# Model-Agnostic Metric for Zero-Shot Learning

Jiayi Shen[1], Haochen Wang[1], Anran Zhang[1], Qiang Qiu[2], Xiantong Zhen[3], Xianbin Cao[1,4,5]*
[1]School of Electronic and Information Engineering, Beihang University, Beijing, China
[2]Duke University, Durham, NC, USA
[3]Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
[4]Key Laboratory of Advanced Technology of Near Space Information System (Beihang University), Ministry of Industry and Information Technology of China, Beijing, China
[5]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beijing, China
shenjiayi@buaa.edu.cn, haochen.hobot@gmail.com, zhanganran@buaa.edu.cn,
qiang.qiu@duke.edu, zhenxt@gmail.com, xbcao@buaa.edu.cn

## Abstract

*Zero-shot Learning (ZSL) aims to learn a classifier to recognize unseen categories without training samples. Most ZSL works based on embedding models handle the visual space and the semantic space through a common metric space and then apply a simple nearest neighbor search which directly leads to the hubness problem, one of the main challenges of ZSL. Contrary to recent works, whose conclusions about hubs are drawn based on Euclidean and specific models like ridge regression, we adopt cosine metric and for the first time prove cosine is model-agnostic to alleviate the hubness problem in ZSL. Assuming that the normalized mapped semantic vectors follow a uniform distribution, we provide theoretical analysis which demonstrates that hubs can be better reduced with a higher-dimensional cosine metric space. Moreover, we introduce a diversity-based regularizer with the cosine metric which underpins the assumption about the uniform distribution and further improves the model's discriminative ability. Extensive experiments on five benchmarks and large-scale Imagenet dataset show that our method can improve the performance, surpassing previous embedding methods by large margins.*

## 1. Introduction

Zero-shot learning (ZSL) has recently gained great popularity in both machine learning and computer vision communities. The aim of ZSL is to recognize the test samples of unseen classes with the help of the shared knowledge explored from the training samples of seen classes. To learn the transferable recognition ability from seen to unseen classes, ZSL introduces the semantic space to bridge the visual space and the label space [17].

*Corresponding author

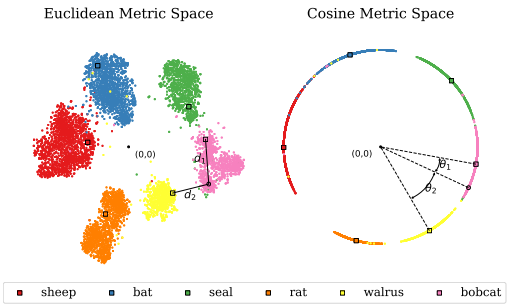



Figure 1. Six unseen classes of AWA2 are visualized by t-SNE [26] by different metric spaces. Dots and squares denote the mapped visual features and attributes respectively. The results intuitively shows that Cosine metric space demonstrates superior performance in bridging the gap between visual features and attributes over Euclidean metric space. The best viewed in color.

Recent works differ in utilizing the semantic space and can be roughly divided into two categories: (1) embedding models [10, 15, 20, 36, 39, 40, 53–56] map visual features extracted from images in the visual space and semantic vectors defined by experts in the semantic space to a common metric space and then apply a non-parameterized classifier, a simple nearest neighbor (NN) search, to recognize new instances of unseen classes during the test stage. (2) generative models [9, 16, 28, 38, 45, 50] most generates fake samples with semantic vectors from unseen classes and then train a parameterized classier with these fake samples to recognize the real unseen test samples. However, they violate the ZSL assumption that the unseen class is prohibitively seen at training [6]. Meanwhile, [31, 46] also point out that due to using unseen classes attributes in model learning most generative models follow the class-transductive setting while embedding models belong to the class-inductive setting.

In this paper, we mainly focus on the embedding models

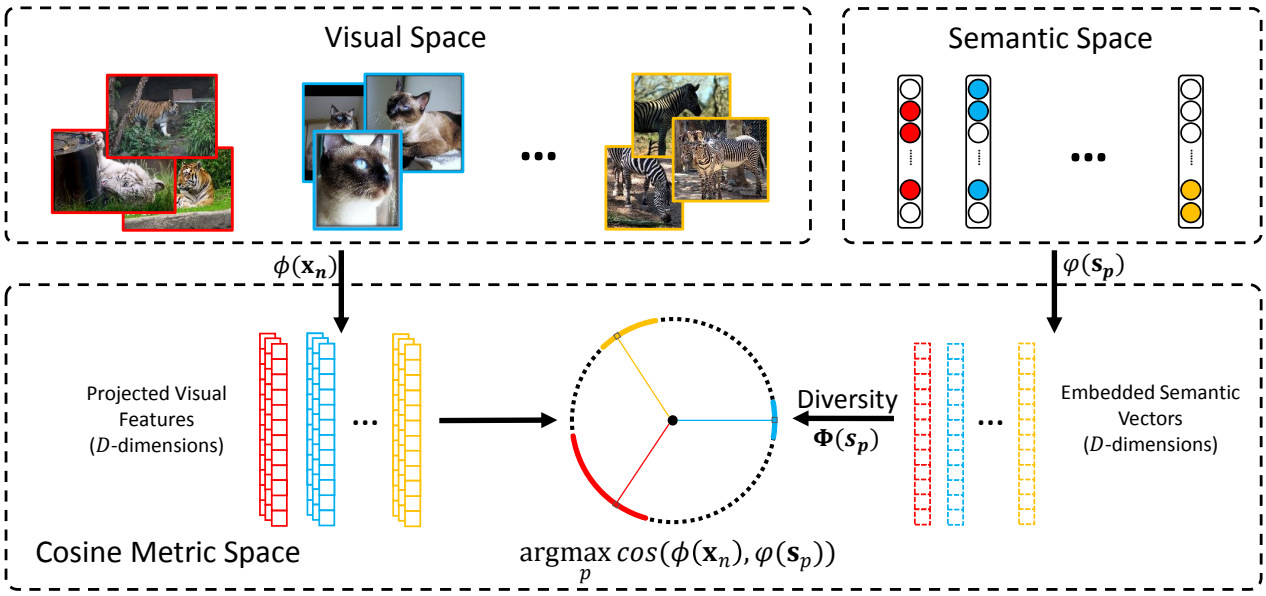



Figure 2. An illustration of our method, including three components: (1) $\varphi(\mathbf{s}_p)$ learns deep embedding for each semantic vector of the semantic space and $\phi(\mathbf{x}_n)$ is a projection for each visual feature of the visual space. (2) a diversity-based regularizer $\mathbf{\Phi}$ which encourages the pairwise embedded semantic vectors different from each other as certain as possible. (3) a prediction classifier with cosine metric instantiated by the nearest neighbor search. The best viewed in color.

under the class-inductive setting and alleviate the hubness problem, one of the main challenges of ZSL. The hubness problem [19, 32, 33, 39] means that some semantic vectors ("hubs") appear in the top neighbor lists of many test samples. Hubs are always harmful to the prediction accuracy of typical ZSL task, due to the fine-grained and huge label space [39]. In order to understand and alleviate the hubness problem in ZSL, some works [19, 39, 56] propose different solutions based on Euclidean metric. [39] proposes using linear ridge regression with a mapping direction opposite to that in existing work. Like [39], [56] applies deep networks to project the semantic vectors. However, these solutions all depend on their specific models, rather than general models. For the greatest applicability, we propose a novel method which can be generalized to different models. The conceptual diagram of our method is demonstrated in Figure 2.

We are the first to prove that cosine is a model-agnostic metric to alleviate the hubness problem in zero-shot learning and draw the conclusion that a higher-dimensional cosine metric space can better suppress the emergence of these hubs. Specifically, assuming the normalized embedded semantic vectors follows a uniform distribution, we conduct a theoretical investigation which shows that the hubness is inversely proportional to the dimensionality of the metric space.

Particularly, our conclusion is the sufficient and unnecessary condition to previous findings in [39, 56]. Due to the dimensionality of the visual features is larger than that of semantic vectors, reasoning from our conclusion we also can deliver the same solution with [39, 56] that mapping from the semantic space to the visual space is better to alleviate hubness than the opposite direction in other works. Therefore, our conclusions have more general applicability for different models and metric spaces' dimensions, rather than being restricted to ridge regression with Euclidean metric or choosing from only two embedding dimensions.

Moreover, we introduce a diversity-based regularization with the cosine metric to underpin satisfy the assumption of theoretical analysis, *the normalized embedded semantic vectors follow a uniform distribution*, and endow the model with great discriminative ability. Our diversity regularizer encourages a smaller mean of the cosine metric between pairwise mapped semantic vectors and a smaller variance of that. A smaller mean indicates that pairwise embedded semantic vectors share less cosine similarity and are more different from each other, while a smaller variance implies that the embedded semantic vectors spread out as uniformly as possible to different directions. Similar idea [51] applies the diversity on the angles between pairwise hidden units to regularize the restricted Boltzmann machine and use the lower bound of the diversity for avoiding the hard optimization. In contrast, we simplify the diversity with the cosine metric rather than angels and directly optimize it by seamlessly incorporating into the embedding process.

The major contributions can be summarized in the following three aspects:

- We are the first to prove that cosine is a model-agnostic metric to alleviate the hubness problem in ZSL. We provide theoretical analysis which demonstrates that hubness can be better handled with co-

sine metric through a higher-dimensional metric space, which brings a better understanding of hubs and gives a useful guide to develop new ZSL models.

- We introduce a diversity-based regularization. It can underpin the theoretical assumption about uniform distribution and further improve the performance of our model by incorporating diversity to increase the discriminative ability.
- We apply a plain embedding model and conduct extensive experiments on five benchmark datasets and the large-scale Imagenet dataset to show that our method can consistently improve the performance by large margins compared to previous embedding methods.

## 2. Related Work

Zero-shot learning has generated widespread research attention in diverse research fields. We review the related work in terms of embedding models, loss functions, the hubness problem and distance metrics.

**Embedding models.** Embedding models in ZSL aims to transfer the shared knowledge from seen to unseen categories by projecting the low-level visual features to their corresponding semantic vectors. Early works were based on the Bayesian formulation [17, 18, 58]. The direct attribute prediction (DAP) and the indirect attribute prediction (IAP) are two-stage approaches, which suffer from the domain shift problem [49]. After that, traditional regression models, e.g., ridge regression, have also been explored [15, 36, 39, 55] as the embedding functions to grasp the relationship between visual space and semantic space. Moreover, deep embedding models [10, 20, 39, 40, 53, 54, 56] have also been investigated for zero-shot learning, and they mainly differ in choosing the metric space.

**Hubness problem.** ZSL suffers from the hubness problem due to the use of nearest neighbor search in the test stage. To alleviate this problem, a few recent approaches [39, 53, 56] have used the visual space as the metric space. [39] discussed the effect of hubness in zero-shot learning when ridge regression is used to find the mapping from the visual space to the semantic space. This indicates it is better to choose the visual space as the metric space to alleviate the hubness problem. [56] argued that in the visual space the nearest neighbor search can suffer much less from the hubness problem and thus become more effective. However, those conclusions are drawn based on the Euclidean and ridge regression, so they may not be applicable to other specific models.

**Loss functions.** Different loss functions have also been investigated in the setting of zero-shot learning [11]. Most previous work [15, 36, 40, 56] straightforwardly adopted the least square loss, which however is not optimal for ZSL because of the huge semantic gap between visual and attribute spaces. The margin-based loss is used by regarding the nearest neighbor search evaluation as a ranking problem [1, 2, 10, 34, 54]. [20] combined three training losses, including the least square loss, margin-based loss and a binary cross-entropy loss.

**Distance metrics.** Existing ZSL methods mainly use the Euclidean distance [13, 15, 39, 40, 52, 56] or dot product [6, 10, 20, 36] as the metric to find the most appropriate metric space. In addition, the cosine similarity has recently been explored in diverse visual recognition tasks [3, 22–25, 43, 44]. Especially, PSRZSL [3] uses cosine metric to measure the semantic relationships in ZSL but does not reveal the relationship between the cosine metric and hubness problem. In contrast, we are the first to prove that cosine is a model-agnostic metric to alleviate the hubness problem in the metric space. The theoretical conclusions further inspire us to introduce a diversity-based regularizer for enhancing the model's discriminative ability.

## 3. Method

In this section, we mainly discuss the effect of hubness in zero-shot learning, which is directly related to embedding models. We first describe a simple and general embedding model for ZSL in Section 3.1. Based on the embedding model, we prove the model-agnostic metric for hubness in Section 3.2, followed by introducing a diversity-based regularizer in Section 3.3.

Following the GBU setting in [49], we begin by defining the three important spaces in zero-shot learning: the visual space $\mathbf{X}$, which consists of visual features extracted by pre-trained models like GoogLeNet [42] and ResNet [12]; the label space $\mathbf{Y}$, which labels each visual feature according to its category; and the semantic space $\mathbf{S}$, which provides high-level semantic vectors for each training or test category. The semantic vector can be represented by either attributes annotated by humans to describe the visual patterns [18], or word embeddings generated by Word2Vec [27]. Let $\mathbf{Y}_{tr} = \{y_p\}_{p=1}^{P}$ and $\mathbf{Y}_{ts} = \{y_q\}_{q=1}^{Q}$ denote disjoint seen and unseen classes, $\mathbf{Y}_{tr} \cap \mathbf{Y}_{ts} = \emptyset$. $P$ denotes the number of seen classes, while $Q$ denotes that of unseen classes. Meanwhile, $\mathbf{S}_{tr} = \{\mathbf{s}_p\}_{p=1}^{P}$ and $\mathbf{S}_{ts} = \{\mathbf{s}_q\}_{q=1}^{Q}$ denote the corresponding seen and unseen semantic vectors. With the help of semantic space $\mathbf{S}$, zero-shot learning aims to learn a hypothesis $f : \mathbf{X} \rightarrow \mathbf{Y}$ and apply it to predict the test sample from unseen classes.

### 3.1. A General Embedding Model

According to the utilization of semantic space, recent works can be roughly divided into embedding models and generative models. In this paper, we mainly focus on traditional embedding models. For clear clarification, we first describe a simple and general embedding model with the backbone architecture, which is further developed in Section 3.2 and Section 3.3.

**Backbone architecture.** The backbone architecture consists of a visual projection and a semantic embedding network. We adopt a basic multi-layer perceptrons (MLP) as our semantic embedding network like [21, 56], which takes a $k$-dimensional semantic vector $\mathbf{s}_p \in \mathbb{R}^k$ as the input and outputs a $D$-dimensional embedded semantic vector $\varphi(\mathbf{s}_p) \in \mathbb{R}^D$. Meanwhile, following [5], we apply the PCA projection as the visual projection, which takes a $d$-dimensional visual feature $\mathbf{x}_n \in \mathbb{R}^d$ as the input and outputs an projected visual feature $\phi(\mathbf{x}_n) = \mathbf{M}\mathbf{x}_n$, $\mathbf{M} \in \mathbb{R}^{d\times D}$ is the PCA projection matrix computed over training data of the seen classes. To sum up, $\phi(\mathbf{x}_n)$ and $\varphi(\mathbf{s}_p)$ are visual projection of visual features and semantic embedding of semantic vectors in the common $D$-dimensional metric space.

**Loss function.** In the train stage, there are $N$ labelled training samples from seen classes. For optimizing the model's parameters, we adopt cross-entropy loss with softmax over training samples from seen classes $\mathbf{Y}_{tr}$:

$$\mathcal{L} = -\sum_{n=1}^{N}\sum_{p=1}^{P} y_{n,p}\log\left(\frac{e^{f(\phi(\mathbf{x}_n),\varphi(\mathbf{s}_p))}}{\sum_{p'=1}^{P} e^{f\left(\phi(\mathbf{x}_n),\varphi\left(\mathbf{s}_{p'}\right)\right)}}\right), \quad (1)$$

where $f(\cdot,\cdot)$ denotes the similarity metric between the projected visual feature and the embedded semantic vector, e.g. Euclidean and cosine; the ground truth of the training example $y_{n,p} = 1$ if $y_n = p$ and 0 otherwise.

**Prediction function.** In the test stage, the nearest neighbor search is applied for making the predicted label $y(\mathbf{x}_t)$ of test visual feature $\mathbf{x}_t$:

$$y(\mathbf{x}_t) = \arg\max_{c} f(\phi(\mathbf{x}_t), \varphi(\mathbf{s}_c)). \quad (2)$$

In the conventional ZSL setting, test visual features only come from unseen classes and thus the search scale is unseen classes, $c \in \mathbf{Y}_{ts}$. In the generalized ZSL setting, the test visual features are not sure from seen or unseen classes. Thus, its search scale is the union set of all seen and unseen classes, $c \in \mathbf{Y}_{tr} \cup \mathbf{Y}_{ts}$.

### 3.2. Model-Agnostic Metric for Hubness

Based on the embedding model in Section 3.1, we prove cosine is the model-agnostic metric for alleviating the hubness problem in ZSL. The theoretical analysis of hubness is closely related to the expectation and variance of the embedded semantic vectors' coordinates. Thus, before providing the analysis, we first introduce the assumption about the distribution of the embedded semantic vectors and infer the statistics.

Before the specific analysis, we first instantiate the similarity metric $f(\cdot,\cdot)$ in the loss function (1) and prediction function (2) as cosine metric and review it by normalization,

$$f(\phi(\mathbf{x}), \varphi(\mathbf{s})) = \cos(\phi(\mathbf{x}), \varphi(\mathbf{s})) = \mathbf{a}^T\mathbf{b}, \quad (3)$$

where $\mathbf{a} = \frac{\varphi(\mathbf{x})}{\|\varphi(\mathbf{x})\|}$ and $\mathbf{b} = \frac{\phi(\mathbf{s})}{\|\phi(\mathbf{s})\|}$ denotes the normalized projected visual feature and the normalized embedded semantic vector in the cosine metric space. Due to the norm of $\mathbf{a} = [a_1, a_2..., a_D]^T$ and $\mathbf{b} = [b_1, b_2..., b_D]^T$ are equal to 1, the cosine metric space can be viewed as a $D$-dimensional unit sphere. We provide the visualization of Euclidian and cosine metric spaces to show the difference between them in the Figure 1. Especially, Cosine measures the angle difference corresponding to the semantic difference, which enables us to better distinguish different semantic classes.

**Assumption about the distribution.** We assume that the normalized embedded semantic vectors follow a uniform distribution on the cosine metric space.

As the prototypes of the nearest neighbor search, the embedded semantic vectors have a great influence on the classifier's discriminative ability. The smaller similarity of pairwise embedded semantic vectors indicates the classifier has the larger discriminative ability for test samples from the two classes. Intuitively, different embedded semantic vectors should be diversified as certain as possible. Therefore, we assume that the normalized embedded semantic vectors follow a uniform distribution on the cosine metric space. Based on this assumption, we can get the marginal probability density of components $b_i (i = 1, ..., D)$ of the normalized embedded semantic vectors like Theorem 1 from previous work [41].

**Theorem 1** ([41]). *If* $\mathbf{b} = [b_1, b_2..., b_D]^T$ *follows an uniform distribution on the D-dimensional unit sphere, the marginal probability density of the component* $b_i$ *is*

$$f_D(b_i) = \kappa_D\left(1 - {b_i}^2\right)^{\frac{(D-3)}{2}} I_{(-1,1)}(b_i),$$

*where* $\kappa_D = \frac{\Gamma\left(\frac{D}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{D-1}{2}\right)}$, *and* $I_{(-1,1)}(b_i)$ *is a indicator function that indicates the definitional domain is* $(-1, 1)$.

**Statistics of the component.** From the Theorem 1, we further infer the expectation and variance of the component $b_i$ of the normalized embedded semantic vector. Due to this marginal probability density is a symmetric function, intuitively the expectation of $b_i$ is zero.

Hereby, we are more interested in the result of the variance, which is also negatively related to the dimension of the unit sphere. Theoretically, the following proposition shows that the variance of the component is indeed only inversely proportional to the dimension of the cosine metric space. The full proof of Proposition 1 is included in the supplementary material.

**Proposition 1.** *If* $\mathbf{b} = [b_1, b_2..., b_D]^T$ *follows an uniform distribution on the D-dimensional unit sphere, the variance of the component* $b_i$ *is*

$$\mathrm{Var}\left[b_i\right] = \int_{-1}^{+1} {b_i}^2 f_D\left(b_i\right) \mathrm{d}b_i = \frac{1}{D}.$$

Therefore, we get the mean and variance of the component of the normalized embedded semantic vectors which is assumed to be uniformly distributed.

**Hubness phenomenon with cosine metric.** The statistics of the component are applied for analyzing the effect of hubness in ZSL. The hubness phenomenon is concerned with the nearest neighbor search during the test stage of ZSL. Hubs, a small number of prototypes in the embedding space, may occur as the nearest neighbor of multiple visual features whose labels are inconsistent with the central hub.

In order to explicitly describe the effect of hubness, following [32, 39] we adopt the expected difference between the squared distances from the two random normalized embedded semantic vectors $\mathbf{b}_1$ and $\mathbf{b}_2$ to the normalized projected visual feature $\mathbf{a}$, which can guide us to understand and alleviate the hubness in ZSL. Specifically, the smaller absolute value the expected difference has, the harder for $\mathbf{b}_1$ and $\mathbf{b}_2$ to be hubs.

**Proposition 2.** *Let* $\mathbf{a} = [a_1, a_2..., a_D]^T$ *be a point sampled from a distribution* $\mathcal{X}$ *on the D-dimensional unit sphere. Let* $\mathbf{b} = [b_1, b_2..., b_D]^T$ *follows an uniform distribution* $\mathcal{S}$ *on the D-dimensional unit sphere.* $\varepsilon$ *is the norm of expectation of* $\mathbf{a}$*, and we assume that* $\varepsilon \neq 0$*. Further, let* $\mathbf{a}^* = \frac{\mathrm{E}[\mathbf{a}]}{\|\mathrm{E}[\mathbf{a}]\|}$ *be the normalized expectation of* $\mathbf{a}$*, and* $\sigma = \sqrt{\mathrm{Var}_{\mathcal{S}}\left[\cos(\mathbf{a}^*, \mathbf{b})\right]}$ *be the standard deviation of* $\cos\left(\mathbf{a}^*, \mathbf{b}\right)$*. Consider two fixed samples* $\mathbf{b}_1$ *and* $\mathbf{b}_2$*, such that the cosine metric* $\cos\left(\mathbf{a}^*, \mathbf{b}_1\right)$ *and* $\cos\left(\mathbf{a}^*, \mathbf{b}_2\right)$ *are* $\gamma\sigma$ *apart. In other words,*

$$\cos\left(\mathbf{a}^*, \mathbf{b}_1\right) - \cos\left(\mathbf{a}^*, \mathbf{b}_2\right) = \gamma\sigma.$$

*The expected difference* $\Delta$ *between the squared distances from* $\mathbf{b_1}$ *and* $\mathbf{b_2}$ *to* $\mathbf{a}$*, i.e.,*

$$\Delta = \mathrm{E}_{\mathcal{X}}\left[\left\|\mathbf{a} - \mathbf{b}_2\right\|^2\right] - \mathrm{E}_{\mathcal{X}}\left[\left\|\mathbf{a} - \mathbf{b}_1\right\|^2\right]$$

*is given as follows:*

$$\Delta = \frac{2\varepsilon\gamma}{\sqrt{D}}. \tag{4}$$

In Proposition 2, $\mathbf{a}$ and $\mathbf{b}$ denotes the normalized projected visual feature and the normalized embedded semantic vector on the cosine metric space, respectively. $\Delta$ represents the expected difference between the chord length from $\mathbf{b}_1$ and $\mathbf{b}_2$ to $\mathbf{a}$.

From Equation (4), we conclude that the absolute value of the expected difference is inversely proportional to the dimensionality of the metric space. This conclusion based on cosine metric indicates that a higher-dimensional embedding space can better alleviate the hubness problem. Compared with commonly-used dimensionality of the metric space ($D = 64, 128, 256, 512, 1024, 2048, 2560, 3072, 4096$), our method sets the dimensionality of the cosine metric space as 2048. Experimental results can be found in Sec.4.4.

Importantly, we would like to highlight that our conclusion based on the cosine metric is general and model-agnostic, which is not restricted to any specific model, e.g., ridge regression [39]. The full proof sketch is provided in the supplementary material.

### 3.3. Diversity regularizer

The assumption in Section 3.2 originates from the intuition, *different embedded semantic vectors should be diversified as certain as possible*. Inspired by this, we introduce a diversity-based regularization to underpin the assumption and further enhance the model's discriminative ability.

In order to clearly state the regularizer, we first review the cosine metric between pairwise mapped semantic vectors by normalization as follows,

$$f\left(\varphi(\mathbf{s}_i), \varphi(\mathbf{s}_j)\right) = \cos\left(\varphi(\mathbf{s}_i), \varphi(\mathbf{s}_j)\right) = {\mathbf{b}_i}^T \mathbf{b}_j. \tag{5}$$

Same with Equation (3), $\mathbf{b}$ represents the normalized embedded semantic vector.

Our diversity regularizer encourages a smaller mean of the cosine metric between pairwise mapped semantic vectors and a smaller variance of that. A smaller mean indicates that pairwise embedded semantic vectors share smaller cosine similarity and more different from each other, while a smaller variance implies that the embedded semantic vectors uniformly spread out to different directions. To be specific, manipulating all semantic vectors from seen classes, the regularizer takes the following form:

$$\Phi = \mathrm{E}_{i,j}\left[{\mathbf{b}_i}^T \mathbf{b}_j\right] + \mathrm{Var}_{i,j}\left[{\mathbf{b}_i}^T \mathbf{b}_j\right], \tag{6}$$

where $i, j \in \mathbf{Y}_{tr}$ and $\mathrm{E}\left[\cdot\right]$ and $\mathrm{Var}\left[\cdot\right]$ denote the mean and variance of the cosine metric between pairwise mapped semantic vectors.

Minimizing $\Phi$ can enhance the diversity of the embedded semantic vectors in the metric space, which further promotes the aforementioned uniform distribution.

Finally, we obtain the final objective function of our method by combining (1) and (6), as follows:

$$\min_{\varphi} \mathcal{L} + \lambda\Phi + \gamma\Psi(\varphi), \tag{7}$$

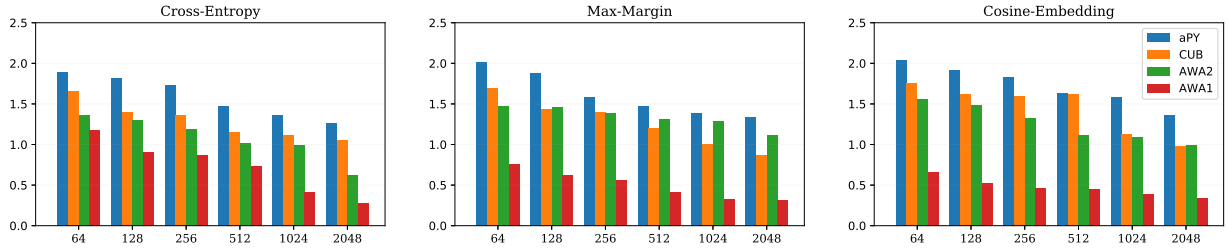



Figure 3. The $N_1$ skewness on four datasets for evaluating three kinds of embedding models based on cross-entropy, max-margin and cosine-embedding objective functions, which all apply the cosine metric in their formula. For fair comparisons, we apply the same backbone architecture for the three embedding models. The x-axis and y-axis represent the skewness values and the metric space's dimensions, respectively. For the three models, the skewness value of measuring hubness decreases consistently for the above datasets when the metric space dimension increases. The best viewed in color.

Table 1. Statistics of datasets used in our experiments.

| Dataset | Semantic/Dim | Image | Seen/Unseen Classes |
|---|---|---|---|
| AWA1 | A/85 | 30475 | 40/10 |
| AWA2 | A/85 | 37322 | 40/10 |
| SUN | A/102 | 14340 | 645/72 |
| CUB | A/312 | 11788 | 150/50 |
| aPY | A/64 | 15339 | 20/12 |
| Imagenet | W/1000 | 254000 | 1000/360 |

where $\lambda$ and $\gamma$ are the hyper-parameters, which are fixed as 1 in our experiments; $\Psi(\varphi)$ represents the weight decay of the learnable semantic network in Section 3.1. We minimize the objective function (7) by applying the Adam optimizer [14] with a learning rate of $10^{-4}$ during the training stage.

# 4. Experiments and Results

## 4.1. Datasets and Experimental Setting

We evaluate our method on five widely-used benchmark datasets and one large-scale dataset in our experimens: Animals with Attributes (**AWA1**) [17], Animals with Attributes 2 (**AWA2**), SUN attribute dataset (**SUN**)[30], Caltech-UCSD-Birds 200-2011 (**CUB**) [47], Attribute Pascal and Yahoo dataset (**aPY**) [8], and Imagenet dataset [37]. Datasets and their statistics are summarized in Table 1.

For fair comparisons, attributes [18] are adopted as the semantic vectors for the five benchmark while word2vec representations [27] of each class are used as the semantic vectors for large-scale Imagenet dataset.

We follow the proposed splits on the evaluation work [49] for the above five datasets. Moreover, we also apply the 2048 dimensional visual features extracted from Resnet-101 [12], which was pre-trained on the ImageNet dataset [35] provided by [49].

For Imagenet dataset, 1000 classes from ILSVRC 2012 [37] are used for training, while non-overlapping 360 classes from the ILSVRC 2010 data are used for test. We follow the settings in [15, 16] adopt GoogLeNet [42] features for this dataset.

## 4.2. Hubness with Different Models

In Section 3.2, we prove cosine is model-agnostic for alleviating the hubness problem and conclude that a higher-dimensional embedding space with cosine metric can better alleviate the hubness problem. Particularly, we provide experimental results about the hubness problem with different models to verify our conclusions. To be specific, we evaluate three kinds of models based on cross-entropy, max-margin and cosine-embedding objective functions, which all apply the cosine metric in their formula. For fair comparisons, we apply the same backbone architecture in Section 3.1 for the three models.

We adopt the widely-used $skewness$ of the (empirical) $N_k$ distribution [33, 39, 56] to measure the degree of hubness in the nearest neighbor search on those metric spaces of different dimensions. The skewness is defined as follows:

$$S_{N_k(\mathbf{s})} = \frac{\mathrm{E}\left[N_k(\mathbf{s}) - \mathrm{E}\left[N_k(\mathbf{s})\right]\right]^3}{\mathrm{Var}\left[N_k(\mathbf{s})\right]^{\frac{3}{2}}},$$

where $N_k(\mathbf{s})$ denotes the number of times each prototype $\mathbf{s}$ occurs among the $k$ nearest neighbors of all test samples, and $\mathrm{E}\left[N_k\right]$ and $\mathrm{Var}\left[N_k\right]$ denote its mean and variance.

The results have been reported in Figure 3. The x-axis and y-axis represent the skewness values and the metric space's dimensions, respectively. We observe that as the dimension of the cosine metric space increases, the value of $N_1$ in different datasets and models decreases consistently which indicates that the emergence of hubs is suppressed. Agreeing with the conclusions of Proposition 2, this observation demonstrates that cosine is a model-agnostic metric for alleviating the hubness problem and a higher dimensional metric space can better reduce the hubs. In addition, cross-entropy based embedding model (in the left of Figure 3) have smaller average skewness values than other models, which indicates that embedding models with cross-entropy are more appropriate for alleviating the hubness problem in ZSL. Thus, we adopt the cross-entropy objective function in the subsequent experiments.

Table 2. Average per-class accuracy (%) of ZSL for the five benchmarks on the proposed split (PS). '-' means that no reported results are available. The best result is marked in **red**, the second best in **blue**.

| Method | AWA1 | AWA2 | SUN | CUB | aPY |
|---|---|---|---|---|---|
| DAP [18] | 44.1 | 46.1 | 39.9 | 40.0 | 33.8 |
| IAP [18] | 35.9 | 35.9 | 19.4 | 24.0 | 36.6 |
| CONSE [29] | 45.6 | 44.5 | 38.8 | 34.3 | 26.9 |
| CMT [40] | 39.5 | 37.9 | 39.9 | 34.6 | 28.0 |
| SSE [57] | 60.1 | 61.0 | 51.5 | 43.9 | 34.0 |
| LATEM [48] | 55.1 | 55.8 | 55.3 | 49.3 | 35.2 |
| ALE [1] | 59.9 | 62.5 | 58.1 | 54.9 | 39.7 |
| DESIVE [10] | 54.2 | 59.7 | 56.5 | 52.0 | 39.8 |
| SJE [2] | 65.6 | 61.9 | 53.7 | 53.9 | 32.9 |
| ESZSL [36] | 58.2 | 58.6 | 54.5 | 53.9 | 38.3 |
| SYNC [4] | 54.0 | 46.6 | 56.3 | 55.6 | 23.9 |
| SAE [15] | 53.0 | 54.1 | 40.3 | 33.3 | 8.3 |
| DEM [56] | 68.4 | 67.1 | 61.9 | 51.7 | 35.0 |
| RN [53] | 68.2 | 64.2 | - | 55.6 | - |
| PSRZSL [3] | - | 63.8 | 61.4 | 56.0 | 38.4 |
| ZSKL [55] | 71.0 | **70.5** | 61.7 | 57.1 | 45.3 |
| SP-ANE [6] | - | 58.5 | 59.2 | 55.4 | 24.1 |
| MLSE [7] | - | 67.8 | **62.8** | **64.2** | **46.2** |
| Ours w/o diveristy | **71.1** | 70.2 | 62.1 | 54.0 | **46.2** |
| **Ours** | **72.7** | **72.0** | **62.6** | **59.6** | **47.3** |

Table 3. Average per-class accuracy (%) of ZSL for the Imagenet dataset.

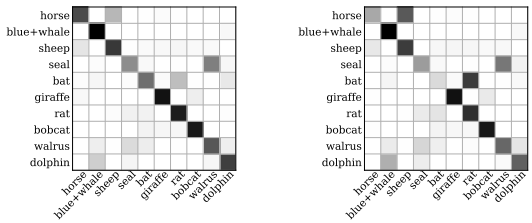
| Method | Accuracy |
|---|---|
| DESIVE [10] | 12.8 |
| CONSE [29] | 15.5 |
| DEM [56] | 25.7 |
| SAE [15] | 27.2 |
| Ours w/o diveristy | 27.3 |
| **Ours** | **27.6** |

### 4.3. Compared with Class-Inductive Methods

We start our evaluations in the conventional ZSL setting followed by the generalized zero-shot learning setting (GZSL). Test samples defined in conventional ZSL all belong to unseen classes and their search space is thus limited to the unseen classes. Meanwhile, for practical considerations, GZSL not only considers the accuracy of samples for the unseen classes but also that of the samples belonging to seen classes. The search space of GZSL is the union set of seen and unseen classes.

**Conventional Zero-Shot Learning.** For the conventional ZSL setting, we first train our simple and general embedding model with the objective function of Equation (7) on all training samples from seen classes. We then apply the nearest neighbor search of Equation (2) using the semantic vectors from unseen classes. The nearest neighbor search is used to predict the test examples from unseen classes. The average per-class accuracy for the five benchmarks and large-scale Imagenet dataset is reported in Table 2 and Table 3, respectively.

Experimental results show that the improvements are consistent across scale in the five middle-scale benchmarks and large-scale Imagenet dataset. Specifically, in Table 2, our proposed method outperforms state-of-the-art embedding models (upper part of the table) on AWA1, AWA2 and aPY by 1.7%, 1.5% and 1.1% Top-1 accuracy, respectively. We observe a large increase in performance when we add the diversity regularization to the model. This shows that the diversity regularizer is beneficial for improving the discriminative ability of the embedding space. Meanwhile, in Table 3, our method outperforms SAE by 0.4% on Imagenet dataset, which verifies our method's applicability for the large-scale dataset.




(a) Ours (Acc. 72.0%) (b) Ours w/o diversity (Acc. 70.2%)

Figure 4. Confusion matrixes evaluated on the AWA2 dataset under the ZSL setting. The x-axis and y-axis represents the predicted labels and the true labels of the test samples from unseen classes, respectively.

In addition, to further assess the effectiveness of the diversity-based regularizer, we provide the visualization about the confusion matrixes for our method with and without diversity on AWA2 dataset in Figure 4. The visualization shows that more test samples from unseen classes are classified corrected by our method with diversity than ours without diversity, which indicates that diversity-based regularizer can enhance our model's discriminative ability.

**Generalized Zero-Shot Learning.** The GZSL setting predicts the test examples from both the seen and unseen classes, with no prior distinction between them. The training process of the model is the same as that of the ZSL setting. Table 4 presents our results for generalized zero-shot learning (GZSL). $ts$ and $tr$ indicate the average per-class accuracies of test samples belonging to unseen classes and seen classes, respectively. $H$ is the harmonic mean of $ts$ and $tr$, which balances their quality of performance.

We first note that our method outperforms state-of-the-art embedding models on AWA2, CUB and aPY by 0.2%, 3.2% and 1.4%, respectively, in terms of the generalized harmonic mean $H$. Meanwhile, our method also outperforms the state-of-the-art embedding approaches on AWA2, CUB and aPY by 0.8%, 3.5% and 0.7%, respectively, in terms of indicator $ts$. In addition, we also observe that our method with diversity is noticeably better than that without diversity on all five benchmarks, especially for the evaluation $ts$. When the search space of GZSL becomes larger than that of the conventional ZSL, the diversity regularizer shows better performance on improving the discriminative ability of our model.

Table 4. Results on GZSL under the GBU setting. '-' means that no reported results are available. The best number is marked in **bold**.

| | AWA1 | | | AWA2 | | | SUN | | | CUB | | | aPY | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | $ts$ | $tr$ | $H$ | $ts$ | $tr$ | $H$ | $ts$ | $tr$ | $H$ | $ts$ | $tr$ | $H$ | $ts$ | $tr$ | $H$ |
| DAP [18] | 0.0 | 88.7 | 0.0 | 0.0 | 84.7 | 0.0 | 4.2 | 25.1 | 7.2 | 1.7 | 67.9 | 3.3 | 4.8 | 78.3 | 8.0 |
| IAP [18] | 2.1 | 78.2 | 4.1 | 0.9 | 87.6 | 1.8 | 1.0 | 37.8 | 1.8 | 0.2 | **72.8** | 0.4 | 5.7 | 65.6 | 10.4 |
| CONSE [29] | 0.4 | 88.6 | 0.8 | 0.5 | 90.6 | 1.0 | 6.8 | 39.9 | 11.6 | 1.6 | 72.2 | 3.1 | 0.0 | **91.2** | 0.0 |
| CMT [40] | 0.9 | 87.6 | 1.8 | 0.5 | 90.0 | 1.0 | 8.1 | 21.8 | 11.8 | 7.2 | 60.1 | 8.7 | 1.4 | 85.2 | 2.8 |
| SSE [57] | 7.0 | 80.5 | 12.9 | 8.1 | 82.5 | 14.8 | 2.1 | 36.4 | 4.0 | 8.5 | 46.9 | 14.4 | 0.2 | 78.9 | 0.4 |
| LATEM [48] | 7.3 | 71.7 | 13.3 | 11.5 | 77.3 | 20.0 | 14.7 | 28.8 | 19.5 | 15.2 | 57.3 | 24 | 0.1 | 73.0 | 0.2 |
| ALE [1] | 16.8 | 76.1 | 27.5 | 14.0 | 81.8 | 23.9 | 21.8 | 33.1 | 26.3 | 23.7 | 62.8 | 34.4 | 4.6 | 73.7 | 8.7 |
| DESIVE [10] | 13.4 | 68.7 | 22.4 | 17.1 | 74.7 | 27.8 | 16.9 | 27.4 | 20.9 | 23.8 | 53 | 32.8 | 4.9 | 76.9 | 9.2 |
| SJE [2] | 11.3 | 74.6 | 19.6 | 8.0 | 73.9 | 14.4 | 14.7 | 30.5 | 19.8 | 23.5 | 59.2 | 33.6 | 3.7 | 55.7 | 6.9 |
| ESZSL [36] | 6.6 | 75.6 | 12.1 | 5.9 | 77.8 | 11.0 | 11.0 | 27.9 | 15.8 | 12.6 | 63.8 | 21 | 2.4 | 70.1 | 4.6 |
| SYNC [4] | 8.9 | 87.3 | 16.2 | 10.0 | 90.5 | 18.0 | 7.9 | **43.3** | 13.4 | 11.5 | 70.9 | 19.8 | 7.4 | 66.3 | 13.3 |
| SAE [15] | 1.8 | 77.1 | 3.5 | 1.1 | 82.2 | 2.2 | 8.8 | 18.0 | 11.8 | 7.8 | 54.0 | 13.6 | 0.4 | 80.9 | 0.9 |
| DEM [56] | **32.8** | 84.7 | **47.3** | 30.5 | 86.4 | 45.1 | 20.5 | 34.3 | 25.6 | 19.6 | 57.9 | 29.2 | 11.1 | 75.1 | 19.4 |
| RN [53] | 31.4 | **91.3** | 46.7 | 30.0 | **93.4** | 45.3 | - | - | - | 38.1 | 61.1 | 47.0 | - | - | - |
| PSRZSL [3] | - | - | - | 20.7 | 73.8 | 32.3 | 20.8 | 37.2 | 26.7 | 24.6 | 54.3 | 33.9 | 13.5 | 51.4 | 21.4 |
| ZSKL [55] | 18.3 | 79.3 | 29.8 | 18.9 | 82.7 | 30.8 | 21.0 | 31.0 | 25.1 | 24.2 | 63.9 | 35.1 | 11.9 | 76.3 | 20.5 |
| SP-ANE [6] | - | - | - | 23.3 | 90.9 | 37.1 | **24.9** | 38.6 | **30.3** | 34.7 | 70.6 | 46.6 | 13.7 | 63.4 | 22.6 |
| MLSE [7] | - | - | - | 23.8 | 83.2 | 37.0 | 20.7 | 36.4 | 26.4 | 22.3 | 71.6 | 34.0 | 12.7 | 74.3 | 21.7 |
| Ours w/o diversity | 25.2 | 82.7 | 38.6 | 23.9 | 84.9 | 37.2 | 19.4 | 38.3 | 25.7 | 22.7 | 57.8 | 32.6 | 13.2 | 68.8 | 22.2 |
| **Ours** | 30.8 | 81.5 | 44.7 | **31.3** | 83.2 | **45.5** | 21.6 | 39.1 | 27.8 | **41.6** | 63.2 | **50.2** | **14.4** | 71.4 | **24.0** |

### 4.4. Detailed analysis

We provide extend experiments to show the advantage of PCA and investigate the upper bound of dimension in the metric space. We choose two different strategies (PCA and MLP) for projecting visual features into the common metric space. We set the dimensionality of the metric space as 64, 128, 256, 512, 1024, 2048, 2560, 3072 and 4096. Especially, PCA(D=2056/3072/4096)-based visual features concatenate the projected features of PCA(D=2048)-based and PCA(D=512/1024/2048)-based.

**Advantage of PCA.** We choose PCA as the dimensional reduction strategy for concreteness. Nevertheless, our method is also applicable to other dimensional reduction strategies. Figure 5 shows that the PCA-based method outperforms the MLP-based method on different-dimensional embedding space by a large margin. Compared with the non-parametric strategy (PCA), the MLP with parameters needs more training times and is more prone to overfitting [5]. Thus, in our paper, we choose PCA as the dimensional reduction strategy. See additional results on other datasets in the supplementary material.

**Upper bound of dimension in the metric space.** Due to various complicated factors, the performance of our method should have the upper bound with the increase of the dimensionality of the metric space. Figure 5 shows that the average per-class accuracy obtains the peak value at the dimensionality equal to 2048. When the dimensionality is less than or equal to 2048, the metric space with higher dimensionality has better performance as also indicated in our theoretical analysis. Meanwhile, when the dimensionality is more than 2048, the performance decreases because the backbone network has more parameters and then are more prone to over-fitting to seen classes. Thus, in our paper, the upper bound of the dimension in the metric space is 2048.

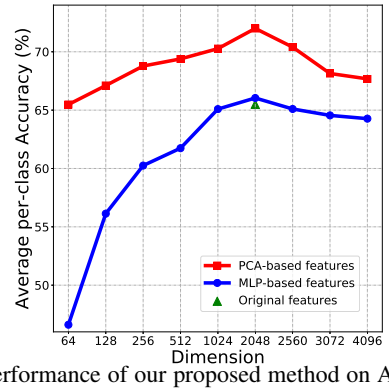



Figure 5. Performance of our proposed method on AWA2 dataset when visual features are projected into different dimensional embedding space by MLP and PCA. The best viewed in color.

## 5. Conclusions

In this paper, we are the first to prove that cosine is a model-agnostic metric to alleviate hubness in ZSL and conclude that a higher dimensional cosine metric space can better suppress the emergence of these hubs, which provides useful guidance in designing ZSL algorithms. Moreover, we introduce a diversity-based regularizer which underpins the theoretical assumption about the uniform distribution and further improves our model's discriminative ability. Extensive experiments on five benchmarks and large-scale Imagenet dataset demonstrate our method consistently delivers high performance across scales and substantially surpasses previous embedding models.

## Acknowledgment

This paper was supported by the Natural Science Foundation of China under Grant No. 61827901, No. 91738301, and No. 61871016.

## References


[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2016. 3, 7, 8

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015. 3, 7, 8

[3] Y. Annadani and S. Biswas. Preserving semantic relations for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7603–7612, 2018. 3, 7, 8

[4] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5327–5336, 2016. 7, 8

[5] S. Changpinyo, W.-L. Chao, and F. Sha. Predicting visual exemplars of unseen classes for zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3476–3485, 2017. 4, 8

[6] L. Chen, H. Zhang, J. Xiao, W. Liu, and S.-F. Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1043–1052, 2018. 1, 3, 7, 8

[7] Z. Ding and H. Liu. Marginalized latent semantic encoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6191–6199, 2019. 7, 8

[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009. 6

[9] R. Felix, V. B. Kumar, I. Reid, and G. Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, 2018. 1

[10] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 1, 3, 7, 8

[11] Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, L. Sigal, and S. Gong. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine*, 35(1):112–125, 2018. 3

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 6

[13] H. Jiang, R. Wang, S. Shan, and X. Chen. Learning class prototypes via structure alignment for zero-shot recognition. In *Proceedings of the European conference on computer vision. ECCV*, pages 118–134, 2018. 3

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[15] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3174–3183, 2017. 1, 3, 6, 7, 8

[16] V. Kumar Verma, G. Arora, A. Mishra, and P. Rai. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4281–4289, 2018. 1, 6

[17] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009. 1, 3, 6

[18] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. 3, 6, 7, 8

[19] A. Lazaridou, G. Dinu, and M. Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 270–280, 2015. 2

[20] 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*, pages 4247–4255, 2015. 1, 3

[21] S. Liu, M. Long, J. Wang, and M. I. Jordan. Generalized zero-shot learning with deep calibration network. In *Advances in Neural Information Processing Systems*, pages 2005–2015, 2018. 4

[22] W. Liu, Z. Liu, Z. Yu, B. Dai, R. Lin, Y. Wang, J. M. Rehg, and L. Song. Decoupled networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2771–2779, 2018. 3

[23] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 1, 2017.

[24] W. Liu, Y.-M. Zhang, X. Li, Z. Yu, B. Dai, T. Zhao, and L. Song. Deep hyperspherical learning. In *Advances in Neural Information Processing Systems*, pages 3950–3960, 2017.

[25] C. Luo, J. Zhan, X. Xue, L. Wang, R. Ren, and Q. Yang. Cosine normalization: Using cosine similarity instead of dot product in neural networks. In *International Conference on Artificial Neural Networks*, pages 382–391. Springer, 2018. 3

[26] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 1

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 3, 6

[28] A. Mishra, S. Krishna Reddy, A. Mittal, and H. A. Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2188–2196, 2018. 1

[29] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens,

A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 7, 8

[30] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE, 2012. 6

[31] A. Paul, N. C. Krishnan, and P. Munjal. Semantically aligned bias reducing zero shot learning. *arXiv: Computer Vision and Pattern Recognition*, 2019. 1

[32] M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531, 2010. 2, 5

[33] M. Radovanović, A. Nanopoulos, and M. Ivanović. On the existence of obstinate results in vector space models. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193. ACM, 2010. 2, 6

[34] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58, 2016. 3

[35] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1641–1648, 2011. 6

[36] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015. 1, 3, 7, 8

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6

[38] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019. 1

[39] Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto. Ridge regression, hubness, and zero-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 135–151. Springer, 2015. 1, 2, 3, 5, 6

[40] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013. 1, 3, 7, 8

[41] M. Spruill et al. Asymptotic distribution of coordinates on high dimensional spheres. *Electronic communications in probability*, 12:234–247, 2007. 4

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 3, 6

[43] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. *arXiv preprint arXiv:1801.09414*, 2018. 3

[44] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin. Deep metric learning with angular loss. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2612–2620. IEEE, 2017. 3

[45] W. Wang, Y. Pu, V. K. Verma, K. Fan, Y. Zhang, C. Chen, P. Rai, and L. Carin. Zero-shot learning via class-conditioned deep generative models. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1

[46] W. Wang, V. W. Zheng, H. Yu, and C. Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):13, 2019. 1

[47] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200. 2010. 6

[48] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016. 7, 8

[49] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 3, 6

[50] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018. 1

[51] P. Xie, Y. Deng, and E. Xing. Diversifying restricted boltzmann machine for document modeling. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1315–1324. ACM, 2015. 2

[52] X. Xu, F. Shen, Y. Yang, D. Zhang, H. T. Shen, and J. Song. Matrix tri-factorization with manifold regularizations for zero-shot learning. In *Proceeding of the IEEE conference on computer vision and pattern recognition. CVPR*, 2017. 3

[53] F. S. Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. 2018. 1, 3, 7, 8

[54] Y. Yang and T. M. Hospedales. A unified perspective on multi-domain and multi-task learning. *arXiv preprint arXiv:1412.7489*, 2014. 3

[55] H. Zhang and P. Koniusz. Zero-shot kernel learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7670–7679, 2018. 3, 7, 8

[56] L. Zhang, T. Xiang, S. Gong, et al. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3, 4, 6, 7, 8

[57] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 4166–4174, 2015. 7, 8

[58] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6042, 2016. 3