Feature Fine-Tuning and Attribute Representation Transformation for Zero-Shot Learning

Shanmin Pang[a], Xin He[a], Wenyu Hao[a], Yang Long[b,*]

[a]School of Software Engineering, Xi'an Jiaotong University, Xi'an, 710049, China
[b]Department of Computer Science, Durham University, Durham, UK

Abstract

Zero-Shot Learning (ZSL) aims to generalise a pretrained classification model to unseen classes with the help of auxiliary semantic information. Recent generative methods are based on the paradigm of synthesizing unseen visual data from class attributes. A mapping is learnt from semantic attributes to visual features extracted by a pre-trained backbone such as ResNet101 by training a generative adversarial network. Considering the domain-shift problem between pre-trained backbone and task ZSL dataset as well as the information asymmetry problem between images and attributes, this manuscript suggests that the visual-semantic balance should be learnt separately from the ZSL models. In particular, we propose a plug-and-play Attribute Representation Transformation (ART) framework to pre-process visual features with a contrastive regression module and an attribute place-holder module. Our contrastive regression loss is a tailored design for visual-attribute transformation, which gains favorable properties from both classification and regression losses. As for the attribute place-holder module, an end-to-end mapping loss function is introduced to build the relationship between transformed features and semantic attributes. Experiments conducted on five popular benchmarks manifest that the proposed ART framework can significantly benefit existing generative models in both ZSL and generalized ZSL settings.

Keywords: Generalized zero-shot learning, Generative adversarial networks, Data distribution, Information asymmetric problem

1. Introduction

Traditional visual recognition task has made remarkable progress with pretrained deep models on large-scale datasets. However, most existing deep models struggle to generalise to new classes while maintain the performance on training classes. Zero-Shot Learning (ZSL) aims to match visual features with auxiliary semantic information so as to achieve the generalisation via inference. Since the distributions of semantic attributes are more consistent compared to those of visual features, most of recent ZSL models adopt the generative paradigm, such as conditional Generative Adversarial Networks (GANs) [1], that first synthesizes unseen class visual features from manually defined attributes [2, 3, 4] or discriminative latent attributes [5, 6, 7] and then trains a unified supervised classifier for both seen and unseen classes. Such a paradigm has shown promising progress to overcome the problems caused by the imbalance between seen and unseen classes.

Most of existing generative methods synthesize unseen class features by learning a mapping from semantic attributes to visual features originally extracted by a pre-trained backbone. Generally, the pre-trained backbone is ResNet101 [8] trained on ImageNet1K. Nevertheless, the distribution of ImageNet1K can be obviously different from the distribution of ZSL datasets. For this reason, the information of extracted visual features is not exactly consistent with the information of data from specific zero-shot learning datasets. Since the attributes are heavily dependent on the specific dataset, it is difficult to build a good correspondence between visual features and attributes if we utilize the pre-trained visual features directly.

In addition, the visual features and attribute vectors always face the information asymmetry problem. More specifically, the attributes are difficult to describe all details of visual features especially for those attributes embedded by sentence description. This causes that generative models are nothing more than over-fitting representations of seen classes. In other words, these models are difficult to generalize to unseen classes.

In order to verify the existence of the information asymmetry problem, we design a simple experiment in Fig. 1. In particular, similar to [9] that utilizes the idea of attention, we train a basic two-layer fully connected network which builds a mapping between input features and attributes. The first layer is the attention layer activated by the 'sigmoid' function, and the output will do dot multiplication with input features. The second layer is an encoder that maps from features with attention to attributes. We visualize the output of the attention layer in Fig. 1, where one can find most dimensions of original features are weakly related to attributes. This implies that the information of original visual features is much richer than attributes. This kind of information gap is a manifestation of information asymmetry problem.

To tackle the information asymmetry and domain-shift problems, we believe it is beneficial to transform original visual features to a new space where we can make a better alignment

*Corresponding author
Email addresses: pangsm@xjtu.edu.cn (Shanmin Pang), xjtu1hexin@gmail.com (Xin He), haowenyu@stu.xjtu.edu.cn (Wenyu Hao), yang.long@ieee.org (Yang Long)

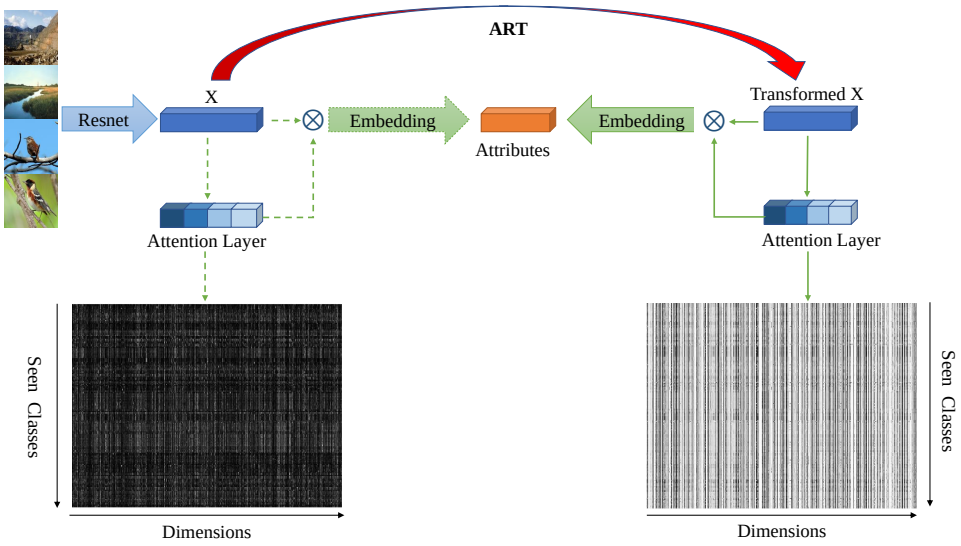


Figure 1: The illustration of Information Asymmetry problem. We suppose that only a few dimensions in visual features are related to semantic attributes. To verify the assumption, we build a mapping from input visual features to attributes, where we utilize an attention layer activated by the 'sigmoid' function to capture the dimensions that are strongly related to attributes. We visualize the average output of attention layer for each seen classes in the figure, where the black to white represents the output values varied from 0 to 1. Obviously, transformed features are more closely related to attributes than original features.

between transformed visual features and provided semantic attributes. In particular, we propose a new Attribute Representation Transformation (**ART**) paradigm that can effectively reduce the attribute irrelevant information while remain the required classification information as much as possible for both seen and unseen visual features. The proposed ART framework contains a contrastive regression module and an attribute placeholder module. A variant form of supervised contrastive loss which can strengthen the classification ability of transformed features is adopted for the former, and an end-to-end mapping loss is adopted for the latter to build the relationship between transformed features and attributes. The similar experiment as that in the Fig.1 left is conducted for transformed features. From the right of Fig. 1, we find that most dimensions of transformed features are closely related to attributes.

In summary, we make the following contributions:

- We suggest a two-stage process for future ZSL paradigm. The visual-semantic balance is implemented in the first pre-processing framework ART and prevent the impacts caused by the two problems described above. ART is a universal framework which can be adapted easily to most of existing zero-shot learning methods.

- We propose a variant loss function of supervised contrastive loss which falls between classification and regression tasks to strengthen the classification ability of transformed features for the first module of ART framework.

- We introduce an end-to-end mapping loss function which considers both attribute prediction and classification tasks to build the relationship between transformed features and semantic attributes for the second module of ART framework.

Extensive experiments on five static benchmarks (CUB, AwA1, AwA2, SUN and FLO) as well as two acitivity recognition datasets (UCF and HMDB) show that the proposed approach greatly improves the performance of our baseline generative model f-VAEGAN [10] under both the traditional ZSL and generalized ZSL (GZSL) tasks.

2. Related Work

Zero-shot learning (ZSL) image classification is a challenging task that image samples from seen classes take part in model training and ones from unseen classes are responsible for model evaluation, where the seen and unseen classes are disjoint. The main strategy for ZSL is to build up the relationship between seen and unseen classes through intermediate semantic attributes, which can be defined in different forms, such as binary vector [2], text descriptions [3] and word2vectors [4]. In this section, we provide an overview of related works on the ZSL problem.

2.1. Visual-Semantic Embedding Methods

A large number of methods address the ZSL problem by learning an embedding/mapping between visual features and semantic attributes so that features from unseen classes can find their class prototypes by the mapping. The embedding space can be roughly grouped into semantic embedding, visual embedding and latent space embedding. Semantic embedding methods learn the mapping from visual features to semantic attributes [11, 12, 13, 14]. For instance, Socher *et al.* [15] utilize a simple linear model to project images into the 50-dimensional word space. To provide a better embedding on the semantic

side, Bucher *et al.* [16] jointly optimize the attribute embedding and the classification metric in a multi-objective framework. In contrast to semantic embedding, visual embedding learns a reverse projection to map semantic attributes back into visual space so as to make the semantic representations close to their corresponding visual features [17, 18, 19]. For example, based on selecting a fixed number of samples from each class across all training classes, SCILM [19] embeds the class semantics into the visual space under the supervision of the class visual prototype, and yields a general semantic-visual interaction model. The aforementioned models force the projection functions in the space of semantic modality or in the space of visual modality. However, it is a challenging issue to learn an explicit projection function between two spaces due to the distinctive properties of different modalities [20]. To tackle the problem, a number of approaches [21, 22, 23, 24] discuss the idea of embedding features and attributes into another intermediate space.

Despite promising in ZSL, some embedding based methods [25, 2, 26, 18] give unsatisfactory performance in the generalized ZSL [27] setting due to over-fitting. In this manuscript, we propose to transform original visual features in a new space so that feature representations in the new space are closely related to attributes. To accomplish the goal, we build a mapping from visual features to attributes and adopt an end-to-end cross entropy loss.

2.2. Generative Methods

To mitigate the data imbalance problem between seen and unseen classes, Generative Adversarial Networks (GANs) [21, 28, 29] have been employed to synthesize features of unseen classes. As such, the ZSL problem can be transformed to a traditional classification task. The first attempt is f-CLSWGAN [30] that uses a conditional Wasserstein-GAN [29, 1] to synthesize features of unseen classes based on semantic attributes. Since then, generative methods become more and more pervasive in zero-shot learning. For example, LisGAN [31] captures several soul samples from different views of an object and guides synthesized samples close to at least one soul sample. Felix *et al.* [32] map synthesized samples back to semantic attributes. CE-GZSL [33] tries to combine the generative methods and semantic embedding methods together. Noting that the instability of the training of GAN, some approaches such as [34, 35] propose to use Variational Autoencoders (VAE) [36] for the ZSL problem. Apart from GANs and VAEs, a large number of studies employ various generative models such as compositional learning [37, 38] and autoencoders [39, 40] to improve GZSL performance. For instance, Ji *et al.* [40] recently propose a simple yet effective meta-learning-based unseen prototype learning framework, which learns visual prototypes from the corresponding class-level semantic prototypes with an autoencoder framework. More types of methods can be found in a recent ZSL review [20].

Although these methods perform better than embedding based methods in GZSL, most of them synthesize original visual features directly, ignoring the problem of information asymmetry, i.e., visual features contain more details than attributes. As a result, there is a gap between features synthesized by utilizing attributes and the real features. We propose a feature transformation model to mitigate the gap.

Recently, there are some generative methods that have paid attention to optimize visual features such as [41] and [33]. However, it should be noted that these methods still synthesize original features and deal with features **after** synthesizing. The proposed pre-processing method can reduce the difficulty of building relationship between visual features and semantic attributes. As a result, we synthesize transformed features directly. Besides the performance improvement, this kind of two-stage framework has two important advantages. On the one hand, the proposed method does not change the cost of generative methods and the cost of pre-processing method is much less than generative methods. On the other hand, the proposed ART method is an universal framework that can be combined with most of existing generative zero-shot learning methods.

2.3. Zero-shot Event Recognition Methods

Although zero-shot image classification has been extensively studied, zero-shot learning for temporal events such as activity and gesture recognition in videos has gained much less attention [42]. As in ZSL, zero-shot event recognition methods can also be roughly grouped into two categories: embedding-based and generative-based methods. Nevertheless, being different from that embedding and generative methods are running neck and neck in ZSL, embedding methods absolutely dominate the field of zero-shot event recognition. The common practice for the embedding methods is to map the visual embedding to a semantic embedding space [42, 43, 44], or project visual and semantic embeddings into a common space [45, 46]. Recently, a few methods attempt to utilize GANs to synthesize unseen class video features. For example, Zhang and Peng [47] synthesize video features of unseen categories with a GAN model to build seen-to-unseen correlation for action recognition. Mandal *et al.* [48] adopt the conditional Wasserstein GAN with additional loss terms to synthesize unseen features for training an out-of-distribution detector.

While the focus of this manuscript is to increase zero-shot image classification performance, we further evaluate the proposed model on two activity datasets, namely, UCF101 [49] and HMDB51 [50], to demonstrate the versatility of our model.

3. Methodology

In this section, we first explain the whole idea of two-stage framework. And then, we describe two modules of proposed pre-processing method ART respectively. Finally, for easy understanding, we give the overall algorithm process.

3.1. Two-Stage Balancing Framework

The basic assumption of ZSL is that the inference can go across seen and unseen classes according to their semantic relationship. Thus, it is significant to build up the relationship between visual features $\mathcal{X} = \{x_i, i = 1, 2, \cdots, N\}$ and semantic

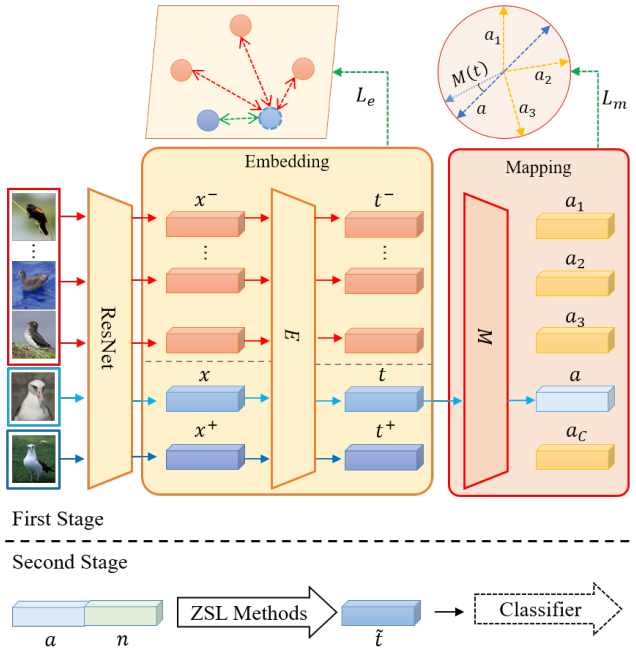


Figure 2: The illustration of ART. The whole transformation framework includes two modules: contrastive regression module E which is shown as "Embedding" and attribute place-holder module M which is shown as "Mapping". Where the former embeds original visual features to new ones, and the latter maps the transformed features to semantic attributes. We split the whole zero-shot learning task into two stages. At the first stage, we train the transformation model and get the transformed visual features t . At the second stage, we synthesize the transformed features instead of original visual features.

attributes $\mathcal{A} = \{a_i, i = 1, 2, \cdots, N\}$. Here $x_i \in \mathbb{R}^D$, $a_i \in \mathbb{R}^K$, and N denotes the number of samples. Note that samples from the same class share the same attribute vector.

The generative ZSL approaches assume that unseen visual features can be synthesized by the semantic attributes of an unseen class. However, due to the information asymmetry and domain-shift problems described in the introduction section, there is some gap between synthesized visual features by generative methods or mapped features by the methods based on mapping which map from attributes to visual features and real features. We propose a two-stage zero-shot learning framework to alleviate the problems. The illustration of this kind of two-stage framework is shown in the Fig. 2. At the first stage, we propose a pre-processing method ART that transforms original features to new ones. Corresponding to the domain-shift and information asymmetry problems, the method includes two modules, *i.e.*, the contrastive regression module E and the attribute place-holder module M. At the second stage, we directly utilize the transformed features as input features for existing zero-shot learning methods. In the experiment section, we consider f-VAEGAN[10] as our baseline generative method and we do not do any changes except input visual features.

3.2. Contrastive Regression E

The first problem that the data distribution between the datasets for feature pre-training and specific zero-shot learning datasets is different, is similar to model fine-tuning problem in the domain of transfer learning. In the domain of transfer learning, it is common to utilize the pre-trained backbone to train a specific task on the new dataset. Despite many limitations, ZSL is still a classification task. As a result, similar to the general practice of fine-tuning the models, we can alleviate the problems caused by the difference of data distribution by enhancing the classification ability of transformed new features.

The contrastive regression module E is designed for above purpose. In practice, we design E as a simple single fully connected layer to avoid over-fitting. Mathematically, $t = E(x)$, where x denotes the original features extracted by the pre-trained backbone. Making transformed features more discriminative is a classification task while constraints on features are generally regression tasks. Motivated by that, we use the L_2 constraint to modify the contrastive loss function, and obtain a new variant as follows:

$$\mathcal{L}_e(t, y) = -\log \frac{\exp(-\|t - t_y^+\|_2)}{\exp(-\|t - t_y^+\|_2) + \sum_{i \neq y}^{c} \exp(-\|t - t_i^-\|_2)}, \quad (1)$$

where y is the corresponding label of the current sample, and c is the number of classes. For every sample t, we get one sample from every seen class and we represent the sample from the same class as t^+ and others as t^-. We can both effectively reduce intra-class distances as well as increase inter-class distances by optimizing the loss function $\mathcal{L}_e$. This loss function falls between regression loss and classification loss. As such, the classification ability of transformed features can be favorably enhanced.

3.3. Attributes Place-holder M

Neural networks can extract related information layer by layer. If we build a mapping between visual features and semantic attributes through a fully connected neural network with several layers, the output of each layer will become more and more abstract and the information will become closer and closer to semantic attributes.

Motivated by this principle, we design the attribute place-holder module M that maps the transformed features to attributes to address the information asymmetry problem. The loss function of M is given as,

$$\mathcal{L}_m(t, y) = -\log \frac{\exp(s(M(t), a_y))}{\sum_i^c \exp(s(M(t), a_i))}. \quad (2)$$

Similarly, y is the corresponding label of current sample t, a_y is the attribute vector of label y, and s means the similarity function which measures the similarity of $M(t)$ and a. In this paper, we define s as standard cosine similarity:

$$s(p, q) = \frac{p^T q}{\|p\|_2 \|q\|_2}. \quad (3)$$

The loss function is the variant form of cross entropy and we replace the inner product by the function s. The final loss function falls between attribute prediction and classification tasks.

During the training stage, the module E and the module M are jointly trained as a whole. As a result, $\mathcal{L}_m$ is an end-to-end

Algorithm 1 ART and Generative Zero-Shot Learning

Input: Seen features X_s, attributes A_s and labels Y_s; Unseen attributes A_u; Training epoch $\mathcal{N}_1$ for ART and $\mathcal{N}_2$ for generative ZSL methods; The number of synthesized samples n.

Output: Visual-to-category classifier C.

Initialize contrastive regression module E and normalizing attribute place-holder module M;
for epoch = 1 to $\mathcal{N}_1$ **do**
  Sample a batch $\{ x, a, y, x^+, x^- \}$;
  Compute $t = E(x)$ and $a' = M(t)$;
  Compute $\mathcal{L}_e$ and $\mathcal{L}_m$ and update the parameters of M and E;
end for
Compute transformed features $T_s = E(X_s)$;
Initialize ZSL generative models including the generator G and the discriminator D;
for epoch = 1 to $\mathcal{N}_2$ **do**
  Sample a batch $\{t, a\}$ from $\{T_s, A_s\}$, a batch z from Gaussian distribution $N(0, 1)$;
  Compute $t' = G(a, z)$, $D(t, a)$ and $D(t', a)$;
  Compute loss function of generative methods and update the parameters of G and D;
end for
Synthesized unseen features T'_u by the generator G and unseen attributes A_u.
Train the final classifier C with both T'_u and T_s.

classification loss function. After training process, we output the intermediate results $t = E(x)$ as transformed features which are more abstract and closer to semantic attributes than original features. In addition, although data from unseen classes can not take part in the training process, we compare the output $M(t)$ with attributes from all classes. This kind of place-holder idea constrains predicted attributes to be far from the attributes from unseen classes. We hope to avoid introducing more over-fitting through the place-holder idea.

3.4. The Whole Classification Process

By combining both E and M, we obtain the total loss of ART as follows.

$$\mathcal{L}_{art} = \mathcal{L}_m + \eta \mathcal{L}_e, \tag{4}$$

where η is a hyper-parameter. For easy understanding, the completed process is listed in Algorithm 1. We train ART at the first stage and extract transformed features. And then, we train the generative method f-VAEGAN to synthesize transformed features directly at the second stage. Finally, the GZSL classifier is trained with synthesized features of unseen classes and real features of seen classes.

4. Experiments

4.1. Datasets and Experimental Settings

Datasets. We validate our method on five widely-used datasets for zero-shot learning, namely, Animal with Attributes1 (AwA1) [2], Animal with Attributes2 (AwA2) [27], Caltech-UCSD-Birds-200-2011 (CUB) [51], Oxford Flowers (FLO) [52] and SUN attributes (SUN) [53].

Table 1: Details for benchmark datasets.

Dataset	AwA1	AwA2	CUB	SUN	FLO
# Images	30,475	37,322	11,788	14,340	8,189
# Attribute Dims.	85	85	312/1024	102	1024
# Seen classes	40	40	150	645	82
# Unseen classes	10	10	50	72	20

AwA1 contains 30,475 images and AwA2 contains 37,322 images. Both of them share the same 50 categories and each category is annotated with 85 attributes. We use the provided attributes as the class-level semantic descriptors.

CUB is an extended version of the CUB-200 dataset, which includes 11,788 images of 200 bird species. There are two kinds of semantic descriptors for CUB. Specifically, the dataset itself contains 312 dimensional binary attributes. The another are 1,024 dimensional class embeddings generated from textual descriptions [54].

FLO consists of 8,189 images which are derived from 102 flowers categories. We adopt the same 1,024 dimensional semantic descriptions which are similar to the 1,024 dimensional semantic descriptions in CUB.

SUN is a large-scale scene attribute dataset, which includes 717 categories and 14,340 images in total. Each category has 102 dimensional semantic attributes.

We divide AwA1, AwA2, CUB, SUN into seen classes and unseen classes according to the benchmark setting of [27], and divide FLO according to [54]. Besides, we use the default attributes included in the datasets as our semantic attributes. To be clear, details about these five datasets are reported in Tab. 1.

Experimental Settings. For a fair comparison, we adopt the general settings in zero-shot learning introduced by [27]. The input visual features are extracted by the pre-trained network ResNet-101. We evaluate the proposed method under two scenarios. As for the conventional ZSL, we only evaluate the per-class Top-1 accuracy on unseen classes. With regard to the GZSL, we evaluate the Top-1 accuracy on both seen and unseen classes, which are denoted as S and U, respectively. The performance of GZSL is measured by their harmonic mean:

$$H = \frac{2 * S * U}{S + U}. \tag{5}$$

Implementation Details. We implement ART with PyTorch. There are mainly two hyper-parameters in ART, namely, the dimensions of transformed features and η in Equation 4. In the experiments, we transform original visual features with 2,048 dimensions to new features with 1,024 dimensions for all five datasets and we activate the output of the module E by the "tanh" function. We set $\eta = 0.1$ for CUB and FLO and $\eta = 0.01$ for AwA1, AwA2 and SUN.

We adopt the same settings as [10] to implement f-VAEGAN. It is worth noting that, due to the change of synthesized visual features, the number of synthesized samples is different from original settings. Briefly, we synthesize 5,000 samples for each

Table 2: The effectiveness of the proposed two-stage framework. We combine the ART architecture with different zero-shot learning methods. "MBM-S" and "MBM-V" are simple mapping based methods. "MBM-S" optimizes $\mathcal{L} = max(0, \Delta - a^T E(x) + (a')^T E(x))$ and "MBM-V" optimizes $\mathcal{L} = max(0, \Delta + \|E(a) - x\|_2 - \|E(a') - x\|_2)$ where a' means the attribute vector from another random class which are different from current sample belonging to. The experiments are conducted on CUB.

Settings	Unseen	Seen	H
MBM-S	**6.1**	**46.0**	**10.7**
MBM-S + ART	5.8	28.7	9.7
MBM-V	12.6	42.0	19.3
MBM-V + ART	**18.3**	**56.1**	**27.6**
f-CLSWGAN[30]	43.3	59.9	50.3
f-CLSWGAN + ART	**53.7**	**61.1**	**57.2**
f-VAEGAN[10]	46.9	59.1	52.3
f-VAEGAN + ART	**55.4**	**62.3**	**58.6**

Table 3: The influence of different parts of ART to final GZSL evaluation results. The experiments are based on CUB.

$\mathcal{L}_e$	$\mathcal{L}_m$	Unseen	Seen	H
☐	☐	46.9	59.1	52.3
☑	☐	54.8	61.0	57.8
☐	☑	50.7	57.7	54.0
☑	☑	**55.4**	**62.3**	**58.6**

class for AwA1 and AwA2, 300 samples for CUB, 100 samples for SUN and 2,000 samples for FLO.

For training details, the settings are same on all five datasets. we use "Adam" as our optimizer with learning rate 0.0001 and beta1 0.5. The batch size is set as 64. These settings are suitable for both ART and f-VAEGAN in our experiments.

4.2. Ablation Study

The Impacts of Our Two-Stage Framework. In Tab. 2, we combine proposed ART with different zero-shot learning methods including two mapping based methods MBM-S and MBM-V, as well as two generative methods f-CLSWGAN and f-VAEGAN. According to the table, one can find that after combining with ART framework, the performance of MBM-S reduces a little and the performance of the other three methods improve obviously. These results show that proposed ART framework is effective to those methods which map from attributes to visual features. However, ART method may lead to a little more serious over-fitting problem for those methods that perform comparisons in attribute space. In addition, this table confirms that the proposed ART method is an universal framework and can be adapted to different methods.

The Impacts of Two Modules. Our ART consists of two modules E and M with their loss function $\mathcal{L}_e$ and $\mathcal{L}_m$. We set $\mathcal{L}_e = 0$ and $\mathcal{L}_m = 0$ in sequence to observe the influence of different parts. The results are shown in the Tab. 3.

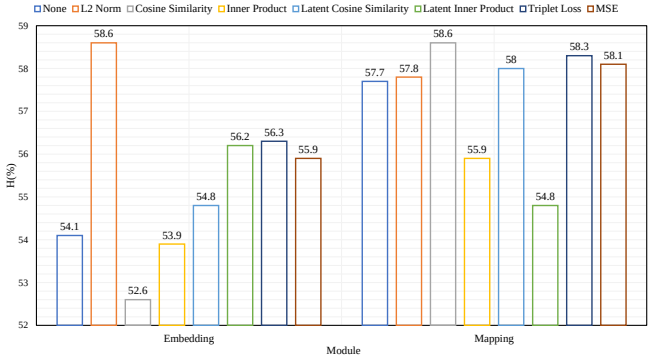
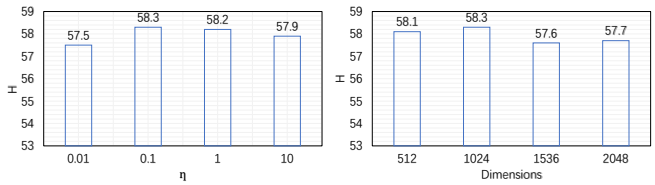


Figure 3: The influence of different forms of $\mathcal{L}_e$ and $\mathcal{L}_m$. In order to verify the effectiveness of proposed $\mathcal{L}_e$ and $\mathcal{L}_m$, we freeze the form of one module and change the form of another module. "None" means the results without corresponding module. Except for "Triplet Loss" and "MSE", different forms mean the different similarity metrics in the equations such as L2 Norm in equation (1) and cosine similarity in equation (2). The experiments are conducted on CUB.



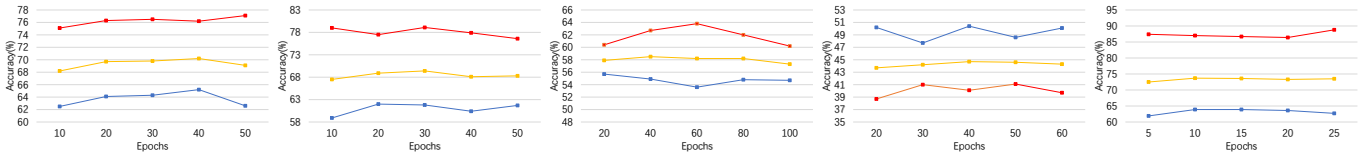
(a) The impact of η.

(b) The impact of the dimensionality of transformed features

Figure 4: The impacts of hyper-parameters η and the dimensionality of transformed features. The experiments are conducted on CUB.

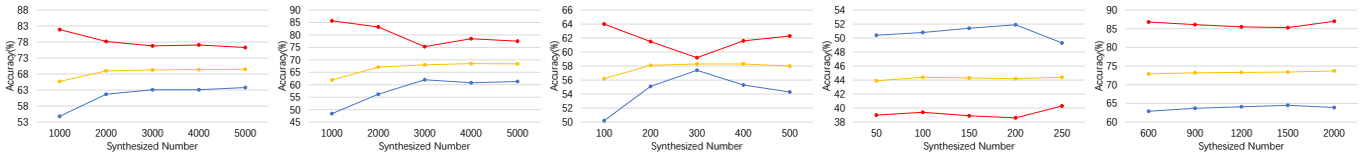
In Tab. 3, we reproduce the experiment of f-VAEGAN firstly. And then, we introduce $\mathcal{L}_m$ and $\mathcal{L}_e$ to the baseline respectively. From Tab. 3, we find that $\mathcal{L}_e$ plays a more important role in learning better transformed features and the performance can be improved from 52.3% to 57.8%. Besides, although $\mathcal{L}_m$ can not improve the performance as much as $\mathcal{L}_e$, it is complementary to $\mathcal{L}_e$ and improves the harmonic mean from 57.8% to 58.3%. The result of $\mathcal{L}_e$ may illustrate that requiring transformed features to be adapted to the distribution of specific datasets is more important to improve the performance of zero-shot learning task. Furthermore, the result of $\mathcal{L}_m$ verifies the existence of information asymmetry problem to a certain extent because enhancing the relationship between transformed features and attributes indeed works to further improve the performance.

The Influence of Different Forms of $\mathcal{L}_e$ and $\mathcal{L}_m$. Both $\mathcal{L}_e$ and $\mathcal{L}_m$ are variants of cross entropy loss. We utilize L2 norm to replace inner product for $\mathcal{L}_e$ because the contrastive regression module falls between classification and regression tasks. Similarly, we utilize cosine similarity to replace inner product for $\mathcal{L}_m$. So in the Fig.3, we verify the performance of different similarity metrics for these two loss function. It is worth noting that, for "Latent inner product" and "Latent cosine similarity", we add two extra fully connected layer to embed transformed visual features or predicted attributes to latent vectors and calculate the similarity with latent vectors. Additionally, we also



(a) AwA1 (b) AwA2 (c) CUB (d) SUN (e) FLO

Figure 5: The experimental results with different number of training epochs. We denote the mean accuracy of unseen classes, the mean accuracy of seen classes, and the harmonic mean in blue, red, and orange, respectively. As shown, the harmonic mean accuracy is relatively stable along with the increase of training epochs.



(a) AwA1 (b) AwA2 (c) CUB (d) SUN (e) FLO

Figure 6: The experimental results by varying the number of synthesized examples for each class. We denote the mean accuracy of unseen classes, the mean accuracy of seen classes, and the harmonic mean in blue, red, and orange, respectively. As shown, the harmonic mean accuracy is relatively stable on all evaluated datasets.

test two classical implement forms "Triplet Loss" and "MSE" for these two modules.

According to the Fig.3, we find that the performance of L2 norm in the equation (1) is much better than other similarity metrics which verifies that contrastive regression module indeed falls between classification and regression tasks as well as the effectiveness of $\mathcal{L}_e$. Besides, the experimental results for $\mathcal{L}_m$ verify that cosine similarity is better choice for attribute comparison in the zero-shot learning problem.

4.3. The Choices of Hyper-Parameters

We study the impacts of hyper-parameters to ART in this section. These include the trade-off scalar η in Eq. (4), the number of dimensions of transformed features, the number of training epochs and the number of synthesized samples for each class.

The impact of η. It is not our key point to find the best hyper-parameters for the ART framework. As a result, we simply try four orders of magnitude for η. To be specific, we set $\eta \in \{0.01, 0.1, 1.0, 10\}$ in the experiment. According to Fig. 4 left, ART is stable enough to η. In this paper, we finally choose $\eta = 0.1$ for five datasets.

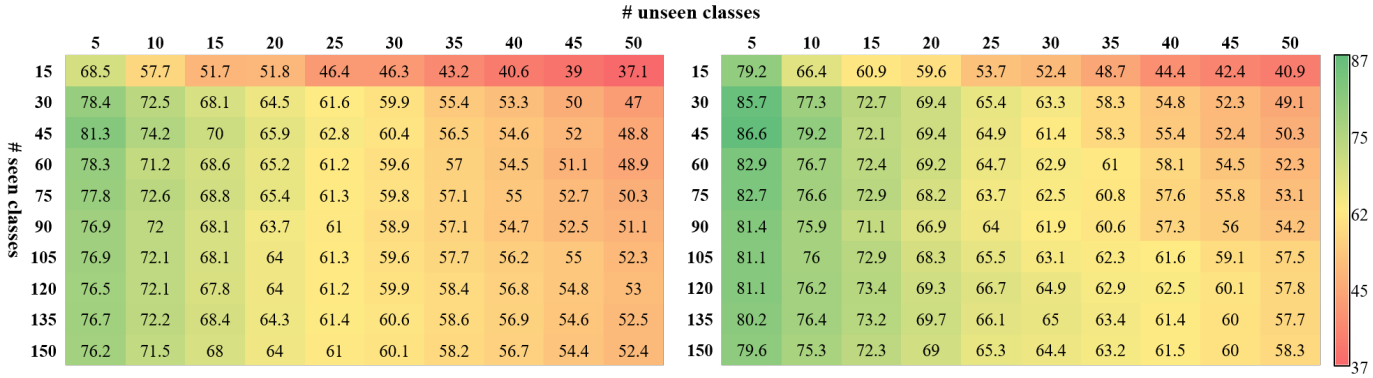
The number of dimensions of Transformed Features. We embed the original visual features to transformed features with a number of different dimensions including 512, 1,024, 1,536 and 2,048. The experimental results are illustrated in Fig. 4 right. From the figure, we can find that the number of dimensions does not effect the performance too much. Since 1,024 dimensions gives slightly better performance compared to the other three dimensions, we therefore set transformed features be 1,024 dimensions for all datasets.

The Number of Training Epochs. The ART framework may likely lead to over-fitting to zero-shot learning since we can only utilize the samples from seen classes during its training process. Therefore, it is necessary to explore the stability of ART training. We study the performance curves with increased the number of training epochs and the experimental results are shown in Fig. 5. From the figure, we can find that although there are some fluctuations in the mean accuracy of seen classes and unseen classes, the harmonic average H remains relatively stable along with the increase of training epochs. This illustrates that the ART framework does not necessarily introduce over-fitting to the feature generation model in practice.

The Number of Synthesized Features for Each Class. In order to increase the diversity of synthesized unseen features, we may need to synthesize more samples than our baseline. So we further explore the impact of the number of synthesized features for each class to recognition performance. The experimental results are shown in the Fig. 6. As shown, the harmonic mean accuracy is relatively stable in a wide range of the number of synthesized features. We finally synthesize 5,000 samples for each class of AwA1 and AwA2, 300 samples for each class of CUB, 100 samples for each class of SUN and 2,000 samples for each class of FLO.

4.4. Effectiveness of Our Approach

The Recognition Accuracy under Different Data Imbalances. We examine the effect of the imbalance of the number of seen classes and unseen classes in Fig. 7 from two perspectives under the GZSL setting on the CUB dataset. In particular, we first fix the training seen classes, and explore the recognition performance under the varied number of unseen classes. During this process, we compute the average cosine similarities among unseen classes, and remove one class from testing if it gets the smallest average cosine similarity with other unseen classes. The corresponding experimental results of the baseline method f-VAEGAN and our two stage method ART+f-VAEGAN are shown in the rows of Fig. 7. As shown, both f-VAEGAN and ART+f-VAEGAN have steadily performance improvement along with the reduced number of unseen classes.



(a) f-VAEGAN (b) ART+f-VAEGAN

Figure 7: The effect of the imbalance of the number of seen classes and unseen classes on the CUB dataset.

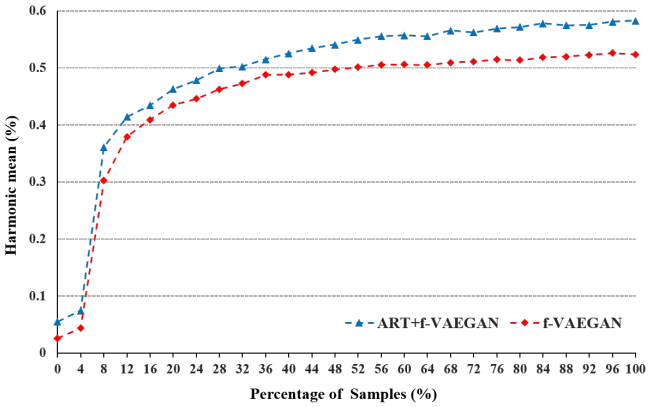


Figure 8: The experimental results of ART+f-VAEGAN and f-VAEGAN in a sequential task on the CUB dataset. Note that the zero value at x-axis means we use 1 observation of 1 class for training.

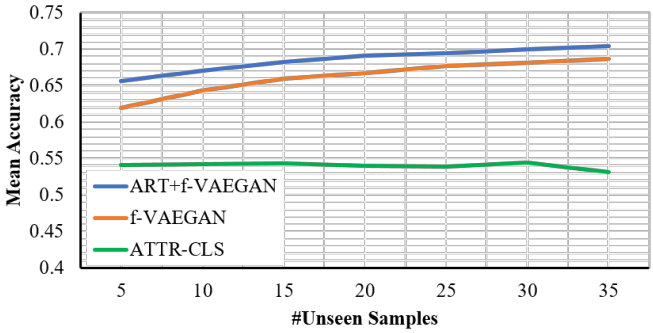


Figure 9: Within-Category Experiments on the CUB dataset. We combine a number of unseen samples varied from 5 to 35 per class with the original training samples to train the classification model.

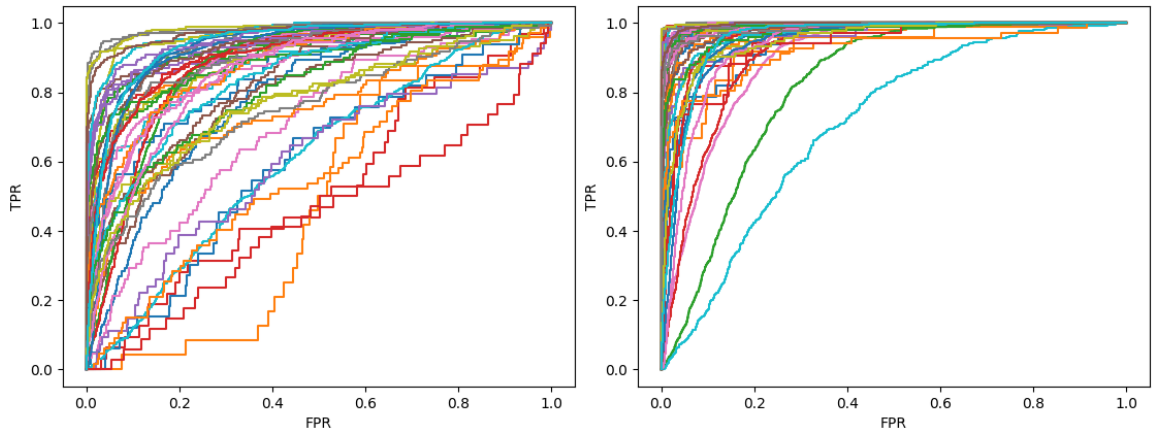
Besides, since ART can better utilize attributes to synthesize visual features, it is always effective to increase the performance of f-VAEGAN.

Second, we keep the unseen classes unchanged, and vary the number of seen classes in the same way. The corresponding results are listed in the columns of Fig. 7. As shown, given the same unseen classes, the performance gap among different number of seen classes is usually not large. This illustrates that generative models are robust to the imbalance of the number of seen classes and unseen classes. In this case, ART also consistently improves the baseline f-VAEGAN. This again demonstrates the effectiveness of our method.

Experimental Results in the Form of Sequential Learning. By gradually adding the number of samples, we evaluate the performance of recognition systems in a sequential task on the CUB dataset. Specifically, we assume that 1 observation of 1 class was given only at the beginning, and use that for training. Then, we sequentially add another one, and so on and so for (In the implementation, considering that the number of samples for different classes is not necessarily the same, we add a fixed portion of samples at each time to perform sequential learning). As such, we obtain the performance curves along with percentage of samples. During the sequential learning process, we keep the testing set unchanged.

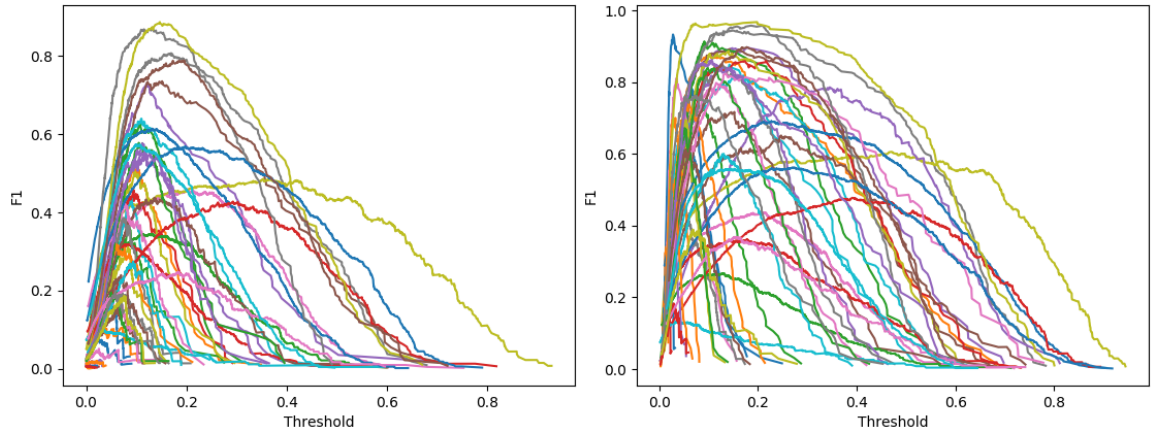
Fig. 8 gives the recognition accuracy in terms of harmonic mean under the GZSL setting. As illustrated, the evaluated systems can make effective recognition when about 8% training samples involved in the sequential training. Besides, the performance curves become relatively stable when the percentage of training samples is greater than 50%. In this sequential task, ART+f-VAEGAN exhibits superior performance the baseline f-VAEGAN, and the improvement gain is large than 3% in most cases. This shows ART favourably reduces the negative effects of domain-shift and information asymmetry.

Within-Category Experiments. We also examine the effectiveness of ART through conducting within-category experiments. In particular, we evaluate our method to recognize categories via attributes. In other words, we assume that samples present in training and test partitions belong to both seen and unseen classes. To the end, we put a small proportion of samples from unseen classes into the training set, and convert the ZSL problem into an imbalanced classification problem. We compare the mean accuracy of all classes of ART+f-VAEGAN and f-VAEGAN. Besides, we implement a simple two-layer MLP called ATTR-CLS, which maps visual features to semantic attributes and considers the most similar attributes as the



(a) f-VAEGAN (b) ART+f-VAEGAN

Figure 10: The experimental results in terms of ROC curves on the AWA1 dataset. Note that FPR is short for false positive rate, and TPR denotes true positive rate. In this figure, each color corresponds to a test class.



(a) f-VAEGAN (b) ART+f-VAEGAN

Figure 11: The experimental results in terms of F1-score on the AWA1 dataset. Note that each color corresponds to a test class.

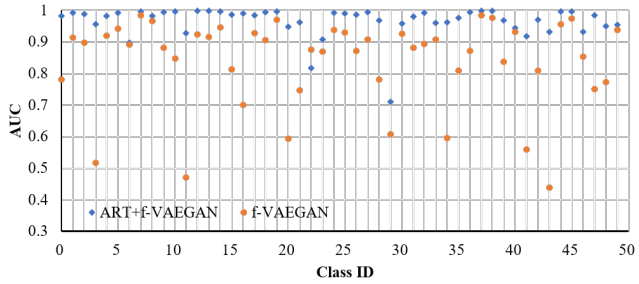


Figure 12: The experimental results in terms of AUC on the AWA1 dataset.

corresponding class label.

We carry out the experiment on the CUB dataset. We randomly select 60 training samples for each seen class, and combine them with a small number of samples from unseen classes to jointly train the classification model. In the test stage, we evaluate the performance with samples not involved in the training process. The experimental results are shown in Fig. 9. As shown, the proposed ART strategy is also beneficial to recognition accuracy. Since the number of seen samples is larger than that of unseen samples, both the generative methods ART+f-VAEGAN and f-VAEGAN significantly outperform the embedding method ATTR-CLS. This explains that generative methods can effectively reduce the class imbalance problem by synthesizing unseen examples.

Experimental Results under More Evaluation Metrics. In order to extensively show the effectiveness of ART, we assess performance with more metrics, namely, F1-score, ROC curves and area under ROC (AUC). In this experiment, given a class, we set samples in this class be true, and samples not in this class be false. Based on this protocol, we train the binary classification models for each class in the dataset, where we utilize original visual features for seen classes and synthesized features for unseen classes. We conduct the experiment on the AWA1 dataset as the number of classes for this dataset is much smaller than that for CUB, SUN and FLO. In addition, for a fair

Table 4: Score comparison on GZSL. U means the average top-1 accuracy of unseen classes, S means the average top-1 accuracy of seen classes, and H is the harmonic mean of U and S. We consider ART is an effective method to improve visual feature quality, and combine it with the previous generative method f-VAEGAN. The prefix "$*$" means that the attributes embedded by textual descriptions are adopted on CUB. The prefix "†" means the fine-tuned backbone is utilized to improve performance. We mark the best results with the bold fonts.

Method	CUB			FLO			SUN			AwA2			AwA1		
	U	S	H	U	S	H	U	S	H	U	S	H	U	S	H
f-CLSWGAN [30]	43.7	57.7	49.7	59.0	73.8	65.6	42.6	36.6	39.4	-	-	-	57.9	61.4	59.6
Cycle-WGAN [32]	47.9	59.3	53.0	61.6	69.2	65.2	47.2	33.8	39.4	-	-	-	59.6	63.4	59.8
LisGAN [31]	46.5	57.9	51.6	57.7	83.8	68.3	42.9	37.8	40.2	-	-	-	52.6	**76.3**	62.3
LsrGAN [55]	48.1	59.1	53.0	-	-	-	44.8	37.7	40.9	-	-	-	54.6	74.6	63.0
f-VAEGAN [10]	48.4	60.1	53.6	56.8	74.9	64.6	45.1	38.0	41.3	57.6	70.6	63.5	-	-	-
RFF(softmax) [56]	52.6	56.6	54.6	65.2	78.2	71.1	45.7	38.6	41.9	-	-	-	59.8	75.1	66.5
TF-VAEGAN [41]	52.8	64.7	58.1	62.5	84.1	71.7	45.6	40.7	43.0	59.8	75.1	66.6	-	-	-
$*$ CE-GZSL [33]	63.9	**66.8**	65.3	**69.0**	78.7	73.5	**48.8**	38.6	43.1	**63.1**	78.6	**70.0**	**65.3**	73.4	69.1
Our ART+f-VAEGAN	55.4	62.3	58.6	63.9	**87.0**	**73.7**	48.6	**41.1**	**44.6**	61.8	**79.1**	69.4	65.2	76.2	**70.2**
$*$ **Our ART+f-VAEGAN**	**68.4**	64.8	**66.5**	-	-	-	-	-	-	-	-	-	-	-	-
† f-VAEGAN [10]	63.2	75.6	68.9	63.3	92.4	75.1	50.1	37.8	43.1	57.1	76.1	65.2	-	-	-
† TF-VAEGAN [41]	63.8	**79.3**	**70.7**	69.5	92.5	79.4	41.8	**51.9**	46.3	55.5	**83.6**	66.7	-	-	-
† **Our ART+f-VAEGAN**	**66.6**	73.7	70.0	**70.1**	**92.6**	**80.1**	**52.4**	42.9	**47.2**	**63.0**	77.9	**69.7**	-	-	-

comparison, we save the trained generative model, and only optimize the classification models.

Fig. 10 shows the receiver operating characteristic (ROC) curves of the 50 classes for the baseline f-VAEGAN and our model ART+f-VAEGAN. As shown, incorporating ART into f-VAEGAN achieves superior ROC when compared to the original baseline model. To be clear, we plot area-under of the ROC curve (AUC) for all 50 categories in Fig. 12. The F1 score curves with different classification thresholds are depicted in Fig. 11. As demonstrated, ART gives F1-score boost on this classification task for most classes.

4.5. Comparison with State-of-the-arts

In this section, we report our final results and compare with other works under the settings of ZSL and GZSL respectively. To make an extensive evaluation, except for the most common settings, we also add experimental results with attributes embedded by textual descriptions on CUB and the experiments with the fine-tuned backbone proposed in [41] on four datasets.

The GZSL Results. We compare f-VAEGAN combined with our ART framework with recent generative methods in Tab. 4. From the table, one can find that our ART framework improves the performance of f-VAEGAN significantly and achieves the state-of-the-art results on CUB, FLO, SUN and AwA1 four datasets, and slightly falls behind the best competitor CE-GZSL [33] on AWA2. It can be seen that the attributes embedded by textual descriptions perform much better than attributes formulated by experts. With textural descriptions, we outperforms CE-GSZL by about 1.2% in harmonic mean on CUB.

Both f-VAEGAN [10] and TF-VAEGAN [41] achieve performance improvements through adopting the fine-tuned backbone. The phenomenon verifies the existence of the first problem that we proposed in introduction section. As fine-tuning can also adapt the backbone to specific zero-shot learning datasets which has the similar function as our ART framework, so in this case, we change the loss $\mathcal{L}_e$ to the classical triplet loss for the purpose of avoiding introducing more over-fitting problem. By utilizing the fine-tuned backbone directly, we achieve the best performance on several ZSL datasets. The improvement of performance compared to f-VAEGAN can further implies the existence of information asymmetry problem.

Table 5: Score comparisons on ZSL. We calculate the average top-1 accuracy of unseen classes. The prefix "$*$" means that the attributes embedded by textual descriptions are adopted on CUB. The prefix "†" means the fine-tuned backbone is utilized to improve performance. We mark the best results with the bold fonts.

Method	CUB	FLO	SUN	AwA2	AwA1
f-CLSWGAN [30]	57.3	67.2	60.8	-	68.2
Cycle-WGAN [32]	58.6	70.3	59.9	-	66.8
LisGAN [31]	58.8	69.7	61.7	-	70.6
LsrGAN [55]	60.3	-	62.5	-	66.4
f-VAEGAN [10]	61.0	67.7	64.7	71.1	-
TF-VAEGAN [41]	64.9	70.8	**66.0**	**72.2**	-
$*$ CE-GZSL [33]	77.5	70.6	63.3	70.4	71.0
Our ART+f-VAEGAN	65.1	**71.1**	64.8	70.1	**72.4**
$*$ **Our ART+f-VAEGAN**	**77.7**	-	-	-	-
† f-VAEGAN [10]	72.9	70.4	65.6	70.3	-
† TF-VAEGAN [41]	**74.3**	74.7	**66.7**	73.4	-
† **Our ART+f-VAEGAN**	73.1	**75.1**	66.3	**74.5**	-

The ZSL Results. Although the introduced ART framework is primarily designed to improve the results of GZSL, it does also work under the ZSL setting. According to Tab. 5, our ART framework achieves the state-of-the-art results on CUB, FLO and AwA1 datasets. Similar to the results of GZSL, attributes embedded by textual descriptions are also better than the ones formulated by experts on CUB. Besides, after fine-tuning the backbone, our method is still better than the baseline method f-VAEGAN and overall is similar to TF-VAEGAN.

Table 6: Score comparison on ZSL and GZSL for action recognition. We mark the best results with the bold fonts.

Method	HMDB51		UCF101	
	ZSL	GZSL	ZSL	GZSL
GGM [57]	20.7	20.1	20.3	17.5
f-CLSWGAN [30]	29.1	32.7	37.5	44.4
CEWGAN [48]	30.2	36.1	38.3	49.4
f-VAEGAN [10]	31.1	35.6	38.2	47.2
ZSVD [44]	-	23.2	-	49.7
TF-VAEGAN [41]	**33.0**	37.6	**41.0**	50.9
JSSE [42]	-	25.7	-	51.7
ART+ f-VAEGAN	32.2	**38.4**	40.7	**51.8**

5. Zero-Shot Action Recognition

We finally evaluate our ART model for zero-shot action recognition under ZSL and GZSL settings. As suggested by CEWGAN [48], we validate our method with I3D (inflated 3D) features. The appearance and flow I3D features are respectively extracted from the pre-trained RGB and Flow I3D networks. We concatenate appearance and flow features to obtain 8,192-d video features. As done in previous works [48, 41], we utilize an out-of-distribution classifier at the classification stage. With respect to semantic descriptions, we use semantic embeddings of size 300 generated by a skip-gram model using action class names as input for HMDB51, and use the class-embedding in form of manually-annotated class attributes of size 115 for UCF101. In regard to the choices of hyper-parameters, different from configs of static datasets, we use ReLU as the activate function instead of tanh and adjust the learning rate of ART to 1e-5. Besides, in consideration that 8,192-d features are much larger than 2,048-d features of static images, we therefore set transformed features be 2,048 dimensions for both datasets.

We give top-1 comparison results of our method against previous approaches in Tab. 6. On HMDB51, by leveraging the ART ingredient, the baseline f-VAEGAN improves the classification score from 31.1% to 32.2% for ZSL, and from 35.6% to 38.4% for GZSL. On UCF101, ART provides 2.5% improvement for ZSL and 4.6% for GZSL. Importantly, we can see our approach outperforms all the existing methods for generalized zero-shot action recognition.

6. Conclusions

In this manuscript, we have underlined that it is necessary to fine-tune visual features before synthesizing unseen visual data for the zero-shot learning problem. This is motivated by the two reasons. The first one is the distribution of the feature backbone training dataset is different from the distribution of ZSL datasets. Hence, the pre-trained visual features should be adapted to zero-shot learning datasets. The second reason is that there is some gap between real and synthesized features due to the information asymmetry problem, which further inspires us to reduce information in visual features irrelevant to attributes. We have successfully designed a simple ART framework that consists of a contrastive regression module and a normalizing attribute place-holder module to fine-tune original visual features. Experimental results on five benchmarks have shown that, combined with our ART framework, the baseline generative model f-VAEGAN sets a new state-of-the-art record under both ZSL and GZSL settings.

7. Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 61972312. We would like to thank Renzhong Zhang for evaluating our algorithm on the task of action recognition and helpful discussions. We are also grateful for valuable suggestions from reviewers.

References

[1] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).

[2] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 951–958.

[3] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, A. Elgammal, A generative adversarial approach for zero-shot learning from noisy texts, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1004–1013.

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

[5] H. Kim, J. Lee, H. Byun, Discriminative deep attributes for generalized zero-shot learning, Pattern Recognition 124 (2022) 108435.

[6] H. Jiang, R. Wang, S. Shan, Y. Yang, X. Chen, Learning discriminative latent attributes for zero-shot classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4223–4232.

[7] H. Zhang, H. Bai, Y. Long, L. Liu, L. Shao, A plug-in attribute correction module for generalized zero-shot learning, Pattern Recognition 112 (2021) 107767.

[8] 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.

[9] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, L. Shao, Attentive region embedding network for zero-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9384–9393.

[10] Y. Xian, S. Sharma, B. Schiele, Z. Akata, f-vaegan-d2: A feature generating framework for any-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10275–10284.

[11] Z. Ding, M. Shao, Y. Fu, Low-rank embedded ensemble semantic dictionary for zero-shot learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2050–2058.

[12] 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) (2013) 453–465.

[13] J. Lei Ba, K. Swersky, S. Fidler, et al., Predicting deep zero-shot convolutional neural networks using textual descriptions, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4247–4255.

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

[15] R. Socher, M. Ganjoo, C. D. Manning, A. Ng, Zero-shot learning through cross-modal transfer, in: Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013, p. 935–943.

[16] M. Bucher, S. Herbin, F. Jurie, Improving semantic embedding consistency by metric learning for zero-shot classiffication, in: Computer Vision - ECCV 2016, Lecture Notes in Computer Science, 2016, pp. 730–746.

[17] C. Xie, T. Zeng, H. Xiang, K. Li, Y. Yang, Q. Liu, Class knowledge overlay to visual feature learning for zero-shot image classification, Computer Vision and Image Understanding 207 (2021) 103206.

[18] V. K. Verma, P. Rai, A simple exponential family framework for zero-shot learning, in: Proceedings of Joint European conference on machine learning and knowledge discovery in databases, 2017, pp. 792–808.

[19] Z. Ji, X. Yu, Y. Yu, Y. Pang, Z. Zhang, Semantic-guided class-imbalance learning model for zero-shot image classification, IEEE Transactions on Cybernetics 52 (7) (2022) 6543–6554.

[20] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, Q. M. J. Wu, A review of generalized zero-shot learning methods, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 1–20doi:10.1109/TPAMI.2022.3191696.

[21] 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.

[22] Y. Guo, G. Ding, J. Han, S. Tang, Zero-shot learning with attribute selection, in: Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 6870–6877.

[23] H. Jiang, R. Wang, S. Shan, X. Chen, Learning class prototypes via structure alignment for zero-shot recognition, in: Proceedings of the European conference on computer vision, 2018, pp. 118–134.

[24] Y. Wang, H. Zhang, Z. Zhang, Y. Long, L. Shao, Learning discriminative domain-invariant prototypes for generalized zero shot learning, Knowledge-Based Systems 196 (105796) (2020).

[25] 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.

[26] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, J. Dean, Zero-shot learning by convex combination of semantic embeddings, arXiv preprint arXiv:1312.5650 (2013).

[27] Y. Xian, C. H. Lampert, B. Schiele, Z. Akata, Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly, IEEE transactions on pattern analysis and machine intelligence 41 (9) (2018) 2251–2265.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of Advances in neural information processing systems, 2014, pp. 2672–2680.

[29] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved training of wasserstein gans, in: Proceedings of Advances in neural information processing systems, 2017, pp. 5767–5777.

[30] Y. Xian, T. Lorenz, B. Schiele, Z. Akata, Feature generating networks for zero-shot learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5542–5551.

[31] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, Z. Huang, Leveraging the invariant side of generative zero-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7402–7411.

[32] R. Felix, V. B. Kumar, I. Reid, G. Carneiro, Multi-modal cycle-consistent generalized zero-shot learning, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 21–37.

[33] Z. Han, Z. Fu, S. Chen, J. Yang, Contrastive embedding for generalized zero-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2371–2381.

[34] V. Kumar Verma, G. Arora, A. Mishra, P. Rai, Generalized zero-shot learning via synthesized examples, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4281–4289.

[35] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, Z. Akata, Generalized zero-and few-shot learning via aligned variational autoencoders, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8247–8255.

[36] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).

[37] D. Huynh, E. Elhamifar, Compositional zero-shot learning via fine-grained dense feature composition, in: Proceedings of the Advances in Neural Information Processing Systems, 2020, pp. 19849–19860.

[38] H. Dong, Y. Fu, S. J. Hwang, L. Sigal, X. Xue, Learning the compositional domains for generalized zero-shot learning, Computer Vision and Image Understanding (2022) 103454.

[39] Y. Yu, Z. Ji, Y. Pang, J. Guo, Z. Zhang, F. Wu, Bi-adversarial auto-encoder for zero-shot learning, ArXiv abs/1811.08103 (2018).

[40] Z. Ji, B. Cui, Y. Yu, Y. Pang, Z. Zhang, Zero-shot classification with unseen prototype learning, Neural Computing and Applications (2021) 1–11doi:10.1007/s00521-021-05746-9.

[41] S. Narayan, A. Gupta, F. S. Khan, C. G. Snoek, L. Shao, Latent embedding feedback and discriminative features for zero-shot classification, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 479–495.

[42] N. Madapana, J. P. Wachs, JSSE: Joint sequential semantic encoder for zero-shot event recognition, IEEE Transactions on Artificial Intelligence (2022) 1–12doi:10.1109/TAI.2022.3208860.

[43] N. Madapana, J. P. Wachs, Zsgl: zero shot gestural learning, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, p. 331–335.

[44] B. Brattoli, J. Tighe, F. Zhdanov, P. Perona, K. Chalupka, Rethinking zero-shot video classification: End-to-end training for realistic applications, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2020, pp. 4613–4623.

[45] J. Gao, T. Zhang, C. Xu, I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 8303–8311.

[46] S. Chen, D. Huang, Elaborative rehearsal for zero-shot action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13618–13627.

[47] C. Zhang, Y. Peng, Visual data synthesis via gan for zero-shot video classification, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018, pp. 1128–1134.

[48] D. Mandal, S. Narayan, S. K. Dwivedi, V. Gupta, S. Ahmed, F. S. Khan, L. Shao, Out-of-distribution detection for generalized zero-shot action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9977–9985.

[49] K. Soomro, A. R. Zamir, M. Shah, Ucf101: A dataset of 101 human actions classes from videos in the wild, ArXiv abs/1212.0402 (2012).

[50] H. Kuehne, H. Jhuang, E. Garrote, T. A. Poggio, T. Serre, Hmdb: A large video database for human motion recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2011, pp. 2556–2563.

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

[52] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: Proceedings of the Sixth Indian Conference on Computer Vision, Graphics & Image Processing, 2008, pp. 722–729.

[53] G. Patterson, J. Hays, Sun attribute database: Discovering, annotating, and recognizing scene attributes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2751–2758.

[54] S. Reed, Z. Akata, H. Lee, B. Schiele, Learning deep representations of fine-grained visual descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 49–58.

[55] M. R. Vyas, H. Venkateswara, S. Panchanathan, Leveraging seen and unseen semantic relationships for generative zero-shot learning, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 70–86.

[56] Z. Han, Z. Fu, J. Yang, Learning the redundancy-free features for generalized zero-shot object recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12865–12874.

[57] A. Mishra, V. K. Verma, M. S. K. Reddy, A. S., P. Rai, A. Mittal, A generative approach to zero-shot and few-shot action recognition, in: Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 372–380. doi:10.1109/WACV.2018.00047.