Creativity Inspired Zero-Shot Learning

Mohamed Elhoseiny[1,2] Mohamed Elfeki[3]
[1]Facebook AI Research (FAIR), [2]King Abdullah University of Science and Technology (KAUST)
[3]University of Central Florida
mohamed.elhoseiny@kaust.edu.sa, elfeki@cs.ucf.edu

Abstract

Zero-shot learning (ZSL) aims at understanding unseen categories with no training examples from class-level descriptions. To improve the discriminative power of zero-shot learning, we model the visual learning process of unseen categories with an inspiration from the psychology of human creativity for producing novel art. We relate ZSL to human creativity by observing that zero-shot learning is about recognizing the unseen and creativity is about creating a likable unseen. We introduce a learning signal inspired by creativity literature that explores the unseen space with hallucinated class-descriptions and encourages careful deviation of their visual feature generations from seen classes while allowing knowledge transfer from seen to unseen classes. Empirically, we show consistent improvement over the state of the art of several percents on the largest available benchmarks on the challenging task or generalized ZSL from a noisy text that we focus on, using the CUB and NABirds datasets. We also show the advantage of our approach on Attribute-based ZSL on three additional datasets (AwA2, aPY, and SUN). Code is available at `https://github.com/mhelhoseiny/CIZSL`.

1. Introduction

With hundreds of thousands of object categories in the real world and countless undiscovered species, it becomes unfeasible to maintain hundreds of examples per class to fuel the training needs of most existing recognition systems. Zipf's law, named after George Zipf (1902−1950), suggests that for the vast majority of the world-scale classes, only a few examples are available for training, validated earlier in language (e.g., [60, 61]) and later in vision (e.g., [43]). This problem becomes even more severe when we target recognition at the fine-grained level. For example, there exists tens of thousands of bird and flower species, but the largest available benchmarks have only a few hundred classes motivating a lot of research on classifying instances of unseen classes, known as Zero-Shot Learning (ZSL).

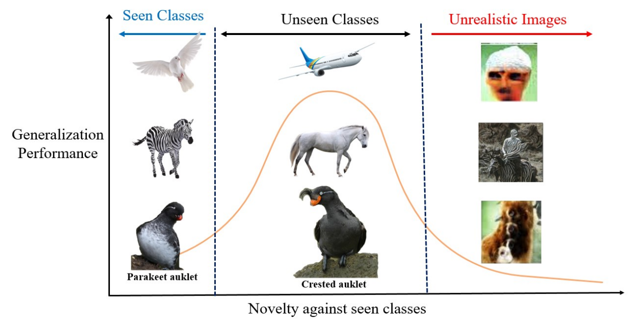


Figure 1: Generalizing the learning of zero-shot models requires a deviation from seen classes to accommodate recognizing unseen classes. We carefully model a learning signal that inductively encourages deviation of unseen classes from seen classes, yet not pushed far that the generation fall in the negative hedonic unrealistic range on the right and loses knowledge transfer from seen classes. Interestingly, this curve is similar to the famous Wundt Curve in the Human Creativity literature (Martindale, 1990) [33].

People have a great capability to identify unseen visual classes from text descriptions like "The crested auklet is subspecies of birds with dark-gray bodies tails and wings and orange-yellow bill. It is known for its forehead crests, made of black forward-curving feathers."; see Fig 1 (bottom). We may *imagine* the appearance of "crested auklet" in different ways yet all are correct and may collectively help us understand it better. This *imagination* notion been modeled in recent ZSL approaches (e.g., [19, 29, 20, 59, 55]) successfully adopting deep generative models to synthesize visual examples of an unseen object given its semantic description. After training, the model generates imaginary data for each unseen class transforming ZSL into a standard classification task with the generated data.

However, these generative ZSL methods do not guarantee the discrimination between seen and unseen classes since the generations are not motivated with a learning signal to deviate from seen classes. For example, "Parakeet Auklet" as a seen class in Fig 1 (left) has a visual text description [52] that significantly overlaps with "Crested Auklet" description, yet one can identify "Crested Auklet"'s unique “’black forward-curving feathers” against “Parakeet Auklet” from text. *The core of our work is to address the question of how to produce discriminative generations of unseen visual classes from text descriptions by explicitly learning to deviate from seen classes while allowing transfer to unseen classes.* Let’s imagine the space of conditional visual generations from class descriptions on an intensity map where light regions implies seen and darker regions implies unseen. These class descriptions are represented in a shared space between the unseen (dark) and the seen (light) classes, and hence the transfer is expected. In existing methods, this transfer signal is formulated by encouraging the generator to produce quality examples conditioned only on the descriptions of the seen classes (light regions only). In this inductive zero-shot learning, class descriptions of unseen classes are not available during training and hence can not used as a learning signal to explicitly encourage the discrimination across unseen and seen classes. *Explicitly modeling an inductive and discriminative learning signal from the dark unseen space is at the heart of our work.*

Creativity Inspiration to Zero-shot Learning. We propose to extend generative zero-shot learning with a discriminative learning signal inspired from the psychology of human creativity. Colin Marindale [33] proposes a psychological theory to explain the perception of human creativity. The definition relates likability of an art piece to novelty by *“the principle of least effort”*. The aesthetic appeal of an art work first increases when it deviates from existing work till some point, then decreases when the deviation goes too far. This means that it gets difficult to connect this art to what we are familiar with, and hence deems it hard to understand and hence appreciate. This principle can be visualized by the Wundt Curve where the X axis represents novelty and Y axis represents likability like an inverted U-shape; similar to the curve in Fig 1. We relate the Wundt curve behavior in producing creative art to a desirable generalized ZSL mode that has a better capability to distinguish the “crested auklet” unseen class from the “parakeet auklet” seen class given how similar they are as mentioned before; see Fig 1. A generative ZSL model that cannot deviate generations of unseen classes from instances of seen classes is expected to underperform in generalized zero-shot recognition due to confusion; see Fig 1(left). As the deviation capability increases, the performance is expected to get better but similarly would decrease when the deviation goes too far producing unrealistic generation and reducing the needed knowledge transfer from seen classes; see Fig 1(middle and right). Our key question is how to properly formulate deviation from generating features similar to existing classes while balancing the desirable transfer learning signal.

Contributions. 1) We propose a zero-shot learning approach that explicitly models generating unseen classes by learning to carefully deviate from seen classes. We examine a parametrized entropy measure to facilitate learning how to deviate from seen classes. Our approach is inspired from the psychology of human creativity; and thus we name it Creativity Inspired Zero-shot Learning (CIZSL).

2) Our creativity inspired loss is unsupervised and orthogonal to any Generative ZSL approach. Thus it can be integrated with any GZSL while adding no extra parameters nor requiring any additional labels.

3) By means of extensive experiments on seven benchmarks encompassing Wikipedia-based and attribute-based descriptions, our approach consistently outperformed state-of-the-art methods on zero-shot recognition, zero-shot retrieval, and generalized zero-shot learning using several evaluation metrics.

2. Related Work

Early Zero-Shot Learning(ZSL) Approaches A key idea to facilitate zero-shot learning is finding a common semantic representation that both seen and unseen classes can share. Attributes and text descriptions are shown to be effective shared semantic representations that allow transferring knowledge from seen to unseen classes. Lampert *et al.* [26] proposed a Direct Attribute Prediction (DAP) model that assumed independence of attributes and estimated the posterior of the test class by combining the attribute prediction probabilities. A parallelly developed, yet similar model was developed by Farhadi *et al.* [13].

Visual-Semantic Embedding ZSL. Relaxing the unrealistic independence assumption, Akata *et al.* [2] proposed an Attribute Label Embedding(ALE) approach that models zero-shot learning as a linear joint visual-semantic embedding. In principal, this model is similar to prior existing approaches that learn a mapping function from visual space to semantic space [57, 46]. This has been also investigated in the opposite direction [57, 46] as well as jointly learning a function for each space that map to a common space [56, 28, 3, 42, 47, 12, 1, 31, 30, 48].

Generative ZSL Approaches The notion of generating artificial examples has been recently proposed to model zero-shot learning reducing it to a conventional classification problem [19, 29, 20, 59]. Earlier approaches assumed a Gaussian distribution prior for visual space to every class and the probability densities for unseen classes are modeled as a linear combination of seen class distributions [19]. Long *et al.* [29] instead proposed a one-to-one mapping approach where synthesized examples are restricted. Recently, Zhu *et al.* [59], Xian *et al.* [55], and Verma *et al.*[25] relaxed this assumption and built on top of generative adversarial networks (GANs) [17, 39] to generate examples from unseen class descriptions. Different from ACGAN [36], Zhu *et al.* added a visual pivot regularizer (VPG) encourages generations of each class to be close to the average of its corresponding real features.

Semantic Representations in ZSL (e.g., Attributes, Description). ZSL requires by definition additional information (e.g., semantic description of unseen classes) to enable their recognition. A considerable progress has been made in studying attribute representation [26, 27, 2, 15, 57, 56, 28, 3, 42, 1]. Attributes are a collection of semantic characteristics that are filled to uniquely describe unseen classes. Another ZSL trend is to use online textual descriptions [11, 12, 38, 40, 28]. Textual descriptions can be easily extracted from online sources like Wikipedia with a minimal overhead, avoiding the need to define hundreds of attributes and filling them for each class/image. Elhoseiny *et al.* [11] proposed an early approach for Wikipedia-based zero-shot learning that combines domain transfer and regression to predict visual classifiers from a TF-IDF textual representation [44]. Qiao *et al.* [38] proposed suppress the noise in the Wikipedia articles by encouraging sparsity of the neural weights to the text terms. Recently, part-based zero-shot learning model [12] was proposed with a capability to connect text terms to its relevant parts of objects without part-text annotations. More recently, Zhu *et al.* [59] showed that suppressing the non-visual information is possible by the predictive power of the their model to synthesize visual features from the noisy Wikipedia text. Our work also focus on the challenging task of recognizing objects based on Wikipedia articles and is also a generative model. Unlike existing, we explicitly model the careful deviation of unseen class generations from seen classes .

Visual Creativity. Computational Creativity studies building machines that generate original items with realistic and aesthetic characteristics [32, 34, 7]. Although GANs [17, 39, 22] are a powerful generative model, yet it is not explicitly trained to create novel content beyond the training data. For instance, a GAN model trained on art works might generate the "Mona Lisa" again, but would not produce a novel content that it did not see. It is not different for some existing style transfer work [16, 8] since there is no incentive in these models to generate a new content. More recent work adopts computational creativity literature to create novel art and fashion designs [9, 45]. Inspired by [33], Elgammal*et al*. [9] adapted GANs to generate unconditional creative content (paintings) by encouraging the model to deviate from existing painting styles. Fashion is a 2.5 trillion dollar industry and has an impact in our everyday life, this motivated [45] to develop a model that can for example create an unseen fashion shape "pants to extended arm sleeves". The key idea behind these models is to add an additional novelty loss that encourage the model to explore the creative space of image generation.

3. Background

GANs [17, 39] train the generator G, with parameters θ_G, to produce samples that the Discriminator D believe they are real. On the other hand, the Discriminator D, with parameters θ_D, is trained to classify samples from the real distribution p_{data} as real (1), and samples produced by the generator as fake (0); see Eq 2.

$$\min_{\theta_G} \mathcal{L}_G = \min_{\theta_G} \sum_{z_i \in \mathbb{R}^n} \log(1 - D(G(z_i))) \quad (1)$$

$$\min_{\theta_D} \mathcal{L}_D = \min_{\theta_D} \sum_{x_i \in \mathcal{D}, z_i \in \mathbb{R}^n} - \log D(x_i) - \log(1 - D(G(z_i)))) \quad (2)$$

where z_i is a noise vector sampled from prior distribution p_z and x is a real sample from the data distribution p_{data}. In order to learn to deviate from seen painting styles or fashion shapes, [9, 45] proposed an additional head for the discriminator D that predicts the class of an image (painting style or shape class). During training, the Discriminator D is trained to predict the class of the real data through its additional head, apart from the original real/fake loss. The generator G is then trained to generate examples that are not only classified as real but more importantly are encouraged to be hard to classify using the additional discriminator head. More concretely,

$$\mathcal{L}_G = \mathcal{L}_{G \text{ real/fake}} + \lambda \mathcal{L}_{G \text{ creativity}} \quad (3)$$

The common objective between [9] and [45] is to produce novel generations with high entropy distribution over existing classes but they are different in the loss function. In [9], $\mathcal{L}_{G \text{ creativity}}$ is defined as the binary cross entropy (BCE) over each painting style produced by the discriminator additional head and the uniform distribution (i.e., $\frac{1}{K}$, K is the number of classes). Hence, this loss is a summation of BCE losses over all the classes. In contrast, Sbai *et al.* [45] adopted the Multiclass Cross Entropy (MCE) between the distribution over existing classes and the uniform distribution. To our knowledge, creative generation has not been explored before conditioned on text and to also facilitate recognizing unseen classe, *two key differences to our work.* Relating computational creativity to zero-shot learning is one of the novel aspects in our work by encouraging the deviation of generative models from seen classes. However, proper design of the learning signal is critical to (1) hallucinate class text-descriptions whose visual generations can help the careful deviation, (2) allow discriminative generation while allowing transfer between seen and unseen classes to facilitate zero-shot learning.

4. Proposed Approach

Problem Definition. We start by defining the zero-shot learning setting. We denote the semantic representations of unseen classes and seen classes as $t_i^u = \phi(T_k^u) \in \mathcal{T}$ and $t_i^s \in \mathcal{T}$ respectively, where $\mathcal{T}$ is the semantic space (e.g., features $\phi(\cdot)$ of a Wikipedia article T_k^u). Let's denote the seen data as $D^s = \{(x_i^s, y_i^s, t_i^s)\}_{i=1}^{N^s}$, where N^s is the number of training(seen) image examples, where $x_i^s \in \mathcal{X}$ denotes the visual features of the i^{th} image in the visual space

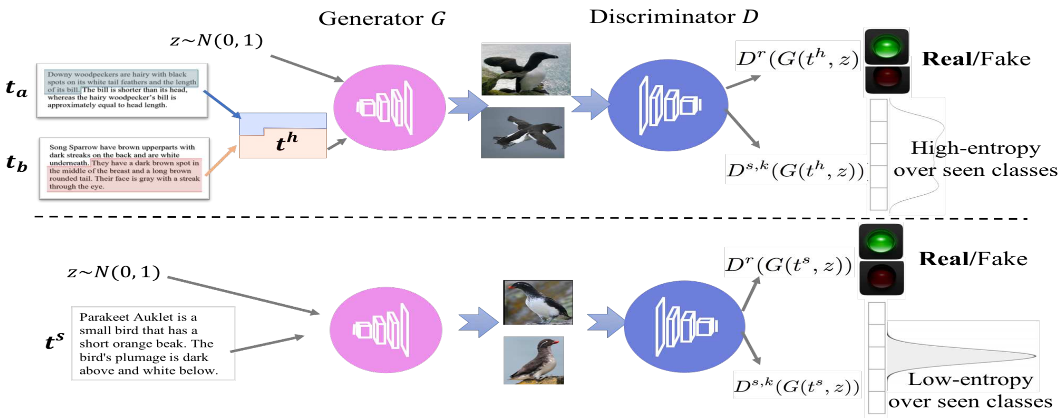


Figure 2: Generator G is trained to carefully deviate from seen to unseen classes without synthesizing unrealstic images. Top part: G is provided with a hallucinated text t^h and trained to trick discriminator to believe it is real, yet it encourages to deviate learning from seen classes by maximizing entropy over seen classes given t^h. Bottom part: G is provided with text of a seen class t^s and is trained to trick discriminator to believe it is real with a corresponding class label(low-entropy).

$\mathcal{X}$, y_i^s is the corresponding category label. We denote the number of unique seen class labels as K^s. We denote the set of seen and unseen class labels as $\mathcal{S}$ and $\mathcal{U}$, where the aforementioned $y_i^s \in \mathcal{S}$. Note that the seen and the unseen classes are disjointed, i.e., $\mathcal{S} \cap \mathcal{U} = \varnothing$. For unseen classes, we are given their semantic representations, one per class, $\{t_i^u\}_{i=1}^{K^u}$, where K^u is the number of unseen classes. The zero-shot learning (ZSL) task is to predict the label $y_u \in \mathcal{U}$ of an unseen class visual example $x^u \in \mathcal{X}$. In the more challenging Generalized ZSL (GZSL), the aim is to predict $y \in \mathcal{U} \cup \mathcal{S}$ given x that may belong to seen or unseen classes.

Approach Overview. Fig. 2 shows an overview of our Creativity Inspired Zero-Shot Learning model(CIZSL). Our method builds on top of GANs [17] while conditioning on semantic representation from raw Wikipedia text describing unseen classes. We denote the generator as G: $\mathbb{R}^Z \times \mathbb{R}^T \xrightarrow{\theta_G} \mathbb{R}^X$ and the discriminator as $D : \mathbb{R}^X \xrightarrow{\theta_D} \{0, 1\} \times \mathbb{L}_{cls}$, where θ_G and θ_D are parameters of the generator and the discriminator as respectively, $\mathbb{L}_{cls}$ is the set of seen class labels (i.e., $\mathcal{S} = \{1 \cdots K^s\}$). For the Generator G and as in [55], the text representation is then concatenated with a random vector $z \in \mathbb{R}^Z$ sampled from Gaussian distribution $\mathcal{N}(0, 1)$; see Fig. 2. In the architecture of [59], the encoded text t_k is first fed to a fully connected layer to reduce the dimensionality and to suppress the noise before concatenation with z. In our work, the discriminator D is trained not only to predict real for images from the training images and fake for generated ones, but also to identify the category of the input image. We denote the real/fake probability produced by D for an input image as $D^r(\cdot)$, and the classification score of a seen class $k \in \mathcal{S}$ given the image as $D^{s,k}(\cdot)$. Hence, the features are generated from the encoded text description t_k, as follows $\tilde{x}_k \leftarrow G(t_k, z)$. The discriminator then has two heads. The first head is an FC layer that for binary real/fake classification. The second head is a K^s-way classifier over the seen classes. Once our generator is trained, it is then used to hallucinate fake generations for unseen classes, where conventional classifier could be trained as we detail later in Sec 4.3.

The generator G is the key imagination component that we aim to train to generalize to unseen classes guided by signals from the discriminator D. In Sec 4.1, we detail the definition of our Creativity Inspired Zero-shot Signal to augment and improve the learning capability of the Generator G. In Sec 4.2, we show how our proposed loss can be easily integrated into adversarial generative training.

4.1. Creativity Inspired Zero-Shot Loss (CIZSL)

We explicitly explore the unseen/creative space of the generator G with a hallucinated text ($t^h \sim p_{text}^h$). We define p_{text}^h, as a probability distribution over hallucinated text description that is likely to be unseen and hard negatives to seen classes. To sample $t^h \sim p_{text}^h$, we first pick two seen text features at random $t_a^s, t_b^s \in \mathcal{S}$. Then we sample t^h by interpolating between them as

$$t^h = \alpha t_a^s + (1 - \alpha) t_b^s \tag{4}$$

where α is uniformally sampled between 0.2 and 0.8. We discard α values close to 0 or 1 to avoid sampling a text feature very close to a seen one. We also tried different ways to sample α which modifies p_{text}^h like fixed $\alpha = 0.5$ or $\alpha \sim \mathcal{N}(\mu = 0.5, \sigma = 0.5/3)$ but we found uniformally sampling from 0.2 to 0.8 is simple yet effective; see ablations at the supplementary(sec 5). We define our *creativity inspired zero-shot loss* L_G^C based on $G(t^h, z)$ as follows

$$\begin{aligned} L_G^C = & -\mathbb{E}_{z \sim p_z, t^h \sim p_{text}^s}[D^r(G(t^h, z))] \\ & + \lambda \mathbb{E}_{z \sim p_z, t^h \sim p_{text}^h}[L_e(\{D^{s,k}(G(t^h, z))\}_{k=1 \to K^s})] \end{aligned} \tag{5}$$

We encourage $G(t^h, z)$ to be real (first term) yet hard to classify to any of the seen classes (second term) and hence achieve more discrimination against seen classes; see Fig. 2

(top). More concretely, the first term encourage the generations given $t^h \sim p^h_{text}$ to trick the discriminator to believe it is real (i.e., maximize $D^r(G(t^h, z))$. This loss encourages the generated examples to stay realistic while deviating from seen classes. In the second term, we quantify the difficulty of classification by maximizing an entropy function L_e that we define later in this section. Minimizing L^C_G connects to the principal of least effort by Martindale et.al. 1990, where exaggerated novelty would decrease the transferability from seen classes (see visualized in Fig. 1). Promoting the aforementioned high entropy distribution incents discriminative generation. However, it does not disable knowledge transfer from seen classes since the unseen generations are encouraged to be an entropic combination of seen classes. We did not model deviation from seen classes as an additional class with label $K^s + 1$ that we always classify $G(t^h, z)$ to, since this reduces the knowledge transfer from seen classes as we demonstrate in our results.

Definition of L_e : L_e is defined over the seen classes' probabilities, produced by the second discriminator head $\{D^{s,k}(\cdot)\}_{k=1\to K^s}$ (i.e., the softmax output over the seen classes). We tried different entropy maximization losses. They are based on minimizing the divergence between the softmax distribution produced by the discriminator given the hallucinated text features and the uniform distribution. Concretely, the divergence, also known as relative entropy, is minimized between $\{D^{s,k}(G(t^h, z))\}_{k=1\to K^s})$ and $\{\frac{1}{K^s}\}_{k=1\to K^s}$; see Eq 6. Note that similar losses has been studied in the context of creative visual generation of art and fashion(e.g., [9, 45]). However, the focus there was mainly unconditional generation and there was no need to hallucinate the input text t^h to the generator, which is necessary in our case; see Sec 3. In contrast, our work also relates two different modalities (i.e., Wikipedia text and images).

$$L^{KL}_e = \sum_{k=1}^{K^s} \frac{1}{K^s} D^{s,k}(G(t^h, z))$$
$$L^{SM}_e(\gamma, \beta) = \frac{1}{\beta - 1}\left[\sum_{k=1}^{K^s}(D^{s,k}(G(t^h, z)^{1-\gamma}(\frac{1}{K^s})^{\gamma})^{\frac{1-\beta}{1-\gamma}} - 1\right] \tag{6}$$

Several divergence/entropy measures has been proposed in the information theory literature [41, 49, 23, 4, 21]. We adopted two divergence losses, the well-known Kullback-Leibler(KL) divergence in L^{KL}_e and the two-parameter Sharma-Mittal(SM) [21] divergence in L^{SM}_e which is relatively less known; see Eq 6. It was shown in [4], that other divergence measures are special case of Sharma-Mittal(SM) divergence by setting its two parameters γ and β. It is equivalent to *Rényi* [41] when $\beta \to 1$ (single-parameter), *Tsallis* divergence [49] when $\gamma = \beta$ (single-parameter), Bhattacharyya divergence when$\beta \to 0.5$ and $\gamma \to 0.5$, and *KL* divergence when $\beta \to 1$ and $\gamma \to 1$ (no-parameter). So, when we implement SM loss, we can also minimize any of the aforementioned special-case measures; see details in the supplementary(sec 2). Note that we also learn γ and β when we train our model with SM loss.

4.2. Integrating CIZSL in Adversarial Training

The integration of our approach is simple that L^C_G defined in Eq 5 is just added to the generator loss; see Eq 7. Similar to existing methods, when the generator G is provided with text describing a seen class t^s, its is trained to trick the discriminator to believe it is real and to predict the corresponding class label (low-entropy for t^s versus hig-entropy for t^h); see Fig 2(bottom). Note that the remaining terms, that we detail here for concreteness of our method, are simlar to existing generative ZSL approaches [55, 59]

Generator Loss The generator loss is an addition of four terms, defined as follows

$$\begin{aligned} L_G = L^C_G - \mathbb{E}_{z\sim p_z,(t^s,y^s)\sim p^s_{text}}[D^r(G(t^s, z))+ \\ \sum_{k=1}^{K^s} y^s_k log(D^{s,k}(G(t^s, z)))] \\ + \frac{1}{K^s}\sum_{k=1}^{K^s} ||\mathbb{E}_{z\sim p_z}[G(t_k, z)] - \mathbb{E}_{x\sim p^k_{data}}[x]||^2 \end{aligned} \tag{7}$$

The first term is our creativity inspired zero-shot loss L^C_G, described in Sec 4.1. Note that seen class text descriptions $\{t_k\}_{k=1\to K^s}$ are encouraged to predict a low entropy distribution since loss is minimized when the corresponding class is predicted with a high probability. Hence, the second term tricks the generator to classify visual generations from seen text t^s as real. The third term encourages the generator to be capable of generating visual features conditioned on a given seen text. The fourth term is an additional visual pivot regularizer that we adopted from [59], which encourages the centers of the generated (fake) examples for each class k (i.e., with $G(t_k, z)$) to be close to the centers of real ones from sampled from p^k_{data} for the same class k. Similar to existing methods, the loss for the discriminator is defined as:

$$\begin{aligned} L_D =& \mathbb{E}_{z\sim p_z,(t^s,y^s)\sim p^s_{text}}[D^r(G(t^s, z))] - \mathbb{E}_{x\sim p_{data}}[D^r(x)] \\ &+ L_{Lip} - \frac{1}{2}\mathbb{E}_{x,y\sim p_{data}}[\sum_{k=1}^{K^s} y_k log(D^{s,k}(x))] \\ &- \frac{1}{2}\mathbb{E}_{z\sim p_z,(t^s,y^s)\sim p^s_{text}}[\sum_{k=1}^{K^s} y^s_k log(D^{s,k}(G(t^s, z)))] \end{aligned} \tag{8}$$

where y is a one-hot vector encoding of the seen class label for the sampled image x, t^s and y^s are features of a text description and the corresponding on-hot label sampled from seen classes p^s_{text}. The first two terms approximate Wasserstein distance of the distribution of real features and fake features. The third term is the gradient penalty to enforce the Lipschitz constraint: $L_{Lip} = (|| \bigtriangledown_{\tilde{x}} D^r(\tilde{x})||_2 - 1)^2$, where $\tilde{x}$ is the linear interpolation of the real feature x and the fake feature $\hat{x}$; see [18]. The last two terms are classification losses of the seen real features and fake features from text descriptions of seen category labels.

Training. We construct two minibatches for training the generator G, one from seen class t^s and from the halllucinated text t^h to minimize L_G (Eq. 7) and in particular L_G^C (Eq. 5). The generator is optimized to fool the discriminator into believing the generated features as real either from hallucinated text t^h or the seen text t^s. In the mean time, we maximize their entropy over the seen classes if the generated features comes from hallucinated text $t^h \sim p^h_{text}$ or to the corresponding class if from a real text t^s. Training the discriminator is similar to existing works; see in the supplementary (sec3) a detailed algorithm and code to show how G and D are alternatively trained with an Adam optimizer. Note that when L_e has parameters like γ and β for Sharma-Mittal(SM) divergence (Eq 6), that we also learn.

4.3. Zero-Shot Recognition Test

After training, the visual features of unseen classes can be synthesized by the generator conditioned on a given unseen text description t_u, as $x_u = G(t_u, z)$. We can generate an arbitrary number of generated visual features by sampling different z for the same text t_u. With this synthesized data of unseen classes, the zero-shot recognition becomes a conventional classification problem. We used nearest neighbor prediction, which we found simple and effective.

5. Experiments

We investigate the performance of our approach on two class-level semantic settings: textual and attribute descriptions. Since the textual based ZSL is a harder problem, we used it to run an ablation study for zero-shot retrieval and generalized ZSL. Then, we conducted experiments for both settings to validate the generality of our work.

Cross-Validation The weight λ of our loss in Eq 5 is a hyperparameter that we found easy to tune on all of our experiments. We start by splitting the data into training and validation split with nearly 80-20% ratio for all settings. Training and validation classes are selected randomly prior to the training. Then, we compute validation performance when training the model on the 80% split every 100 iterations out of 3000 iterations. We investigate a wide range of values for λ, and the value that scores highest validation performance is selected to be used at the inference time. Finally, we combine training and validation data and evaluate the performance on testing data.

Zer-Shot Performance Metrics. We use two metrics widely used in in evaluating ZSL recognition performance: Standard Zero-shot recognition with the Top-1 unseen class accuracy and Seen-Unseen Generalized Zero-shot performance with Area under Seen-Unseen curve [6]. The Top-1 accuracy is the average percentage of images from unseen classes classifying correctly to one of unseen class labels. However, this might be incomplete measure since it is more realistic at inference time to encounter also seen classes. Therefore, We also report a generalized zero-shot recognition metric with respect to the seen-unseen curve, proposed by Chao *et al.* [6]. This metric classifies images of both seen $\mathcal{S}$ and unseen classes $\mathcal{U}$ at test time. Then, the performance of a ZSL model is assessed by classifying these images to the label space that covers both seen classes and unseen labels $\mathcal{T} = \mathcal{S} \cup \mathcal{U}$. A balancing parameter is used sample seen and unseen class test accuracy-pair. This pair is plotted as the (x, y) co-ordinate to form the Seen-Unseen Curve(SUC). We follow [59] in using the Area Under SUC to evaluate the generalization capability of class-level text zero-shot recognition, and the haromnic mean of SUC for attribute-based zero-shot recognition. In our model, we use the trained GAN to synthesize the visual features for both training and testing classes.

5.1. Wikipedia based ZSL Results (4 benchmarks)

Text Representation. Textual features for each class are extracted from corresponding raw Wikipedia articles collected by [11, 12]. We used Term Frequency-Inverse Document Frequency (TF-IDF) [44] feature vector of dimensionality 7551 for CUB and 13217 for NAB.

Visual Representation. We use features of the part-based FC layer in VPDE-net [57]. The image are fed forward to the VPDE-net after resizing to 224×224, and the feature activation for each detected part is extracted which is of 512 dimensionality. The dimensionalities of visual features for CUB and NAB are 3583 and 3072 respectively. There are six semantic parts shared in CUB and NAB:"head", "back", "belly", "breast", "leg", "wing", "tail". Additionally, CUB has an extra part which is "leg" which makes its feature representation 512D longer compared to NAB (3583 vs 3072). More details in the supplementary(sec 6).

Datasets. We use two common fine-grained recognition datasets for textual descriptions: *Caltech UCSD Birds-2011* (CUB) [51] and *North America Birds* (NAB) [50]. CUB dataset contains 200 classes of bird species and their Wikipedia textual description constituting a total of 11,788 images. Compared to CUB, NAB is a larger dataset of birds, containing a 1011 classes and 48,562 images.

Splits. For both datasets, there are two schemes to split the classes into training/testing (in total four benchmarks): Super-Category-Shared (SCS) or *easy* split and Super-Category-Exclusive Splitting (SCE) or *hard* split, proposed in [12]. Those splits represents the similarity of the seen to unseen classes, such that the former represents a higher similarity than the latter. For SCS (easy), unseen classes are deliberately picked such that for every unseen class, there is at least one seen class with the same super-category. Hence, the relevance between seen and unseen classes is very high,

Metric	Top-1 Accuracy (%)				Seen-Unseen AUC (%)			
Dataset	CUB		NAB		CUB		NAB	
Split-Mode	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
CIZSL SM-Entropy (ours final)	**44.6**	**14.4**	36.5	**9.3**	**39.2**	**11.9**	**24.5**	**6.4**
CIZSL SM-Entropy (replace 2^{nd} term in Eq 5 by Classifying t^h as new class)	43.2	11.31	35.6	8.5	38.3	9.5	21.6	5.6
CIZSL SM-Entropy (minus 1^{st} term in Eq 5)	43.4	10.1	35.2	8.3	35.0	8.2	20.1	5.4
CIZSL SM-Entropy: (minus 2^{nd} term in Eq 5)	41.7	11.2	33.4	8.1	33.3	10.1	21.3	5.1
CIZSL Bachatera-Entropy ($\gamma = 0.5, \beta = 0.5$)	44.1	13.7	35.9	8.9	38.9	10.3	24.3	6.2
CIZSL Renyi-Entropy ($\beta \to 1$)	44.1	13.3	35.8	8.8	38.6	10.3	23.7	6.3
CIZSL KL-Entropy ($\gamma \to 1, \beta \to 1$)	44.5	14.2	36.3	8.9	38.9	11.6	24.3	6.2
CIZSL Tsallis-Entropy ($\beta = \gamma$)	44.1	13.8	**36.7**	8.9	38.9	11.3	24.5	6.3
CIZSL SM-Entropy: (minus 1^{st} and 2^{nd} terms in Eq 5)= GAZSL [59]	43.7	10.3	35.6	8.6	35.4	8.7	20.4	5.8

Table 1: Ablation Study using Zero-Shot recognition on **CUB** & **NAB** datasets with two split settings each. CIZSL is GAZSL [59]+ our loss

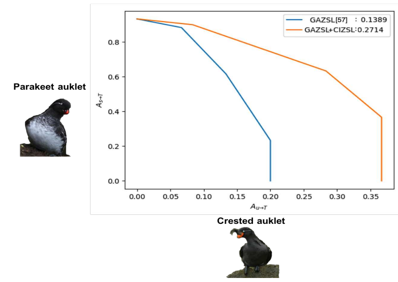
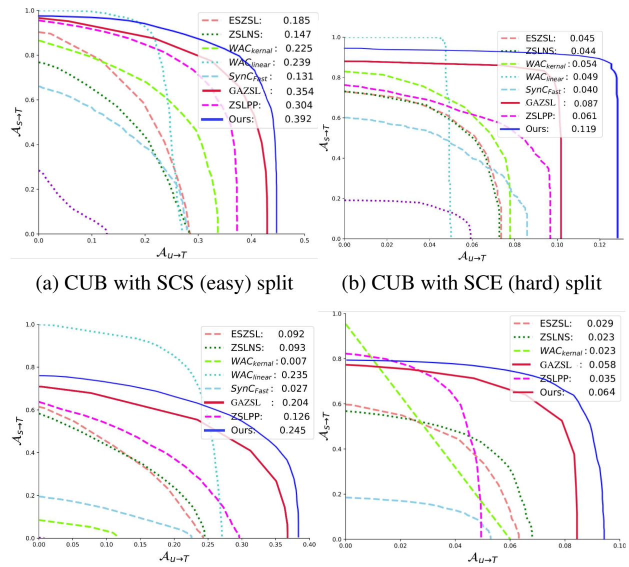


Figure 3: Seen Unseen Curve for Parakeet Auklet (Seen, y-axis) vs Crested Auklet (Unseen, x-axis) for GAZSL[57] and GAZSL[57]+ CIZSL.

deeming the zero-shot recognition and retrieval problems relatively easier. On the other end of the spectrum, SCE (hard) scheme, the unseen classes do not share the super-categories with the seen classes. Hence, there is lower similarity between the seen and unseen classes making the problem harder to solve. Note that the easy split is more common in literature since it is more Natural yet the deliberately designed hard-split shows the progress when the super category is not seen that we also may expect.

Ablation Study (Table 1). Our loss is composed of two terms shown that encourage the careful deviation in Eq 5. The first term encourages that the generated visual features from the hallucinated text t^h to deceive the discriminator believing it is real, which restricts synthesized visual features to be realistic. The second term maximizes the entropy using a deviation measure. In our work, Shama-Mittal(SM) entropy parameters γ and β are learnt and hence adapt the corresponding data and split mode to a matching divergence function, leading to the best results especially in the generalized SUAUC metric; see first row in Table 1. We first investigate the effect of deviating the hallucinated text by classifying it t a new class $K^s + 1$, where K^s is the number of the seen classes. We found the performance is significantly worse since the loss would significantly increase indecencies against seen classes and hence reduces seen knowledge transfer to unseen classes; see row 2 in Table 1. When we remove the first term (realistic constraints), the performance degrades especially under the generalized Seen-Unseen AUC metric because generated visual features became unrealistic; see row 3 in Table 1 (e.g., 39.2% to 35.0% AUC drop for CUB Easy and 11.9%-8.2% drop for CUB Hard). Alternatively, when we remove the second term (entropy), we also observe a significant drop in performance showing that both losses are complementary to each other; see row 4 in Table 1 (e.g., 39.2% to 33.5% AUC drop for CUB Easy and 11.9%-10.1% drop for CUB Hard). In our ablation, applying our approach without both terms (our loss) is equivalent to [59], shown is the last row in Table 1 as one of the least performing baselines. Note that our loss is applicable to other generative ZSL methods as we show in our state-of-the-art comparisons later in this section.

We also compare different entropy measures to encourage the deviation from the seen classes: *Kullback-Leibler (KL)*, *Rényi* [41], *Tsallis* [49], *Bhattacharyya* [23]; see rows 5-8 in Table 1. All these divergences measure are special cases of the two parameter (γ , β) Sharma-Mittal(SM) [21] divergence that we implemented. For instance, Renyi [41] and Tsallis [49] on the other hand only learns one parameter and achieves comparable yet lower performance. Bhattacharyya [23] and KL have no learnable parameters an achieves lower performance compared to SM.



(a) CUB with SCS (easy) split

(b) CUB with SCE (hard) split

(c) NAB with SCS (easy) split

(d) NAB with SCE (hard) split

Figure 4: Seen-Unseen accuracy Curve with two splits: SCS(easy) and SCE(hard). Ours indicates GAZSL+ CIZSL

Zero-Shot Recognition and Generality on [55] and [59]. Fig 3 shows the key advantage of our CIZSL loss, doubling the capability of [57] from 0.13 AUC to 0.27 AUC to distinguish between two very similar birds: Parakeet Auklet (Seen class) and Crested Auklet (unseen class), in 200-way classification; see supplementary (sec1) for details. Table 2 shows state-of-the-art comparison on CUB and NAB datasets for both their SCS(easy) and SCE(hard) splits (total of four benchmarks). Our method shows a significant advantage compared to the state of the art especially in generalized Seen-Unseen AUC metric ranging from 1.0-4.5%

Metric	Top-1 Accuracy (%)				Seen-Unseen AUC (%)			
Dataset	CUB		NAB		CUB		NAB	
Split-Mode	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
WAC-Linear [11]	27.0	5.0	–	–	23.9	4.9	23.5	–
WAC-Kernel [10]	33.5	7.7	11.4	6.0	14.7	4.4	9.3	2.3
ESZSL [42]	28.5	7.4	24.3	6.3	18.5	4.5	9.2	2.9
ZSLNS [38]	29.1	7.3	24.5	6.8	14.7	4.4	9.3	2.3
SynC$_{fast}$ [5]	28.0	8.6	18.4	3.8	13.1	4.0	2.7	3.5
ZSLPP [12]	37.2	9.7	30.3	8.1	30.4	6.1	12.6	3.5
FeatGen [55]	43.9	9.8	36.2	8.7	34.1	7.4	21.3	5.6
FeatGen[55]+ **CIZSL**	44.2 $^{+0.3}$	12.1 $^{+2.3}$	36.3 $^{+0.1}$	**9.8** $^{+1.1}$	37.4 $^{+2.7}$	9.8 $^{+2.4}$	24.7 $^{+3.4}$	6.2 $^{+0.6}$
GAZSL [59]	43.7	10.3	35.6	8.6	35.4	8.7	20.4	5.8
GAZSL [59] + **CIZSL**	**44.6** $^{+0.9}$	**14.4** $^{+4.1}$	**36.6** $^{+1.0}$	9.3 $^{+0.7}$	**39.2** $^{+3.8}$	**11.9** $^{+3.2}$	**24.5** $^{+4.1}$	**6.4** $^{+0.6}$

Table 2: Zero-Shot Recognition on class-level textual description from **CUB** and **NAB** datasets with two-split setting.

	Top-1 Accuracy(%)			Seen-Unseen H		
	AwA2	aPY	SUN	AwA2	aPY	SUN
DAP [27]	46.1	33.8	39.9	–	9.0	7.2
SSE [58]	61.0	34.0	51.5	14.8	0.4	4.0
SJE [3]	61.9	35.2	53.7	14.4	6.9	19.8
LATEM [53]	55.8	35.2	55.3	20.0	0.2	19.5
ESZSL [42]	58.6	38.3	54.5	11.0	4.6	15.8
ALE [2]	62.5	39.7	58.1	23.9	8.7	26.3
CONSE [35]	44.5	26.9	38.8	1.0	–	11.6
SYNC [5]	46.6	23.9	56.3	18.0	13.3	13.4
SAE [24]	54.1	8.3	40.3	2.2	0.9	11.8
DEM [57]	67.1	35.0	61.9	25.1	19.4	25.6
DEVISE [15]	59.7	39.8	56.5	**27.8**	9.2	20.9
GAZSL [59]	58.9	41.1	61.3	15.4	24.0	26.7
GAZSL [59] + **CIZSL**	**67.8** $^{+8.9}$	42.1 $^{+1.0}$	63.7 $^{+2.4}$	24.6 $^{+9.2}$	25.7 $^{+1.7}$	**27.8** $^{+1.1}$
FeatGen [55]	54.3	42.6	60.8	17.6	21.4	24.9
FeatGen [55] + **CIZSL**	60.1 $^{+5.8}$	43.8 $^{+1.2}$	59.4 $^{-0.6}$	19.1 $^{+1.5}$	24.0 $^{+2.6}$	26.5 $^{+1.6}$
cycle-(U)WGAN [14]	56.2	44.6	60.3	19.2	23.6	24.4
cycle-(U)WGAN [14] + **CIZSL**	63.6 $^{+7.4}$	**45.1** $^{+0.5}$	**64.2** $^{+3.9}$	23.9 $^{+4.7}$	**26.2** $^{+2.6}$	27.6 $^{+3.2}$

Table 3: Zero-Shot Recognition on class-level attributes of **AwA2**, **aPY** and **SUN** datasets.

	CUB			NAB		
	25%	50%	100%	25%	50%	100%
ESZSL [42]	27.9	27.3	22.7	28.9	27.8	20.9
ZSLNS [38]	29.2	29.5	23.9	28.8	27.3	22.1
ZSLPP [12]	42.3	42.0	36.3	36.9	35.7	31.3
GAZSL [59]	49.7	48.3	40.3	**41.6**	37.8	31.0
GAZSL [59]+ CIZSL	**50.3** $^{+0.6}$	**48.9** $^{+0.6}$	**46.2** $^{+5.9}$	41.0 $^{-0.6}$	**40.2** $^{+2.4}$	**34.2** $^{+3.2}$

Table 4: Zero-Shot Retrieval using mean Average Precision(mAP) (%) on CUB and NAB with SCS(easy) splits.

improvement. Fig 4 visualizes Seen-Unseen curves for our four benchmarks CUB (east and hard splits) and NABirds (easy and hard splits) where ours has a significant advantage compared to state-of-the-art on recognizing unseen classes; see our area under SU curve gain in Fig 4 against the runner-up GAZSL. The average relative SU-AUC improvement on the easy splits is 15.4% and 23.56% on the hard split. Meaning, the advantage of our loss becomes more clear as splits get harder, showing a better capability of discriminative knowledge transfer. We show the generality of our method by embedding it with another feature generation method, FeatGen [55], causing a consistent improvement. All the methods are using same text and visual representation.

Zero-Shot Retrieval. We investigate our model's performance for zero-shot retrieval task given the Wikipedia article of the class using mean Average Precision (mAP), the common retrieval metric. In table 4, we report the performance of different settings: retrieving 25%, 50%, 100% of the images at each class. We follow [59] to obtain the visual center of unseen classes by generating 60 examples for the given text then computing the average. Thus, given the visual center, the aim is to retrieve images based on the nearest neighbor strategy in the visual features space. Our model is the best performing and improves the MAP (100%) over the runner-up ([59]) by 14.64% and 9.61% on CUB and NAB respectively. Even when the model fails to retrieve the exact unseen class, it tends to retrieve visually similar images; see qualitative examples in supplementary.

5.2. Attribute-based Zero-Shot Learning

Datasets. Although it is not our focus, we also investigate the performance of our model's zero-shot recognition ability using different semantic representation. We follow the GBU setting [54], where images are described by their attributes instead of textual describtion deeming the problem to be relatively easier than textual-description zero-shot learning. We evaluated our approach on the following datasets: Animals with Attributes (AwA2) [26], aPascal/aYahoo objects(aPY) [13] and the SUN scene attributes dataset [37]. They consist of images covering a variety of categories in different scopes: animals, objects and scenes respectively. AwA contains attribute-labelled classes but aPY and SUN datasets have their attribute signature calculated as the average of the instances belonging to each class.

Zero-Shot Recognition. On AwA2, APY, and SUN datasets, we show in Table 3 that our CIZSL loss improves three generative zero-shot learning models including GAZSL [59], FeatGen [55], and cycle-(U)WGAN [14]. The table also shows our comparison to the state-of-the-art where we mostly obtain a superior performance. Even when obtaining a slightly lower score than state-of-the-art on AWA2, our loss adds a 9.2% Seen-Unseen H absolute improvement to the non-creative GAZSL [59]. We also evaluated our loss on *CUB-T1(Attributes) benchmark [54]*, where the Seen-Unseen H for GAZSL [59]] and GAZSL [59]+CIZSL are 55.8 and 57.4, respectively.

6. Conclusion

We draw an inspiration from the psychology of human creativity to improve the capability of unseen class imagination for zero-shot recognition. We adopted GANs to discrimnatively imagine visual features given a hallucinated text describing an unseen visual class. Thus, our generator learns to synthesize unseen classes from hallucinated texts. Our loss encourages deviating generations of unseen from seen classes by enforcing a high entropy on seen class classification while being realistic. Nonetheless, we ensure the realism of hallucinated text by synthesizing visual features similar to the seen classes to preserve knowledge transfer to unseen classes. Comprehensive evaluation on seven benchmarks shows a consistent improvement over the state-of-the-art for both zero-shot learning and retrieval with class description defined by Wikipedia articles and attributes.

References

[1] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele. Multi-cue zero-shot learning with strong supervision. In *CVPR*, 2016.

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *PAMI*, 38(7):1425–1438, 2016.

[3] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015.

[4] E. Akturk, G. Bagci, and R. Sever. Is sharma-mittal entropy really a step beyond tsallis and rényi entropies? *arXiv preprint cond-mat/0703277*, 2007.

[5] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, 2016.

[6] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, pages 52–68. Springer, 2016.

[7] S. DiPaola and L. Gabora. Incorporating characteristics of human creativity into an evolutionary art algorithm. *Genetic Programming and Evolvable Machines*, 10(2):97–110, 2009.

[8] V. Dumoulin, J. Shlens, M. Kudlur, A. Behboodi, F. Lemic, A. Wolisz, M. Molinaro, C. Hirche, M. Hayashi, E. Bagan, et al. A learned representation for artistic style. *ICLR*, 2017.

[9] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. In *International Conference on Computational Creativity*, 2017.

[10] M. Elhoseiny, A. Elgammal, and B. Saleh. Write a classifier: Predicting visual classifiers from unstructured text. *PAMI*, 2016.

[11] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, 2013.

[12] M. Elhoseiny, Y. Zhu, H. Zhang, and A. Elgammal. Link the head to the "beak": Zero shot learning from noisy text description at part precision. In *CVPR*, July 2017.

[13] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR 2009.*, pages 1778–1785. IEEE, 2009.

[14] R. Felix, V. B. Kumar, I. Reid, and G. Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, pages 21–37, 2018.

[15] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[16] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[18] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

[19] Y. Guo, G. Ding, J. Han, and Y. Gao. Synthesizing samples for zero-shot learning. In *IJCAI*, 2017.

[20] Y. Guo, G. Ding, J. Han, and Y. Gao. Zero-shot learning with transferred samples. *IEEE Transactions on Image Processing*, 2017.

[21] H. Gupta and B. D. Sharma. On non-additive measures of inaccuracy. *Czechoslovak Mathematical Journal*, 26(4):584–595, 1976.

[22] D. Ha and D. Eck. A neural representation of sketch drawings. *ICLR*, 2018.

[23] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE transactions on communication technology*, 15(1):52–60, 1967.

[24] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. *arXiv preprint arXiv:1704.08345*, 2017.

[25] V. Kumar Verma, G. Arora, A. Mishra, and P. Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, 2018.

[26] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958. IEEE, 2009.

[27] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *PAMI*, 36(3):453–465, March 2014.

[28] J. Lei Ba, K. Swersky, S. Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, 2015.

[29] Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, and J. Han. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. In *CVPR*, 2017.

[30] Y. Long and L. Shao. Describing unseen classes by exemplars: Zero-shot learning using grouped simile ensemble. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 907–915. IEEE, 2017.

[31] Y. Long and L. Shao. Learning to recognise unseen classes by a few similes. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 636–644. ACM, 2017.

[32] P. Machado and A. Cardoso. Nevar–the assessment of an evolutionary art tool. In *Proc. of the AISB00 Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science*, volume 456, 2000.

[33] C. Martindale. *The clockwork muse: The predictability of artistic change.* Basic Books, 1990.

[34] A. Mordvintsev, C. Olah, and M. Tyka. Inceptionism: Going deeper into neural networks. *Google Research Blog. Retrieved June*, 2015.

[35] 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.

[36] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017.

[37] 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.

[38] R. Qiao, L. Liu, C. Shen, and A. v. d. Hengel. Less is more: Zero-shot learning from online textual documents with noise suppression. In *CVPR*, June 2016.

[39] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016.

[40] S. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016.

[41] A. Rényi. On Measures Of Entropy And Information. In *Berkeley Symposium on Mathematics, Statistics and Probability*, 1960.

[42] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, pages 2152–2161, 2015.

[43] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, 2011.

[44] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[45] O. Sbai, M. Elhoseiny, A. Bordes, Y. LeCun, and C. Couprie. Design: Design inspiration from generative networks. In *ECCV workshop*, 2018.

[46] 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.

[47] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, pages 935–943, 2013.

[48] Y.-H. H. Tsai, L.-K. Huang, and R. Salakhutdinov. Learning robust visual-semantic embeddings. In *ICCV*, 2017.

[49] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *J. Statist. Phys.*, 1988.

[50] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, 2015.

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

[52] Wikipedia. Crested auklet. `https://en.wikipedia.org/wiki/Crested_auklet`, 2009. [Online; accessed 19-March-2019].

[53] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016.

[54] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *PAMI*, 2018.

[55] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018.

[56] Y. Yang and T. M. Hospedales. A unified perspective on multi-domain and multi-task learning. In *ICLR*, 2015.

[57] L. Zhang, T. Xiang, and S. Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2016.

[58] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, pages 4166–4174, 2015.

[59] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, 2018.

[60] G. K. Zipf. The psycho biology of language an introduction to dynamic philology. 1935.

[61] G. K. Zipf. Human behavior and the principle of least effort. 1949.