same class, and $l(\tilde{\mathbf{x}}_i, \tilde{\mathbf{s}}_j) = 0$ when $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{s}}_j$ belong to different classes.

3.9. Overall Objective

With the objective functions introduced above, the overall objective of the proposed model is given by:

$$\ell = \ell_{rel} + \alpha\ell_{tri} + \beta\ell_{rec} + \gamma\ell_{piv}, \tag{11}$$

where α, β and γ are trade-off parameters chosen based on the validation dataset.

Given a testing image $\mathbf{x}_k^u$, its label can be inferred by:

$$j^* = \arg\max_j g(\tilde{\mathbf{x}}_k^u, \tilde{\mathbf{s}}_j). \tag{12}$$

For ZSL task, $\tilde{\mathbf{s}}_j$ refers to the semantic embeddings of only the unseen classes, *i.e.*, $j\epsilon\{1, 2, \cdots, c_u\}$, and for GZSL task, $\tilde{\mathbf{s}}_j$ refers to the semantic embeddings of both seen as well as unseen classes, *i.e.*, $j\epsilon\{1, 2, \cdots, c + c_u\}$. c and c_u are the numbers of seen and unseen classes, respectively.

4. Experiments

In this section, firstly, we introduce five popular ZSL datasets SUN, CUB, AWA1, AWA2, aPY and a large-scale ImageNet dataset. Secondly, we provide the implementation details of the architecture. Thirdly, some experimental results with related analysis on the traditioal zero-shot setting and the more realistic generalized zero-shot setting are given. Finally, we show some visualized results for the proposed model.

4.1. Datasets Descriptions

Five Small-scale Attribute datasets: SUN Attribute (SUN) [46] is a fine-grained dataset and it consists of 14,340 images belonging to 717 classes annotated with 102 attributes. Following the ZSL setting in [6], 72 out of 717 classes are as unseen categories and the rest of 645 categories as seen categories.

CUB-200-2011 Birds (CUB) [47] is a fine-grained and medium scale dataset which has in total 11,788 images distributed in 200 bird categories. Each class is annotated with a 312-dim attribute vector. We follow the standard ZSL split with 150 categories for seen classes and 50 for unseen classes as in [6].

Animals with Attributes 1 (AWA1) [6] is a kind of coarse-grained dataset, which includes a total of 30,475 images belonging to 50 classes. Each class is annotated with a 85-dim attribute vector, where 40 categories (seen) are used for training and rest 10 categories (unseen) for testing. Animals with Attributes 2 (AWA2) [13] has the same 50 categories as AWA1 dataset. However, AWA2 dataset contains 37,322 images. Similar to AWA1, 40 categories are used for seen classes and 10 categories are used for unseen classes.

A Pascal and Yahoo (aPY) [48] is a kind of small-scale coarse-grained dataset. Each category is annotated with a 64-dim attribute vector. Among the total number of 32 classes, 20 Pascal classes are used for training and 12 Yahoo classes are used for testing.

One large-scale dataset: ImageNet [49] has a total of 218,000 images. 21,841 classes with more than 10 million images, where 1k classes containing 1.2 million images are used for training the mapping. There are different splits in the test. Specifically, 2-hops/3-hops refers to test classes belonging to 2/3 tree hops away from 1k train classes in the WordNet hierarchy, which contains 1,509/7,678 unseen classes. Such classes that contain the top 500/1k/5k maximum images and top 500/1k/5k minimum images are given for test splits respectively. At last, all 20K classes are given for testing, which is a challenging task.

4.2. Implementation Details

We use ReLU activation for all layers except for the output of the encoder and the decoder, which adopt LeakyReLU activations with the negative slope of 0.3. A single-layer FC network compared with K parallel single-layer FC network are given for embedding the original and cooperated semantic vectors, respectively. Parameters α and β in our objective function are fine-tuned in the range $\left[5 \times 10^{-6}, 10^{-2}\right]$ and γ from $\left[10^{-7}, 5 \times 10^{-4}\right]$. Moreover, the K value is given in the range $[3, 12]$. For relation module, the discriminative visual and semantic features are concatenated with a hidden layer before passing relation network, We adopt Adam optimizer [50] with a initialized learning rate of 10^{-3} and a weight decay of 5×10^{-5}. For fair comparison, we follow the settings in [13] to split each dataset for training and testing. Moreover, each image is represented by 2048-dim vector extracted by 101-layered ResNet, *i.e.,* ResNet101 [51].

4.3. Zero-Shot Learning (ZSL) Experiments

In this work, the average per-class top-1 accuracy is adopted as the evaluation criteria for zero-shot classification, *i.e.,* we average the correct predictions independently for each class by follows:

$$acc\Upsilon = \frac{1}{\|\Upsilon\|} \sum_{c=1}^{\|\Upsilon\|} \frac{\#correct\ predictions\ in\ c}{\#samples\ in\ c} \tag{13}$$

where Υ and $\|\Upsilon\|$ indicate the set of classes and number of classes with corresponding dataset, respectively. So Υ includes all the test classes *i.e.,* the unseen classes for ZSL task.

The results of the different ZSL models on five popular small-scale datasets is given in Table 1. We can see that the proposed RDCN consistently performs better than compared methods, and RDCN gets the state-of-the-art on four datasets: SUN, AWA1, AWA2 and CUB. Specifically, the accuracies increase of 2.7% and 4.2% compared to the strongest competitor on SUN dataset and AWA2 dataset, respectively. We also observe a significant increase when we include all of the ℓ_{tri}, ℓ_{rec} and ℓ_{piv} in our model. This indicates that the reconstruction term makes a contribution to vary levels

Table 1: Zero-shot learning (ZSL) results on five small-scale attribute datasets. The results report average per-class Top-1 accuracy in %.

Method	SUN	CUB	AWA1	AWA2	aPY
DeViSE [24]	56.5	52.0	54.2	59.7	39.8
CONSE [52]	38.8	34.3	45.6	44.5	26.9
CMT [9]	39.9	34.6	39.5	37.9	28.0
SP-AEN [53]	59.2	55.4	-	58.5	24.1
PSR [23]	61.4	56.0	-	63.8	38.4
DCN [54]	61.8	56.2	65.2	-	**43.6**
CCSS [55]	56.8	44.1	56.3	63.7	35.5
DAP [6]	39.9	40.0	44.1	46.1	33.8
IAP [6]	19.4	24.0	35.9	35.9	36.6
SSE [37]	51.5	43.9	60.1	61.0	34.0
LATEM [56]	55.3	49.3	55.1	55.8	35.2
ALE [57]	58.1	54.9	59.9	62.5	39.7
SJE [58]	53.7	53.9	65.6	61.9	32.9
ESZSL [20]	54.5	53.9	58.2	58.6	38.3
SYNC [59]	56.3	55.6	54.0	46.6	23.9
SAE [19]	53.4	42.0	58.1	50.3	32.9
f-CLSWGAN [38]	58.5	57.7	64.1	-	-
TVN [60]	59.3	54.9	64.7	-	40.9
DVN [61]	62.4	57.8	67.7	-	41.2
Zhang's [62]	60.4	53.2	67.4	-	42.8
RDCN ($\alpha = 0$)	58.9	55.3	66.0	65.3	36.1
RDCN ($\beta = 0$)	60.6	56.5	67.9	66.1	37.9
RDCN ($\gamma = 0$)	59.3	56.1	69.3	65.2	39.5
RDCN	**65.1**	**60.5**	**71.6**	**68.0**	42.1

of gain fatures and the discriminative information among visual and semantic features is also essential.

4.4. Generalized Zero-Shot Learning (GZSL) Experiments

GZSL means that the search space includes both test classes (Υ^{ts}) and training classes (Υ^{tr}). At first, the average per-class top-1 accuracy on Υ^{tr} and Υ^{ts} can be obtained by Eq. (13), then the harmonic mean is computed by follows:

$$H = \frac{2 \times acc\Upsilon^{tr} \times acc\Upsilon^{ts}}{acc\Upsilon^{tr} + acc\Upsilon^{ts}} \tag{14}$$

where $acc\Upsilon^{tr}$ and $acc\Upsilon^{ts}$ are accuracies of samples from Υ^{tr} and Υ^{ts}, respectively.

The GZSL results on five popular attribute datasets is given in Table 2. We have following observations according to the results:

Table 2: Generalized Zero-Shot Learning (GZSL) results on five small-scale attribute datasets. ts = $acc\left(\Upsilon^{ts}\right)$, tr = $acc\left(\Upsilon^{tr}\right)$, H = harmonic mean. We measure Top-1 accuracy in %.

	SUN			CUB			AWA1			AWA2			aPY		
Method	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H
DeViSE [24]	16.9	27.4	20.9	23.8	53.0	32.8	13.4	68.7	22.4	17.1	74.7	27.8	4.9	76.9	9.2
CMT [9]	8.1	21.8	11.8	7.2	49.8	12.6	0.9	87.6	1.8	0.5	90.0	1.0	1.4	85.2	2.8
CONSE [52]	6.8	39.9	11.6	1.6	72.2	3.1	0.4	88.6	0.8	0.5	90.6	1.0	0.0	**91.2**	0.0
DAP [6]	4.2	25.1	7.2	1.7	67.9	3.3	0.0	88.7	0.0	0.0	84.7	0.0	4.8	78.3	9.0
IAP [6]	1.0	37.8	1.8	0.2	**72.8**	0.4	2.1	78.2	4.1	0.9	87.6	1.8	5.7	65.6	10.4
GFZSL [34]	0.0	39.6	0.0	0.0	45.7	0.0	1.8	80.3	3.5	2.5	80.1	4.8	0.0	83.3	0.0
SSE [37]	2.1	36.4	4.0	8.5	46.9	14.4	7.0	80.5	12.9	8.1	82.5	14.8	0.2	78.9	0.4
LATEM [56]	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	11.5	77.3	20.0	0.1	73.0	0.2
ALE [57]	21.8	33.1	26.3	23.7	62.8	34.4	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7
SJE [58]	14.7	30.5	19.8	23.5	59.2	33.6	11.3	74.6	19.6	8.0	73.9	14.4	3.7	55.7	6.9
ESZSL [20]	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
SYNC [59]	7.9	**43.3**	13.4	11.5	70.9	19.8	8.9	87.3	16.2	10.0	90.5	18.0	7.4	66.3	13.3
SAE [19]	17.8	32.0	22.8	18.8	58.5	28.5	14.2	81.2	24.1	16.7	82.5	27.8	9.9	74.7	17.5
ZSKL [29]	19.8	29.1	23.6	19.9	52.5	28.9	18.3	79.3	29.8	17.6	80.9	29.0	11.9	76.3	20.5
f-CLSWGAN [38]	42.6	36.6	39.4	43.7	57.7	49.7	57.9	61.4	59.6	-	-	-	-	-	-
CVAE-ZSL [40]	-	-	26.7	-	-	34.5	-	-	47.2	-	-	51.2	-	-	-
SE-GZSL [42]	40.9	30.5	34.9	41.5	53.3	46.7	56.3	67.8	61.5	58.3	68.1	62.8	-	-	-
TVN [60]	18.2	28.9	22.3	21.6	47.5	29.7	18.2	87.5	30.2	-	-	-	8.8	59.1	15.4
DVN [61]	25.3	34.6	29.2	26.2	55.1	35.5	34.9	73.4	48.5	-	-	-	13.7	72.2	23.1
RN [12]	-	-	-	38.1	61.1	47	31.4	**91.3**	46.7	30	**93.4**	45.3	-	-	-
CRnet [14]	34.1	36.5	35.3	45.5	56.8	50.5	58.1	74.7	65.4	52.6	78.8	63.1	32.4	68.4	44
Zhang's [62]	**39.7**	38.9	39.3	37.8	58.2	45.9	37.0	84.7	51.4	-	-	-	25.9	79.5	39.1
RDCN	37.3	37.7	**37.5**	**45.5**	58.1	**51.0**	**60.2**	79.0	**68.3**	**56.6**	72.3	**63.5**	**34.0**	75.6	**46.9**

(1) Compared with Table 1, ZSL results are higher than GZSL results ("ts" value). The main reason is that all seen classes are included in the search space and these seen classes confuse the test images. That is to say, an image from unseen class is more likely to be mistaken for a seen class when it is projected into the semantic space in GZSL task.

(2) The "tr" value in Table 2 just represents the classification performance in the seen dataset. Moreover, high accuracy on "tr" is often accompanied by low accuracy on "ts" and "H" such as IAP and SYNC, which indicates that these models perform well most seen classes but fails to generalize for unseen classes, *i.e.*, overfitting.

(3) With respect to the state-of-the-art, RDCN gets best "H" value almost on all

Table 3: GZSL comparisons (ts) in ImageNet dataset. The results report Top-10 accuracy in %.

	Hierarchy All		Most populated			Least populated			All
Method	2-hops	3-hops	500	1k	5k	500	1k	5k	20k
CONSE [52]	0.86	7.14	23.47	18.38	9.92	0.00	0.00	0.66	3.43
CMT [9]	7.80	2.77	9.65	7.73	3.83	3.37	2.71	1.45	1.25
LATEM [56]	16.99	6.28	23.61	18.65	8.73	8.73	7.60	3.50	2.71
ALE [57]	17.79	6.34	24.93	19.37	9.12	10.38	8.46	3.63	2.77
DeViSE [24]	17.59	6.28	24.66	19.11	8.99	10.11	8.26	3.63	2.71
SJE [58]	17.46	6.21	23.61	18.45	8.79	9.85	8.00	3.50	2.71
ESZSL [20]	19.24	6.81	26.52	20.56	9.72	9.12	7.73	3.76	3.10
SYNC [59]	14.55	5.62	16.33	13.82	7.87	2.77	2.44	1.78	2.64
SAE [19]	13.55	4.82	20.76	16.60	7.60	3.43	2.57	1.58	2.24
PQZSL [63]	21.80	7.41	29.30	23.75	11.3	9.42	7.87	3.72	3.45
RDCN	**25.69**	**10.15**	**33.71**	**26.91**	**15.65**	**11.33**	**10.37**	**7.83**	**7.31**

datasets. In detail, RDCN obtains 68.3% on AWA1 dataset and 46.9% on aPY dataset, which is better than the next best model CRnet by 2.9%. On AWA2 dataset, RDCN gets a best accuracy of 56.6% on the first setting ("ts" value) and 63.5% overall. In addition, RDCN achieves better results compared to some synthesis-based models like f-CLSWGAN, CVAE-ZSL, SE-GZSL and so on.

Moreover, Table 3 reports the result of GZSL ("ts" value) on ImageNet. Compared with some baselines, RDCN obtains the best performance in most splits, which proves the superiority of the proposed method on large datasets. The whole GZSL experimental results supports our hypothesis that discriminative visual and semantic information are beneficial for generalized zero-shot recognition.

4.5. Ablation Studies

In this subsection, we compare the RDCN with its different variants to study the role of each item in the objective function 11. The experimental results are shown in Table 1. We analyze the following three cases: 1). "RDCN ($\alpha = 0$)" means there is no triplet loss ℓ_{tri} in the objective function 11; 2). "RDCN ($\beta = 0$)" means there is no reconstruction loss ℓ_{rec} in the objective function 11; 3). "RDCN ($\gamma = 0$)" means there is no semantic pivot regularization ℓ_{piv} in the objective function 11.

We observe in Table 1 that each kind of strategy of RDCN can improve the ZSL classification performances effectively. In addition, The role of the triplet loss ℓ_{tri} is

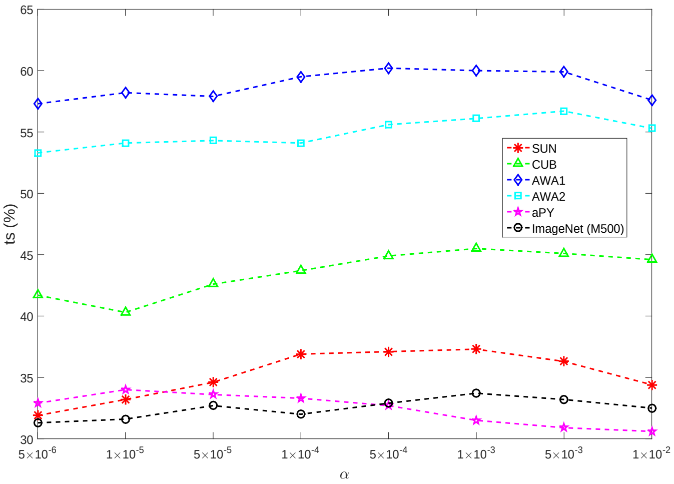


Figure 2: The influence of α on six datasets. α is the parameter of the triplet loss ℓ_{tri} in the objective function.

more important than that of reconstruction loss ℓ_{rec} and semantic pivot regularization ℓ_{piv}, which is based on the fact that the result of "RDCN ($\alpha = 0$)" is worse than that of "RDCN ($\beta = 0$)" and "RDCN ($\gamma = 0$)". According to the results of the last four rows in Table 1, we can be see that each item in the objective function plays a positive role in the ZSL classification task.

For the RDCN, there are three parameters, *i.e.,* α, β and γ in the objective function. By varying one of the parameters while fixing the other parameters, we run the model for 100 epochs and produce the GZSL results ("ts" value). Specifically, we conduct experiments varying α and β from $\left[5 \times 10^{-6}, 10^{-2}\right]$ and γ from $\left[10^{-7}, 5 \times 10^{-4}\right]$. The influence of α, β and γ on each dataset are illustrated in Figure 2, Figure 3 and Figure 4, respectively. For the ImageNet, due to the large number of testing samples (20K classes) in the complete dataset, we selected top 500 maximum images (M500) as test splits for analysis. According to the "ts" results under different values of three parameters α, β and γ, we conclude that the RDCN can obtain promising performance within a small range of parameters.

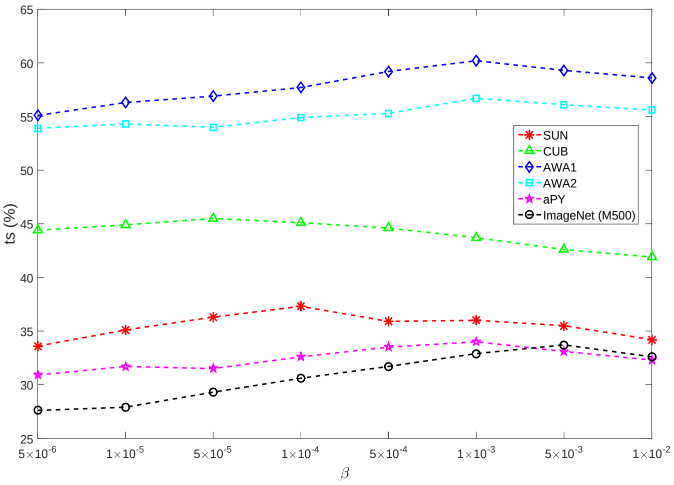


Figure 3: The influence of β on six datasets. β is the parameter of the reconstruction loss ℓ_{rec} in the objective function.

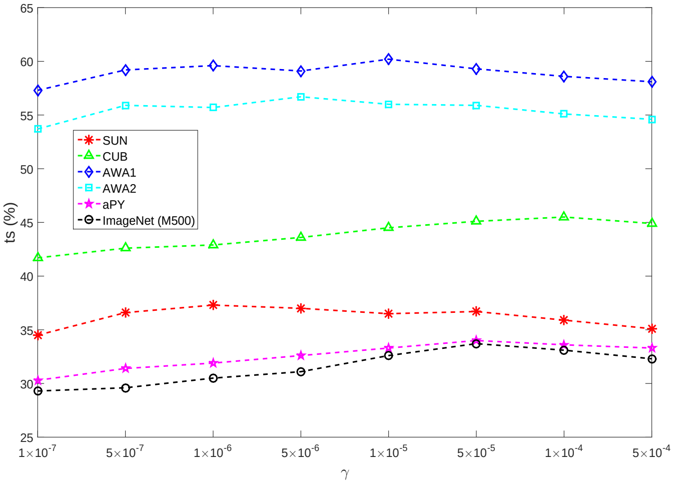


Figure 4: The influence of γ on six datasets. γ is the parameter of the semantic pivot regularization ℓ_{piv} in the objective function.

4.6. Visualized results

We further provide some visualized results for the proposed method. Figure 5 shows the confusion matrices of unseen classes on aPY dataset.

According to the results on Figure 5, we can see the proposed model RDCN can

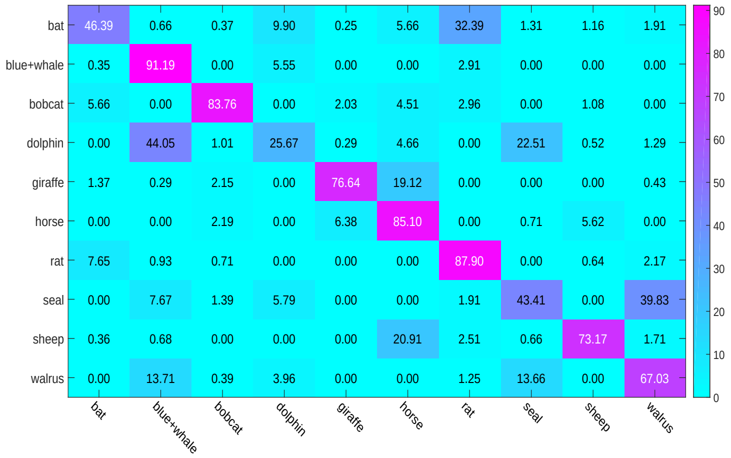


Figure 5: Confusion matrices for unseen classes of the proposed model on the AWA2 dataset.

identify most of unseen classes, except “bat” (46.39%), “dolphin” (25.67%) and “seal” (43.41%) on AWA2 dataset. We also observe that RDCN achieves appealing results on some classes, such as “blue+whale” (91.19%), “rat” (87.90%) and “horse” (85.10%). Considering the unseen samples are unavailable in training process, it strongly supports the superiority of the proposed method for ZSL task.

The t-SNE model [64] is used to project samples and prototypes from the semantic space to the 2D plane. Its main function is to visualize the distance between the sample and the corresponding class prototype. We selected seven seen classes and five unseen classes from the AWA2 dataset to check whether the prototype was learned correctly. Figure 6 and Figure 7 give the visualization results. It can be seen intuitively that most of the samples are located near the prototype of the corresponding class, which indicates that the RDCN can learn proper mapping from the feature space to the semantic space.

5. Conclusion

In this work, we have proposed a relation-based discriminative cooperation network to address the zero-shot classification problem. It keeps the discriminative information by separating the inter-classes and cluster the intra-classes with a margin. In addition, a

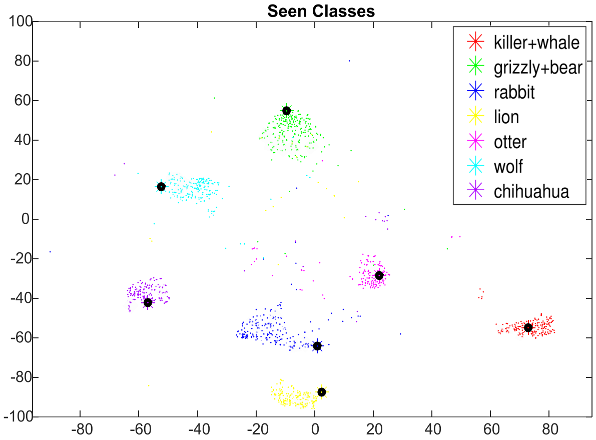


Figure 6: The tSNE visualisation of the visual features of training seen class samples from the AwA2 dataset together with the projected class prototypes for the proposed model. Prototypes is denoted by ”*” and black circles are used to mark them visible.

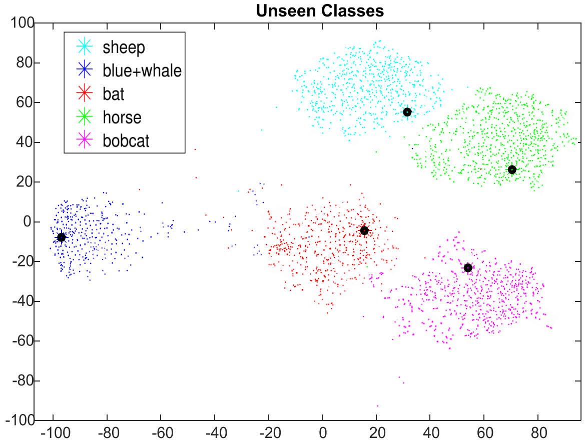


Figure 7: The tSNE visualisation of the visual features of test unseen class samples from the AwA2 dataset together with the projected class prototypes for the proposed model. Prototypes is denoted by ”*” and black circles are used to mark them visible.

pivot regularization is utilized to ensure the cooperated semantic structures discriminative. Finally, relation module is introduced to measure the relationship between visual and semantic features. Experimental results on six benchmarks with multiple settings including both ZSL and GZSL demonstrated the superiority of the proposed model for zero-shot classification.

In the future, since the acquisition of attributes requires prior knowledge, we plan to exploit some other semantic information to construct the common space, e.g., click-through data. Moreover, we will exploit GAN based generative methods to establish a more robust representation in RDCN for zero-shot and few-shot classification.

Acknowledgment

This work is supported by National Natural Science Foundation of China under Grant 61906141, the National Natural Science Foundation of Shaanxi Province under Grant No. 2020JQ-317, China Postdoctoral Science Foundation (Grant No. 2019M653564), the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2021B02), Open Project Program of the State Key Lab of CAD&CG (Grant No. A2018), Zhejiang University and the Fundamental Research Funds for the Central Universities.

References

[1] C. Geng, L. Tao, S. Chen, Guided cnn for generalized zero-shot and open-set recognition using visual and semantic prototypes, Pattern Recognition 102 (2020) 107263.

[2] Z. Cao, J. Lu, S. Cui, C. Zhang, Zero-shot handwritten chinese character recognition with hierarchical decomposition embedding, Pattern Recognition 107 (2020) 107488.

[3] X. Li, M. Fang, H. Li, J. Wu, Zero shot learning based on class visual prototypes and semantic consistency, Pattern Recognition Letters 135 (2020) 368–374.

[4] Y. Liu, X. Gao, Q. Gao, J. Han, L. Shao, Label-activating framework for zero-shot learning, Neural Networks 121 (2020) 1–9.

[5] R. Gao, X. Hou, J. Qin, J. Chen, L. Liu, F. Zhu, Z. Zhang, L. Shao, Zero-vae-gan: Generating unseen features for generalized and transductive zero-shot learning, IEEE Transactions on Image Processing 29 (2020) 3665–3680.

[6] C. H. Lampert, H. Nickisch, S. Harmeling, Attribute-based classification for zero-shot visual object categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (3) (2014) 453–465.

[7] R. Del Chiaro, A. D. Bagdanov, A. Del Bimbo, Webly-supervised zero-shot learning for artwork instance recognition, Pattern Recognition Letters 128 (2019) 420–426.

[8] Z. Li, L. Yao, X. Chang, K. Zhan, J. Sun, H. Zhang, Zero-shot event detection via event-adaptive concept relevance mining, Pattern Recognition 88 (2019) 595–603.

[9] R. Socher, M. Ganjoo, C. D. Manning, A. Ng, Zero-shot learning through cross-modal transfer, in: Advances in neural information processing systems, 2013, pp. 935–943.

[10] E. H. Huang, R. Socher, C. D. Manning, A. Y. Ng, Improving word representations via global context and multiple word prototypes, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Association for Computational Linguistics, 2012, pp. 873–882.

[11] Radovanovi, Nanopoulos, Alexandros, Ivanovi, Mirjana, Hubs in space: Popular nearest neighbors in high-dimensional data, Journal of Machine Learning Research 11 (5) (2010) 2487–2531.

[12] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, T. M. Hospedales, Learning to compare: Relation network for few-shot learning, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 1199–1208.

[13] Y. Xian, C. H. Lampert, B. Schiele, Z. Akata, Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly, arXiv preprint arXiv:1707.00600.

[14] F. Zhang, G. Shi, Co-representation network for generalized zero-shot learning, in: ICML, 2019.

[15] J. W. Zhang, J. Yu, D. Tao, Local deep-feature alignment for unsupervised dimension reduction, IEEE Transactions on Image Processing 27 (2018) 2420–2432.

[16] J. Yu, M. Tan, H. Zhang, D. Tao, Y. Rui, Hierarchical deep click feature prediction for fine-grained image recognition., IEEE transactions on pattern analysis and machine intelligence.

[17] M. Hou, W. Xia, X. Zhang, Q. Gao, Discriminative comparison classifier for generalized zero-shot learning, Neurocomputing 414 (2020) 10–17.

[18] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, in: NIPS, 2005.

[19] E. Kodirov, T. Xiang, S. Gong, Semantic autoencoder for zero-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3174–3183.

[20] B. Romera-Paredes, P. Torr, An embarrassingly simple approach to zero-shot learning, in: International Conference on Machine Learning, 2015, pp. 2152–2161.

[21] Y. Liu, D. Xie, Q. Gao, J. Han, S. Wang, X. Gao, Graph and autoencoder based feature extraction for zero-shot learning, in: IJCAI, 2019.

[22] Y. Shi, W. Wei, Discriminative embedding autoencoder with a regressor feedback for zero-shot learning, IEEE Access 8 (2020) 11019–11030.

[23] Y. Annadani, S. Biswas, Preserving semantic relations for zero-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7603–7612.

[24] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al., Devise: A deep visual-semantic embedding model, in: Advances in neural information processing systems, 2013, pp. 2121–2129.

[25] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, 1998.

[26] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient estimation of word representations in vector space, CoRR abs/1301.3781.

[27] R. Socher, M. Ganjoo, C. D. Manning, A. Y. Ng, Zero-shot learning through cross-modal transfer, in: NIPS, 2013.

[28] H. Zhang, J. Liu, Y. Yao, Y. Long, Pseudo distribution on unseen classes for generalized zero shot learning, Pattern Recognition Letters 135 (2020) 451–458.

[29] H. Zhang, P. Koniusz, Zero-shot kernel learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7670–7679.

[30] Y. Liu, J. Li, X. Gao, A simple discriminative dual semantic auto-encoder for zero-shot classification, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020) 4053–4057.

[31] H. Zhang, L. Liu, Y. Long, Z. Zhang, L. Shao, Deep transductive network for generalized zero shot learning, Pattern Recognition 105 (2020) 107370.

[32] S. M. Shojaee, M. S. Baghshah, Semi-supervised zero-shot learning by a clustering-based approach, ArXiv abs/1605.09016.

[33] Q. Wang, K. Chen, Zero-shot visual recognition via bidirectional latent embedding, International Journal of Computer Vision 124 (3) (2017) 356–383.

[34] V. K. Verma, P. Rai, A simple exponential family framework for zero-shot learning, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2017, pp. 792–808.

[35] J. Li, X. Lan, Y. Long, Y. Liu, X. Chen, L. Shao, N. Zheng, A joint label space for generalized zero-shot classification, IEEE Transactions on Image Processing 29 (2020) 5817–5831.

[36] N. Xing, Y. Liu, H. Zhu, J. Wang, J. Han, Zero-shot learning via discriminative dual semantic auto-encoder, IEEE Access 9 (2021) 733–742.

[37] Z. Zhang, V. Saligrama, Zero-shot learning via semantic similarity embedding, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4166–4174.

[38] Y. Xian, T. Lorenz, B. Schiele, Z. Akata, Feature generating networks for zero-shot learning, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 5542–5551.

[39] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan, ArXiv abs/1701.07875.

[40] A. Mishra, M. S. K. Reddy, A. Mittal, H. A. Murthy, A generative model for zero shot learning using conditional variational autoencoders, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2018) 2269–22698.

[41] K. Sohn, H. Lee, X. Yan, Learning structured output representation using deep conditional generative models, in: NIPS, 2015.

[42] G. Arora, V. K. Verma, A. Mishra, P. Rai, Generalized zero-shot learning via synthesized examples, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 4281–4289.

[43] D. P. Kingma, M. Welling, Auto-encoding variational bayes, CoRR abs/1312.6114.

[44] M. Chen, Z. E. Xu, K. Q. Weinberger, F. Sha, Marginalized denoising autoencoders for domain adaptation, ArXiv abs/1206.4683.

[45] Y. Li, J. Zhang, J. Zhang, K. Huang, Discriminative learning of latent features for zero-shot recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[46] G. Patterson, C. Xu, H. Su, J. Hays, The sun attribute database: Beyond categories for deeper scene understanding, International Journal of Computer Vision 108 (1-2) (2014) 59–81.

[47] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset.

[48] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1778–1785.

[49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[50] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980.

[51] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[52] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, J. Dean, Zeroshot learning by convex combination of semantic embeddings, in: In Proceedings of ICLR, Citeseer, 2014.

[53] L. Chen, H. Zhang, J. Xiao, W. Liu, S.-F. Chang, Zero-shot visual recognition using semantics-preserving adversarial embedding networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1043–1052.

[54] S. Liu, M. Long, J. Wang, M. I. Jordan, Generalized learning with deep calibration network, in: Advances in Neural Information Processing Systems, 2018, pp. 2009–2019.

[55] J. Liu, X. Li, G. Yang, Cross-class sample synthesis for zero-shot learning, in: British Machine Vision Conference, 2019.

[56] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, B. Schiele, Latent embeddings for zero-shot classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 69–77.

[57] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for image classification, IEEE transactions on pattern analysis and machine intelligence 38 (7) (2016) 1425–1438.

[58] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2927–2936.

[59] S. Changpinyo, W.-L. Chao, B. Gong, F. Sha, Synthesized classifiers for zero-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5327–5336.

[60] H. Zhang, Y. Long, Y. Guan, L. Shao, Triple verification network for generalized zero-shot learning, IEEE Transactions on Image Processing 28 (1) (2019) 506–517.

[61] H. Zhang, Y. Long, W. Yang, L. Shao, Dual-verification network for zero-shot learning, Information Sciences 470 (2019) 43–57.

[62] H. Zhang, H. Mao, Y. Long, W. kou Yang, L. Shao, A probabilistic zero-shot learning method via latent nonnegative prototype synthesis of unseen classes., IEEE transactions on neural networks and learning systems.

[63] J. Li, X. Lan, Y. Liu, L. Wang, N. Zheng, Compressing unknown images with product quantizer for efficient zero-shot classification, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 5458–5467.

[64] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (Nov) (2008) 2579–2605.

Revision Notes

We would like to thank the associate editor and anonymous reviewers for their constructive advices. We have greatly benefited from the ideas and recommendations. Based on the reviewer team's inputs, we have been able to prepare a much better manuscript. Indeed, in revising this paper, we have made great efforts to address each single comment made by the review team, which are elaborated below.

Response to EiC:

1. Most of the references are to conference paper, please improve relevance to the readership of PRJ by better grounding the paper in the recent archival pattern recognition field journal literature.

Response: We sincerely thank the reviewer for these constructive advices, which help to improve the quality of the paper. We have added the following seven pattern recognition field journal literature in the revised manuscript (See [1], [2], [3], [7], [8], [28] and [31] in the revised paper). More details about these literature can refer to the Introduction and Related Work in the revised manuscript.

[1] C. Geng, L. Tao, S. Chen, Guided cnn for generalized zero-shot and open-set recognition using visual and semantic prototypes, Pattern Recognition 102 (2020) 107263.

[2] Z. Cao, J. Lu, S. Cui, C. Zhang, Zero-shot handwritten chinese character recognition with hierarchical decomposition embedding, Pattern Recognition 107 (2020) 107488

[3] X. Li, M. Fang, H. Li, J. Wu, Zero shot learning based on class visual prototypes and semantic consistency, Pattern Recognition Letters 135 (2020) 368–374.

[4] R. Del Chiaro, A.D. Bagdanov, A. Del Bimbo, Webly-supervised zero-shot learning for artwork instance recognition, Pattern Recognition Letters 128 (2019) 420–426.

[5] Z. Li, L. Yao, X. Chang, K. Zhan, J. Sun, H. Zhang, Zero-shot event detection via

event-adaptive concept relevance mining, Pattern Recognition 88 (2019) 595–603.

[6] H. Zhang, J. Liu, Y. Yao, Y. Long, Pseudo distribution on unseen classes for generalized zero shot learning, Pattern Recognition Letters 135 (2020) 451–458.

[7] H. Zhang, L. Liu, Y. Long, Z. Zhang, L. Shao, Deep transductive network for generalized zero shot learning, Pattern Recognition 105 (2020) 107370.

Response to Reviewer #1:

1. The revised manuscript has been significantly improved and I have no further concern.

Response: We appreciate the comments from the reviewer, including the confirmation on the novelty and promising results.

Response to Reviewer #2:

1. I do not have other questions.

Response: We appreciate the comments from the reviewer, including the confirmation on the novelty and promising results.

Relation-based Discriminative Cooperation Network for Zero-Shot Classification

Yang Liu[a], Xinbo Gao[a,b], Quanxue Gao[a], Jungong Han[c], Ling Shao[d]

[a] *State Key Laboratory of Integrated Services Networks, Xidian University, Shaanxi 710071, China.*
[b] *Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing, China*
[c] *Computer Science Department, Aberystwyth University, Aberystwyth SY23 3FL, U.K.*
[d] *Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates.*

Abstract

Zero-shot learning (ZSL) aims to assign the category corresponding to the relevant semantic as the label of the unseen sample based on the relationship between the learned visual and semantic features. However, most typical ZSL models faced with the domain bias problem, which leads to unseen or test samples being easily misclassified into seen or training categories. To handle this problem, we propose a relation-based discriminative cooperation network (RDCN) model for ZSL in this work. The proposed model effectively utilize the robust metric space spanned by the cooperated semantics with the help of a set of relations. On the other hand, we devise the relation network to measure the relationship between the visual features and embedded semantics, and the validation information will guide the embedding module to learn more discriminative information. At last, the proposed RDCN model is validated on six benchmarks, and extensive experiments demonstrate the superiority of proposed method over most existing ZSL models on the traditional zero-shot setting and the more realistic generalized zero-shot setting.

Keywords: Zero-shot learning, Bias, Discriminative, Relation

*Corresponding author
Email address: yangl@xidian.edu.cn (Yang Liu)

1. Introduction

Humans can recognize many categories, including about 30,000 basic categories and even more subcategories. On the other hand, humans are also very good at recognizing objects even if they have never seen any of their examples. For example, if a child has seen cattle before, he/she can easily recognize a cow and learn that a cow looks like cattle with black-and-white color. Inspired by the ability of humans to identify unseen categories, that the research area of Zero-Shot Learning (ZSL) [1, 2, 3, 4] aims to recognize classes whose samples haven't been available during training time has received increasing interests.

Different from traditional supervised learning, ZSL considers an extreme case where testing classes is unavailable during training, *i.e.,* the training (seen) classes and testing (unseen) classes are disjoint [5]. ZSL links the seen classes and unseen classes through the semantic information to complete the recognition task. The semantic is defined as a high dimensional vector space where unseen and seen classes are connected together, which can be a semantic attribute space [6, 7, 8] or a semantic word vector space [9, 10]. In adddition, test images may come from both unseen and seen classes, which is named Generalized Zero-Shot Learning (GZSL). In real-world applications, since we cannot predict whether a new sample comes from an unseen class or a seen class, GZSL is more practical and challenging than ZSL.

Regarding the bridge between visual space and semantic space, most traditional ZSL methods tend to learn a mapping that project samples from the visual space to the semantic space with the labelled training set including seen classes only. When classifying unseen images, the learned embedding is used to project the visual representation of unseen samples into the semantic space including unseen and seen classes. Then the Nearest Neighbor (NN) search method is used to recognize the sample of unseen class, which is the testing process. However, the NN search method is easy to cause hubness problem [11]. To solve this problem, Sung *et al.* [12] recently proposed a model named Relation Network (RN) to compare the test samples with the embedded semantics in a self-adaptive way. Different from NN search, RN tries to measure the relation score between unseen samples and semantics by learning a distance metric.

Unfortunately, there is a strong domain bias problem [13] when applying almost all standard ZSL models to deal with a GZSL task, which leads to unseen images being misclassified into seen classes. To alleviate this problem, Zhang and Shi [14] proposed a Co-Representative network (CRnet) based on RN. CRnet tries to learn a uniform embedding space by a single-layer module with parallel structure and high local linearity.

However, CRnet still relies heavily on obtaining human-defined semantics for knowledge transfer. Similar to other embedding models, CRnet focuses on the original visual features and semantics, but completely ignores the discriminative information among them. Although a few works [15, 16, 17] have been proposed to maintain the relationship between semantics by using complementary features or extracting deep local features, they are rarely used in ZSL task. To address this point, we formulate a novel framework named Relation-based Discriminative Cooperation Network (RDCN) for ZSL task in this paper.

The RDCN model aims to preserve the discriminative information of the visual features and semantics. At first, RDCN adopts the encoder-decoder paradigm to obtain discriminative visual features. Specifically, the encoder aims to learn a mapping from the visual space to the embedding space where the distance between classes is adjusted by the triplet loss [18], while decoder reconstructs the original input features. On the other hand, RDCN uses a decomposition structure to alleviate the bias problem in the semantic space, and adopts a novel semantic pivot regularization to obtain discriminative semantic features. At last, RDCN adopts a relation network as the similarity function to measure the relationship between the discriminative visual features and semantics. In summary, our contributions are concluded into the following three-fold:

- The discriminative visual embedding preserves the discriminative information of the input image features by separating inter-classes and clustering intra-classes with a margin.

- The discriminative semantic embedding acts as a pivot regularization to ensure the cooperated structures representative by utilizing semantic relations between classes.

- The experimental evaluation on several popular datasets, including the ImageNet demonstrates that the proposed RDCN performs favorably against state-of-the-art ZSL models.

The remainder of this paper is organized as follows: Section 2 reviews the related ZSL work followed by Section 3 which describes the proposed model in detail. Experimental results with some detailed analysis are given in Section 4. At last, the conclusion of the work is given in Section 5.

2. Related Work

In this section, we introduce the related works of zero-shot learning from two aspects: relation-based models and synthesis-based models.

2.1. Relation-based Models

Relation-based models aims to learn the relationship between images and the semantics. In early works of ZSL, most algorithms focus on building this relationship by Visual-Semantic Embedding (VSE) framework. According to different directions of mappings, the VSE framework is divided into three types as follows.

(1) *Visual→Semantic Embedding* tries to learn a mapping from the visual space to the semantic space either using linear function [19, 20, 21, 22] or by deep neural network regression [23]. For example, Deep Visual-Semantic Embedding (DeViSE) model [24] is one of the earlier attempts. It inputs Convolutional Neural Networks (C-NN) [25] and Word2Vec [26] features to learn an end-to-end deep classification model. Socher *et al.* [27] mapped the improved visual features to the semantic space by the two-layer or three-layer neural network, and used the least squares loss to train the network.

(2) *Semantic→Visual Embedding* tries to learn a mapping from the semantic space to the visual space, such as [28, 29, 30]. The training and test process are similar with the first mapping manner. For example, Deep Transductive Network (DTN) [31] exploits the high confidence assignments with the assistance of an auxiliary target distribution to reduce the impact of the hubness problem [11]. Shojaee *et al.* [32] proposed a

semi-supervised ZSL method based on the visual features of similar samples clustered together in the visual space.

(3) *Visual$\rightarrow$Common Space$\leftarrow$Semantic* learns a common space where both the visual features and the semantics are embedded to, such as [33, 34, 35, 36]. For the test phase, the visual features and attributes are embedded into the common space for final classification task. Semantic Similarity Embedding (SSE) [37] is a typical method which learns an embedding for visual features and semantics to find mixture features that used to measure the similarity.

Different from traditional VSE framework, Sung *et al.* [12] recently proposed a Relation Network (RN) that aims to compare the query visual features with the target embedded semantics through a self-adaptive way. Inspired by RN, Zhang and Shi [14] proposed a Co-Representative network (CRnet) with the help of high local linearity. In this paper, the proposed model extract discriminative visual and semantic features by VSE framework and measure their relationship with RN framework.

2.2. Synthesis-based Models

Synthesis-based models tries to learn a generator that generates samples from the semantics, and to then train a classifier to predict classes with the generated synthesis samples. For example, f-CLSWGAN [38] generates sufficiently discriminative CNN features to train softmax classifiers with the help of a Wasserstein Generative Adversarial Networks (WGAN) [39]. CVAE-ZSL [40] implements the generation by a Conditional Variational AutoEncoder (CVAE) [41]. SE-GZSL [42] also designs a generator based on a Variational AutoEncoder (VAE) [43] but generates synthesis samples in a feedback-driven way. Recently, Zero-VAE-GAN [5] is proposed to convert ZSL problems into supervised tasks by a combination of VAE and GAN.

The advantage of synthesis-based models is that both seen and unseen samples may be obtained by generator, which contributes to significantly alleviate the domain bias problem. However, synthesis-based methods require the generator to generate a large number of high-quality and diverse samples (including unseen ones) for each class, which is costly and requires additional classifier learning. In this work, we aim to

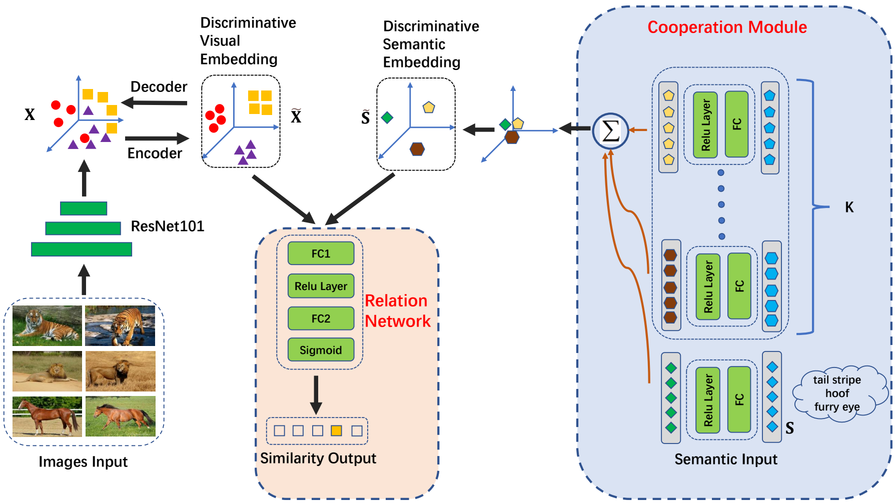


Figure 1: The framework of RDCN.

achieve high performance in both ZSL and GZSL tasks by the end-to-end relation-based model.

3. Proposed Approach

120 In this section, firstly, we provide the problem definition of ZSL by mathematical notation. Secondly, we give the model structure and detailed description of different modules. At last, we conclude the overall objective function.

3.1. Problem Definition

Suppose there are n labeled samples with c seen classes $\{\mathbf{X}, \mathbf{S}, \mathbf{Y}\}$ and n_u unlabeled samples with c_u unseen classes $\{\mathbf{X}^u,\mathbf{S}^u,\mathbf{Y}^u\}$, where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbf{R}^{d\times n}$ and $\mathbf{X}^u = [\mathbf{x}^u_1, \mathbf{x}^u_2, \cdots, \mathbf{x}^u_{n_u}] \in \mathbf{R}^{d\times n_u}$ are d-dimensional visual features, while the corresponding labels are $\mathbf{Y}$ and $\mathbf{Y}^u$, respectively. It is noteworthy that the labels of seen and unseen samples are disjoint, *i.e.*, $\mathbf{Y} \cap \mathbf{Y}^u = \emptyset$. $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_c] \in \mathbf{R}^{k\times c}$ and $\mathbf{S}^u = [\mathbf{s}^u_1, \mathbf{s}^u_2, \cdots, \mathbf{s}^u_{c_u}] \in \mathbf{R}^{k\times c_u}$ are k-dimensional semantic features of seen and 125 130 unseen samples. The ZSL task aims to learn a classifier $f : \mathbf{X}^u \rightarrow \mathbf{Y}^u$, where classes of testing data $\mathbf{X}^u$ are unavailable in training phrase.

3.2. Model Architecture

Figure 1 shows the framework of proposed model RDCN. Specifically, the RDCN consists of four parts: (1) The visual features $\mathbf{X}$ extracted by ResNet101 are encoded with the help of the discriminative visual embeddings $\tilde{\mathbf{X}}$. (2) The discriminative visual embeddings are decoded to reconstruct the input visual features $\mathbf{X}$. (3) The cooperated semantic features obtained by cooperation network are send into a discriminative semantic embedding space, where $\tilde{\mathbf{S}}$ is generated. (4) A relation network is adopted as the similarity function to measure the relationship between the discriminative visual features $\tilde{\mathbf{X}}$ and discriminative semantics $\tilde{\mathbf{S}}$.

3.3. Encoder

The deep image features $\mathbf{X}$ are trained by following function to obtain the discriminative visual embeddings:

$$\tilde{\mathbf{X}} = f_e\left(\mathbf{X}; \theta_e\right), \tag{1}$$

where f_e indicates the operation of the encoder whose parameters are denoted by θ_e. In detail, the deep visual features $\mathbf{X}$ extracted by RessNet101 pass through multilayer perceptron (MLP) with two hidden layers (h1 = 1024-D and h2 = 512-D), followed by a dense layer with the LeakyReLU [44] activation. The output discriminative visual embeddings $\tilde{\mathbf{X}}$ have the same dimension with the semantic embeddings.

3.4. Discriminative Visual Embedding

Most embedding models that solve the ZSL problem focus on calculating a typical description of images in all classes, which makes the encoder non-discriminatory. Motivated by [45], adding discriminative embedding operations in the encoding process can make the learned low-dimensional features more discriminative, which is helpful for classification.

In order to ensure that an embedding visual feature $\tilde{\mathbf{x}}_i$ $(\tilde{\mathbf{x}}_i \in \mathbf{X})$ is closer to each image feature $\tilde{\mathbf{x}}_j$ from the same class than any image feature $\tilde{\mathbf{x}}_k$ from different classes. We use the triplet loss [18] to learn a discriminative embedding by adjusting the intra-class and inter-class distance between the learned features:

$$\ell_{tri} = \frac{1}{n} \sum_{i=1}^{n} \max\left(0, m + d_{intra} - d_{inter}\right), \tag{2}$$

Where $m > 0$ is a margin that is enforced between positive (same class) and negative (different classes) pairs. d_{intra} denotes the squared Euclidean distance between $\tilde{\mathbf{x}}_i$ and visual features from the same class, meanwhile, d_{inter} denotes the squared Euclidean distance between $\tilde{\mathbf{x}}_i$ and visual features from the different class, *i.e.,*

$$d_{intra} = \sum_{\tilde{\mathbf{x}}_j \in c_i} \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\|_2^2, \tag{3}$$

$$d_{inter} = \sum_{\tilde{\mathbf{x}}_k \notin c_i} \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_k\|_2^2. \tag{4}$$

3.5. Decoder

The encoder and the decoder are connected by the discriminative visual embeddings. The discriminative visual feature $\tilde{\mathbf{X}}$ is the input of the decoder, then the reconstructed visual feature is denoted by following equation:

$$\hat{\mathbf{X}} = f_d\left(\tilde{\mathbf{X}}; \theta_d\right), \tag{5}$$

where f_d indicates the operation of the decoder whose parameters are denoted by θ_d. Similar with the encoder, the decoder is a multilayer perceptron (MLP) includes two hidden layers (h1 = 512-D and h2 = 1024-D) and a dense layer with the LeakyReLU activation.

Since the proposed framework involves a decoder which reconstructs the original visual features, there is an accompanying reconstruction loss:

$$\ell_{rec} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2, \;\; \hat{\mathbf{x}}_i \in \hat{\mathbf{X}}. \tag{6}$$

3.6. Cooperation Module

Motivated by the CRnet [14], we use a decomposition structure to alleviate the bias problem in cooperation module. Specifically, We first adopt the unsupervised K-means clustering method to divide the semantic features $\mathbf{S}$ into $K(K < c)$ subsets. The clustering center of k-th subset is denoted by $\bar{\mathbf{s}}_k(k = 1, 2, \cdots, K)$. Then, the semantic feature $\mathbf{s}_i(i = 1, 2, \cdots, c)$ with a combination of K clustering center are trained by following cooperation module to obtain the discriminative semantic embedding:

$$\tilde{\mathbf{s}}_i = \sum_{k=0}^{K} f_c\left(\left[\mathbf{s}_i - \bar{\mathbf{s}}_k\right]; \theta_c\right), \tag{7}$$

where f_c indicates the operation of the cooperation module whose parameters are denoted by θ_c. In detail, the vector calculated by $\mathbf{s}_i - \bar{\mathbf{s}}_k$ is fed into a single FC layer with a ReLU activation. Moreover, different from the CRnet, we also add the original semantic features as the input to retain its own sparsity feature, *i.e.,* $\bar{\mathbf{s}}_0 = \mathbf{0}$.

3.7. Discriminative Semantic Embedding

Intuitively, maximizing the distance of the semantic embeddings by following function can maintain the discriminative information between different semantic features,

$$\max \sum_{i=1}^{c} \sum_{j=1}^{c} \|\tilde{\mathbf{s}}_i - \tilde{\mathbf{s}}_j\|_2^2, \tag{8}$$

where $\tilde{\mathbf{s}}_i$ and $\tilde{\mathbf{s}}_j$ are different semantic embeddings. In order to reduce the computational complexity of the above calculation methods, a semantic pivot is used to simplify the calculation. The semantic pivot of semantics is defined as the center of semantic embedding. It can be solved by the average embedded feature, or it can be calculated by the mean shift technique. In fact, there is almost no difference in performance between these two calculations [17]. For simplicity, in this paper, the semantic pivot $\bar{\mathbf{s}}$ is calculated by the center of the semantic embeddings, *i.e.,* $\bar{\mathbf{s}} = \frac{1}{c} \sum\limits_{i=1}^{c} \tilde{\mathbf{s}}_i$. Then, we get the following loss function:

$$\ell_{piv} = -\sum_{i=1}^{c} \|\tilde{\mathbf{s}}_i - \bar{\mathbf{s}}\|_2^2. \tag{9}$$

3.8. Relation Module

After obtaining discriminative visual feature $\tilde{\mathbf{x}}_i$ and discriminative semantic feature $\tilde{\mathbf{s}}_j$, we adopt the RN [12] to measure their relationship. Specifically, the relation module is a two-layer neural network and the input is the concatenation of $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{s}}_j$; the hidden layer increase nonlinearity and the output of the network is a scalar in range of 0 to 1 representing the similarity between discriminative visual and semantic features, which is called relation score.

In this module, we adopt RN as the similarity function $g(\cdot)$ and follow the original settings. Thus the output relation score of training pairs $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{s}}_j$ be denoted as

$g(\tilde{\mathbf{x}}_i, \tilde{\mathbf{s}}_j)$. In the training process, we randomly sample the entire training set to generate training pairs of $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{s}}_j$, and control the ratio of matched pairs(discriminative visual and semantic features belong to the same class) to mismatched pairs(discriminative visual and semantic features come form different classes) at about $1 : 30$. The similarities of matched pairs and mismatched pairs are set to 1 and 0, respectively. The relation module is trained by mean square error loss:

$$\ell_{rel} = \sum_{j=1}^{c} \sum_{i=1}^{n} \left[g(\tilde{\mathbf{x}}_i, \tilde{\mathbf{s}}_j) - l(\tilde{\mathbf{x}}_i, \tilde{\mathbf{s}}_j)\right]^2 \tag{10}$$

where $l(\cdot)$ is the similarity ground-truth, $l(\tilde{\mathbf{x}}_i, \tilde{\mathbf{s}}_j) = 1$ when $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{s}}_j$ belong to the same class, and $l(\tilde{\mathbf{x}}_i, \tilde{\mathbf{s}}_j) = 0$ when $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{s}}_j$ belong to different classes.

3.9. Overall Objective

With the objective functions introduced above, the overall objective of the proposed model is given by:

$$\ell = \ell_{rel} + \alpha\ell_{tri} + \beta\ell_{rec} + \gamma\ell_{piv}, \tag{11}$$

where α, β and γ are trade-off parameters chosen based on the validation dataset.

Given a testing image $\mathbf{x}_k^u$, its label can be inferred by:

$$j^* = \arg\max_j g(\tilde{\mathbf{x}}_k^u, \tilde{\mathbf{s}}_j). \tag{12}$$

For ZSL task, $\tilde{\mathbf{s}}_j$ refers to the semantic embeddings of only the unseen classes, *i.e.*, $j\epsilon\{1, 2, \cdots, c_u\}$, and for GZSL task, $\tilde{\mathbf{s}}_j$ refers to the semantic embeddings of both seen as well as unseen classes, *i.e.*, $j\epsilon\{1, 2, \cdots, c + c_u\}$. c and c_u are the numbers of seen and unseen classes, respectively.

4. Experiments

In this section, firstly, we introduce five popular ZSL datasets SUN, CUB, AWA1, AWA2, aPY and a large-scale ImageNet dataset. Secondly, we provide the implementation details of the architecture. Thirdly, some experimental results with related analysis on the traditioal zero-shot setting and the more realistic generalized zero-shot setting are given. Finally, we show some visualized results for the proposed model.

4.1. Datasets Descriptions

Five Small-scale Attribute datasets: SUN Attribute (SUN) [46] is a fine-grained dataset and it consists of 14,340 images belonging to 717 classes annotated with 102 attributes. Following the ZSL setting in [6], 72 out of 717 classes are as unseen categories and the rest of 645 categories as seen categories.

CUB-200-2011 Birds (CUB) [47] is a fine-grained and medium scale dataset which has in total 11,788 images distributed in 200 bird categories. Each class is annotated with a 312-dim attribute vector. We follow the standard ZSL split with 150 categories for seen classes and 50 for unseen classes as in [6].

Animals with Attributes 1 (AWA1) [6] is a kind of coarse-grained dataset, which includes a total of 30,475 images belonging to 50 classes. Each class is annotated with a 85-dim attribute vector, where 40 categories (seen) are used for training and rest 10 categories (unseen) for testing. Animals with Attributes 2 (AWA2) [13] has the same 50 categories as AWA1 dataset. However, AWA2 dataset contains 37,322 images. Similar to AWA1, 40 categories are used for seen classes and 10 categories are used for unseen classes.

A Pascal and Yahoo (aPY) [48] is a kind of small-scale coarse-grained dataset. Each category is annotated with a 64-dim attribute vector. Among the total number of 32 classes, 20 Pascal classes are used for training and 12 Yahoo classes are used for testing.

One large-scale dataset: ImageNet [49] has a total of 218,000 images. 21,841 classes with more than 10 million images, where 1k classes containing 1.2 million images are used for training the mapping. There are different splits in the test. Specifically, 2-hops/3-hops refers to test classes belonging to 2/3 tree hops away from 1k train classes in the WordNet hierarchy, which contains 1,509/7,678 unseen classes. Such classes that contain the top 500/1k/5k maximum images and top 500/1k/5k minimum images are given for test splits respectively. At last, all 20K classes are given for testing, which is a challenging task.

4.2. Implementation Details

We use ReLU activation for all layers except for the output of the encoder and the decoder, which adopt LeakyReLU activations with the negative slope of 0.3. A single-layer FC network compared with K parallel single-layer FC network are given for embedding the original and cooperated semantic vectors, respectively. Parameters α and β in our objective function are fine-tuned in the range $\left[5 \times 10^{-6}, 10^{-2}\right]$ and γ from $\left[10^{-7}, 5 \times 10^{-4}\right]$. Moreover, the K value is given in the range $[3, 12]$. For relation module, the discriminative visual and semantic features are concatenated with a hidden layer before passing relation network, We adopt Adam optimizer [50] with a initialized learning rate of 10^{-3} and a weight decay of 5×10^{-5}. For fair comparison, we follow the settings in [13] to split each dataset for training and testing. Moreover, each image is represented by 2048-dim vector extracted by 101-layered ResNet, *i.e.,* ResNet101 [51].

4.3. Zero-Shot Learning (ZSL) Experiments

In this work, the average per-class top-1 accuracy is adopted as the evaluation criteria for zero-shot classification, *i.e.,* we average the correct predictions independently for each class by follows:

$$acc\Upsilon = \frac{1}{\|\Upsilon\|} \sum_{c=1}^{\|\Upsilon\|} \frac{\#correct\ predictions\ in\ c}{\#samples\ in\ c} \tag{13}$$

where Υ and $\|\Upsilon\|$ indicate the set of classes and number of classes with corresponding dataset, respectively. So Υ includes all the test classes *i.e.,* the unseen classes for ZSL task.

The results of the different ZSL models on five popular small-scale datasets is given in Table 1. We can see that the proposed RDCN consistently performs better than compared methods, and RDCN gets the state-of-the-art on four datasets: SUN, AWA1, AWA2 and CUB. Specifically, the accuracies increase of 2.7% and 4.2% compared to the strongest competitor on SUN dataset and AWA2 dataset, respectively. We also observe a significant increase when we include all of the ℓ_{tri}, ℓ_{rec} and ℓ_{piv} in our model. This indicates that the reconstruction term makes a contribution to vary levels

Table 1: Zero-shot learning (ZSL) results on five small-scale attribute datasets. The results report average per-class Top-1 accuracy in %.

Method	SUN	CUB	AWA1	AWA2	aPY
DeViSE [24]	56.5	52.0	54.2	59.7	39.8
CONSE [52]	38.8	34.3	45.6	44.5	26.9
CMT [9]	39.9	34.6	39.5	37.9	28.0
SP-AEN [53]	59.2	55.4	-	58.5	24.1
PSR [23]	61.4	56.0	-	63.8	38.4
DCN [54]	61.8	56.2	65.2	-	**43.6**
CCSS [55]	56.8	44.1	56.3	63.7	35.5
DAP [6]	39.9	40.0	44.1	46.1	33.8
IAP [6]	19.4	24.0	35.9	35.9	36.6
SSE [37]	51.5	43.9	60.1	61.0	34.0
LATEM [56]	55.3	49.3	55.1	55.8	35.2
ALE [57]	58.1	54.9	59.9	62.5	39.7
SJE [58]	53.7	53.9	65.6	61.9	32.9
ESZSL [20]	54.5	53.9	58.2	58.6	38.3
SYNC [59]	56.3	55.6	54.0	46.6	23.9
SAE [19]	53.4	42.0	58.1	50.3	32.9
f-CLSWGAN [38]	58.5	57.7	64.1	-	-
TVN [60]	59.3	54.9	64.7	-	40.9
DVN [61]	62.4	57.8	67.7	-	41.2
Zhang's [62]	60.4	53.2	67.4	-	42.8
RDCN ($\alpha = 0$)	58.9	55.3	66.0	65.3	36.1
RDCN ($\beta = 0$)	60.6	56.5	67.9	66.1	37.9
RDCN ($\gamma = 0$)	59.3	56.1	69.3	65.2	39.5
RDCN	**65.1**	**60.5**	**71.6**	**68.0**	42.1

of gain fatures and the discriminative information among visual and semantic features is also essential.

4.4. Generalized Zero-Shot Learning (GZSL) Experiments

GZSL means that the search space includes both test classes (Υ^{ts}) and training classes (Υ^{tr}). At first, the average per-class top-1 accuracy on Υ^{tr} and Υ^{ts} can be obtained by Eq. (13), then the harmonic mean is computed by follows:

$$H = \frac{2 \times acc\Upsilon^{tr} \times acc\Upsilon^{ts}}{acc\Upsilon^{tr} + acc\Upsilon^{ts}} \tag{14}$$

where $acc\Upsilon^{tr}$ and $acc\Upsilon^{ts}$ are accuracies of samples from Υ^{tr} and Υ^{ts}, respectively.

The GZSL results on five popular attribute datasets is given in Table 2. We have following observations according to the results:

Table 2: Generalized Zero-Shot Learning (GZSL) results on five small-scale attribute datasets. ts = $acc\left(\Upsilon^{ts}\right)$, tr = $acc\left(\Upsilon^{tr}\right)$, H = harmonic mean. We measure Top-1 accuracy in %.

	SUN			CUB			AWA1			AWA2			aPY		
Method	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H
DeViSE [24]	16.9	27.4	20.9	23.8	53.0	32.8	13.4	68.7	22.4	17.1	74.7	27.8	4.9	76.9	9.2
CMT [9]	8.1	21.8	11.8	7.2	49.8	12.6	0.9	87.6	1.8	0.5	90.0	1.0	1.4	85.2	2.8
CONSE [52]	6.8	39.9	11.6	1.6	72.2	3.1	0.4	88.6	0.8	0.5	90.6	1.0	0.0	**91.2**	0.0
DAP [6]	4.2	25.1	7.2	1.7	67.9	3.3	0.0	88.7	0.0	0.0	84.7	0.0	4.8	78.3	9.0
IAP [6]	1.0	37.8	1.8	0.2	**72.8**	0.4	2.1	78.2	4.1	0.9	87.6	1.8	5.7	65.6	10.4
GFZSL [34]	0.0	39.6	0.0	0.0	45.7	0.0	1.8	80.3	3.5	2.5	80.1	4.8	0.0	83.3	0.0
SSE [37]	2.1	36.4	4.0	8.5	46.9	14.4	7.0	80.5	12.9	8.1	82.5	14.8	0.2	78.9	0.4
LATEM [56]	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	11.5	77.3	20.0	0.1	73.0	0.2
ALE [57]	21.8	33.1	26.3	23.7	62.8	34.4	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7
SJE [58]	14.7	30.5	19.8	23.5	59.2	33.6	11.3	74.6	19.6	8.0	73.9	14.4	3.7	55.7	6.9
ESZSL [20]	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
SYNC [59]	7.9	**43.3**	13.4	11.5	70.9	19.8	8.9	87.3	16.2	10.0	90.5	18.0	7.4	66.3	13.3
SAE [19]	17.8	32.0	22.8	18.8	58.5	28.5	14.2	81.2	24.1	16.7	82.5	27.8	9.9	74.7	17.5
ZSKL [29]	19.8	29.1	23.6	19.9	52.5	28.9	18.3	79.3	29.8	17.6	80.9	29.0	11.9	76.3	20.5
f-CLSWGAN [38]	42.6	36.6	39.4	43.7	57.7	49.7	57.9	61.4	59.6	-	-	-	-	-	-
CVAE-ZSL [40]	-	-	26.7	-	-	34.5	-	-	47.2	-	-	51.2	-	-	-
SE-GZSL [42]	40.9	30.5	34.9	41.5	53.3	46.7	56.3	67.8	61.5	58.3	68.1	62.8	-	-	-
TVN [60]	18.2	28.9	22.3	21.6	47.5	29.7	18.2	87.5	30.2	-	-	-	8.8	59.1	15.4
DVN [61]	25.3	34.6	29.2	26.2	55.1	35.5	34.9	73.4	48.5	-	-	-	13.7	72.2	23.1
RN [12]	-	-	-	38.1	61.1	47	31.4	**91.3**	46.7	30	**93.4**	45.3	-	-	-
CRnet [14]	34.1	36.5	35.3	45.5	56.8	50.5	58.1	74.7	65.4	52.6	78.8	63.1	32.4	68.4	44
Zhang's [62]	**39.7**	38.9	39.3	37.8	58.2	45.9	37.0	84.7	51.4	-	-	-	25.9	79.5	39.1
RDCN	37.3	37.7	**37.5**	**45.5**	58.1	**51.0**	**60.2**	79.0	**68.3**	**56.6**	72.3	**63.5**	**34.0**	75.6	**46.9**

(1) Compared with Table 1, ZSL results are higher than GZSL results ("ts" value). The main reason is that all seen classes are included in the search space and these seen classes confuse the test images. That is to say, an image from unseen class is more likely to be mistaken for a seen class when it is projected into the semantic space in GZSL task.

(2) The "tr" value in Table 2 just represents the classification performance in the seen dataset. Moreover, high accuracy on "tr" is often accompanied by low accuracy on "ts" and "H" such as IAP and SYNC, which indicates that these models perform well most seen classes but fails to generalize for unseen classes, *i.e.*, overfitting.

(3) With respect to the state-of-the-art, RDCN gets best "H" value almost on all

Table 3: GZSL comparisons (ts) in ImageNet dataset. The results report Top-10 accuracy in %.

	Hierarchy All		Most populated			Least populated			All
Method	2-hops	3-hops	500	1k	5k	500	1k	5k	20k
CONSE [52]	0.86	7.14	23.47	18.38	9.92	0.00	0.00	0.66	3.43
CMT [9]	7.80	2.77	9.65	7.73	3.83	3.37	2.71	1.45	1.25
LATEM [56]	16.99	6.28	23.61	18.65	8.73	8.73	7.60	3.50	2.71
ALE [57]	17.79	6.34	24.93	19.37	9.12	10.38	8.46	3.63	2.77
DeViSE [24]	17.59	6.28	24.66	19.11	8.99	10.11	8.26	3.63	2.71
SJE [58]	17.46	6.21	23.61	18.45	8.79	9.85	8.00	3.50	2.71
ESZSL [20]	19.24	6.81	26.52	20.56	9.72	9.12	7.73	3.76	3.10
SYNC [59]	14.55	5.62	16.33	13.82	7.87	2.77	2.44	1.78	2.64
SAE [19]	13.55	4.82	20.76	16.60	7.60	3.43	2.57	1.58	2.24
PQZSL [63]	21.80	7.41	29.30	23.75	11.3	9.42	7.87	3.72	3.45
RDCN	**25.69**	**10.15**	**33.71**	**26.91**	**15.65**	**11.33**	**10.37**	**7.83**	**7.31**

datasets. In detail, RDCN obtains 68.3% on AWA1 dataset and 46.9% on aPY dataset, which is better than the next best model CRnet by 2.9%. On AWA2 dataset, RDCN gets a best accuracy of 56.6% on the first setting (“ts” value) and 63.5% overall. In addition, RDCN achieves better results compared to some synthesis-based models like f-CLSWGAN, CVAE-ZSL, SE-GZSL and so on.

Moreover, Table 3 reports the result of GZSL (“ts” value) on ImageNet. Compared with some baselines, RDCN obtains the best performance in most splits, which proves the superiority of the proposed method on large datasets. The whole GZSL experimental results supports our hypothesis that discriminative visual and semantic information are beneficial for generalized zero-shot recognition.

4.5. Ablation Studies

In this subsection, we compare the RDCN with its different variants to study the role of each item in the objective function 11. The experimental results are shown in Table 1. We analyze the following three cases: 1). “RDCN ($\alpha = 0$)” means there is no triplet loss ℓ_{tri} in the objective function 11; 2). “RDCN ($\beta = 0$)” means there is no reconstruction loss ℓ_{rec} in the objective function 11; 3). “RDCN ($\gamma = 0$)” means there is no semantic pivot regularization ℓ_{piv} in the objective function 11.

We observe in Table 1 that each kind of strategy of RDCN can improve the ZSL classification performances effectively. In addition, The role of the triplet loss ℓ_{tri} is

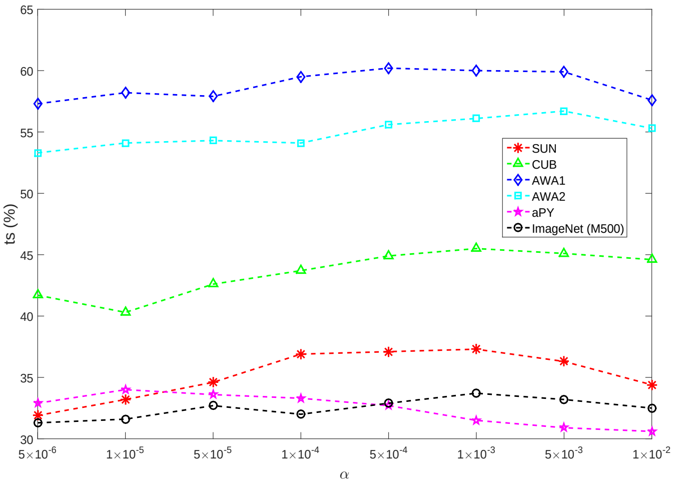


Figure 2: The influence of α on six datasets. α is the parameter of the triplet loss ℓ_{tri} in the objective function.

more important than that of reconstruction loss ℓ_{rec} and semantic pivot regularization ℓ_{piv}, which is based on the fact that the result of "RDCN ($\alpha = 0$)" is worse than that of "RDCN ($\beta = 0$)" and "RDCN ($\gamma = 0$)". According to the results of the last four rows in Table 1, we can be see that each item in the objective function plays a positive role in the ZSL classification task.

For the RDCN, there are three parameters, *i.e.,* α, β and γ in the objective function. By varying one of the parameters while fixing the other parameters, we run the model for 100 epochs and produce the GZSL results ("ts" value). Specifically, we conduct experiments varying α and β from $\left[5 \times 10^{-6}, 10^{-2}\right]$ and γ from $\left[10^{-7}, 5 \times 10^{-4}\right]$. The influence of α, β and γ on each dataset are illustrated in Figure 2, Figure 3 and Figure 4, respectively. For the ImageNet, due to the large number of testing samples (20K classes) in the complete dataset, we selected top 500 maximum images (M500) as test splits for analysis. According to the "ts" results under different values of three parameters α, β and γ, we conclude that the RDCN can obtain promising performance within a small range of parameters.

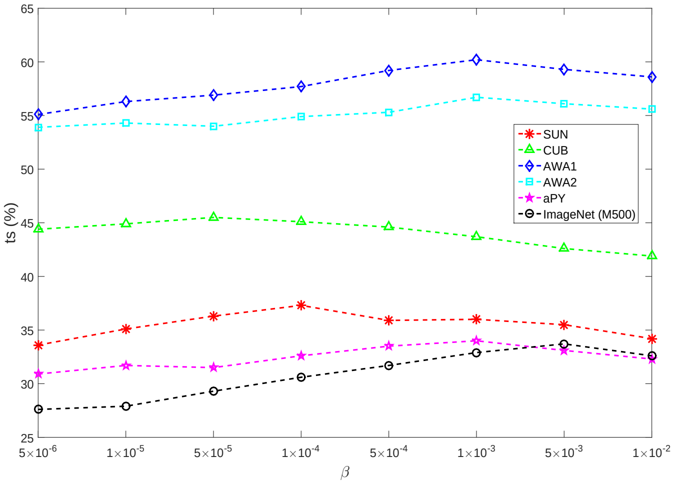


Figure 3: The influence of β on six datasets. β is the parameter of the reconstruction loss ℓ_{rec} in the objective function.

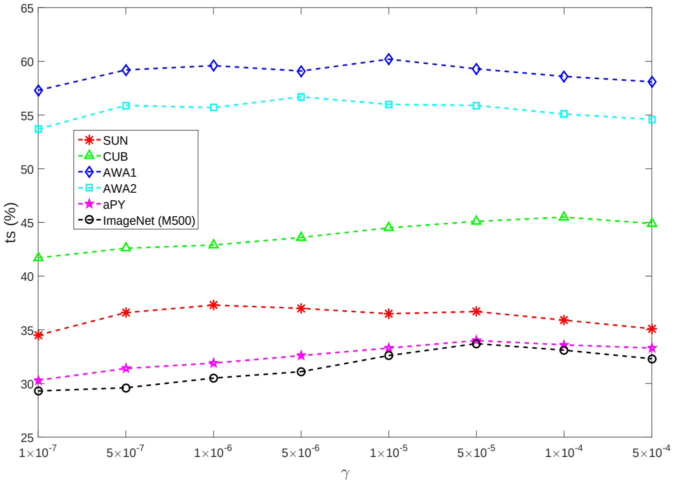
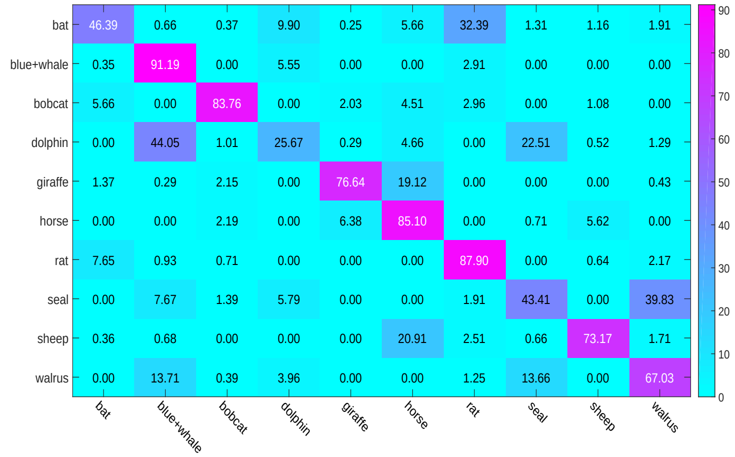


Figure 4: The influence of γ on six datasets. γ is the parameter of the semantic pivot regularization ℓ_{piv} in the objective function.

4.6. Visualized results

We further provide some visualized results for the proposed method. Figure 5 shows the confusion matrices of unseen classes on aPY dataset.

According to the results on Figure 5, we can see the proposed model RDCN can



	bat	blue+whale	bobcat	dolphin	giraffe	horse	rat	seal	sheep	walrus
bat	46.39	0.66	0.37	9.90	0.25	5.66	32.39	1.31	1.16	1.91
blue+whale	0.35	91.19	0.00	5.55	0.00	0.00	2.91	0.00	0.00	0.00
bobcat	5.66	0.00	83.76	0.00	2.03	4.51	2.96	0.00	1.08	0.00
dolphin	0.00	44.05	1.01	25.67	0.29	4.66	0.00	22.51	0.52	1.29
giraffe	1.37	0.29	2.15	0.00	76.64	19.12	0.00	0.00	0.00	0.43
horse	0.00	0.00	2.19	0.00	6.38	85.10	0.00	0.71	5.62	0.00
rat	7.65	0.93	0.71	0.00	0.00	0.00	87.90	0.00	0.64	2.17
seal	0.00	7.67	1.39	5.79	0.00	0.00	1.91	43.41	0.00	39.83
sheep	0.36	0.68	0.00	0.00	0.00	20.91	2.51	0.66	73.17	1.71
walrus	0.00	13.71	0.39	3.96	0.00	0.00	1.25	13.66	0.00	67.03

Figure 5: Confusion matrices for unseen classes of the proposed model on the AWA2 dataset.

identify most of unseen classes, except “bat” (46.39%), “dolphin” (25.67%) and “seal” (43.41%) on AWA2 dataset. We also observe that RDCN achieves appealing results on some classes, such as “blue+whale” (91.19%), “rat” (87.90%) and “horse” (85.10%). Considering the unseen samples are unavailable in training process, it strongly supports the superiority of the proposed method for ZSL task.

The t-SNE model [64] is used to project samples and prototypes from the semantic space to the 2D plane. Its main function is to visualize the distance between the sample and the corresponding class prototype. We selected seven seen classes and five unseen classes from the AWA2 dataset to check whether the prototype was learned correctly. Figure 6 and Figure 7 give the visualization results. It can be seen intuitively that most of the samples are located near the prototype of the corresponding class, which indicates that the RDCN can learn proper mapping from the feature space to the semantic space.

5. Conclusion

In this work, we have proposed a relation-based discriminative cooperation network to address the zero-shot classification problem. It keeps the discriminative information by separating the inter-classes and cluster the intra-classes with a margin. In addition, a

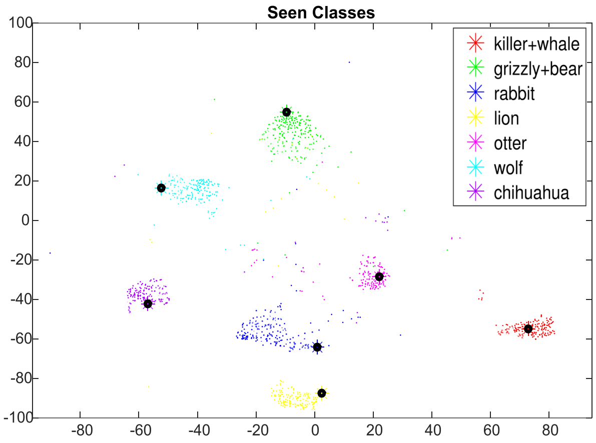


Figure 6: The tSNE visualisation of the visual features of training seen class samples from the AwA2 dataset together with the projected class prototypes for the proposed model. Prototypes is denoted by "*" and black circles are used to mark them visible.

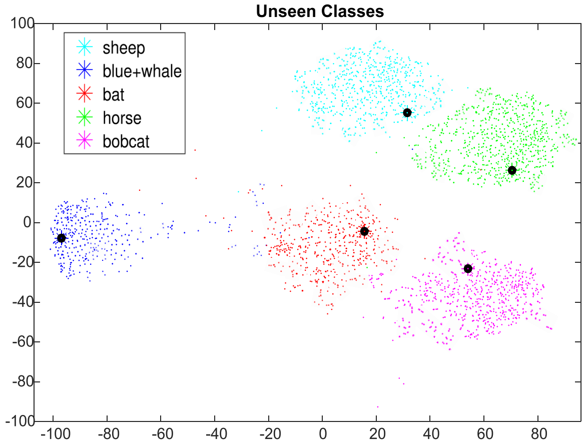


Figure 7: The tSNE visualisation of the visual features of test unseen class samples from the AwA2 dataset together with the projected class prototypes for the proposed model. Prototypes is denoted by "*" and black circles are used to mark them visible.

pivot regularization is utilized to ensure the cooperated semantic structures discriminative. Finally, relation module is introduced to measure the relationship between visual and semantic features. Experimental results on six benchmarks with multiple settings including both ZSL and GZSL demonstrated the superiority of the proposed model for zero-shot classification.

In the future, since the acquisition of attributes requires prior knowledge, we plan to exploit some other semantic information to construct the common space, e.g., click-through data. Moreover, we will exploit GAN based generative methods to establish a more robust representation in RDCN for zero-shot and few-shot classification.

Acknowledgment

This work is supported by National Natural Science Foundation of China under Grant 61906141, the National Natural Science Foundation of Shaanxi Province under Grant No. 2020JQ-317, China Postdoctoral Science Foundation (Grant No. 2019M653564), the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2021B02), Open Project Program of the S-tate Key Lab of CAD&CG (Grant No. A2018), Zhejiang University and the Fundamental Research Funds for the Central Universities.

References

[1] C. Geng, L. Tao, S. Chen, Guided cnn for generalized zero-shot and open-set recognition using visual and semantic prototypes, Pattern Recognition 102 (2020) 107263.

[2] Z. Cao, J. Lu, S. Cui, C. Zhang, Zero-shot handwritten chinese character recognition with hierarchical decomposition embedding, Pattern Recognition 107 (2020) 107488.

[3] X. Li, M. Fang, H. Li, J. Wu, Zero shot learning based on class visual prototypes and semantic consistency, Pattern Recognition Letters 135 (2020) 368–374.

[4] Y. Liu, X. Gao, Q. Gao, J. Han, L. Shao, Label-activating framework for zero-shot learning, Neural Networks 121 (2020) 1–9.

[5] R. Gao, X. Hou, J. Qin, J. Chen, L. Liu, F. Zhu, Z. Zhang, L. Shao, Zero-vae-gan: Generating unseen features for generalized and transductive zero-shot learning, IEEE Transactions on Image Processing 29 (2020) 3665–3680.

[6] C. H. Lampert, H. Nickisch, S. Harmeling, Attribute-based classification for zero-shot visual object categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (3) (2014) 453–465.

[7] R. Del Chiaro, A. D. Bagdanov, A. Del Bimbo, Webly-supervised zero-shot learning for artwork instance recognition, Pattern Recognition Letters 128 (2019) 420–426.

[8] Z. Li, L. Yao, X. Chang, K. Zhan, J. Sun, H. Zhang, Zero-shot event detection via event-adaptive concept relevance mining, Pattern Recognition 88 (2019) 595–603.

[9] R. Socher, M. Ganjoo, C. D. Manning, A. Ng, Zero-shot learning through cross-modal transfer, in: Advances in neural information processing systems, 2013, pp. 935–943.

[10] E. H. Huang, R. Socher, C. D. Manning, A. Y. Ng, Improving word representations via global context and multiple word prototypes, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Association for Computational Linguistics, 2012, pp. 873–882.

[11] Radovanovi, Nanopoulos, Alexandros, Ivanovi, Mirjana, Hubs in space: Popular nearest neighbors in high-dimensional data, Journal of Machine Learning Research 11 (5) (2010) 2487–2531.

[12] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, T. M. Hospedales, Learning to compare: Relation network for few-shot learning, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 1199–1208.

[13] Y. Xian, C. H. Lampert, B. Schiele, Z. Akata, Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly, arXiv preprint arXiv:1707.00600.

[14] F. Zhang, G. Shi, Co-representation network for generalized zero-shot learning, in: ICML, 2019.

[15] J. W. Zhang, J. Yu, D. Tao, Local deep-feature alignment for unsupervised dimension reduction, IEEE Transactions on Image Processing 27 (2018) 2420–2432.

[16] J. Yu, M. Tan, H. Zhang, D. Tao, Y. Rui, Hierarchical deep click feature prediction for fine-grained image recognition., IEEE transactions on pattern analysis and machine intelligence.

[17] M. Hou, W. Xia, X. Zhang, Q. Gao, Discriminative comparison classifier for generalized zero-shot learning, Neurocomputing 414 (2020) 10–17.

[18] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, in: NIPS, 2005.

[19] E. Kodirov, T. Xiang, S. Gong, Semantic autoencoder for zero-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3174–3183.

[20] B. Romera-Paredes, P. Torr, An embarrassingly simple approach to zero-shot learning, in: International Conference on Machine Learning, 2015, pp. 2152–2161.

[21] Y. Liu, D. Xie, Q. Gao, J. Han, S. Wang, X. Gao, Graph and autoencoder based feature extraction for zero-shot learning, in: IJCAI, 2019.

[22] Y. Shi, W. Wei, Discriminative embedding autoencoder with a regressor feedback for zero-shot learning, IEEE Access 8 (2020) 11019–11030.

[23] Y. Annadani, S. Biswas, Preserving semantic relations for zero-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7603–7612.

[24] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al., Devise: A deep visual-semantic embedding model, in: Advances in neural information processing systems, 2013, pp. 2121–2129.

[25] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, 1998.

[26] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient estimation of word representations in vector space, CoRR abs/1301.3781.

[27] R. Socher, M. Ganjoo, C. D. Manning, A. Y. Ng, Zero-shot learning through cross-modal transfer, in: NIPS, 2013.

[28] H. Zhang, J. Liu, Y. Yao, Y. Long, Pseudo distribution on unseen classes for generalized zero shot learning, Pattern Recognition Letters 135 (2020) 451–458.

[29] H. Zhang, P. Koniusz, Zero-shot kernel learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7670–7679.

[30] Y. Liu, J. Li, X. Gao, A simple discriminative dual semantic auto-encoder for zero-shot classification, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020) 4053–4057.

[31] H. Zhang, L. Liu, Y. Long, Z. Zhang, L. Shao, Deep transductive network for generalized zero shot learning, Pattern Recognition 105 (2020) 107370.

[32] S. M. Shojaee, M. S. Baghshah, Semi-supervised zero-shot learning by a clustering-based approach, ArXiv abs/1605.09016.

[33] Q. Wang, K. Chen, Zero-shot visual recognition via bidirectional latent embedding, International Journal of Computer Vision 124 (3) (2017) 356–383.

[34] V. K. Verma, P. Rai, A simple exponential family framework for zero-shot learning, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2017, pp. 792–808.

[35] J. Li, X. Lan, Y. Long, Y. Liu, X. Chen, L. Shao, N. Zheng, A joint label space for generalized zero-shot classification, IEEE Transactions on Image Processing 29 (2020) 5817–5831.

[36] N. Xing, Y. Liu, H. Zhu, J. Wang, J. Han, Zero-shot learning via discriminative dual semantic auto-encoder, IEEE Access 9 (2021) 733–742.

[37] Z. Zhang, V. Saligrama, Zero-shot learning via semantic similarity embedding, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4166–4174.

[38] Y. Xian, T. Lorenz, B. Schiele, Z. Akata, Feature generating networks for zero-shot learning, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 5542–5551.

[39] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan, ArXiv abs/1701.07875.

[40] A. Mishra, M. S. K. Reddy, A. Mittal, H. A. Murthy, A generative model for zero shot learning using conditional variational autoencoders, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2018) 2269–22698.

[41] K. Sohn, H. Lee, X. Yan, Learning structured output representation using deep conditional generative models, in: NIPS, 2015.

[42] G. Arora, V. K. Verma, A. Mishra, P. Rai, Generalized zero-shot learning via synthesized examples, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 4281–4289.

[43] D. P. Kingma, M. Welling, Auto-encoding variational bayes, CoRR abs/1312.6114.

[44] M. Chen, Z. E. Xu, K. Q. Weinberger, F. Sha, Marginalized denoising autoencoders for domain adaptation, ArXiv abs/1206.4683.

[45] Y. Li, J. Zhang, J. Zhang, K. Huang, Discriminative learning of latent features for zero-shot recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[46] G. Patterson, C. Xu, H. Su, J. Hays, The sun attribute database: Beyond categories for deeper scene understanding, International Journal of Computer Vision 108 (1-2) (2014) 59–81.

[47] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset.

[48] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1778–1785.

[49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[50] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980.

[51] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[52] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, J. Dean, Zeroshot learning by convex combination of semantic embeddings, in: In Proceedings of ICLR, Citeseer, 2014.

[53] L. Chen, H. Zhang, J. Xiao, W. Liu, S.-F. Chang, Zero-shot visual recognition using semantics-preserving adversarial embedding networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1043–1052.

[54] S. Liu, M. Long, J. Wang, M. I. Jordan, Generalized learning with deep calibration network, in: Advances in Neural Information Processing Systems, 2018, pp. 2009–2019.

[55] J. Liu, X. Li, G. Yang, Cross-class sample synthesis for zero-shot learning, in: British Machine Vision Conference, 2019.

[56] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, B. Schiele, Latent embeddings for zero-shot classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 69–77.

[57] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for image classification, IEEE transactions on pattern analysis and machine intelligence 38 (7) (2016) 1425–1438.

[58] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2927–2936.

[59] S. Changpinyo, W.-L. Chao, B. Gong, F. Sha, Synthesized classifiers for zero-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5327–5336.

[60] H. Zhang, Y. Long, Y. Guan, L. Shao, Triple verification network for generalized zero-shot learning, IEEE Transactions on Image Processing 28 (1) (2019) 506–517.

[61] H. Zhang, Y. Long, W. Yang, L. Shao, Dual-verification network for zero-shot learning, Information Sciences 470 (2019) 43–57.

[62] H. Zhang, H. Mao, Y. Long, W. kou Yang, L. Shao, A probabilistic zero-shot learning method via latent nonnegative prototype synthesis of unseen classes., IEEE transactions on neural networks and learning systems.

[63] J. Li, X. Lan, Y. Liu, L. Wang, N. Zheng, Compressing unknown images with product quantizer for efficient zero-shot classification, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 5458–5467.

[64] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (Nov) (2008) 2579–2605.

Author biography

Yang Liu received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 2013, 2015 and 2018, respectively. He is currently a Post-Doctoral Researcher in Xidian University, Xi'an, China. He has authored nearly 20 technical articles in refereed journals and proceedings, including IEEE Trans. Image, IEEE Trans. Cybernetics, PR, CVPR, AAAI, and IJCAI. His research interests include dimensionality reduction, pattern recognition, and deep learning.

Xinbo Gao received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. He was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan, from 1997 to 1998. From 2000 to 2001,he was a Post-Doctoral Research Fellow with the Department of Information Engineering, Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is currently a Professor of pattern recognition and intelligent systems, and the Director of the State Key Laboratory of Integrated Services Networks, Xidian University. He has authored five books and around 150 technical articles in refereed journals and proceedings, including IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Systems, Man and Cybernetics, and Pattern Recognition in his areas of expertise. His current research interests include computational intelligence, machine learning, computer vision, pattern recognition and wireless communications.

Quanxue Gao received the B.Eng. degree from Xi’an Highway University, Xi’an, China, in 1998, the M.S. degree from the Gansu University of Technology, Lanzhou, China, in 2001, and the Ph.D. degree from Northwestern Polytechnical University, Xi’an, in 2005. He was an Associate Research with the Biometrics Center, The Hong Kong Polytechnic University, Hong Kong, from 2006 to 2007. From 2015 to 2016, he was a Visiting Scholar with the Department of Computer Science, The University of Texas at Arlington, Arlington, USA. He is currently a Professor with the School of Telecommunications Engineering, Xidian University, and also a Key Member of the State Key Laboratory of Integrated Services Networks. His current research interests include pattern recognition and machine learning.

Jungong Han is a tenured faculty member with the School of Computing and Communications at Lancaster University, Lancaster, UK. His current research interests include multimedia content identification, multisensor data fusion, computer vision, and multimedia security. Dr. Han is an Associate Editor of Neurocomputing (Elsevier), and an Editorial Board Member of Multimedia Tools and Applications (Springer).

Ling Shao (M’09–SM’10) was a Professor with the School of Computing Sciences, University of East Anglia, Norwich, U.K. He is currently the CEO and Chief Scientist of the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and several other journals.

Declaration of interest statement

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted