LETTER **Communicated by Michiel Stock**

Improving Generalization via Attribute Selection on Out-of-the-Box Data

Xiaofeng Xu
csxuxiaofeng@njust.edu.cn
School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China, and Centre for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW 2007, Australia

Ivor W. Tsang
ivor.tsang@uts.edu.au
Centre for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW 2007, Australia

Chuancai Liu*
chuancailiu@njust.edu.cn
School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China, and Collaborative Innovation Center of IoT Technology and Intelligent Systems, Minjiang University, Fuzhou, Fujian 350000, China

Zero-shot learning (ZSL) aims to recognize unseen objects (test classes) given some other seen objects (training classes) by sharing information of attributes between different objects. Attributes are artificially annotated for objects and treated equally in recent ZSL tasks. However, some inferior attributes with poor predictability or poor discriminability may have negative impacts on the ZSL system performance. This letter first derives a generalization error bound for ZSL tasks. Our theoretical analysis verifies that selecting the subset of key attributes can improve the generalization performance of the original ZSL model, which uses all the attributes. Unfortunately, previous attribute selection methods have been conducted based on the seen data, and their selected attributes have poor generalization capability to the unseen data, which is unavailable in the training stage of ZSL tasks. Inspired by learning from pseudo-relevance feedback, this letter introduces out-of-the-box data—pseudo-data generated by an attribute-guided generative model—to mimic the unseen data. We then present an iterative attribute selection (IAS) strategy that iteratively selects key attributes based on the out-of-the-box data. Since the distribution of the generated out-of-the-box data is similar to that of the

*Corresponding author.

Neural Computation **32, 485–514** (2020)
https://doi.org/10.1162/neco_a_01256

test data, the key attributes selected by IAS can be effectively generalized to test data. Extensive experiments demonstrate that IAS can significantly improve existing attribute-based ZSL methods and achieve state-of-the-art performance.

1 Introduction

With the rapid development of machine learning technologies, especially the rise of deep neural networks, visual object recognition has made tremendous recent progress (Zheng, Li, Yan, Tang, & Tan, 2018; Shen, Ji, Wang, Li, & Li, 2018). These recognition systems even outperform humans when provided with a massive amount of labeled data. However, it is expensive to collect sufficient labeled samples for all natural objects, especially for the new concepts and many subordinate categories (Zhou, Fang et al., 2019). Therefore, how to achieve an acceptable recognition performance for objects with limited or even no training samples is a challenging but practical problem (Palatucci, Pomerleau, Hinton, & Mitchell, 2009). Inspired by a human cognition system that can identify new objects when provided with a description in advance (Murphy, 2004), zero-shot learning (ZSL) has been proposed to recognize unseen objects with no training samples (Cheng, Qiao, Wang, & Yu, 2017; Ji et al., 2019). Since a labeled sample is not given for the target classes, we need to collect some source classes with sufficient labeled samples and find the connection between the target classes and the source classes.

As a kind of semantic representation, attributes are widely used to transfer knowledge from the seen classes (source) to the unseen classes (target) (Ma et al., 2017). Attributes play a key role in sharing information between classes and govern the performance of zero-shot classification. In previous ZSL work, all attributes are assumed to be effective and treated equally. However, as Guo, Ding, Han, and Tang (2018) pointed out, different attributes have different properties, such as distributive entropy and predictability. The attributes with poor predictability or poor discriminability may have negative impacts on the ZSL system performance. Poor predictability means that the attributes are hard to be correctly recognized from the feature space, and poor discriminability means that the attributes are weak in distinguishing different objects. Hence, it is obvious that not all the attributes are necessary and effective for zero-shot classification.

Based on these observations, selecting the key attributes instead of using all attributes is significant and necessary for constructing ZSL models. Guo et al. (2018) proposed the zero-shot learning with attribute selection (ZSLAS) model, which selects attributes by measuring the distributive entropy and predictability of attributes based on the training data. ZSLAS can improve the performance of attribute-based ZSL methods, though it suffers from the drawback of generalization. Since the training classes and test classes are disjoint in ZSL tasks, the training data are bounded by the box

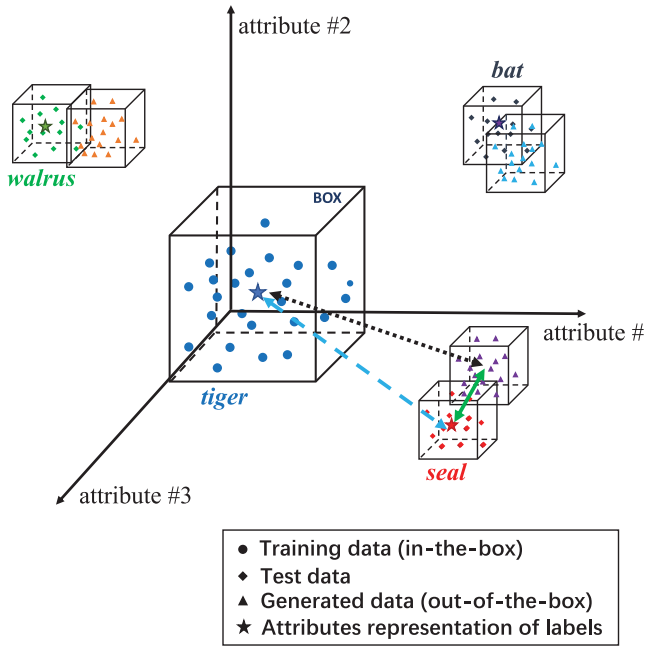


Figure 1: Illustration of out-of-the-box data. The distance between the out-of-the-box data and the test data (green solid arrow) is much less than the distance between the training data and the test data (blue dashed arrow).

cut by attributes (illustrated in Figure 1). Therefore, the attributes selected based on the training data have poor generalization capability to the unseen test data.

To address the drawback, this letter derives a generalization error bound for the ZSL problem. Since attributes for the ZSL task are literally like the code words in the error-correcting output code (ECOC) model (Dietterich & Bakiri, 1994), we analyze the bound from the perspective of ECOC. Our analyses reveal that the key attributes need to be selected based on the data out of the box (i.e., the distribution of the training classes). Considering that test data are unavailable during the training stage for ZSL tasks, inspired by learning from pseudo-relevance feedback (Miao, Huang, & Zhao, 2016), we introduce out-of-the-box data to mimic the unseen test classes.[1] These data are generated by an attribute-guided generative model using the same attribute representation as the test classes. Therefore, the out-of-the-box data have similar distributions to the test data.

[1] Out-of-the-box data are generated based on the training data and the attribute representation without extra information, which follows the standard zero-shot learning setting.

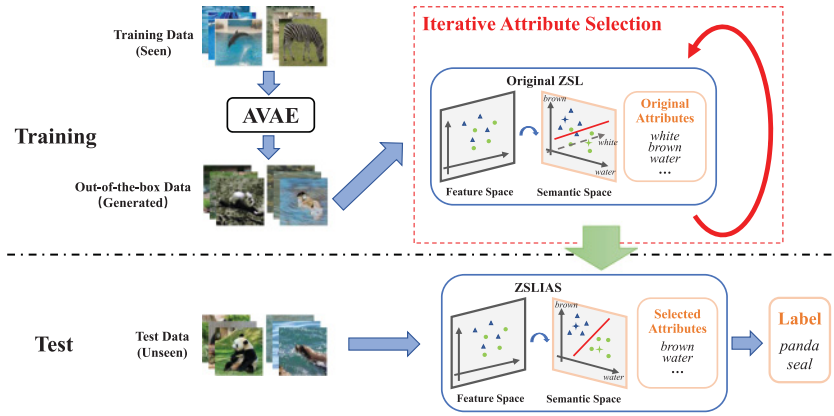


Figure 2: The pipeline of the ZSLIAS framework. In the training stage, we first generate out-of-the-box data by a tailor-made generative model (AVAE) and then iteratively select attributes based on out-of-the-box data. In the test stage, the selected attributes are exploited to build the ZSL model for categorizing unseen objects.

Guided by the performance of the ZSL model on out-of-the-box data, we propose a novel iterative attribute selection (IAS) model to select the key attributes in an iterative manner. Figure 2 illustrates the procedures of the proposed ZSL with iterative attribute selection (ZSLIAS). Unlike the previous ZSLAS that uses training data to select attributes at once, our IAS first generates out-of-the-box data to mimic the unseen classes, and subsequently iteratively selects key attributes based on the generated out-of-the-box data. During the test stage, selected attributes are employed as a more efficient semantic representation to improve the original ZSL model. By adopting the proposed IAS, the improved attribute embedding space is more discriminative for the test data and, hence, improves the performance of the original ZSL model.

The main contributions of this letter are summarized as follows:

- We present a generalization error analysis for the ZSL problem. Our theoretical analyses prove that selecting the subset of key attributes can improve the generalization performance of the original ZSL model, which uses all the attributes.
- Based on our theoretical findings, we propose a novel iterative attribute selection strategy to select key attributes for ZSL tasks.
- Since test data are unseen during the training stage for ZSL tasks, we introduce out-of-the-box data to mimic test data for attribute selection. Such data generated by a designed generative model have a similar distribution to the test data. Therefore, attributes selected

based on out-of-the-box data can be effectively generalized to the unseen test data.
- Extensive experiments demonstrate that IAS can effectively improve the attribute-based ZSL model and achieve state-of-the-art performance.

The rest of the letter is organized as follows. Section 2 reviews related work. Section 3 gives preliminary information and motivation. Section 4 presents the theoretical analyses on generalization bound for attribute selection. Section 5 proposes the iterative attribute selection model. Experimental results are reported in section 6, and conclusions are drawn in section 7.

2 Related Work

In this section, we review some related work on zero-shot learning, attribute selection, and deep generative models.

2.1 Zero-Shot Learning. ZSL can recognize new objects using attributes like the intermediate semantic representation. Some researchers adopt the probability-prediction strategy to transfer information. Lampert, Nickisch, and Harmeling (2013) proposed a popular baseline: direct attribute prediction (DAP). DAP learns probabilistic attribute classifiers using the seen data and infers the label of the unseen data by combining the results of pretrained classifiers.

Most recent work adopts a label-embedding strategy that learns a mapping function directly from the input features space to the semantic embedding space. One line of work is to learn linear compatibility functions. For example, Akata, Perronnin, Harchaoui, and Schmid (2015) presented an attribute label embedding (ALE) model, which learns a compatibility function combined with ranking loss. Romera-Paredes and Torr (2015) proposed an approach that models the relationships among features, attributes, and classes as a two-linear-layers network. Another direction is to learn nonlinear compatibility functions. Xian et al. (2016) presented a nonlinear embedding model that augments bilinear compatibility model by incorporating latent variables. Airola and Pahikkala (2017) proposed a first general Kronecker product kernel-based learning model for ZSL tasks. In addition to the classification task, Ji, Sun, Yu, Pang, and Han (2019) proposed an attribute network for a zero-shot hashing retrieval task.

2.2 Attribute Selection. Attributes, as a popular semantic representation of visual objects, can be the appearance, a part, or a property of objects (Farhadi, Endres, Hoiem, & Forsyth, 2009). For example, the object *elephant* has the attribute *big* and *long nose*, and the object *zebra* has the attribute *striped*. Attributes are widely used to transfer information to

recognize new objects in ZSL tasks (Sun, Schiele, & Fritz, 2017; Xu, Tsang, & Liu, 2019). As shown in Figure 1, using attributes as the semantic representation, data of different categories locate in different boxes bounded by the attributes. Since the attribute representations of the seen classes and the unseen classes are different, the boxes with respect to the seen data and the unseen data are disjoint.

In previous ZSL work, all the attributes are assumed to be effective and treated equally. However, as Guo et al. (2018) pointed out, not all the attributes are effective for recognizing new objects. Therefore, we should select the key attributes to improve the semantic presentation. Liu, Wiliem, Chen, and Lovell (2014) proposed a novel greedy algorithm that selects attributes based on their discriminating power and reliability. Guo et al. (2018) proposed selecting attributes by measuring the distributive entropy and predictability of attributes based on the training data. In short, previous attribute selection models have been conducted based on the training data, which makes the selected attributes have poor generalization capability to the unseen test data. Our IAS iteratively selects attributes based on out-of-the-box data, which has a similar distribution to the test data, and thus the key attributes selected by our model can be more effectively generalized to the unseen test data.

2.3 Attribute-Guided Generative Models. Deep generative models (Ma, Chang, Xu, Sebe, & Hauptmann, 2017) aim to estimate the joint distribution $p(y; x)$ of samples and labels by learning the class prior probability $p(y)$ and the class-conditional density $p(x|y)$ separately. The generative model can be extended to a conditional generative model if the generator is conditioned on some extra information, such as attributes in the proposed method. Odena, Olah, and Shlens (2017) introduced conditional generative adversarial nets (CGAN), which can be constructed by simply feeding the data label. CGAN is conditioned on both the generator and discriminator and can generate samples conditioned on class labels. Conditional variational autoencoder (CVAE) (Sohn, Lee, & Yan, 2015), as an extension of a variational autoencoder, is a deep conditional generative model for structured output prediction using gaussian latent variables. We modify CVAE with the attribute representation to generate out-of-the-box data for the attribute selection.

3 Preliminary Information and Motivation

3.1 ZSL Task Formulation. We consider zero-shot learning as a task that recognizes unseen classes that have no labeled samples available. Given a training set $D_s = \{(x_n, y_n), n = 1, \ldots, N_s\}$, the task of traditional ZSL is to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ from the image feature space to the label embedding space by minimizing the following regularized empirical risk,

Table 1: Notations and Descriptions.

Notation	Description	Notation	Description
D_s	Training data (seen)	N_s	Number of training samples
D_u	Test data (unseen)	N_u	Number of test samples
D_g	Out-of-the-box data	N_g	Number of generated samples
$\mathcal{X}$	Image features	d	Number of dimension of features
$\mathcal{Y}_s$	Training classes (seen)	K	Number of training classes
$\mathcal{Y}_u$	Test classes (unseen)	L	Number of test classes
$\mathbf{A}$	Attribute matrix	$\mathbf{a}_y$	Attribute vector of label y
N_a	Number of all the attributes	$\mathcal{A}$	Set of original attributes
$\mathbf{s}$	Selection vector	$\mathcal{S}$	Subset of selected attributes

$$L\left(y, f\left(x; \mathbf{W}\right)\right) = \frac{1}{N_s} \sum_{n=1}^{N_s} l\left(y_n, f\left(x_n; \mathbf{W}\right)\right) + \Omega\left(\mathbf{W}\right), \tag{3.1}$$

where $l\left(\cdot\right)$ is the loss function, which can be square loss $1/2(f(x) - y)^2$, logistic loss $\log(1 + \exp(-yf(x)))$, or hinge loss $\max(0, 1 - yf(x))$. $\mathbf{W}$ is the parameter of mapping f, and $\Omega\left(\cdot\right)$ is the regularization term.

The mapping function f is defined as follows,

$$f\left(x; \mathbf{W}\right) = \arg\max_{y \in \mathcal{Y}} F\left(x, y; \mathbf{W}\right), \tag{3.2}$$

where the function $F : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$ is the bilinear compatibility function to associate image features and label embeddings defined as follows,

$$F\left(x, y; \mathbf{W}\right) = \theta\left(x\right)^T \mathbf{W} \varphi\left(y\right), \tag{3.3}$$

where $\theta\left(x\right)$ is the image features and $\varphi\left(y\right)$ is the label embedding (i.e., attribute representation).

We summarize some frequently used notations in Table 1.

3.2 Interpretation of the ZSL Task. In traditional ZSL models, all attributes are assumed to be effective and treated equally, though, some researchers have pointed out that not all the attributes are useful and significant for zero-shot classification (Jiang, Wang, Shan, Yang, & Chen, 2017). To the best of our knowledge, there is no theoretical analysis for the generalization performance of ZSL tasks, let alone selecting informative attributes for unseen classes. To fill this gap, we first derive the generalization error bound for ZSL models.

The intuition of our theoretical analysis is to simply treat the attributes as error-correcting output codes; then the prediction of ZSL tasks can be

deemed as the assignment of class labels with respective predefined ECOC, which is the closest to the predicted ECOC problem (Rocha & Goldenstein, 2014). Based on this novel interpretation, we derive a theoretical generalization error bound of the ZSL model as shown in section 4. From the generalization bound analyses, we find that the discriminating power of attributes governs the performance of the ZSL model.

3.3 Deficiency of ZSLAS. Some attribute selection work has been proposed in recent years. Guo et al. (2018) proposed a ZSLAS model that selects attributes based on distributive entropy and the predictability of attributes using training data. Simultaneously considering the ZSL model loss function and attribute properties in a joint optimization framework, they selected attributes by minimizing the following loss function:

$$L(y, f(x; \mathbf{s}, \mathbf{W})) = \frac{1}{N_s} \sum_{n=1}^{N_s} \{l_{\text{ZSL}}(y_n, f(x_n; \mathbf{s}, \mathbf{W})) + \alpha l_p(\theta(x_n), \varphi(y_n); \mathbf{s}) - \beta l_v(\theta(x_n), \mu; \mathbf{s})\}, \tag{3.4}$$

where $\mathbf{s}$ is the weight vector of the attributes, which will be used for attribute selection. $\theta(\cdot)$ is the attribute classifier, $\varphi(y_n)$ is the attribute representation, and μ is an auxiliary parameter. l_{ZSL} is the model-based loss function for ZSL, that is, $l(\cdot)$ as defined in equation 3.1. l_p is the attribute prediction loss, which can be defined based on specific ZSL models, and l_v is the loss of variance, which measures the distributive entropy of attributes (Guo et al., 2018). After getting the weight vector $\mathbf{s}$ by optimizing equation 3.4, attributes can be selected according to $\mathbf{s}$ and then used to construct the ZSL model.

From our theoretical analyses in section 4, ZSLAS can improve the original ZSL model to some extent (Guo et al., 2018). However, ZSLAS suffers the drawback that the attributes are selected based on the training data. Since the training and test classes are disjoint in ZSL tasks, it is difficult to measure the quality and contribution of attributes regarding discriminating the unseen test classes. Thus, the selected attributes by ZSLAS have poor generalization capability to the test data due to the domain shift problem.

3.4 Definition of Out-of-the-Box. Since previous attribute selection models have been conducted based on bounded in-the-box data, the selected attributes have poor generalization capability to the test data. However, the test data are unavailable during the training stage. Inspired by learning from pseudo-relevance feedback (Miao et al., 2016), we introduce pseudo-data, which are outside the box of the training data, to mimic test classes to guide attribute selection. Considering that the training data are bounded in the box by attributes, we generate the out-of-the-box data using an attribute-guided generative model. Since the out-of-the-box data are

generated based on the same attribute representation as test classes, the box of the generated data will overlap with the box of the test data. Consequently, the key attributes selected by the proposed IAS model based on the out-of-the-box data can be effectively generalized to the unseen test data.

4 Generalization Bound Analysis

In this section, we first derive the generalization error bound of the original ZSL model and then analyze the bound changes after attribute selection. In previous work, some generalization error bounds have been presented for the ZSL task. Romera-Paredes and Torr (2015) transformed the ZSL problem to the domain adaptation problem and then analyzed the risk bounds for domain adaptation. Stock, Pahikkala, Airola, De Baets, and Waegeman (2018) considered the ZSL problem as a specific setting of pairwise learning and analyzed the bound by the kernel ridge regression model. However, these bound analysis are not suitable for the ZSL model due to their assumptions. In this work, we derive the generalization bound from the perspective of ECOC model, which is more similar to the ZSL problem.

4.1 Generalization Error Bound of ZSL. Zero-shot classification is an effective way to recognize new objects that have no training samples available. The basic framework of the ZSL model is using attribute representation as the bridge to transfer knowledge from seen objects to unseen objects. To simplify the analysis, we consider ZSL as a multiclass classification problem. Therefore, the ZSL task can be addressed using an ensemble method that combines many binary attribute classifiers. Specifically, we pretrained a binary classifier for each attribute separately in the training stage. To classify a new sample, all the attribute classifiers are evaluated to obtain an attribute code word (a vector in which each element represents the output of an attribute classifier). Then we compare the predicted code word to the attribute representations of all the test classes to retrieve the label of the test sample.

To analyze the generalization error bound of ZSL, we first define some distances in the attribute space and then present a proposition of the error-correcting ability of attributes.

Definition 1. Generalized attribute distance: *Given the attribute matrix* $\boldsymbol{A}$ *for associating labels and attributes, let* $\boldsymbol{a}_i$, $\boldsymbol{a}_j$ *denote the attribute representation of label* y_i *and* y_j *in matrix* $\boldsymbol{A}$ *with length* N_a*, respectively. Then the generalized attribute distance between* $\boldsymbol{a}_i$ *and* $\boldsymbol{a}_j$ *can be defined as*

$$d(\boldsymbol{a}_i, \boldsymbol{a}_j) = \sum_{m=1}^{N_a} \Delta(\boldsymbol{a}_i^{(m)}, \boldsymbol{a}_j^{(m)}), \tag{4.1}$$

where N_a is the number of attributes, and $\boldsymbol{a}_i^{(m)}$ is the m^{th} element in the attribute representation $\boldsymbol{a}_i$ of the label y_i. $\Delta(\boldsymbol{a}_i^{(m)}, \boldsymbol{a}_j^{(m)})$ is equal to 1 if $\boldsymbol{a}_i^{(m)} \neq \boldsymbol{a}_j^{(m)}$, and 0 otherwise.

We further define the minimum distance between any two attribute representations in the attribute space.

Definition 2. Minimum attribute distance: *The minimum attribute distance τ of matrix $\boldsymbol{A}$ is the minimum distance between any two attribute representations $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$:*

$$\tau = \min_{i \neq j} d(\boldsymbol{a}_i, \boldsymbol{a}_j), \qquad \forall\, 1 \leq i,\, j \leq N_a. \tag{4.2}$$

Given the definition of distance in the attribute space, we can prove the following proposition:

Proposition 1. (Zhou, Tsang, Ho, & Muller, 2019): *Error-correcting ability: Given is the label-attribute correlation matrix $\boldsymbol{A}$ and a vector of predicted attribute representation $f(x)$ for an unseen test sample x with known label y. If x is incorrectly classified, the distance between the predicted attribute representation $f(x)$ and the correct attribute representation $\boldsymbol{a}_y$ is greater than half of the minimum attribute distance τ:*

$$d(f(x), \boldsymbol{a}_y) \geq \frac{\tau}{2}. \tag{4.3}$$

Proof. Suppose that the predicted attribute representation for test sample x with correct attribute representation $\mathbf{a}_y$ is $f(x)$, and the sample x is incorrectly classified to the mismatched attribute representation $\mathbf{a}_r$, where $r \in \mathcal{Y}_u \setminus \{y\}$. Then the distance between $f(x)$ and $\mathbf{a}_y$ is greater than the distance between $f(x)$ and $\mathbf{a}_r$:

$$d(f(x), \mathbf{a}_y) \geq d(f(x), \mathbf{a}_r). \tag{4.4}$$

Here, the distance between attribute representation can be expanded as the element-wise summation based on equation 4.1 as follows:

$$\sum_{m=1}^{N_a} \Delta(f^{(m)}(x), \mathbf{a}_y^{(m)}) \geq \sum_{m=1}^{N_a} \Delta(f^{(m)}(x), \mathbf{a}_r^{(m)}). \tag{4.5}$$

Then we have:

$$d(f(x), \mathbf{a}_y) = \sum_{m=1}^{N_a} \Delta(f^{(m)}(x), \mathbf{a}_y^{(m)})$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{m=1}^{N_a}\left\{\Delta(f^{(m)}(x), \mathbf{a}_y^{(m)}) + \Delta(f^{(m)}(x), \mathbf{a}_y^{(m)})\right\} \\
&\overset{(i)}{\geq} \frac{1}{2}\sum_{m=1}^{N_a}\left\{\Delta(f^{(m)}(x), \mathbf{a}_y^{(m)}) + \Delta(f^{(m)}(x), \mathbf{a}_r^{(m)})\right\} \\
&\overset{(ii)}{\geq} \frac{1}{2}\sum_{m=1}^{N_a}\Delta(\mathbf{a}_y^{(m)}, \mathbf{a}_r^{(m)}) \\
&= \frac{1}{2}d(\mathbf{a}_y, \mathbf{a}_r) \overset{(iii)}{\geq} \frac{\tau}{2},
\end{aligned} \tag{4.6}
$$

where (i) follows equation 4.5, (ii) is based on the triangle inequality of distance metric (Zhou, Tsang et al., 2019), and (iii) follows equation 4.2. □

From proposition 1, we can find that the predicted attribute representation is not required to be exactly the same as the ground truth for each unseen test sample. As long as the distance is less than $\tau/2$, ZSL models can correct the error committed by some attribute classifiers and make an accurate prediction.

Based on the proposition of the error-correcting ability of attributes, we can derive the theorem of generalization error bound for ZSL:

Theorem 1. Generalization error bound of ZSL: *Given N_a attribute classifiers, $f^{(1)}$, $f^{(2)}, \ldots, f^{(N_a)}$, trained on training set D_s with label-attribute matrix $\boldsymbol{A}$, the generalization error rate for the attribute-based ZSL model is upper-bounded by*

$$
\frac{2N_a\bar{B}}{\tau}, \tag{4.7}
$$

where $\bar{B} = \frac{1}{N_a}\sum_{m=1}^{N_a} B_m$ and B_m is the upper bound of the prediction loss for the m^{th} attribute classifier $f^{(m)}$.

Proof. According to proposition 1, for any incorrectly classified test sample x with label y, the distance between the predicted attribute representation $f(x)$ and the true attribute representation $\mathbf{a}_y$ is greater than $\tau/2$:

$$
d(f(x), \mathbf{a}_y) = \sum_{m=1}^{N_a}\Delta(f^{(m)}(x), \mathbf{a}_y^{(m)}) \geq \frac{\tau}{2}. \tag{4.8}
$$

Let k be the number of incorrect image classifications for unseen test data set $D_u = \{(x_i, y_i), i = 1, \ldots, N_u\}$. We can obtain

$$k\frac{\tau}{2} \leq \sum_{i=1}^{N_u}\sum_{m=1}^{N_a} \Delta(f^{(m)}(x_i), \mathbf{a}_{y_i}^{(m)}) \leq \sum_{i=1}^{N_u}\sum_{m=1}^{N_a} B_m = N_u N_a \bar{B}, \tag{4.9}$$

where $\bar{B} = \frac{1}{N_a}\sum_{m=1}^{N_a} B_m$ and B_m is the upper bound of attribute prediction loss.

Hence, the generalized error rate k/N_u is bounded by $2N_a\bar{B}/\tau$. □

Remark 1. Generalization error bound is positively correlated to the average attribute prediction loss: From theorem 1, we can find that the generalization error bound of the attribute-based ZSL model depends on the number of attributes N_a, minimum attribute distance τ, and average prediction loss $\bar{B}$ for all the attribute classifiers. According to definitions 1 and 2, the minimum attribute distance τ is positively correlated to the number of attributes N_a. Therefore, the generalization error bound is mainly affected by the average prediction loss $\bar{B}$. Intuitively, the inferior attributes with poor predictability cause greater prediction loss $\bar{B}$; consequently, these attributes will have a negative effect on the ZSL performance and increase the generalization error rate.

4.2 Improvement of Generalization after Attribute Selection. We proved in the previous section that the generalization error bound of the ZSL model is affected by the average prediction loss $\bar{B}$. In this section, we prove that attribute selection can reduce the average prediction loss $\bar{B}$ and, consequently, reduce the generalization error bound of ZSL from the perspective of PAC-style (Valiant, 1984) analysis.

Lemma 1. (Palatucci, Pomerleau, Hinton, & Mitchell, 2009). PAC bound of ZSL: *Given N_a attribute classifiers, to obtain an attribute classifier with $(1-\delta)$ probability that has at most k_a incorrect predicted attributes, the PAC bound D of the attribute-based ZSL model is*

$$D \propto \frac{N_a}{k_a}[4\,log(2/\delta) + 8(d+1)\,log(13N_a/k_a)], \tag{4.10}$$

where d is the dimension of the image features.

Remark 2. The average attribute prediction loss is positively correlated to the PAC bound. Here, k_a/N_a is the tolerable prediction error rate of attribute classifiers. According to the definition of the average attribute prediction loss $\bar{B}$, it is obvious that the ZSL model with smaller $\bar{B}$ could tolerate a

greater k_a/N_a. From lemma 1, we can find that the PAC bound D is monotonically increasing with respect to N_a/k_a. Hence, the PAC bound D decreases when the N_a/k_a decreases, and consequently the average prediction loss $\bar{B}$ decreases.

Lemma 2. (Vapnik, 2013). Test error bound: *Suppose that the PAC bound of the attribute-based ZSL model is D. The probability of the test error distancing from an upper bound is given by*

$$p\left(e_{ts} \leq e_{tr} + \sqrt{\frac{1}{N_s}\left[D\left(log\left(\frac{2N_s}{D}\right)+1\right) - log\left(\frac{\eta}{4}\right)\right]}\right) = 1-\eta, \tag{4.11}$$

where N_s is the size of the training set, $0 \leq \eta \leq 1$, and e_{ts}, e_{tr} are the test error and the training error, respectively.

Remark 3. PAC bound is positively correlated to the test error bound. From lemma 2, we can find that the PAC bound can affect the probabilistic upper bound on the test error. Specifically, to obtain a high probability with a small test error, the PAC bound should be small. In other words, the model with a smaller PAC bound would have a smaller test error bound.

Proposition 2. Bound change after attribute selection: *For the attribute-based ZSL model, attribute selection can decrease the generalization error bound.*

Proof. In attribute selection, the key attributes are selected by minimizing the loss function in equation 3.1 on out-of-the-box data. Since the generated out-of-the-box data have a similar distribution to the test data, the test error of ZSL will decrease after attribute selection; that is, ZSLIAS has a smaller test error bound than the original ZSL model. Therefore, we can infer that ZSLIAS has a smaller PAC bound based on remark 3. According to remark 2, we can infer that the average prediction error $\bar{B}$ decreases after attribute selection. As a consequence, the generalization error bound of ZSLIAS is smaller than the original ZSL model based on remark 1. □

From proposition 2, we can observe that the generalization error of the ZSL model will decrease after adopting the proposed IAS. In other words, a ZSLIAS has a smaller classification error rate comparing to the original ZSL method when generalizing to the unseen test data.

5 IAS with Out-of-the-Box Data

Motivated by the generalization bound analyses, we select the key attributes based on the out-of-the-box data. In this section, we first present the proposed iterative attribute selection model. Then we introduce the attribute-guided generative model designed to generate the out-of-the-box data, followed by the complexity analysis of IAS.

5.1 Iterative Attribute Selection Model. Inspired by the idea of iterative machine teaching (Liu et al., 2017), we propose a novel selection model that iteratively selects attributes based on the generated out-of-the-box data. First, we generate the out-of-the-box data to mimic test classes by an attribute-based generative model. Then the key attributes are selected in an iterative manner based on these data. After obtaining the selected attributes, we can consider them as a more efficient semantic representation to improve the original ZSL model.

Suppose, given the generated out-of-the-box data $D_g = \{(x_n, y_n), n = 1, \ldots, N_g\}$, we can combine the empirical risk in equation 3.1 with the attribute selection model. Then the loss function is rewritten as

$$L\left(y, f\left(x; \mathbf{s}, \mathbf{W}\right)\right) = \frac{1}{N_g} \sum_{n=1}^{N_g} l\left(y_n, f\left(x_n; \mathbf{s}, \mathbf{W}\right)\right) + \Omega\left(\mathbf{W}\right), \tag{5.1}$$

where $\mathbf{s} \in (0, 1)^{N_a}$ is the indicator vector for the attribute selection, in which $s_i = 1$ if the ith attribute is selected or 0 otherwise. N_a is the number of all the attributes.

Correspondingly, the mapping function f in equation 3.2 and the compatibility function F in equation 3.3 can be rewritten as

$$f\left(x; \mathbf{s}, \mathbf{W}\right) = \arg\max_{y \in \mathcal{Y}} F\left(x, y; \mathbf{s}, \mathbf{W}\right), \tag{5.2}$$

$$F\left(x, y; \mathbf{s}, \mathbf{W}\right) = \theta\left(x\right)^T \mathbf{W}\left(\mathbf{s} \circ \varphi\left(y\right)\right), \tag{5.3}$$

where $\circ$ is the element-wise product operator (Hadamard product), and $\mathbf{s}$ is the selection vector defined in equation 5.1.

To solve the optimization problem in equation 5.1, we need to specify the choice of the loss function $l\left(\cdot\right)$. The loss function in equation 5.1 for a single sample (x_n, y_n) is expressed as follows (Xian, Lampert, Schiele, & Akata, 2018):

$$\begin{aligned} & l(y_n, f((x_n; \mathbf{s}, \mathbf{W}))) \\ & = \sum_{y \in \mathcal{Y}_g} r_{ny}[\triangle(y_n, y) + F(x_n, y; \mathbf{s}, \mathbf{W}) - F(x_n, y_n; \mathbf{s}, \mathbf{W})]_+ \\ & = \sum_{y \in \mathcal{Y}_g} r_{ny}[\triangle(y_n, y) + \theta(x_n)^T \mathbf{W}(\mathbf{s} \circ \varphi(y)) - \theta(x_n)^T \mathbf{W}(\mathbf{s} \circ \varphi(y_n))]_+, \end{aligned} \tag{5.4}$$

where $\mathcal{Y}_g$ is the label of generated out-of-the-box data, which is the same as $\mathcal{Y}_u$.

$\triangle(y_n; y) = 0$ if $y_n = y$; 1 otherwise. $r_{ny} \in [0, 1]$ is the weight defined in specific ZSL methods.

Since the dimension of the optimal attribute subset (i.e., l_0-norm of $\mathbf{s}$) is agnostic, finding the optimal $\mathbf{s}$ is an NP-complete (Garey, Johnson, & Stockmeyer, 1974) problem. Therefore, inspired by the idea of iterative machine teaching (Liu et al., 2017), we adopt the greedy algorithm (Cormen, Leiserson, Rivest, & Stein, 2009) to optimize the loss function in an iterative manner. Equation 5.1 gets updated during each iteration as follows:

$$\begin{aligned} L^{t+1} = \frac{1}{N_g}\sum_{n=1}^{N_g} & \; l^{t+1}(y_n, f(x_n; \mathbf{s}^{t+1}, \mathbf{W}^{t+1})) + \Omega(\mathbf{W}^{t+1}), \\ & s.t. \sum_{s_i \in \mathbf{s}^{t+1}} s_i = t+1, \\ & \sum_{s_j \in (\mathbf{s}^{t+1} - \mathbf{s}^t)} s_j = 1. \end{aligned} \tag{5.5}$$

The constraints on $\mathbf{s}$ ensure that $\mathbf{s}^t$ updates one element (from 0 updates to 1) during each iteration, which indicates that only one attribute is selected each time. $\mathbf{s}^0$ is the initial vector of all 0's.

Correspondingly, the loss function in equation 5.5 for a single sample (x_n, y_n) gets updated during each iteration as follows:

$$\begin{aligned} l^{t+1} = \sum_{y \in \mathcal{Y}_g} r_{ny}[\triangle(y_n, y) + \theta(x_n)^T \mathbf{W}^{t+1}(\mathbf{s}^{t+1} \circ \varphi(y)) \\ - \theta(x_n)^T \mathbf{W}^{t+1}(\mathbf{s}^{t+1} \circ \varphi(y_n))]_+. \end{aligned} \tag{5.6}$$

Here l^{t+1} is subjected to the same constraints as equation 5.5.

To minimize the loss function in equation 5.5, we can alternatively optimize $\mathbf{W}^{t+1}$ and $\mathbf{s}^{t+1}$ by optimizing one variable while fixing the other one. In each iteration, we first optimize $\mathbf{W}^{t+1}$ via the gradient descent algorithm (Burges et al., 2005). The gradient of equation 5.5 is calculated as follows:

$$\frac{\partial L^{t+1}}{\partial \mathbf{W}^{t+1}} = \frac{1}{N_g}\sum_{n=1}^{N_g} \frac{\partial l^{t+1}}{\partial \mathbf{W}^{t+1}} + \frac{1}{2}\alpha \mathbf{W}^{t+1}, \tag{5.7}$$

where

$$\frac{\partial l^{t+1}}{\partial \mathbf{W}^{t+1}} = \sum_{y \in \mathcal{Y}_g} r_{ny} \theta(x_n)^T (\mathbf{s}^t \circ (\varphi(y) - \varphi(y_n))), \tag{5.8}$$

where α is the regularization parameter.

After updating $\mathbf{W}^{t+1}$, we can traverse all the elements equal to 0 in $\mathbf{s}^t$ and turn them into 1, respectively. Then $\mathbf{s}^{t+1}$ is updated by the optimal $\mathbf{s}^{t+1}$, which achieves the minimal loss of equation 5.5:

$$\mathbf{s}^{t+1} = \arg\min_{\mathbf{s}^{t+1}} \frac{1}{N_g} \sum_{n=1}^{N_g} l^{t+1}(y_n, f(x_n; \mathbf{s}^{t+1}, \mathbf{W}^{t+1})) + \Omega(\mathbf{W}^{t+1}). \tag{5.9}$$

When iterations end and $\mathbf{s}$ is obtained, we can easily get the subset of key attributes by selecting the attributes corresponding to the elements equal to 1 in the selection vector $\mathbf{s}$.

The procedure of the proposed IAS model is given in algorithm 1.

5.2 Generation of Out-of-the-Box Data. In order to select the discriminative attributes for test classes, we should do attribute selection on the test data. Since the training data and the test data are located in the different boxes bounded by the attributes, we adopt an attribute-based generative model (Bucher, Herbin, & Jurie, 2017) to generate out-of-the-box data to mimic test classes. Compared to the ZSLAS, the key attributes selected by IAS based on the out-of-the-box data can be more efficiently generalized to test data.

Conditional variational autoencoder (CVAE) (Sohn et al., 2015) is a conditional generative model in which both the latent codes and generated data are conditioned on some extra information. In this work, we propose the attribute-based variational autoencoder (AVAE), a special version of CVAE with tailor-made attributes, to generate the out-of-the-box data.

VAE (Kingma & Welling, 2013) is a directed graphical model with certain types of latent variables. The generative process of VAE is as follows. A set of latent codes z is generated from the prior distribution $p(z)$, and the data x are generated by the generative distribution $p(x|z)$ conditioned on $z : z \sim p(z)$, $x \sim p(x|z)$. The empirical objective of VAE is expressed as (Sohn et al., 2015)

$$L_{\text{VAE}}(x) = -\text{KL}(q(z|x) \parallel p(z)) + \frac{1}{L} \sum_{l=1}^{L} \log p(x|z^{(l)}), \tag{5.10}$$

where $z^{(l)} = g(x, \epsilon^{(l)})$ and $\epsilon^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. $q(z|x)$ is the recognition distribution, which is reparameterized with a deterministic and differentiable function $g(\cdot, \cdot)$ (Sohn et al., 2015). KL denotes the Kullback-Leibler divergence (Kullback, 1987) between the incorporated distributions. L is the number of samples.

Combined with the condition (i.e., the attribute representation of labels), the empirical objective of the AVAE is defined as

Algorithm 1: Iterative Attribute Selection Model.

Input:

The generated out-of-the-box data D_g;

Original attribute set $\mathcal{A}$;

Iteration stop threshold ε.

Output:

Subset of selected attributes $\mathcal{S}$.

Initialization: $\mathbf{s}^0 = \mathbf{0}$, randomize $\mathbf{W}^0$;

for $t = 0$ to $N_a - 1$ **do**

$L^t = \frac{1}{N_g} \sum_{n=1}^{N_g} l^t(y_n, f(x_n; \mathbf{s}^t, \mathbf{W}^t)) + \Omega(\mathbf{W}^t)$ (equation 5.5)

$\frac{\partial L^t}{\partial \mathbf{W}^t} = \frac{1}{N_g} \sum_{n=1}^{N_g} \frac{\partial l^t}{\partial \mathbf{W}^t} + \frac{1}{2}\alpha \mathbf{W}^t$ (equation 5.7)

// Update $\mathbf{W}$

$\mathbf{W}^{t+1} = \mathbf{W}^t - \eta_t \frac{\partial L^t}{\partial \mathbf{W}^t}$

// Update $\mathbf{s}$

$\mathbf{s}^{t+1} = \arg\min_{\mathbf{s}^{t+1}} L^{t+1}$ (equation 5.9)

if $|L^{t+1} - L^t| \leq \varepsilon$

Break;

end if

end for

Obtain the subset of selected attributes: $\mathcal{S} = \mathbf{s} \circ \mathcal{A}$.

$$L_{\text{AVAE}}(x, \varphi(y)) = -\text{KL}(q(z|x, \varphi(y)) \parallel p(z|\varphi(y))) + \frac{1}{L} \sum_{l=1}^{L} \log p(x|\varphi(y), z^{(l)}), \tag{5.11}$$

where $z^{(l)} = g(x, \varphi(y), \epsilon^{(l)})$, $\varphi(y)$ is the attribute representation of label y.

In the encoding stage, for each training data point $x^{(i)}$, we estimate the $q(z^{(i)}|x^{(i)}, \varphi(y^{(i)})) = Q(z)$ using the encoder. In the decoding stage, after

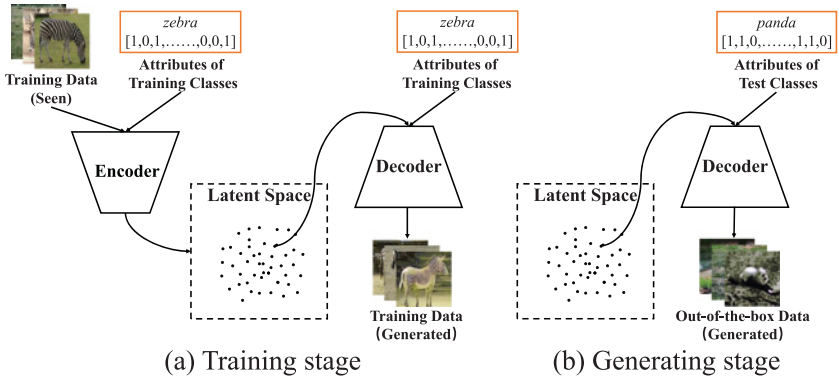


Figure 3: The framework of AVAE. (a) Training stage. (b) Generating stage.

inputting the concatenation of the $\tilde{z}$ sampled from the $Q(z)$ and the attribute representation $\varphi(y_u)$, the decoder will generate a new sample x_g with the same attribute representation as the unseen class $\varphi(y_u)$.

The procedure of AVAE is illustrated in Figure 3. At training time, the attribute representation (of training classes) whose image is being fed in is provided to the encoder and decoder. To generate an image of a particular attribute representation (of test classes), we can just feed this attribute vector along with a random point in the latent space sampled from a standard normal distribution. The system no longer relies on the latent space to encode what object you are dealing with. Instead, the latent space encodes attribute information. Since the attribute representations of test classes are fed into the decoder at the generating stage, the generated out-of-the-box data D_g have a similar distribution to the test data.

5.3 Complexity Analysis. Suppose that there are N_u unseen samples belonging to L test classes and the number of all the attributes is N_a. The complexity of the original ZSL model is $\mathcal{O}_{\text{ZSL}} \sim \mathcal{O}(N_u N_a L^2)$. For the proposed ZSLIAS, the complexity the of training stage is $\mathcal{O}_{\text{ZSLIAS}} \sim N_a(N_a + 1)/2 \cdot \mathcal{O}_{\text{ZSL}}$, that is, $\mathcal{O}(N_u N_a^3 L^2)$, and the complexity of test stage is equal to $\mathcal{O}_{\text{ZSL}}$, that is, $\mathcal{O}(N_u N_a L^2)$.

6 Experiments

To evaluate the performance of the proposed iterative attribute selection model, extensive experiments are conducted on four standard data sets with a ZSL setting. In this section, we first compare the proposed approach with the state-of-the-art and then give detailed analyses.

Table 2: Statistic Information of Four Data Sets with Two Data Set Splits.

Data Set	Number of Attributes	Number of Classes			Number of Images (SS)		Number of Images (PS)	
		Total	Training	Test	Training	Test	Training	Test
AwA	85	50	40	10	24,295	6180	19,832	5685
aPY	64	32	20	12	12,695	2644	5932	7924
CUB	312	200	150	50	8855	2933	7057	2967
SUN	102	717	645	72	12,900	1440	10,320	1440

6.1 Experimental Settings

6.1.1 Data Set. We conduct experiments on four standard ZSL data sets: (1) Animal with Attribute (AwA; Lampert et al., 2013), (2) attribute-Pascal-Yahoo (aPY; Farhadi et al., 2009), (3) Caltech-UCSD Bird 200-2011 (CUB; Wah, Branson, Welinder, Perona, & Belongie, 2011), and (4) SUN Attribute Database (SUN; Patterson, & Hays, 2012). Information on these data sets is summarized in Table 2.

6.1.2 Data Set Split. Zero-shot learning assumes that training classes and test classes are disjoint. Actually, ImageNet, the data set exploited to extract image features via deep neural networks, may include some test classes. Therefore, Xian et al. (2018) proposed a new data set split (PS) ensuring that none of the test classes appear in the data set used to train the extractor model. In this letter, we evaluate the proposed model using both splits: the original standard split (SS) and the proposed split (PS).

6.1.3 Image Feature. Deep neural network features are extracted for the experiments. Image features are extracted from the entire images for the AwA, CUB, and SUN data sets and from the bounding boxes mentioned in Farhadi et al. (2009) for the aPY data set, respectively. The original ResNet-101 (He, Zhang, Ren, & Sun, 2016), pretrained on ImageNet with 1000 classes, is used to calculate 2048-dimensional top-layer pooling units as image features.

6.1.4 Attribute Representation. Attributes are used as the semantic representation to transfer information from training classes to test classes. We use 85-, 64-, 312-, and 102-dimensional continuous-value attributes for the AwA, aPY, CUB, and SUN data sets, respectively.

6.1.5 Evaluation Protocol. The unified data set splits shown in Table 2 are used for all the compared methods to get fair comparison results. Since the data set is not well balanced with respect to the number of images per class

(Xian et al., 2018), we use the mean class accuracy (i.e., per class averaged top-1 accuracy) as the criterion of assessment. Mean class accuracy is calculated as follows,

$$acc = \frac{1}{L} \sum_{y \in \mathcal{Y}_u} \frac{\#\text{correct predictions in } y}{\#\text{samples in } y}, \tag{6.1}$$

where L is the number of test classes and $\mathcal{Y}_u$ is the set comprising of all the test labels.

6.2 Comparison with the State of the Art. To evaluate the efficiency of the proposed iterative attribute selection model, we modify several latest ZSL baselines by the proposed IAS and compare them with the state of the art.

We modify seven representative ZSL baselines to evaluate the IAS model, including three popular ZSL baselines—DAP (Lampert et al., 2013), LatEm (Xian et al., 2016), and SAE (Kodirov, Xiang, & Gong, 2017)—and four latest ZSL baselines—MFMR (Xu et al., 2017), GANZrl (Tong et al., 2018), fVG (Xian et al., 2019), and LLAE (Li et al., 2019).

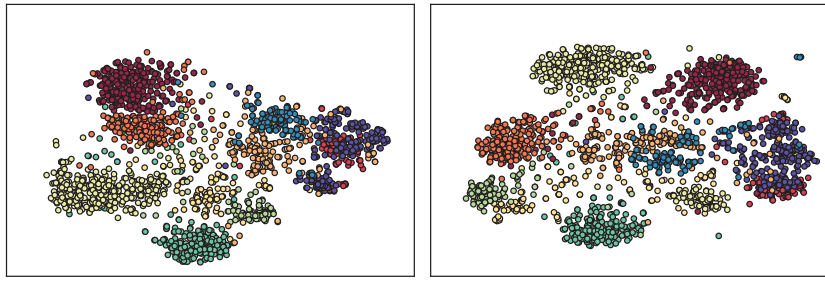
The improvement achieved on these ZSL baselines is summarized in Table 3. It can be observed that IAS can significantly improve the performance of attribute-based ZSL methods. Specifically, the mean accuracies of these ZSL methods on four data sets (AwA, aPY, CUB, and SUN) are increased by 11.09%, 15.97%, 9.10%, and 5.11%, respectively (10.29% on average), after using IAS. For DAP on the AwA and aPY data sets and LatEm on the AwA data set, IAS can improve their accuracy by greater than 20%, which demonstrates that IAS can significantly improve the performance of ZSL models. Interestingly, SAE performs poorly on the aPY and CUB data sets, while the accuracy rises to an acceptable level (from 8.33% to 38.53% and from 24.65% to 42.85%, respectively) by using IAS. Although the performance of the state-of-the-art baselines is quite good, IAS can still improve them to some extent (5.48%, 3.24%, 2.80%, and 3.64% on average for MFMR, GANZrl, fVG, and LLAE, respectively). These results demonstrate that the proposed iterative attribute selection model makes sense and can effectively improve existing attribute-based ZSL methods. This also proves the necessity and effectiveness of attribute selection for ZSL tasks.

Similar to our work, ZSLAS selects attributes based on the distributive entropy and the predictability of attributes. Thus, we compare the improvement of IAS and ZSLAS on DAP and LatEm, respectively. In Table 3, it can be observed that ZSLAS can improve existing ZSL methods, and IAS can improve them even more (2.15% versus 10.61% on average). Compared to ZSLAS, the advantages of ZSLIAS can be interpreted in two ways. First, ZSLIAS selects attributes in an iterative manner; hence, it can select a more optimal subset of key attributes than ZSLAS, which selects attributes at once.

Table 3: Zero-Shot Classification Accuracy Comparison on Benchmarks.

	AwA		aPY		CUB		SUN	
Methods	SS	PS	SS	PS	SS	PS	SS	PS
DAP[b]	64.44	46.22	35.73	39.67	43.47	40.23	41.25	45.83
DAP+AS[a]	—	48.29	—	34.87	—	41.55	—	42.27
DAP+IAS	86.65(+*22.21*)	71.88(+*25.66*)	57.12(+*21.39*)	43.06(+*3.39*)	55.35(+*11.88*)	54.22(+*13.99*)	47.85(+*6.60*)	50.56(+*4.73*)
LatEm[b]	71.51	48.33	24.43	34.66	50.38	48.57	58.75	55.13
LatEm+AS[a]	—	59.07	—	38.82	—	52.82	—	58.09
LatEm+IAS	81.83(+*10.32*)	67.13(+*18.80*)	47.22(+*22.79*)	48.36(+*13.70*)	56.05(+*5.67*)	52.14(+*3.57*)	59.03(+*0.28*)	56.18(+*1.05*)
SAE[b]	79.19	48.48	8.33	8.33	26.41	24.65	36.94	32.78
SAE+IAS	87.95(+*8.76*)	70.36(+*21.88*)	45.90(+*37.57*)	38.53(+*30.20*)	48.21(+*21.80*)	42.85(+*18.20*)	45.14(+*8.20*)	42.22(+*9.44*)
MFMR[b]	86.06	68.04	52.16	34.09	43.09	39.55	50.49	53.33
MFMR+IAS	87.10(+*1.04*)	71.37(+*3.33*)	58.51(+*6.35*)	37.67(+*3.58*)	51.40(+*8.31*)	47.89(+*8.34*)	58.47(+*7.98*)	58.26(+*4.93*)
GANZrl[a]	86.23	—	—	—	62.56	—	—	—
GANZrl+IAS	88.51(+*2.28*)	—	—	—	66.76(+*4.20*)	—	—	—
fVG[a]	—	70.30	—	—	—	72.90	—	65.60
fVG+IAS	—	74.28(+*3.98*)	—	—	—	74.53(+*1.63*)	—	68.39(+*2.79*)
LLAE[a]	85.24	—	56.16	—	61.93	—	—	—
LLAE+IAS	88.95(+*3.71*)	—	60.88(+*4.72*)	—	64.42(+*2.49*)	—	—	—

Notes: Italicized numbers in brackets are relative performance gains. — indicates that no reported results are available. [a]Results published in the paper. [b]Results reproduced.



(a) Real data of AwA (b) Out-of-the-box data

Figure 4: T-SNE visualization of the generated out-of-the-box data and real test data of AwA.

Second, ZSLAS is conducted based on the training data, while ZSLIAS is conducted based on the out-of-the-box data, which have similar distributions to the test data. Therefore, the attributes selected by ZSLIAS are more applicable and discriminative for test data. Experimental results demonstrate the significant superiority of the proposed IAS model over previous attribute selection models.

6.3 Detailed Analysis. In order to further understand the promising performance, we analyze the following experimental results in detail.

6.3.1 Evaluation on the Out-of-the-Box Data. In the first experiment, we evaluate the out-of-the-box data generated by a tailor-made, attribute-based, deep generative model. Figure 4 shows the distribution of the out-of-the-box data and the real test data sampled from the AwA data set using t-SNE. Note that the out-of-the-box data in Figure 4b are generated based on the attribute representation of unseen classes and without extra information on any test images. It can be observed that the generated out-of-the-box data can capture a similar distribution to the real test data, which guarantees that the selected attributes can be effectively generalized to test data.

We also quantitatively evaluate the out-of-the-box data by calculating various distances between three distributions: the generated out-of-the-box data ($\mathcal{X}_g$), unseen test data ($\mathcal{X}_u$), and seen training data ($\mathcal{X}_s$), in pairs. Table 4 shows the distribution distances measured by Wasserstein distance (Vallender, 1974), KL divergence (Kullback, 1987), Hellinger distance (Beran, 1977), and Bhattacharyya distance (Kailath, 1967), respectively. It is obvious that the distance between $\mathcal{X}_g$ and $\mathcal{X}_u$ is much less than the distance between $\mathcal{X}_u$ and $\mathcal{X}_s$, which means that the generated out-of-the-box data have a similar distribution to the unseen test data compared to the seen data. Therefore, attributes selected based on the out-of-the-box data are more discriminative for test data compared to attributes selected based on training data.

Table 4: Distances between Different Data Distributions.

Metrics	$\mathcal{X}_g \sim \mathcal{X}_u$	$\mathcal{X}_g \sim \mathcal{X}_s$	$\mathcal{X}_s \sim \mathcal{X}_u$
Wasserstein distance	**5.99**	19.09	18.97
KL divergence	**0.321**	0.630	0.703
Hellinger distance	**7.78**	16.87	17.15
Bhattacharyya distance	**0.0808**	0.159	0.176

Notes: $\mathcal{X}_g$ indicates the generated out-of-the-box data, $\mathcal{X}_u$ indicates the unseen test data and $\mathcal{X}_s$ indicates the seen training data. Numbers in bold indicate the minimum distance.

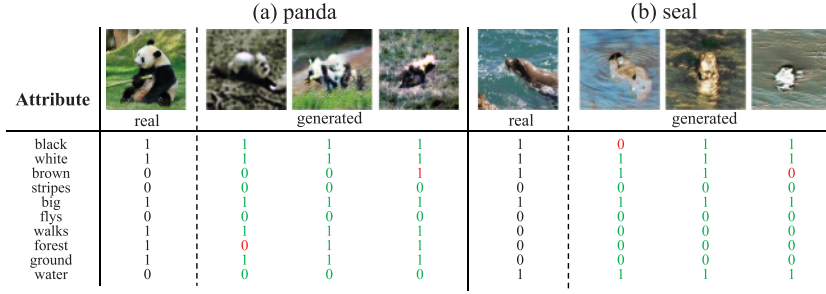


Figure 5: Visualization of generated out-of-the-box images and their attribute representation. The first column of panels a and b is the real image derived from the AwA data set. The remaining three columns of both parts are randomly selected from the generated data. Numbers in black are the ground-truth attributes of the real image. Numbers in green and red are the correct and the incorrect attribute values of the generated images, respectively.

We illustrate some generated images of unseen classes (*panda* and *seal*) and annotate the corresponding attribute representations shown in Figure 5. Numbers in black indicate the attribute representations of the labels of real test images. Numbers in red and green are the correct and the incorrect attribute values of generated images, respectively. We can see that the generated images have a similar attribute representation as test images. Therefore, the tailor-made, attribute-based, deep generative model can generate out-of-the-box data that capture a similar distribution as the unseen data.

6.3.2 Effectiveness of IAS. In the second experiment, we compare the performance of three ZSL methods (DAP, LatEm, and SAE) after using IAS on four data sets, respectively. The accuracies with respect to the number of selected attributes are shown in Figure 6. On the AwA, aPY, and SUN data sets, we can see that the performance of these three ZSL methods increases sharply when the number of selected attributes grows from 0 to about 20% and then reaches a peak. These results suggest that only about a quarter

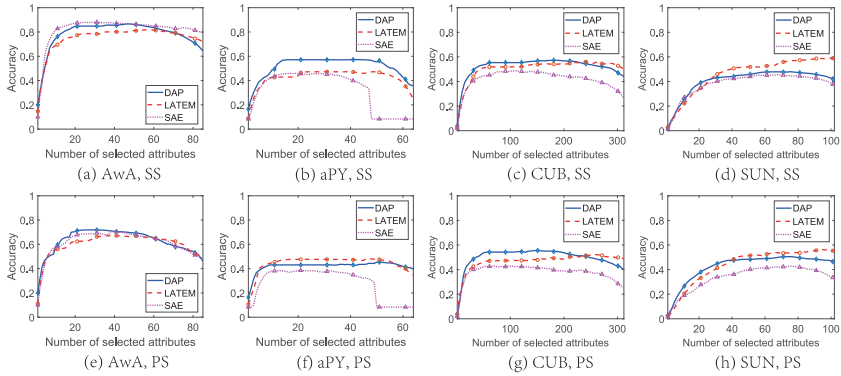


Figure 6: Performance of IAS for DAP, LatEm, and SAE. The performance of baselines without IAS is shown on the right-most side of the curves.

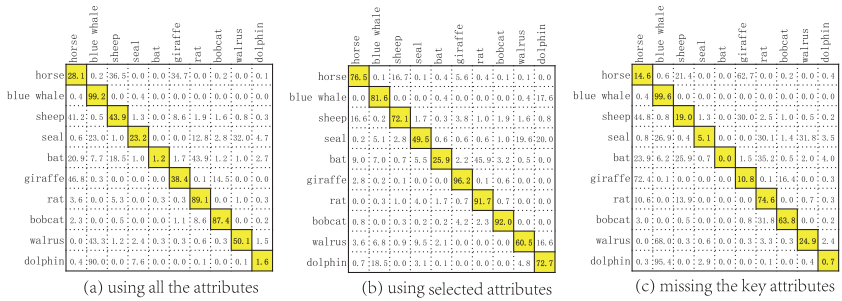


Figure 7: Confusion matrices (in %) between 10 test classes on the AwA data set with the proposed split. (a) DAP using all the original attributes. (b) DAP using the key attributes selected by IAS. (c) DAP using the remaining attributes after selection.

of attributes are the key ones necessary and effective for classifying test objects. In Figures 6b and 6f, there is an interesting result: SAE performs poorly on the aPY data set with both SS and PS (the accuracy is less than 10%), while the performance is acceptable after using IAS (the accuracy is about 40%). These results demonstrate the effectiveness and robustness of IAS for ZSL tasks.

Furthermore, we modify DAP by using all 84 attributes, the 20 selected attributes, and the remaining 64 attributes after attribute selection, respectively. The resulting confusion matrices of these three variants evaluated on the AwA data set with the proposed split setting are illustrated in Figure 7. The numbers in the diagonal area (yellow patches) of confusion matrices indicate the classification accuracy per class. It is obvious that IAS can

significantly improve DAP performance on most of the test classes, and the accuracies on some classes nearly doubled after using IAS, such as *horse*, *seal*, and *giraffe*. Although some objects are hard to be recognized by DAP, like *dolphin* (the accuracy of DAP is 1.6%), we can get acceptable performance after using IAS (the accuracy of DAPIAS is 72.7%). The original DAP performs better than IAS only with regard to the object *blue whale*; this is because in the original DAP, most of the marine creatures (such as *blue whale*, *walrus*, and *dolphin*) are classified as the blue whale, which increases the classification accuracy and also the false-positive rate. More important, the confusion matrix of DAPIAS contains less noise (i.e., smaller numbers in the side regions—the white patches—of confusion matrices apart from the diagonal area) than DAP, which suggests that DAPIAS has fewer prediction uncertainties. In other words, adopting IAS can improve the robustness of attribute-based ZSL methods.

In Figure 7, the accuracy of using the selected attributes (71.88% on average) is significantly improved compared to the accuracy of using all the attributes (46.23% on average), and the accuracy of using the remaining attributes (31.32% on average) is terrible. These results suggest that the selected attributes are key for discriminating test data. The missing attributes are useless and even have a negative impact on the ZSL system. Therefore, not all the attributes are effective for ZSL tasks; clearly, we should select the key attributes to improve performance.

6.3.3 Interpretability of Selected Attributes. In the third experiment, we present the visualization results of attribute selection. We find that ZSL methods obtain the best performance when selecting about 20% attributes, as shown in Figure 6. Therefore, we illustrate the top 20% key attributes selected by DAP, LatEm, and SAE on four data sets in Figure 8. The three rows in each panel are DAP, LatEm, and SAE from top to bottom; the yellow bars indicate the attributes selected by the corresponding methods. We can see that the attribute subsets selected by different ZSL methods are highly coincident for the same data set, which demonstrates that the selected attributes are key for discriminating test data. Specifically, we enumerate the key attributes selected by three ZSL methods on the AwA data set in Table 5. The attributes in bold were simultaneously selected by all three ZSL methods, and those in italics indicate that they were selected by any two of these three methods. Thirteen attributes (65%) were selected by all three ZSL methods. These three attribute subsets selected by diverse ZSL models are very similar, additional evidence that IAS is reasonable and useful for zero-shot classification.

7 Conclusion

We present a novel and effective iterative attribute selection model to improve existing attribute-based ZSL methods. In most previous work on ZSL,

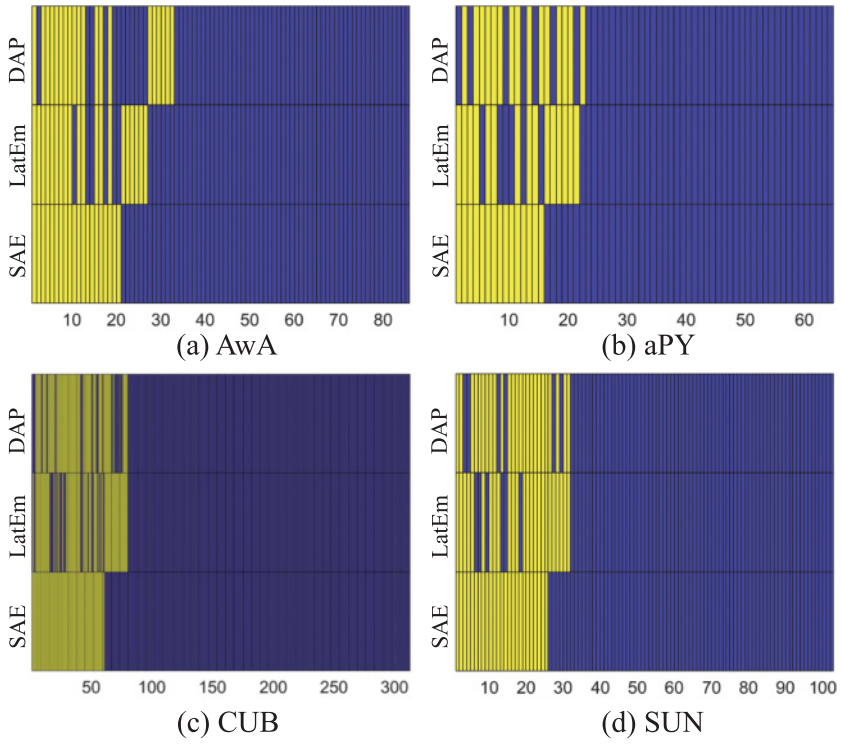


Figure 8: Visualization of the attribute subsets selected by three different ZSL methods on four data sets. Three rows in each figure are DAP, LatEm, and SAE from top to bottom. The horizontal axis represents the attribute, and the yellow bars indicate the attributes selected by the corresponding methods.

all the attributes are assumed to be effective and treated equally. However, we noticed that the attributes have different predictability and discriminability for diverse objects. Motivated by this observation, we propose to select the key attributes to build a ZSL model. Since training classes and test classes are disjoint in ZSL tasks, we introduce out-of-the-box data to mimic test data to guide the progress of attribute selection. These data, generated by a tailor-made, attribute-based, deep generative model, have a similar distribution to the test data. Hence, the attributes selected by IAS based on the out-of-the-box data can be effectively generalized to the test data. To evaluate the effectiveness of IAS, we conduct extensive experiments on four standard ZSL data sets. Experimental results demonstrate that IAS can effectively select the key attributes for ZSL tasks and significantly improve state-of-the-art ZSL methods.

Table 5: Subsets of the Key Attributes Selected by DAP, LatEm and SAE on the AwA Data Set.

DAP		LatEm		SAE	
ground	**fish**	**hands**	**pads**	**black**	**paws**
hands	**fields**	**ground**	**forest**	**ground**	ocean
plains	smelly	bipedal	**gray**	**pads**	**yellow**
tunnels	**pads**	claws	coastal	**gray**	group
forest	**yellow**	**black**	**yellow**	**hands**	*tunnels*
tail	**scavenger**	**fish**	strainteeth	**hooves**	*white*
gray	swims	**fields**	horns	domestic	**fish**
hibernate	**black**	**paws**	**scavenger**	**tail**	**fields**
hooves	**paws**	blue	**tail**	skimmer	**forest**
jungle	weak	**hooves**	*white*	arctic	**scavenger**

Notes: We selected 20 attributes out of 85. The attributes that appear in all three methods are in bold, and those that appear in two methods are in italics.

In this work, we select the same attributes for all the unseen test classes. Obviously, this is not the global optimal solution to select attributes for diverse categories. In the future, we will consider a tailor-made attribute selection model that can identify the special subset of key attributes for each test class.

Acknowledgments

This work is supported in part by ARC under grants LP150100671 and DP180100106; in part by NSFC under grants 61373063 and 61872188; in part by the Project of MIIT under grant E0310/1112/02-1; in part by the Collaborative Innovation Center of IoT Technology and Intelligent Systems of Minjiang University under grant IIC1701; and in part by the China Scholarship Council.

References

Airola, A., & Pahikkala, T. (2017). Fast Kronecker product kernel methods via generalized VEC trick. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(8), 3374–3387.

Akata, Z., Perronnin, F., Harchaoui, Z., & Schmid, C. (2015). Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(7), 1425–1438.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Annals of Statistics*, *5*(3), 445–463.

Bucher, M., Herbin, S., & Jurie, F. (2017). Generating visual representations for zero-shot classification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2666–2673). Piscataway, NJ: IEEE.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 89–96). New York: ACM.

Cheng, Y., Qiao, X., Wang, X., & Yu, Q. (2017). Random forest classifier for zero-shot learning based on relative attribute. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(5), 1662–1674.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms*. Cambridge, MA: MIT Press.

Dietterich, T. G., & Bakiri, G. (1994). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, *2*, 263–286.

Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. (2009). Describing objects by their attributes. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1778–1785). Piscataway, NJ: IEEE.

Garey, M. R., Johnson, D. S., & Stockmeyer, L. (1974). Some simplified NP-complete problems. In *Proceedings of the Sixth Annual ACM Symposium on Theory of Computing* (pp. 47–63). New York: ACM.

Guo, Y., Ding, G., Han, J., & Tang, S. (2018). Zero-shot learning with attribute selection. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. Palo Alto, CA: AAAI.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). Piscataway, NJ: IEEE.

Ji, Z., Sun, Y., Yu, Y., Pang, Y., & Han, J. (2019). Attribute-guided network for cross-modal zero-shot hashing. *IEEE Transactions on Neural Networks and Learning Systems*, forthcoming.

Jiang, H., Wang, R., Shan, S., Yang, Y., & Chen, X. (2017). Learning discriminative latent attributes for zero-shot classification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4223–4232). Piscataway, NJ: IEEE.

Kailath, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, *15*(1), 52–60.

Kingma, D. P., & Welling, M. (2013). *Auto-encoding variational Bayes.* arXiv:1312.6114.

Kodirov, E., Xiang, T., & Gong, S. (2017). Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3174–3183). Piscataway, NJ: IEEE.

Kullback, S. (1987). Letter to the editor: The Kullback-Leibler distance. *American Statistician*, *41*, 340–341.

Lampert, C. H., Nickisch, H., & Harmeling, S. (2013). Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(3), 453–465.

Li, J., Jing, M., Lu, K., Zhu, L., Yang, Y., & Huang, Z. (2019). From zero-shot learning to cold-start recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Palo Alto, CA: AAAI.

Liu, L., Wiliem, A., Chen, S., & Lovell, B. C. (2014). Automatic image attribute selection for zero-shot learning of object categories. In *Proceedings of the 2014 22nd*

International Conference on Pattern Recognition (pp. 2619–2624). Piscataway, NJ: IEEE.

Liu, W., Dai, B., Humayun, A., Tay, C., Yu, C., Smith, L. B., . . . Song, L. (2017). Iterative machine teaching. In *Proceedings of the 34th International Conference on Machine Learning, vol. 70* (pp. 2149–2158).

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., & Van Gool, L. (2017). Pose guided person image generation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems, 30* (pp. 406–416). Red Hook, NY: Curran.

Ma, Z., Chang, X., Xu, Z., Sebe, N., & Hauptmann, A. G. (2017). Joint attributes and event analysis for multimedia event detection. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(7), 2921–2930.

Miao, J., Huang, J. X., & Zhao, J. (2016). TopPRF: A probabilistic framework for integrating topic space into pseudo relevance feedback. *ACM Transactions on Information Systems*, *34*(4), 22.

Murphy, G. (2004). *The big book of concepts*. Cambridge, MA: MIT Press.

Odena, A., Olah, C., & Shlens, J. (2017). Conditional image synthesis with auxiliary classifier GANS. In *Proceedings of the 34th International Conference on Machine Learning, vol. 70* (pp. 2642–2651).

Palatucci, M., Pomerleau, D., Hinton, G. E., & Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems, 22* (pp. 1410–1418). Red Hook, NY: Curran.

Patterson, G., & Hays, J. (2012). Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2751–2758). Piscataway, NJ: IEEE.

Rocha, A., & Goldenstein, S. K. (2014). Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(2), 289–302.

Romera-Paredes, B., & Torr, P. (2015). An embarrassingly simple approach to zero-shot learning. In *Proceedings of the International Conference on Machine Learning* (pp. 2152–2161).

Shen, Y., Ji, R., Wang, C., Li, X., & Li, X. (2018). Weakly supervised object detection via object-specific pixel gradient. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(12), 5960–5970.

Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems*, *28* (pp. 3483–3491). Red Hook, NY: Curran.

Stock, M., Pahikkala, T., Airola, A., De Baets, B., & Waegeman, W. (2018). A comparative study of pairwise learning methods based on kernel ridge regression. *Neural Computation*, *30*(8), 2245–2283.

Sun, Q., Schiele, B., & Fritz, M. (2017). A domain based approach to social relation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3481–3490). Piscataway, NJ: IEEE.

Tong, B., Klinkigt, M., Chen, J., Cui, X., Kong, Q., Murakami, T., & Kobayashi, Y. (2018). Adversarial zero-shot learning with semantic augmentation. In

Proceedings of the 32nd AAAI Conference on Artificial Intelligence (pp. 3483–3491). Palo Alto, CA: AAAI.

Valiant, L. G. (1984). A theory of the learnable. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing* (pp. 436–445). New York: ACM.

Vallender, S. (1974). Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability and Its Applications*, *18*(4), 784–786.

Vapnik, V. (2013). *The nature of statistical learning theory*. New York: Springer Science & Business Media.

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The Caltech-UCSD BIRDS-200-2011 dataset.

Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., & Schiele, B. (2016). Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 69–77). Piscataway, NJ: IEEE.

Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2018). Zero-shot learning: A comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xian, Y., Sharma, S., Schiele, B., & Akata, Z. (2019). f-VAEGAN-D2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 10275–10284). Piscataway, NJ: IEEE.

Xu, X., Shen, F., Yang, Y., Zhang, D., Shen, H., & Song, J. (2017). Matrix trifactorization with manifold regularizations for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3798–3807). Piscataway, NJ: IEEE.

Xu, X., Tsang, I. W., & Liu, C. (2019). Complementary attributes: A new clue to zero-shot learning. *IEEE Transactions on Cybernetics.*

Zheng, Y., Li, S., Yan, R., Tang, H., & Tan, K. C. (2018). Sparse temporal encoding of visual features for robust object recognition by spiking neurons. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(12), 5823–5833.

Zhou, J. T., Fang, M., Zhang, H., Gong, C., Peng, X., Cao, Z., & Goh, R. S. M. (2019). Learning with annotation of various degrees. *IEEE Transactions on Neural Networks and Learning Systems*, *30*, 2794–2804.

Zhou, J. T., Tsang, I. W., Ho, S., & Muller, K. (2019). N-ary decomposition for multi-class classification. *Machine Learning*, *108*(5), 809–830.

Received July 26, 2019; accepted October 2, 2019.