Leveraging Seen and Unseen Semantic Relationships for Generative Zero-Shot Learning

Maunil R Vyas[0000−0003−3906−1540], Hemanth
Venkateswara[0000−0002−3832−0881], and Sethuraman
Panchanathan[0000−0002−8769−6340]

Arizona State University, Tempe AZ 85281, USA
{mrvyas, hemanthv, panch}@asu.edu

Abstract. Zero-shot learning (ZSL) addresses the unseen class recognition problem by leveraging semantic information to transfer knowledge from seen classes to unseen classes. Generative models synthesize the unseen visual features and convert ZSL into a classical supervised learning problem. These generative models are trained using the seen classes and are expected to implicitly transfer the knowledge from seen to unseen classes. However, their performance is stymied by overfitting, which leads to substandard performance on Generalized Zero-Shot learning (GZSL). To address this concern, we propose the novel LsrGAN, a generative model that Leverages the Semantic Relationship between seen and unseen categories and explicitly performs knowledge transfer by incorporating a novel Semantic Regularized Loss (SR-Loss). The SR-loss guides the LsrGAN to generate visual features that mirror the semantic relationships between seen and unseen classes. Experiments on seven benchmark datasets, including the challenging Wikipedia text-based CUB and NABirds splits, and Attribute-based AWA, CUB, and SUN, demonstrates the superiority of the LsrGAN compared to previous state-of-the-art approaches under both ZSL and GZSL. Code is available at *https://github.com/Maunil/LsrGAN*

Keywords: Generalized zero-shot Learning, Generative Modeling (GANs), Seen and Unseen relationship

1   Introduction

Consider the following discussion between a kindergarten teacher and her student.
Teacher: *Today we will learn about a new animal that roams the grasslands of Africa. It is called the Zebra.*
Student: *What does a Zebra look like ?*
Teacher: *It looks like a short white horse but has black stripes like a tiger.*

That description is nearly enough for the student to recognize a zebra the next time she sees it. The student is able to take the verbal (textual) description and
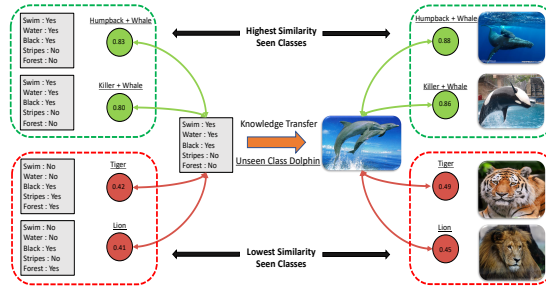
Fig. 1: Driving motivation behind leveraging the semantic relationship between seen and unseen classes to infer the visual characteristics of unseen classes. Notice that though the feature representations are different, the class similarity values are almost the same. e.g. "Dolphin" has almost identical similarity values in visual and semantic space with other seen classes. The similarity values are mentioned in the circles, and computed using the cosine distance.

relate it to the visual understanding of a horse and a tiger and generate a zebra in her mind. In this paper, we propose a zero-shot learning model that transfers knowledge from text to the visual domain to learn and recognize previously unseen image categories.

Collecting and curating large labeled datasets for training deep neural networks is both labor-intensive and nearly impossible for many of the classification tasks, especially for the fine-grained categories in specific domains. Hence, it is desirable to create models that can mitigate these difficulties and learn not to rely on large labeled training sets. Inspired by the human ability to recognize object categories solely based on class descriptions and previous visual knowledge, the research community has extensively pursued the area of "Zero-shot learning" (ZSL) [11, 22, 23, 29, 37, 38]. Zero-shot learning aims to recognize objects that are not seen during the training process of the model. It leverages textual descriptions/attributes to transfer knowledge from seen to unseen classes.

Generative models are the most popular approach to solve zero-shot learning. Despite the recent progress, generative models for zero-shot learning still have some key limitations. These models show a large quality gap between the synthesized and the actual unseen image features [15, 24, 31, 35, 43]. As a result, the performance of generalized zero-shot learning (GZSL) suffers, since many synthesized features of unseen classes are classified as seen classes. The second major concern behind the current generative model based approaches is the assumption that the semantic features are available in the desired form for a class category, e.g., clean attributes. However, in reality, they are hard to get. Getting clean semantic features requires a domain expert to annotate the attributes manually. Moreover, collecting a large number of attributes for all the class categories is again labor-intensive and expensive. Considering this, our generative model learns to transfer semantic knowledge from both noisy text descriptions

(like Wikipedia articles) as well as semantic attributes for zero-shot learning and generalized zero-shot learning.

In this paper we propose a novel generative model called the LsrGAN. The LsrGAN leverages semantic relationships between seen and unseen categories and transfers the same to the generated image features. We implement this transfer through a unique semantic regularization framework called the Semantic Regularized Loss (SR-Loss). In Fig. 1, "Dolphin", an unseen class, has a high semantic similarity with classes such as "Killer whale" and "Humpback whale" from the seen class set. These two seen classes are the potential neighbors of the Dolphin in the visual space. Therefore, when we do not have the real visual features for the Dolphin class, we can use these neighbors to form indirect visual references for the Dolphin class. This illustrates the intuition behind the proposed SR-loss. The LsrGAN also trains a classifier to guide the feature generation. The main contributions of our work are:

1. A generative model leveraging the semantic relationship (LsrGAN) between seen and unseen classes overcoming the overfitting concern towards seen classes.
2. A novel semantic regularization framework that enables knowledge transfer across the visual and semantic domain. The framework can be easily integrated into any generalized zero-shot learning model.
3. We have conducted extensive experiments on seven widely used standard benchmark datasets to demonstrate that our model outperforms the previous state-of-the-art approaches.

2   Related work

Earlier work on ZSL was focused on learning the mapping between visual and semantic space in order to compensate for the lack of visual representation of the unseen classes. These approaches are known as *Embedding methods*. The initial work was focused on two-stage approaches where the attributes of an input image are estimated in the first stage, and the category label is determined in the second phase using highest attribute similarity. DAP and IAP [22] are examples of such an approach. Later, the use of bi-linear compatibility function led to promising ZSL results. Among these, ALE [1] and DEVISE [16] use ranking loss to learn the bilinear compatibility function. ESZSL [30] applies the square loss and explicitly regularizes the objective w.r.t. the Frobenius norm. Unlike standard embedding methods, [8,40] propose reverse mapping from the semantic to the visual space and perform nearest neighbor classification in the visual space. The hybrid models such as CONSE [26], SSE [41], and SYNC [7], discuss the idea of embedding both visual and semantic features into another intermediate space. These methods perform well in the ZSL setting. However, in GZSL, they show a high bias towards seen classes.

Most of the mentioned embedding methods have a bias towards seen classes leading to substandard GZSL performance. Recently, *Generative methods* [10, 15, 18, 24, 31, 33, 35, 43] have attempted to address this concern by synthesizing
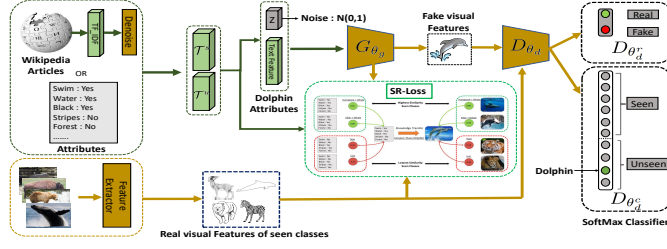
Fig. 2: Conceptual illustration of our LsrGAN model. The basis of our model is a conditional WGAN. The novel SR-Loss is introduced to help the G_{θ_g} to understand the semantic relationship between classes and guide it for applying the same during visual feature generation. The G_{θ_g} will use $\mathcal{T}^s$ and $\mathcal{T}^u$ to generate visual features. The D_{θ_d} has two branches used to perform real/fake game and classification. Notice that when we train the G_{θ_g} using $\mathcal{T}^u$, only the classification branch remains active in D_{θ_d} as the unseen visual features are not available.

unseen class features, leading to state of the art ZSL and GZSL performance. Among these, [10, 15, 24, 35, 43] are based on Generative Adversarial Networks (GAN), while [5, 31] use Variational Autoencoders (VAE) for the ZSL. F-GAN [35] uses a Wasserstein GAN [4, 17] to synthesize samples based on the class conditioned attribute. LisGAN [24] proposes a *soul-sample* regularizer to guide the sample generation process to stay close to a generic example of the seen class. Inspired by the cycle consistency loss [42], Felix et al. [15], propose a generative model that maps the generated visual features back to the original semantic feature, thus addressing the unpaired training issue during generation. These generative methods use the annotated attribute as semantic information for the feature generation. However, in reality, such a desired form of the semantic feature representation is hard to obtain. Hence, [13] suggests the use of Wikipedia descriptions for ZSL and GZSL. GAZSL [43] proposes a very first generative model that handles the Wikipedia description to generate features.

Although generative methods have been quite successful in ZSL, the unseen feature generation is biased towards seen classes leading to poor generalization in GZSL. For better generalization, we have proposed a novel SR-loss to leverage the inter-class relationships between seen and unseen classes in GANs. The idea of utilizing inter-class relationships has been addressed before with Triplet loss-based approaches [3, 6] and using contrastive networks [20]. In this work, we propose a related approach using generative models. To best of our knowledge, such an approach using GANs has not been investigated.

3   Proposed Approach

3.1   Problem Settings

The zero-shot learning problem consists of seen (observed) and unseen (unobserved) categories of images and their corresponding text information. Images

belonging to the seen categories are passed through a feature extractor (for e.g., ResNet-101) to yield features $\{\boldsymbol{x}_i^s\}_{i=1}^{n_s}$, where $\boldsymbol{x} \in \mathcal{X}^s$. The corresponding labels for these features are $\{y_i^s\}_{i=1}^{n_s}$, where $y^s \in \mathcal{Y}^s = \{1, \ldots, C_s\}$ with C_s seen categories. The image features for the unseen categories are denoted as $\{\boldsymbol{x}_i^u\}_{i=1}^{n_u}$, where $\boldsymbol{x} \in \mathcal{X}^u$, and the space of all image features is $\mathcal{X} := \mathcal{X}^s \cup \mathcal{X}^u$. As the name indicates, unseen categories are not observed, and the zero-shot learning model attempts to hallucinate these features with the rest of the information provided. Although we do not have the image features for the unseen categories, we are privy to the C_u unseen categories, where the corresponding labels for the unseen image features would be $\{y_i^u\}_{i=1}^{n_u}$, with $y^u \in \mathcal{Y}^u = \{C_s + 1, \ldots, C\}$, where $C = C_s + C_u$. From the text domain, we have the semantic features for all categories which are either binary attribute vectors or Term-Frequency-Inverse-Document-Frequency (TF-IDF) vectors. The category-wise semantic features are denoted as $\{\boldsymbol{t}_c^s\}_{c=1}^{C_s}$, for seen categories and $\{\boldsymbol{t}_c^u\}_{c=C_s+1}^{C}$ with $\boldsymbol{t} \in \mathcal{T} := \mathcal{T}^s \cup \mathcal{T}^u$. The goal of zero-shot learning is to build a classifier $\mathcal{F}_{zsl} : \mathcal{X}^u \rightarrow \mathcal{Y}^u$, mapping image features to unseen categories, and the goal of the more difficult problem of generalized zero-shot learning is to build a classifier $\mathcal{F}_{gzsl} : \mathcal{X} \rightarrow \mathcal{Y} := \mathcal{Y}^s \cup \mathcal{Y}^u$, mapping image features to seen and unseen categories.

3.2 Adversarial Image Feature Generation

Using the image features of the seen categories and the semantic features of the seen and unseen categories, we propose a generative adversarial network to hallucinate the unseen image features for each of the unseen categories. We apply a conditional Wasserstein Generative Adversarial Network (WGAN) to generate image features for the unseen categories using semantic features as input [4]. The WGAN aligns the real and generated image feature distributions. In addition, we have a feature classifier that is trained to classify image features into C categories of seen and unseen classes. The components of the WGAN are described in the following.

Feature Generator: The conditional generator in the WGAN has parameters θ_g and is represented as $G_{\theta_g} : \mathcal{Z} \times \mathcal{T} \rightarrow \mathcal{X}$, where $\mathcal{Z}$ is the space of random normal vectors $\mathcal{N}(0, I)$ of $|\mathcal{Z}|$ dimensions. Since the TF-IDF features from the Wikipedia articles may contain repetitive and non-discriminating feature information, we apply a denoising transformation upon the TF-IDF vector using a fully-connected neural network layer as proposed by [43]. The WGAN takes as input a random noise vector $\boldsymbol{z} \in \mathcal{Z}$ concatenated with the semantic feature vector $\boldsymbol{t}_c$ for a category c, and generates an image feature $\tilde{\boldsymbol{x}}_c \leftarrow G_{\theta_g}(\boldsymbol{z}, \boldsymbol{t}_c)$. The generator is trained to generate image features for both seen categories $(\tilde{\boldsymbol{x}}_c^s \leftarrow G_{\theta_g}(\boldsymbol{z}, \boldsymbol{t}_c^s))$ and unseen categories $(\tilde{\boldsymbol{x}}_c^u \leftarrow G_{\theta_g}(\boldsymbol{z}, \boldsymbol{t}_c^u))$. In order to generate image features that are structurally similar to the real image features, we implement visual pivot regularization, $\mathcal{L}_{vp}$ that aligns the cluster centers of the real image features with the cluster centers of the generated image features for each of the C_s categories [43]. This is implemented only for the seen categories

where we have real image features.

$$\mathcal{L}_{vp} = \min_{\theta_g} \frac{1}{C_s} \sum_{c=1}^{C_s} \big|\big|\mathbb{E}_{(\boldsymbol{x},y=c)\sim(\mathcal{X}^s,\mathcal{Y}^s)}[\boldsymbol{x}] - \mathbb{E}_{(\boldsymbol{z},\boldsymbol{t}_c^s)\sim(\mathcal{Z},\mathcal{T}^s)}[G_{\theta_g}(\boldsymbol{z},\boldsymbol{t}_c^s)]\big|\big|. \qquad (1)$$

Feature Discriminator: We train the WGAN with an adversarial discriminator having two branches to perform the real/fake game and classification. The discriminator has parameters θ_d^r and θ_d^c for two branches respectively and is denoted as D_{θ_d}. The real/fake branch of the discriminator learns a mapping $D_{\theta_d^r} : \mathcal{X} \to \mathbb{R}$ using the generated and real image features to output $D_{\theta_d^r}(\tilde{\boldsymbol{x}}^s)$ and $D_{\theta_d^r}(\boldsymbol{x}^s)$ that are used to estimate the objective term $\mathcal{L}_d$. The objective $\mathcal{L}_d$ is maximized w.r.t. the discriminator parameters θ_d^r and minimized w.r.t. the generator parameters θ_g.

$$\mathcal{L}_d = \min_{\theta_g} \max_{\theta_d} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}^s}\big[D_{\theta_d}(\boldsymbol{x})\big] - \mathbb{E}_{(\boldsymbol{z},\boldsymbol{t})\sim(\mathcal{Z},\mathcal{T}^s)}\big[D_{\theta_d}(G_{\theta_g}(\boldsymbol{z},\boldsymbol{t}))\big] - \lambda_{gp}\mathcal{L}_{gp} \quad (2)$$

where, the first two terms control the alignment of real image feature and the generated image feature distributions. The third term is the gradient penalty to enforce the Lipschitz constraint with $\mathcal{L}_{gp} = (||\nabla_{\boldsymbol{x}}D_{\theta_d}(\boldsymbol{x})||_2 - 1)^2$ where, input $\boldsymbol{x}$ are real image features, generated image features and random samples on a straight line connecting real image features and generated image features [17]. The parameter λ_{gp} controls the importance of the Lipschitz constraint. The discriminator parameters are trained using only seen category image features since $\boldsymbol{x}^u$ are unavailable.

Feature Classifier: The category classifier has parameters θ_d^c and is denoted as $D_{\theta_d^c}$. It is a softmax cross-entropy classifier for the generated and real image features $\boldsymbol{x}^s, \tilde{\boldsymbol{x}}^s, \tilde{\boldsymbol{x}}^u$ and is trained to minimize the loss $\mathcal{L}_c$. For ease of notation we represent the real/generated image features as $\boldsymbol{x}$ and corresponding labels y

$$\mathcal{L}_c = \min_{\theta_g,\theta_d^c} -\mathbb{E}_{(\boldsymbol{x},y)\sim(\mathcal{X},\mathcal{Y})}\bigg[\sum_{c=1}^{C} 1(y=c)\log(D_{\theta_d^c}(\boldsymbol{x}))_c\bigg], \qquad (3)$$

where $(D_{\theta_d^c}(\boldsymbol{x}))_c$ is the c-th component of the C-dimension softmax output and $1(y=c)$ is the indicator function. While the discriminator performs a marginal alignment of real and generated features, the classifier performs category based conditional alignment.

So far, in the proposed WGAN model the *Generator* generates image features from seen and unseen categories. The *Discriminator* aligns the image feature distributions and the *Classifier* performs image classification. The LsrGAN model is illustrated in Fig. 2. In the following we introduce a novel regularization technique that transfers knowledge across the semantic and image feature spaces for the unseen and seen categories respectively.

3.3   Semantic Relationship Regularization

Conventional generative zero-shot learning approaches [24,35,43], have poor generalization performance in GZSL since the generated visual features are biased towards the seen classes. We address this issue by proposing a novel regularization procedure that will explicitly transfer knowledge of unseen classes from the semantic domain to guide the model in generating distinct seen and unseen image features. We term this the "Semantic Regularized Loss (SR-Loss)".

   We argue that the visual and semantic feature spaces share a common underlying latent space that generates the visual and semantic features. We propose to exploit this relationship by transferring knowledge from the semantic space to the visual space to generate image features. Knowing the inter-class relationships in the semantic space can help us impose the same relationship constraints among the generated visual features. This is the idea behind the SR-Loss in the WGAN where we transfer inter-class relationships from the semantic domain to the visual domain. Fig. 1 illustrates this concept. The visual similarity between class c_i and c_j is represented as $\mathcal{X}_{sim}(\boldsymbol{\mu}_{c_i}, \boldsymbol{\mu}_{c_j})$, where $\boldsymbol{\mu}_c$ is the mean of the image features of class c. Note that for visual similarity we are considering the relationship between the class centers and not between individual image features. Likewise, the semantic similarity between class c_i and c_j is represented as $\mathcal{T}_{sim}(\boldsymbol{t}_{c_i}, \boldsymbol{t}_{c_j})$. For semantic similarity, we do not have the mean, since we only have one semantic vector $\boldsymbol{t}_c$ for every category, although the proposed approach can be extended to include multiple semantic feature vectors. We impose the following semantic relationship constraint for the image features,

$$\mathcal{T}_{sim}(\boldsymbol{t}_{c_i}, \boldsymbol{t}_{c_j}) - \epsilon_{ij} \ \leq \ \mathcal{X}_{sim}(\boldsymbol{\mu}_{c_i}, \boldsymbol{\mu}_{c_j}) \ \leq \ \mathcal{T}_{sim}(\boldsymbol{t}_{c_i}, \boldsymbol{t}_{c_j}) + \epsilon_{ij}, \qquad (4)$$

where, hyper-parameter $\epsilon_{ij} \geq 0$ is a soft margin enforcing the similarity between semantic and image features for classes i and j. Large values of ϵ_{ij} allow more deviation between semantic similarities and visual similarities. We incorporate the constraint into the objective by applying the penalty method [25],

$$p_{c_{ij}} \big[|| \max(0, \mathcal{X}_{sim}(\boldsymbol{\mu}_{c_i}, \boldsymbol{\mu}_{c_j}) - (\mathcal{T}_{sim}(\boldsymbol{t}_{c_i}, \boldsymbol{t}_{c_j}) + \epsilon_{ij})) ||^2$$
$$+ || \max(0, (\mathcal{T}_{sim}(\boldsymbol{t}_{c_i}, \boldsymbol{t}_{c_j}) - \epsilon_{ij}) - \mathcal{X}_{sim}(\boldsymbol{\mu}_{c_i}, \boldsymbol{\mu}_{c_j})) ||^2 \big], \qquad (5)$$

where, $p_{c_{ij}}$ is the penalty for violating the constraint. The penalty becomes zero when the constraints are satisfied and is non-zero otherwise. We use cosine distance to compute the similarities.

   We intend to transfer semantic inter-class relationships to enhance the image feature representations that are output from the generator. Consider a seen class c_i. We estimate its semantic similarity $\mathcal{T}_{sim}(\boldsymbol{t}_{c_i}, \boldsymbol{t}_{c_j})$ with all the other seen semantic features $\boldsymbol{t}_{c_j}$ where $j \in \{1, \ldots, C_s\} \wedge j \neq i$. Not all the similarities are important and for the ease of implementation we select the highest n_c similarities that we wish to transfer. Let I_{c_i} represent the set of n_c seen categories with the highest semantic similarity with c_i. We apply Eq 5 to train the generator to output image features that satisfy the semantic similarity constraints against

the top n_c similarities from the seen categories. For the seen image categories, the objective function is,

$$\mathcal{L}_{sr}^s = \min_{\theta_g} \frac{1}{C_s} \sum_{i=1}^{C_s} \sum_{j \in I_{c_i}} \big[\| \max(0, \mathcal{X}_{sim}(\boldsymbol{\mu}_{c_j}^s, \tilde{\boldsymbol{\mu}}_{c_i}^s) - (\mathcal{T}_{sim}(\boldsymbol{t}_{c_j}^s, \boldsymbol{t}_{c_i}^s) + \epsilon)) \|^2$$
$$+ \| \max(0, (\mathcal{T}_{sim}(\boldsymbol{t}_{c_j}^s, \boldsymbol{t}_{c_i}^s) - \epsilon) - \mathcal{X}_{sim}(\boldsymbol{\mu}_{c_j}^s, \tilde{\boldsymbol{\mu}}_{c_i}^s)) \|^2 \big], \quad (6)$$

where, penalty $p_{ij} = 1$, and $\boldsymbol{\mu}_{c_j}^s := \mathbb{E}_{(\boldsymbol{x}, y=c_j) \sim (\mathcal{X}^s, \mathcal{Y}^s)}[\boldsymbol{x}]$ is the mean of the image features for seen class c_j, and $\tilde{\boldsymbol{\mu}}_{c_i}^s := \mathbb{E}_{\boldsymbol{z} \sim \mathcal{Z}}[G_{\theta_g}(\boldsymbol{z}, \boldsymbol{t}_{c_i}^s)]$ is the mean of the generated image features of seen class c_i. We use a constant value of ϵ as the soft margin to simplify the solution. Similarly, the objective function for the unseen categories is,

$$\mathcal{L}_{sr}^u = \min_{\theta_g} \frac{1}{C_u} \sum_{i=C_s+1}^{C} \sum_{j \in I_{c_i}} \big[\| \max(0, \mathcal{X}_{sim}(\boldsymbol{\mu}_{c_j}^s, \tilde{\boldsymbol{\mu}}_{c_i}^u) - (\mathcal{T}_{sim}(\boldsymbol{t}_{c_j}^s, \boldsymbol{t}_{c_i}^u) + \epsilon_{ij})) \|^2$$
$$+ \| \max(0, (\mathcal{T}_{sim}(\boldsymbol{t}_{c_j}^s, \boldsymbol{t}_{c_i}^u) - \epsilon_{ij}) - \mathcal{X}_{sim}(\boldsymbol{\mu}_{c_j}^s, \tilde{\boldsymbol{\mu}}_{c_i}^u)) \|^2 \big], \quad (7)$$

where, $\tilde{\boldsymbol{\mu}}_{c_i}^u := \mathbb{E}_{\boldsymbol{z} \sim \mathcal{Z}}[G_{\theta_g}(\boldsymbol{z}, \boldsymbol{t}_{c_i}^u)]$ is the mean of the generated image features of unseen class c_i. The time complexity of the proposed SR-loss is $\mathcal{O}(Bn_c|\mathcal{X}|E) + \mathcal{O}(C^2|\mathcal{T}|E) + \mathcal{O}(C^2 \log n_c E)$ where B, n_c, $|\mathcal{X}|$, E, C and $|\mathcal{T}|$ denote the batch size, neighbor size, number of visual features, epoch, number of total classes and the size of the semantic features respectively. This shows the computation cost is manageable.

3.4   LsrGAN Objective Function

The LsrGAN leverages the semantic relationship between seen and unseen categories to generate robust image features for unseen categories using the objective function defined in Eq. 6 and 7. The model generates robust seen image features conditioned by the regularizer in Eq. 1. The LsrGAN trains a classifier over all the C categories as outlined in Eq. 3. The LsrGAN is based on a WGAN model that aligns the image feature distributions using the objective function defined in Eq. 2. The overall objective function of the LsrGAN model is given by,

$$\lambda_c \mathcal{L}_c + \mathcal{L}_d + \lambda_{vp} \mathcal{L}_{vp} + \lambda_{sr}(\mathcal{L}_{sr}^s + \mathcal{L}_{sr}^u) \quad (8)$$

where, λ_c, λ_{vp} and λ_{sr} are hyper parameters controlling the importance of each of the loss terms. Unlike standard zero-shot learning models that generate image features and then have to train a supervised classifier [24, 35, 43], the LsrGAN model has an inbuilt classifier that can also be used for evaluating zero-shot learning and generalized zero-shot learning.

4   Experiments

4.1   Datasets

Attribute-based datasets : We conduct experiments on three widely used attribute-based datasets : *Animal with Attributes* (AWA) [22], *Caltech-UCSD-*

Table 1: Dataset Information. For the attribute-based datasets, the (number) in seen classes denotes the number of classes used for test in GZSL.

	attribute-based			Wikipedia descriptions			
	AWA	CUB	SUN	CUB (Easy)	CUB (Hard)	NAB (Easy)	NAB (Hard)
No. of Samples	30,475	11,788	14,340	11,788	11,788	48,562	48,562
No. of Features	85	312	102	7551	7551	13217	13217
No. of Seen classes	40(13)	150(50)	645(65)	150	160	323	323
No. of Unseen classes	10	50	72	50	40	81	81

Birds 200-2011 (CUB) [34] and Scene UNderstanding (SUN) [27]. AWA is a medium scale coarse-grained animal dataset having 50 animal classes with 85 attributes annotated. CUB is a fine-grained, medium-scale dataset having 200 bird classes annotated with 312 attributes. SUN is a medium scale dataset having 717 types of scenes with 102 annotated attributes. We followed the split mentioned in [36] to have a fair comparison with existing approaches.

Wikipedia descriptions-based datasets: In order to address a more challenging ZSL problem with Wikipedia descriptions as auxiliary information, we have used two common fine-grained datasets with textual descriptions: CUB and *North America Birds* (NAB) [32]. The NAB dataset is larger compared to CUB having 1011 classes in total. We have used two splits, suggested by [14] in our experiments to have a fair comparison with other methods. The splits are termed as *Super-Category-Shared* (SCS, Easy split) and *Super-Category-Exclusive* (SCE, Hard split). These splits represent the similarity between seen and unseen classes. The SCS-split has at least one seen class for every unseen class belonging to the same parent. For example, "Harris's Hawk" in the unseen set and "Cooper's Hawk" in the seen set belong to the same parent category, "Hawks." On the other hand, in the SCE-split, the parent categories are disjoint for the seen and unseen classes. Therefore, SCS and SCE splits are considered as Easy and Hard splits. The details for each dataset and class splits are given in Table 1.

4.2 Implementation Details and Performance Metrics

The 2048-dimensional ResNet-101 [19] features are considered as a real visual feature for attribute-based datasets, and part-based features (e.g., belly, leg, wing, etc.) from VPDE-net [39] are used for the Wikipedia based datasets, as suggested by [35, 43]. We have utilized the TF-IDF to extract the features from the Wikipedia descriptions. For a fair comparison, all of our experiment settings are kept the same as reported in [35, 36, 43].

The base block of our model is GAN, which is implemented using a multilayer perceptron. Specifically, the feature generator G_{θ_g} has one hidden unit having 4096 neurons and LeakyReLU as an activation function. For attribute-based datasets, we intend to get the top max-pooling units of ResNet - 101 (visual features). Hence, the output layer has ReLU activation in the feature generator. On the other hand, for the Wikipedia based datasets, we have used Tanh as an output activation for the feature generator since the VPDE-net feature varies from -1 to 1. $\mathcal{Z}$ is sampled from the normal Gaussian distribution.

Table 2: ZSL and GZSL results on AWA, CUB, and SUN with attributes as semantic information. T1 indicates the Top-1 % accuracy in the ZSL setting. On the other hand, "U", "S" and "H" denotes the Top-1% accuracy for the unseen, seen, and Harmonic mean (seen + unseen).

Methods	zero-shot Learning			Generalized zero-shot Learning								
	AWA	CUB	SUN	AwA			CUB			SUN		
	T1	T1	T1	U	S	H	U	S	H	U	S	H
DAP [22]	44.1	40.0	39.9	0.0	88.7	0.0	1.7	67.9	3.3	4.2	25.2	7.2
CONSE [26]	45.6	34.3	38.8	0.4	88.6	0.8	1.6	72.2	3.1	6.8	39.9	11.6
SSE [41]	60.1	43.9	51.5	7.0	80.5	12.9	8.5	46.9	14.4	2.1	36.4	4.0
DeViSE [16]	54.2	50.0	56.5	13.4	68.7	22.4	23.8	53.0	32.8	16.9	27.4	20.9
SJE [2]	65.6	53.9	53.7	11.3	74.6	19.6	23.5	59.2	33.6	14.7	30.5	19.8
ESZSL [30]	58.2	53.9	54.5	5.9	77.8	11.0	2.4	70.1	4.6	11.0	27.9	15.8
ALE [1]	59.9	54.9	58.1	14.0	81.8	23.9	4.6	73.7	8.7	21.8	33.1	26.3
SYNC [7]	54.0	55.6	56.3	10.0	**90.5**	18.0	7.4	66.3	13.3	7.9	**43.3**	13.4
SAE [21]	53.0	33.3	40.3	1.1	82.2	2.2	0.4	**80.9**	0.9	8.8	18.0	11.8
DEM [40]	68.4	51.7	61.9	30.5	86.4	45.1	11.1	75.1	19.4	20.5	34.3	25.6
PSR [3]	63.8	56.0	61.4	20.7	73.8	32.3	24.6	54.3	33.9	20.8	37.2	26.7
TCN [20]	70.3	59.5	61.5	49.4	76.5	60.0	52.6	52.0	52.3	31.2	37.3	34.0
GAZSL [43]	68.2	55.8	61.3	19.2	86.5	31.4	23.9	60.6	34.3	21.7	34.5	26.7
F-GAN [35]	68.2	57.3	60.8	**57.9**	61.4	59.6	43.7	57.7	49.7	42.6	36.6	39.4
cycle-CLSWGAN [15]	66.3	58.4	60.0	56.9	64.0	60.2	45.7	61.0	52.3	**49.4**	33.6	40.0
LisGAN [24]	**70.6**	58.8	61.7	52.6	76.3	62.3	46.5	57.9	51.6	42.9	37.8	40.2
LsrGAN [ours]	66.4	**60.3**	**62.5**	54.6	74.6	**63.0**	**48.1**	59.1	**53.0**	44.8	37.7	**40.9**

To perform the denoising and dimensionality reduction from Wikipedia descriptions, we have employed a fully connected layer with a feature generator. Also, notice that the semantic similarity for the SR-Loss is computed using the denoiser's output in Wikipedia based datasets. We will discard this layer when dealing with attribute-based datasets. In our model, the discriminator D_{θ_d} has two branches. One is used to play the real/fake game, and the other performs the classification on the generated/real visual feature. The discriminator also has 4096 units in the hidden layer with ReLU as an activation. Since the cosine distance is less prone to the curse of dimensionality when the features are sparse (semantic features), we have considered it in the SR-loss.

To perform the zero-shot recognition we have used nearest neighbor prediction on datasets having Wikipedia descriptions, and the classifier attached to the discriminator for the attributes based recognition. Notice that the classifier is not re-trained for the recognition part. The Top-1 accuracy is used to assess the ZSL setting. Furthermore, to capture the more realistic scenario, we have examine the Generalized zero-shot recognition as well. As suggested by [9], we report the area under the seen and unseen curve (AUC score) as GZSL performance metric for Wikipedia based datasets, and the harmonic mean of the seen and unseen Top-1 accuracies for attribute based dataset. Notice that the choice of these measures and predictions models is to make a fair comparison with existing methods.

4.3   ZSL and GZSL Performance

The results for the ZSL are provided in the left part of Table 2 and 3. It can be seen that our LsrGAN achieves superior performance in both attribute and Wikipedia based datasets compared to the previous state of the art models, especially with generative models GAZSL, F-GAN, cycle-CLSWGAN, and LisGAN. It is worth noticing that all the mentioned generative models have the same base

Table 3: ZSL and GZSL results on CUB and NAB datasets with Wikipedia descriptions as semantic information on the two-split setting. We have used Top-1 % accuracy and Seen-Unseen AUC (%) for ZSL and GZSL, respectively.

Methods	zero-shot Learning				Generalized zero-shot Learning			
	CUB		NAB		CUB		NAB	
	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
WAC-Linear [13]	27.0	5.0	-	-	23.9	4.9	23.5	-
WAC-Kernal [12]	33.5	7.7	11.4	6.0	14.7	4.4	9.3	2.3
ESZSL [30]	28.5	7.4	24.3	6.3	18.5	4.5	9.2	2.9
ZSLNS [28]	29.1	7.3	24.5	6.8	14.7	4.4	9.3	2.3
Sync-fast [7]	28.0	8.6	18.4	3.8	13.1	4.0	2.7	3.5
ZSLPP [14]	37.2	9.7	30.3	8.1	30.4	6.1	12.6	3.5
GAZSL [43]	43.7	10.3	35.6	8.6	35.4	8.7	20.4	5.8
LsrGAN [ours]	**45.2**	**14.2**	**36.4**	**9.04**	**39.5**	**12.1**	**23.2**	**6.4**

architecture. e.g., WGAN. Hence, the superiority of our model suggests that our motivation is realistic, and our experiments are effective. In summary, we achieve, 1.5 %, 3.9 %, 0.8 %, 0.44% improvement on CUB (Easy), CUB (Hard), NAB (Easy) and NAB (Hard) respectively for the Wikipedia based datasets under ZSL. On the other hand, for the attribute-based ZSL, we attain 1.5% and 0.8% improvement on CUB and SUN, respectively. ZSL under performs on AWA probably due to the high feature correlation between similar unseen classes having a common neighbor among the seen classes, e.g. Dolphin and Blue whale. The availability of similar unseen classes slightly affects the prediction capability of the classifier for ZSL as the SR-loss tends to cluster them together due to high semantic similarity with common seen classes. However, it is worth noticing that the GZSL result for the same dataset is superior.

Our primary focus is to elevate the GZSL performance in this work, which is apparent from the right side of Table 2 and 3. Following [35, 36, 43], we report the harmonic mean and AUC score for the attribute and Wikipedia based datasets, respectively. The mentioned metrics help us to showcase the approach's generalizability as the harmonic mean and AUC scores are only high when the performance on seen and unseen classes is high. From the results, we can see that LsrGAN outperforms the previous state of the art for the GZSL. In terms of numbers, we achieve, 4.1%, 3.4% 2.8% and 0.6% gain on CUB (Easy), CUB (Hard), NAB (Easy) and NAB (Hard) respectively for the Wikipedia based datasets and 0.7 %, 1.4% and 0.7 % improvement on attribute-based AWA, CUB and SUN respectively. It is worth noticing that the LsrGAN improves the unseen Top-1 performance in the GZSL setting for the attribute-based CUB and SUN by 1.6% and 1.9% with the previous state of the art LisGAN [24].

The majority of the conventional approaches, including the generative models overfit the seen classes, which results in lower GZSL performance. Notice that most of the approaches mentioned in Table 2 achieve very high performance on seen classes compared to the unseen classes in the GZSL setting. For example, SYNC [7] has around 90% recognition capability on the seen classes, and it drops to only 10% (80% difference) for the unseen classes on AWA dataset. It is also evident from Tables 2 and 3 that the performance on the unseen categories drops drastically when the search space includes both seen and unseen classes in the

Fig. 3: Average class confidence score (Average SoftMax Probability) comparison for classifier trained with F-GAN and LsrGAN. Top 3 average guesses are mentioned here. The red marked label showcase the top 1 average guess. The class names with underline represent seen classes.

GZSL setting. For instance, DAP [22] drops from 40 % to 1.7 %, GAZSL [43] drops from 55.8% to 23.9% and F-GAN [35] drops from 57.3% to 43.7 % on attribute-based CUB dataset. This indicates that the previous approaches are easy to overfit the seen classes. Although the generative models have achieved significant progress compared to the previous embedding methods, they still possess the overfitting issue towards seen classes by having substandard GZSL performance. On the contrary, our model incorporates the novel SR-Loss that enables the explicit knowledge transfer from the similar seen classes to the unseen classes. Therefore, the proposed LsrGAN alleviates the overfitting concern and helps us to achieve a state of the art GZSL performance. It is worth noticing that our model not only outperforms the generative zero-shot models having single GAN but also proves its worth against cycle-GAN [15] in ZSL and GZSL setting. Lastly, to have fair comparisons, we have taken the performance numbers from [35, 36, 43].

4.4   Effectiveness of SR-Loss

Utilizing the semantic relationship between seen and unseen classes to infer the visual characteristics of an unseen class is at the heart of the proposed SR-Loss. Contrary to other generative approaches, it enables explicit knowledge transfer in the generative model to make it learn from the unseen classes together with seen classes during the training process itself. As a result, our LsrGAN will become more robust towards the unseen classes leading to address the seen class overfitting concern. To demonstrate such ability, we have computed the average class confidence score (Average Softmax probabilities) of the classifier trained with the generated features from the LsrGAN or F-GAN model. Since the confidence scores are computed under the GZSL, the classifier's training set contains real visual features from the seen classes and generated unseen features from LsrGAN or F-GAN. To have a fair comparison, we have used the same F-GAN's Softmax classifier in LsrGAN. Since LsrGAN learns from the seen and unseen classes during the training process itself, the Softmax classifier associated with it is not trained in an offline fashion like in the F-GAN model.

Fig. 3 depicts the confidence results on the AWA dataset under the GZSL setting. We have taken mainly four confusing seen and unseen classes for the

Table 4: Comparison of average class confidence score across all seen or unseen classes (SoftMax Probability) between F-GAN and LsrGAN for attribute-based AWA, CUB and SUN

	F-GAN [35]		LrsGAN [ours]	
	Unseen	Seen	Unseen	Seen
AWA	0.29	0.86	0.69	0.83
CUB	0.33	0.65	0.60	0.64
SUN	0.32	0.35	0.65	0.39

comparison with classifier's top 3 guesses. It is evident from the figure that the classifier trained with the F-GAN has lower confidence for the unseen classes, and it mainly showcases very high confidence towards the similar seen classes even if the test image comes from the unseen classes. Also, during the seen class classification, the classifier's confidence values are mainly distributed among the seen classes in F-GAN. For instance, "mouse" has its confidence spread between mouse and hamster only. On the other hand, LsrGAN showcases decent confidence values for the seen and unseen, both leading to better GZSL performance. It is worth noticing that the LsrGAN fails for the "mouse" classification. However, the confidence is well spread among all the three categories showing it has considered an unseen class "rat" with other seen classes "mouse" and "Hamster". These observations reflect the fact that F-GAN has an overfitting issue towards the seen classes. On the other hand, the balanced performance of LsrGAN manifests that explicit knowledge transfer from SR-loss helps it to overcome the overfitting concern towards the seen classes. To bolster our claim, we have also computed the average class confidence across all the seen and unseen classes for these two models on attribute-based AWA, CUB, and SUN datasets. Table 4 reports average confidence values for seen and unseen classes. Clearly, it shows the superiority of LsrGAN in terms of generalizability compared to F-GAN.

4.5   Model Analysis

Parameter Sensitivity : We have tuned our parameters following the conventional grid search approach. We mainly tune the SR-Loss parameters ϵ, λ_{sr} and n_c. For the fair comparison, we have adopted other parameters λ_{vp}, λ_{gp} from [35, 43], also the λ_c is considered between $(0, 1)$, specifically, 0.01 for the majority of our experiments. Fig. 4(b)-(d) show the parameter sensitivity for the SR-Loss. Notice that we estimate the ϵ value from the seen class visual and semantic relations. It can be seen that a lower and higher value of ϵ affects the performance. On the other hand, λ_{sr} and n_c maintain consistent performance after reaching a certain threshold value.

Training Stability : Since GANs are notoriously hard to train, and our proposed LsrGAN model not only uses GAN but also optimizes the similarity constraints from the SR-loss. Therefore, we also report the training stability for ZSL and GZSL for attribute and Wikipedia based datasets. Specifically, for the ZSL, we report the unseen Top-1 accuracy and Epoch behavior in Fig. 5(a) and Fig. 5(c). We have used harmonic mean of the seen and unseen Top-1 accuracy
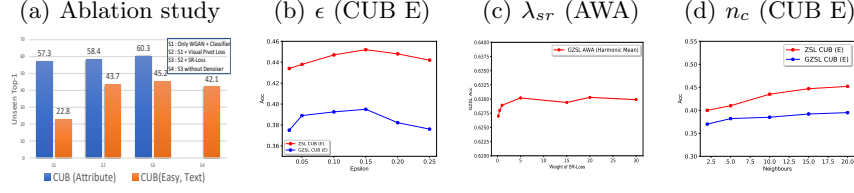
(a) Ablation study   (b) ϵ (CUB E)   (c) λ_{sr} (AWA)   (d) n_c (CUB E)

Fig. 4: Ablation study under ZSL(a), Parameter sensitivity for the SR-loss (b-d).

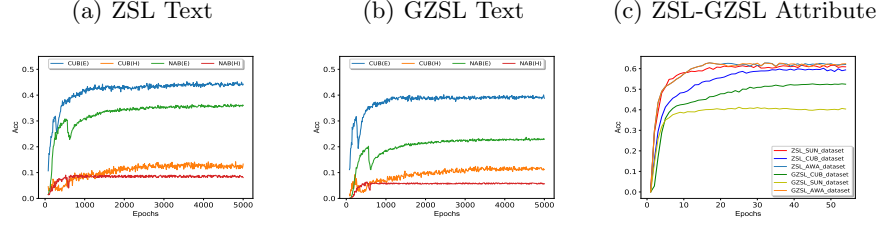(a) ZSL Text   (b) GZSL Text   (c) ZSL-GZSL Attribute

Fig. 5: Training stability (a-c) across all datasets under ZSL and GZSL

and Epoch to showcase the GZSL stability, mentioned in Fig. 5(b) and Fig. 5(c). Mainly we see the stable performance across all the datasets.

Ablation study : We have reported the Ablation study of our model in Fig. 4 (a) under ZSL for both attribute and Wikipedia based CUB. Primarily, we have used CUB (Easy) split for the Wikipedia based dataset. The $S1$ - WGAN with a classifier is considered as a baseline model. $S2$ reflects the effect of the visual pivot regularizer in our model. Finally, $S3$ showcase the performance of a complete LsrGAN with SR-loss. To highlight the effect of the denoiser used to process the Wikipedia based features, we have also reported the LsrGAN without denoiser in $S4$. In summary, Fig. 4 (a) showcases the importance of each component used in our model.

5   Conclusions

In this paper, we have proposed a novel generative zero-shot model named Lsr-GAN that Leverages the Semantic Relationship between seen and unseen classes to address the seen class overfitting concern in GZSL. Mainly, LsrGAN employs a novel Semantic Regularized Loss (SR-Loss) to perform explicit knowledge transfer from seen classes to unseen ones. The SR-Loss explores the semantic relationships between seen and unseen classes to guide the LsrGAN to generate visual features that mirror the same relationship. Extensive experiments on seven benchmarks, including attribute and Wikipedia description based datasets, verify that our LsrGAN effectively addresses the overfitting concern and demonstrates a superior ZSL and GZSL performance.

References

1. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. IEEE transactions on pattern analysis and machine intelligence **38**(7), 1425–1438 (2015)
2. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2927–2936 (2015)
3. Annadani, Y., Biswas, S.: Preserving semantic relations for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7603–7612 (2018)
4. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. arXiv preprint arXiv:1701.07875 (2017)
5. Bucher, M., Herbin, S., Jurie, F.: Generating visual representations for zero-shot classification. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 2666–2673 (2017)
6. Cacheux, Y.L., Borgne, H.L., Crucianu, M.: Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10333–10342 (2019)
7. Changpinyo, S., Chao, W.L., Gong, B., Sha., F.: Synthesized classifiers for zero-shot learning. CVPR pp. 5327–5336 (2016)
8. Changpinyo, S., Chao, W.L., Sha, F.: Predicting visual exemplars of unseen classes for zero-shot learning. In: Proceedings of the IEEE international conference on computer vision. pp. 3476–3485 (2017)
9. Chao, W.L., Changpinyo, S., Gong, B., Sha, F.: An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: European Conference on Computer Vision. pp. 52–68. Springer (2016)
10. Chen, L., Zhang, H., Xiao, J., Liu, W., Chang, S.F.: Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1043–1052 (2018)
11. Ding, Z., Shao, M., Fu, Y.: Low-rank embedded ensemble semantic dictionary for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2050–2058 (2017)
12. Elhoseiny, M., Elgammal, A., Saleh, B.: Write a classifier: Predicting visual classifiers from unstructured text. PAMI (2016)
13. Elhoseiny, M., Saleh, B., Elgammal, A.: Write a classifier: Zero-shot learning using purely textual descriptions. In: ICCV (2013)
14. Elhoseiny, M., Zhu, Y., Zhang, H., Elgammal, A.: Link the head to the "beak": Zero shot learning from noisy text description at part precision. In: CVPR (July 2017)
15. Felix, R., Kumar, V., Reid, I., Carnerio, G.: Multi-model cycle-consistent generalized zero-shot leanring. ECCV (2018)
16. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al: Devise: A deep visual-semantic embedding model. NIPS pp. 2121–2129 (2013)
17. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein GANs. In: Advances in neural information processing systems. pp. 5767–5777 (2017)
18. Guo, Y., Ding, G., Han, J., Gao, Y.: Synthesizing samples fro zero-shot learning. IJCAI (2017)

19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
20. Jiang, H., Wang, R., Shan, S., Chen, X.: Transferable contrastive network for generalized zero-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9765–9774 (2019)
21. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3174–3183 (2017)
22. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE transactions on pattern analysis and machine intelligence **36**(3), 453–465 (2013)
23. Larochelle, H., Erhan, D., Bengio, Y.: Zero-data learning of new tasks. AAAI (2008)
24. Li, J., Jin, M., Lu, K., Ding, Z., Zhu, L., Huang, Z.: Leveraging the invariant side of generative zero-shot learning. CVPR (2019)
25. Lillo, W.E., Loh, M.H., Hui, S., Zak, S.H.: On solving constrained optimization problems with neural networks : A penalty method approach. IEEE Transactions on neural networks **4**(6), 931–940 (1993)
26. Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G.S., Dean, J.: Zero-shot learning by convex combination of semantic embeddings. arXiv preprint arXiv:1312.5650 (2013)
27. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. CVPR (2012)
28. Qiao, R., Liu, L., Shen, C., v. d. Hengel, A.: Less is more: Zero-shot learning from online textual documents with noise suppression. In: CVPR (June 2016)
29. Rohrbach, M., Stark, M., Schiele, B.: Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In: CVPR 2011. pp. 1641–1648. IEEE (2011)
30. Romera-Paredes, B., Torr, P.: An embarrassingly simple approach to zero-shot learning. ICML pp. 2152–2161 (2015)
31. V. Verma, G. Arora, A.M., Rai, P.: Generalized zero-shot learning via synthesized examples. CVPR (2018)
32. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 595–604 (2015)
33. Verma, V.K., Rai, P.: A simple exponential family framework for zero-shot learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 792–808. Springer (2017)
34. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200 (2010)
35. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero shot learning. CVPR (2018)
36. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. TPAMI (2018)
37. Xu, X., Shen, F., Yang, Y., Zhang, D., Tao Shen, H., Song, J.: Matrix tri-factorization with manifold regularizations for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3798–3807 (2017)
38. Yu, X., Aloimonos, Y.: Attribute-based transfer learning for object categorization with zero or one training example. ECCV (2010)

39. Zhang, H., Xu, T., Elhoseiny, M., Huang, X., Zhang, S., Elgammal, A., Metaxas, D.: Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1143–1152 (2016)
40. Zhang, L., Xiang, T., Gong, S.: Learning a deep embedding model for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2021–2030 (2017)
41. Zhang, Z., Saligrama, V.: Zero-shot learning via semantic similarity embedding. ICCV pp. 4166–4174 (2015)
42. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
43. Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., Elgammal, A.: A generative adversarial approach for zero-shot learning from noisy texts. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1004–1013 (2018)