

Zero-Shot Learning and its Applications from Autonomous Vehicles to COVID-19 Diagnosis: A Review

Mahdi Rezaei

University of Leeds, UK, m.rezaei@leeds.ac.uk

Mahsa Shahidi

Syntech Research Centre, m.shahidi@mrl.ir

The challenge of learning a new concept without receiving any examples beforehand is called zero-shot learning (ZSL). One of the major issues in deep learning based methodologies is the requirement of feeding a vast amount of annotated and labelled images by a human to train the network model. ZSL is known for having minimal human intervention by relying only on previously known concepts and auxiliary information. It is an ever-growing research area since it has human-like characteristics in learning new concepts, which makes it applicable in many real-world scenarios, from autonomous vehicles to surveillance systems to medical imaging and COVID-19 CT scan-based diagnosis. In this paper, we present the definition of the problem, we review over fundamentals, and the challenging steps of Zero-shot learning, including recent categories of solutions, motivations behind each approach, and their advantages over other categories. Inspired from different settings and extensions, we have a broaden solution called one/few-shot learning. We then review thorough datasets, the variety of splits, and the evaluation protocols proposed so far. Finally, we discuss the recent applications and possible future directions of ZSL. We aim to convey a useful intuition through this paper towards the goal of handling computer vision learning tasks more similar to the way humans learn.

Keywords: Zero-shot learning; Semantic Embedding; Machine Learning; Deep Learning; COVID-19 Pandemic; Autonomous Vehicles; Supervised Annotation.

1 INTRODUCTION

Object recognition is one of the highly researched areas of computer vision. Recent recognition models have led to great performance through established techniques and large annotated datasets. To this date, the attention over this topic has not dimmed as there are still ways to refine models by eliminating various issues exist in this area. The number of newly emerging unknown objects are growing. Some examples of these unseen or rarely-previously-seen objects are the next generation of concept cars, futuristic-looking object designs, other existing concepts but with restricted access to them (such as licensed or private medical imag-

ing datasets), or rarely seen objects (such a traffic signs with graffiti on them), or fine-grained categories of objects (such as detection of a Caspian tiger comparing to the easier task of detecting a common Bengal tiger). This brings the necessity of developing a fresh way of solving object recognition problems that concern lesser human supervision. Several approaches have tried to gather web images to train the developed deep learning models, but aside from the problem of the noisy images, the keywords are still a form of human supervision. One-Shot learning (OSL) and Few-shot learning (FSL) are two other popular solutions that are able to learn new categories via one or a few images, respectively [1], [2], [3]. Then, Zero-shot learning (ZSL) [4], [5], [6], [7], [8] emerged which is completely free of any laborious task of data collection and annotation by experts. Zero-shot learning is a novel concept learning technique without accessing any exemplars of the unseen categories during training, yet it is able to build recognition models with the help of transferring knowledge from previously seen categories and auxiliary information. One of the interesting facts about ZSL is its similarity with the way human learns a concept without seeing them. This makes it capable of recognizing them later, in case of the appearance. For example, a ZSL-based model would be able to recognize a Persian fallow deer, based on the information available for it and the similarities and differences with other previously known deer. For instance, it belongs to a subgroup of the fallow deer, with a larger body, bigger antlers, white spots around the neck, and also flat antlers for the male type. A similar concept or approach is applicable in autonomous vehicles [9] where a self-driving car is responsible for recognition of a novel unseen Toyota Concept-i car (Figure 1) based on the subgroup of classic sedan cars; or COVID-19 patient diagnosis, based on the chest X-ray symptoms from the subgroup of asthma and lungs inflammatory diseases.

2 ZSL TEST & TRAINING PHASES

ZSL vs. Generalised ZSL Test Settings: ZSL models can be seen from various point of views in terms of training and test phases. In classic ZSL settings, the model

only detects the presence of new classes at the test phase, while in Generalized Zero-Shot Learning (GZSL) settings, the model predicts both unseen and seen classes at the test time; hence, GZSL is more applicable for real-world scenarios. In the next paragraphs, we discuss different types of training approaches.

Inductive vs. Transductive Training Settings: There are two training approaches for feeding the training data: Inductive and Transductive. The inductive approach only uses the seen class of information to learn a new concept, whereas in transductive learning, either unlabelled visual or textual information, or both for unseen classes are being used together with the seen class data, at the training phase. The training data for inductive learning is $\{(x, y, c(y)) | x \in X^S, y \in Y^S, c(y) \in C^S\}$, where x represents image features, y is the class labels, and $c(y)$ denotes the class embeddings. Moreover, X^S and Y^S indicate seen class images and seen class labels respectively. Inductive learning accounts for the majority of the settings used in ZSL and Generalized Zero-Shot Learning (GZSL). i.e. in [5], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19].

Although the original idea of zero-shot learning is more related to the inductive setting, however, in many scenarios, the transductive setting is frequently used. i.e. in [20], [21], [22], [23], [24], [25], [13], [16], [17], [18], [26], [27], [19], [28]. The training data for transductive learning is $\{(x, y, c(y)) | x \in X^{S \cup U}, y \in Y^{S \cup U}, c(y) \in C^{S \cup U}\}$ where $X^{S \cup U}$ denotes that images come from the union of seen and unseen classes. Similarly, $Y^{S \cup U}$ and $C^{S \cup U}$ indicate that train labels and class embeddings belong to both seen and novel categories.

According to [29], any approach that relies on label propagation will fall into the category of transductive learning. Feature generating network with labelled source data and unlabelled target data [19] are also considered as transductive methods. The transductive setting is seen as one of the solutions to the domain shift problem as the provided unseen labelled information during training reduces the discrepancy between the two domains.

There is a slight nuance between the transductive learning and semi-supervised learning in that with the transductive setting, the unlabelled data solely belong to the unseen test classes, while in semi-supervised setting, unseen test classes might not be present in unlabelled data.

2.1 Embedding Spaces

As shown in Figure 1, there are three main categories of zero-shot learning approaches. Such systems either map the visual data to the semantic space (Figure 1(a)) or embed both visual and semantic data to a common latent space

(Figure 1(b)), or see the task as a missing data problem and map the semantic information to the visual space (Figure 1(c)). Two or all of these approaches can be combined to boost up the benefits of each individual categories.

From a different point of view, semantic spaces can be categorized into euclidean and non-euclidean spaces. The intrinsic relationship between data points is better preserved when the geometrical relation between them is considered. These spaces are commonly based on clusters or graph networks. Many of the methods choose manifold learning for the ZSL challenge. i.e. [20], [30], [16], [25], [31], [32], [33], [34], [35], [36]. Euclidean spaces are more conventional and simpler as the data has a flat representation in such spaces. However, the loss of information is a common issue of these spaces, as well. Examples of methods using euclidean spaces are [4], [10], [12], [37], [38], and [39].

2.2 Side Information

Zero-shot learning is the challenge of learning novel classes without seeing their exemplars when training. Instead, freely available auxiliary information is used to compensate for the lack of visually labelled data. Such information can be categorized into two groups:

Human annotated attributes. The supervised way of annotating each image is an arduous process. There are sources in which side information in the form of attributes can be attained. i.e. aPY, AWA1, AWA2, CUB, and SUN which are attribute-based datasets. Several ZSL methods leverage attributes as their side information [5], [12], [40], or visual attributes [41], [42].

Unsupervised auxiliary information. There are several forms of auxiliary information that have minimum supervision and are widely used in the ZSL setting. i.e. human gazes [43], WordNet which is a large-scale lexical database of 1000 English words [44], [45], [46], [5], [47], [48], [49], [50], [51], [33], [34]. Textual descriptions like web search [45] or Wikipedia articles [10], [11], [52], [5], [53], [49], [50], [7], [54], [55], and sentence descriptions [56]. Textual side information needs to be transformed into class embeddings to be used at the training stage. Different class embeddings are discussed later in this paper.

Contributions. The contributions in this review paper are as follows. (1) As shown in Figure 1, We propose to categorize the reviewed approaches by the embedding spaces each model uses to learn/infer unseen labels and describe all of the variations to the embedding of data inside those spaces. (2) We report the evaluation of the state-of-the-art

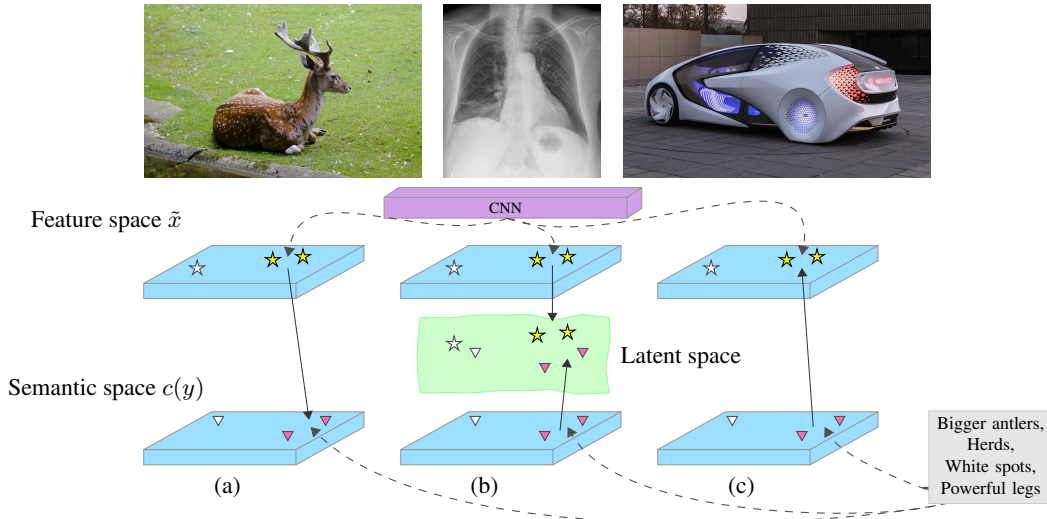


Figure 1: Overview of ZSL models. Typical approaches use one of the three embedding types or a combination of them. (a) Semantic embedding models that map visual features to the semantic space. (b) Models that map visual and semantic features to an intermediate latent space. (c) Visual embedding models that map semantic features to the visual space.

models on benchmark datasets. To the best of our knowledge, we are the first to include the evaluation of data-synthesizing methods in the survey of Zero-shot learning alongside other methods. (3) We study the motivation behind leveraging each space as a way to solve the ZSL challenge by reviewing current issues and solutions to them.

The rest of the paper is organized as follows. In Sec. 3 we review the state-of-the-art approaches to ZSL and categorize them into space-wise groups. We then conduct a survey over popular datasets used for this problem and the variety of different splits they have and the evaluation protocol in Sec. 4. Next, in Sec. 5, we report the overall results of recent approaches. Sec. 6 studies the main issues around the ZSL challenge, the superiority of each category, as well as discussing the possible motivations behind each method and approaches to solve specific problems. Sec. 7 introduces the extension of the problem into few-shot learning, and finally Sec. 8 reviews the applications and future work directions of the zero-shot learning setting.

3 ZSL EMBEDDING APPROACHES

In this section, we first provide the task definition for ZSL and GZSL, then we review the recent embedding methods, categorized into space-wise groups. Let $S = (x, y, c(y)) | x \in X, y \in Y^S, c(y) \in C$ be a training set. The goal is to learn the classifier $f : X \rightarrow Y^U$ for ZSL and $f : X \rightarrow Y^U \cup Y^S$ for the GZSL challenge.

The objective function to be minimized is as follows:

$$\frac{1}{N} \sum_{n=1}^N L(y_n, f(x_n; W)) + \Omega(W) \quad (1)$$

where $f(x, y; W) = \operatorname{argmax}_{y \in Y} F(x, y)$ is the mapping function.

3.1 Semantic Embedding

3.1.1 Attribute Classifiers

Primitive approaches of Zero-Shot learning leverage manually annotated attributes in a two-stage learning schema. Attributes in an image are predicted in the first stage and labels of unseen classes would be chosen using similarity measures in the second stage. [41] uses a probabilistic classifier to learn the attributes and then estimates posteriors for test classes. [44] proposes a method to avoid manual supervision with mining the attributes in an unsupervised manner. [45] adopts DAP together with a hierarchy-based knowledge transfer for large-scale settings. [57] is based on IAP, and uses Self-Organizing and Incremental Neural Networks (SOINN) to learn and update attributes online. Later in IAP-SS [57], an online incremental learning approach is used for faster learning of the new attributes. The Direct Attribute Prediction (DAP) model [4] first learns the posteriors of the attributes, then estimates the posteriors of unseen categories by summation of the learned probabilistic attributes:

$$f(x) = \operatorname{argmax}_{U=1, \dots, N_{y^U}} \prod_{m=1}^M \frac{p(a_m^{y^U} | x)}{p(a_m^{y^U})} \quad (2)$$

where the number of classes y^U and attributes a are N_{y^U} and M respectively. Here, $a_m^{y^U}$ is the m^{th} attribute of the class y^U , $p(a_m^{y^U} | x)$ is the estimated attribute via attribute classifier for image x , and $p(a_m^{y^U})$ is the prior attributes computed for training classes with the MAP. On the other hand, IAP [4] is an indirect approach as it first learns the posteriors for seen classes then uses them to compute the

posteriors for the attributes:

$$p(a_m|x) = \sum_{y^S=1}^{N_{y^S}} p(a_m|y^S)p(y^S|x) \quad (3)$$

where N_{y^S} is the number of training classes, $p(a_m|y^S)$ is the pre-trained attribute of the classes and $p(y^S|x)$ is probabilistic multi-class classifier to be learned. [58] uses a unified probabilistic model based on the Bayesian Network (BN) [59] that discovers and captures both object-dependent and object-independent relationships to overcome the problem of relating the attributes. CONSE [11] takes a probabilistic approach and predicts an unseen class by the convex combination of the class label embedding vectors. It first learns the probability of the training samples:

$$f(x, t) = \underset{Y \in y^S}{\operatorname{argmax}} p_S(y|x) \quad (4)$$

in which y is the most probable label for the training sample. It then computes a weighted combination of the semantic embedding to its probability to find a label for a given unseen image.

$$\frac{1}{Z} \sum_{i=1}^{N_T} p_S(f(x, t)|x) \cdot s(f(x, t)) \quad (5)$$

In this function, Z is the normalization factor and s combines N_T semantic vectors to infer unseen labels. [60] uses a random forest approach for learning more discriminative attributes. Hierarchy and Exclusion (HEX) [61] considers relations between objects and attributes and maps the visual features [62] of the images to a set of scores to estimate labels for unseen categories. [63] takes on an unsupervised approach where they capture the relations between the classes and attributes with a three-dimensional tensor while using a DAP-based scoring function to infer the labels. LAGO [64] also follows the DAP model. It learns soft and-or logical relations between attributes. Using soft-OR, the attributes are divided into groups, and the label class from unseen samples is predicted via a soft-AND within these groups. If each attribute comes from a singleton group, the all-AND will be used.

3.1.2 Label Embedding

Instead of using an intermediate step, more recent approaches learn to map images to the structured euclidean semantic space automatically which would be the implicit way of representing knowledge. The compatibility function for linear mapping is:

$$F(x, y; w) = \theta(x)^T w c(y) \quad (6)$$

where w is parameters in vector form to be learned. In the case of bilinear projection where it is more common, w takes the form of matrix:

$$F(x, y; W) = \theta(x)^T W c(y) \quad (7)$$

SOC [65] first maps the image features to the semantic embedding space, it then estimates the correct class using nearest neighbour. DeVise [10] uses linear corresponding function with a combination of dot-product similarity and hinge rank loss used in [66]. ALE [23] optimizes the ranking loss in [67] with a bilinear mapping compatibility function. The objective function used in ALE is similar to unregularized structured SVM (SSVM) [68].

$$\frac{1}{N} \sum_{n=1}^N \max_{y \in y^S} \Delta(y_n, y) + F(x_n, y; W) - F(x_n, y_n; W) \quad (8)$$

$F(\cdot)$ is the compatibility function, W is the matrix with dimensions of image and label embeddings and Δ is the loss of the mapping function. Although they have different losses, the inspiration comes from WSABIE algorithm [66]. In ALE, rank 1 loss with a multi-class objective is used instead of all of the weighted ranks. SJE [5] learns a bi-linear compatibility function using the structural SVM objective function [68]. ESZSL [12] introduces a better regularizer and optimizes a close form solution objective function in a linear manner. ZSLNS [50] proposes a $l_{1,2}$ -norm based loss function.[69] takes on a metric learning approach and linearly embeds the visual features to the attribute space. LAGO [64] is a probabilistic model that depicts soft and- or relations between groups of attributes. In a case where all attributes form all-OR group, It becomes similar to ESZSL [12] and learns a bilinear compatibility function. AREN [70] uses attentive region embedding while learning the bilinear mapping to the semantic space in order to enhance the semantic transfer. ZSLPP [7] combines two networks VPDE-net for detecting bird parts from images and PZSC-net that trains a part-based Zero-Shot classifier from the noisy text of the Wikipedia. DSRL [29] uses non-negative sparse matrix factorization to align vector representations with the attribute-based label representation vectors so that more relevant visual features are passed to the semantic space.

Some approaches to ZSL use non-linear compatibility functions. CMT [71] uses a two-layer neural network, similar to common MLP networks [72] that minimizes the objective function

$$\sum_{y \in y^S} \sum_{x^{(i)} \in X_y} \|c(y) - \theta^{(2)} \tanh(\theta^{(1)} x^{(i)})\|^2 \quad (9)$$

$\theta = (\theta^{(1)}, \theta^{(2)})$. In UDA [21] a non-linear projection from feature space to semantic space (word vector and attribute) is proposed in an unsupervised domain adaptation problem based on regularised sparse coding. [53] uses a deep neural network [73] regression which generates pseudo attributes for each visual category via Wikipedia. LATEM [48] constructs a piece-wise non-linear compatibility function alongside a ranking loss. To optimize the solution, they used the SGD-based method

$$F(x, y; W_i) = \max_{1 \leq i \leq K} \theta(x)^T W_i c(y) \quad (10)$$

i is the number of the latent indexes over each parameter W_i of linear components ($K \geq 2$). [74] regularizes the model using structural relations of the cluster by which cluster centres characterize visual features. QFSL [26] solves the problem in a transductive setting. It projects both source and target images into several specified points to fight bias problem.

GFZSL [18] introduces both linear and non-linear regression models in a generative approach as it produces a probability distribution for each class. It then uses MLE for estimating seen class parameters and two regression functions for unseen categories.

$$\mu_y = f_\mu(c(y)) \quad (11)$$

$$\sigma_y^2 = f_\sigma^2(c(y)) \quad (12)$$

where μ is the Gaussian mean vector and σ is the diagonal covariance matrix of the attribute vector. In its transductive setting, it uses Expectation-Maximization (EM) that works like estimation a Gaussian Mixture Model (GMM) of unlabelled data in an iterative manner. The inferred labels will be included in the next iterations.

Leveraging the non-euclidean spaces to capture the manifold structure of the data is another approach to the problem. Together with the knowledge graphs, the explicit relations between the labels will be demonstrated. In this setting, the side information mainly comes from a hierarchy ontology like WordNet. The mapping function will have the following form:

$$F(x, y; W) = \theta(X, A)^T W c(y) \quad (13)$$

where X is the $n \times k$ feature matrix and A is the adjacency matrix of graph.

Propagated Semantic Transfer (PST) [20] first uses DAP model to transfer knowledge to novel categories, following the graph-based learning schema, it improves local neighbourhood in them. DMaP [32] jointly optimizes the projecting of the visual features and the semantic space to improve the transferability of the visual features to the semantic space manifold. MFMR [31] decomposes the visual feature matrix into three matrices to further facilitate the mapping of visual features to the semantic spaces. To improve

the representation of the geometrical manifold structure of the visual and semantic features, manifold regularization is used. In [34] a Graph Search Neural Network (GSNN) [51] is used in the semantic space based on the WordNet knowledge graph to predict multiple labels per image using the relations between them. [33] distils both auxiliary information in forms of word embedding and knowledge graph to learn novel categories. DGP [35] proposes dense graph propagation to propagate knowledge directly through dense connections. In [36] a graphical model with a low dimensional visually semantic space is utilized which has a chain-like structure to close the gap between the high-dimensional features and the semantic domain.

3.2 Intermediate-Space Embedding

Measuring the similarity in a joint space to the visual and semantic features is another approach.

3.2.1 Fusion-based Models

Considering unseen classes as a fusion of previously learned seen concepts is called hybrid learning. Standard scoring function for hybrid models is:

$$f(x, y; W) = \sum_{s \in S} (W, \theta_s(X)) c(y) \quad (14)$$

SSE [14] learns two embedding functions, one being ψ which is learned from seen class auxiliary information and the other one from seen data which is target class π embedding and predicts unseen labels via maximizing the similarity between histograms:

$$\operatorname{argmax}_{y \in y^u} \pi(\theta(x))^T \psi(c(y)) \quad (15)$$

In SYNC [15] the mapping is between the semantic space of the external information and the model space. They introduced phantom classes to align the two spaces. The classifier is trained with the sparse linear combination of the classifiers for the phantom classes:

$$\min_{w_c, v_r} \|w_c - \sum_{r=1}^R s_{cr} v_r\|_2^2 \quad (16)$$

where w_c and v_r are weighted graphs of the real and phantom classes respectively. While s_{cr} is the bipartite graph of those to previously graph combinations. TVSE [25] learns a latent space using collective matrix factorization with graph regularization to incorporate the manifold structure between source and target instances, moreover, it represents each sample as a mixture of seen class scores. LDF [75] combines the prototypes of seen classes and jointly learns embeddings for both user-defined attributes and latent attributes.

3.2.2 Joint Representation Space Models

Inferring unseen labels via measuring similarity between cross-modal data in a shared latent space is another workaround to the ZSL challenge. The first term in the objective function for standard cross-modal alignment approaches is:

$$\min_{c(y)_s} \|X_s - c(y)_s Y_s\|_F^2 \quad (17)$$

with Y being a one-hot vector of corresponding class labels and $\|\cdot\|_F^2$ is the Frobenius norm. Approaches to joint space learning are grouped into two categories, Parametric which follows a slow learning via optimizing a problem and Non-parametric that leverage data points extracted from neural networks in a shared space. In parametric methods including [22] a multi-view alignment space is proposed for embedding low-level visual features. The learning procedure is based on the multi-view Canonical Correlation Analysis (CCA) [76]. [47] applies PCA and ICA embeddings to reveal the visual similarity across the classes and obtains the semantic similarity with the WordNet graph, followed by embedding the two outputs into a common space. MCZSL [49] combines compatibility learning with Deep Fragment embeddings [77] in a joint space. Their visual part and multi-cue language embedding are defined as follows, respectively:

$$\theta_i = E^{\text{visual}}[\text{CNN}_{\theta}(I_b) + b^{\text{visual}}] \quad (18)$$

$$c(y)_j = f\left(\sum_m E_m^{\text{language}} l_m + b^{\text{language}}\right) \quad (19)$$

In this equation, l_m and E_m^{language} are the language encoder for each modality. $f(\cdot)$ is the language token from the m modality and ReLU, respectively. Also, E^{visual} is the visual encoder and $\text{CNN}_{\theta}(I_b)$ is the part descriptor extracted from bounding box I_b for the image part annotation b . Hence the complete objective function is as follows:

$$\sum_i \sum_j \max(0, 1 - y_{ij} \theta_i^T c(y)_j) + \alpha \|w\|_2^2 \quad (20)$$

where w is the parameters of the two encoders and α is the hyperparameter.

In [78] both images and words are represented by Gaussian distribution embeddings. JLSE [24] decides on a dictionary learning approach to learn the parameters of source and target domains across two separate latent spaces where the similarity is computed by the likelihood of similarity independent to the class label. CDL [79] uses a coupled dictionary to align the structure of visual-semantic space using discriminative information of the visual space. In [80] and [81] a coupled sparse dictionary is leveraged to relate visual and attribute features together. It uses entropy regularization to alleviate the domain shift problem.

There are several non-parametric methods. ReViSE [82] that combines auto-encoders with Maximum Mean Discrepancy (MMD) loss [83] in order to align the visual and textual features. DMAE [84] introduces a latent alignment matrix with representations from auto-encoders optimized by kernel target alignment (KTA) [85] and squared-loss mutual information (SMI) [86]. DCN [87] proposes a novel Deep Calibration Network in which an entropy minimization principle is used to calibrate the uncertainty of unseen classes as well as seen classes.

To narrow the semantic gap, BiDiLEL [88] introduces a sequential bidirectional learning strategy and creates a latent space using the visual data, then the semantic representations of unseen classes are embedded in the previously created latent space. This method comprises both parametric and non-parametric models.

3.3 Visual Embedding

Visual embedding is the other type of ZSL methods that performs classification in the original feature space and is orthogonal to semantic space projection. This is done by learning a linear or non-linear projection function. For linear corresponding functions, WAC-Linear [52] uses textual description for seen and unseen categories and projects them to the visual feature space with a linear classifier. [16] follows a transductive setting in which it refines unseen data distributions using unseen image data. To approximate manifold structure of data, they used a global linear mapping for synthesizing virtual cluster centres. [13] assigns pseudo labels to samples using reliability (with robust SVM) and diversity (via diversity regularization). For learning a Non-linear corresponding function, In WAC-Kernel [89] in order to leverage any kind of side information, a kernel method is proposed to predict a kernel-based on the representer theorem [90]. DEM [91] uses the least square embedding loss to minimize the discrepancy between the visual features and their class representation embedding vector in the visual feature space. OSVE [92] reversely maps from attribute space to visual space then trains the classifier using SVM [93]. In [94] the authors introduce a stacked attention network that corporates both global and local visual features weighted by relevance along with the semantic features. In [27] visual constraint is used in class centres in the visual space to avoid the domain shift problem.

3.3.1 Visual Data Augmentation

There are a variety of generative networks that augment unseen data, taking GAN [95] as an example, the first term in

objective function would be:

$$\max E[\log D(x, c(y))] + \min E[\log(1 - D(\tilde{x}, c(y)))] \quad (21)$$

$\tilde{x} = G(z, c(y))$ is the synthesized data of the generator and $z \in R^{d_z}$ is a random Gaussian noise. The role of discriminator D and generator G contradicts in loss function as the first one attempts to maximize the loss while the latter tries to minimize it.

RKT [30] leverages relational knowledge of the manifold structure in the semantic space to generate virtually labelled data for unseen classes from Gaussian distributions generated from sparse coding. Then it uses them alongside the seen data and projected to the semantic space via a linear mapping. GLaP [17] generates virtual instances of an unseen class with the assumption that each representation obeys a prior distribution where one can draw samples from. To ease the embedding to the semantic space, GANzrl [96] proposes to increase the visual diversity by generating samples with specified semantics using GAN models. SE-GZSL [97] uses a feedback-driven mechanism for their discriminator that learns to map the produced images to the corresponding class attribute vectors. To enforce the similarity of the distribution of the sample and generated sample, a loss component was added to the VAE objective [98] function:

$$\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z)) - \mathbb{E}_{\hat{x}} KL[q_\phi(z|\hat{x})||q(z)] \quad (22)$$

where $q_\phi(z|x)$ is the encoder and $p_\theta(x|z)$ is a decoder. z corresponds to a random unstructured component from the prior $p(z)$ and $q(z)$ can be either the prior $p(z)$ or the posterior of a labelled sample $p(z|x)$.

Synthesized images often suffer from looking unrealistic since they lack intricate details. A way around this issue is to generate features instead. [99] uses a GMMN model [100] to generate visual features for unseen classes. In [101] a multi-modal cycle consistency loss is used in training the generator for better reconstruction of the original semantic features. CVAE-ZSL [38] takes attributes and generates features for the unseen categories via a Conditional Variational Autoencoder (CVAE) [102]. L_2 norm is used as the reconstruction loss. GAZSL [55] utilizes noisy textual descriptions from Wikipedia to generate visual features. A visual pivot regularizer is introduced to help generate features with better qualities:

$$\Omega = \frac{1}{N} \sum_{n=1}^N \|\mathbb{E}_{\tilde{x} \sim p_g}[\tilde{x}] - \mathbb{E}_{x \sim p_{data}}[x]\|^2 \quad (23)$$

Due to the inaccessibility of data, empirical expectations $\mathbb{E}_{x \sim p_g}[\hat{x}] = \frac{1}{N} \sum_{i=1}^{N_S} x^i$ and $\mathbb{E}_{x \sim p}[\tilde{x}] =$

$\frac{1}{N^U} \sum_{i=1}^{N^U} G_\theta(T_U, z_i)$ are used instead; where N^S and N^U are the number of samples in class y^S and number of synthesized features in class y^U , respectively. f-CLSWGAN [37] combines three conditional GAN variants.

$$\min_G \max_D \mathcal{L}_{WGAN} + \beta \mathcal{L}_{CLS}$$

The classification loss is like a regularizer for the enhancement of the generated features and β is a hyperparameter.

$$\mathcal{L}_{WGAN} = E[D(x, c(y))] - E[D(\tilde{x}, c(y))] - \lambda E[(\|\nabla_{\hat{x}} D(\hat{x}, c(y))\|_2 - 1)^2] \quad (24)$$

and

$$\mathcal{L}_{CLS} = -E_{\tilde{x} \sim p_{\tilde{x}}}[\log P(y|\tilde{x}; \theta)] \quad (25)$$

This model adds $c(y)$ to both generator and discriminator. The first two terms are the Wasserstein distance and the third term is the gradient penalty of D to have a unit norm between pairs of real and generated features. λ and α are the penalty coefficient and a hyperparameter respectively. $\hat{x} = \beta x + (1 - \beta)\tilde{x}$ with β being a random number. f-VAEGAN-D2 [19] combines the architectures of conditional VAE [102], GAN [95] and a non-conditional discriminator for the transductive setting. LisGAN [103] generates unseen features from random noises using conditional Wasserstein GANs [104]. For regularization, they introduced semantically meaningful soul samples for each class and forced the generated features to be close to at least one of the soul samples. Gradient Matching Network (GMN) [28] trains an improved version of the conditional WGAN [105] to produce image features for the novel classes. It also introduces Gradient Matching (GM) loss to improve the quality of the synthesized features. In order to synthesize unseen features, SPF-GZSL [106] selects similar instances and combines them to form pseudo features using a centre loss function [107]. In Don't Even Look Once (DELO) [108] a detection algorithm is conducted to synthesize unseen visual features to gain high confidence predictions for unseen concepts while maintaining low confidence for backgrounds with vanilla detectors.

Instead of augmenting data using synthesizing methods, data can be acquired by gathering web images. [54] jointly uses web data which are considered weakly-supervised categories alongside the fully-supervised auxiliary labelled categories. It then learns a dictionary for the two categories.

3.4 Hybrid Models

Several works make use of both visual and semantic projections to reconstruct better semantics to confront domain shift issue by alleviating the contradiction between the two domains. Semantic AutoEncoder (SAE) [109] adds a visual feature reconstruction constraint. It combines linear

visual-to-semantic (encoder) and linear semantic-to-visual (decoder). SP-AEN [110] is a supervised Adversarial Autoencoder [111] which improves preserving the semantics by reconstructing the images from the raw 256 x 256 x 3 RGB colour space. BSR [112] uses two different semantic reconstructing regressors to reconstruct the generated samples into semantic descriptions. CANZSL [113] combines feature-synthesis with semantic embedding by using a GAN for generating visual features and an inverse GAN to project them into semantic space. In this way, the produced features are consistent with their corresponding semantics.

Some of the synthesizing approaches utilize a common latent space to align the generated features space with the semantic space to facilitate capturing the relations between the two spaces. [40] introduces a latent-structure-preserving space where synthesized features from given attributes would suffer less from bias and variance decay with the help of Diffusion Regularisation. CADA-VAE [39] learns latent space features and class embedding by training VAE [98] for both visual and semantic modalities. used Cross-Alignment (CA) Loss to align latent distributions in cross-modal reconstruction:

$$\mathcal{L}_{CA} = \sum_i^M \sum_{j \neq i}^M |x^{(j)} - D_j(E_i(x^{(i)}))| \quad (26)$$

Here, i and j are two different modalities. Wasserstein distance [114] is used between the latent distributions i and j to align the Latent distribution (LDA):

$$\mathcal{L}_{DA} = \sum_i^M \sum_{j \neq i}^M W_{ij} \quad (27)$$

where $W_{ij} = (\|\mu_i - \mu_j\|_2^2 + \|\eta_i^{\frac{1}{2}} - \eta_j^{\frac{1}{2}}\|_{\text{Frobenius}}^2)^{\frac{1}{2}}$. μ and η are predictions of the encoder. GDAN [115] combines all three approaches and designs a dual adversarial loss so for regressor and discriminator to learn from each other.

A summary of the different approaches is reported in Table 1. The number of methods are growing with time and we can interpret that some areas like direct learning, common space learning and visual data synthesizing are more popular in solving the task, while models combining different approach are fairly newer techniques thus have fewer works that are reported here.

4 EVALUATION PROTOCOLS

4.1 Datasets

There are several well-known benchmark datasets for Zero-shot learning which can be categorized into attribute datasets and ImageNet, which is a WordNet hierarchy

dataset.

Attribute datasets. SUN Attribute [116] is a medium-scale and fine-grained attribute database consisting 102 attributes, 717 categories and a total of 14,340 images of different scenes. CUB-200-2011 Birds (CUB) [117] is a 200 category fine-grained attribute dataset with 11,788 images of bird species that includes 312 attributes. Animals with Attributes (AWA1) [4] is another attribute dataset of 30,475 images with 50 categories and 85 attributes, the image features in this dataset are licensed and not available publicly. later, Animals with Attributes2 (AWA2) is presented by [6] which is a free version of AWA1 with more images than the previous one (37,322 images), and the same number of classes and attributes but different images. aPascal and Yahoo (aPY) [42] is a dataset with a combination of 32 classes, including 20 pascal and 12 yahoo attribute classes with 15,339 images and 64 attributes in total. North America Birds (NAB) [118] is another fine-grained dataset of birds consisting of 1,011 classes and 48,562 images. A new version of this dataset is proposed by [7] in which the identical leaf nodes are merged to their parent nodes where their only differences were genders and resulted in final 404 classes. Summaries of the statics for the attribute datasets are gathered in Table 2.

ImageNet. ImageNet [119] is a large-scale dataset that contains 14 million images, shared between 21k categories with each image having one label that makes it a popular benchmark to evaluate models in real-world scenarios. Its organization is based on WordNet hierarchy[120]. ImageNet is imbalanced between classes as the number of samples in each class vary greatly and is partially fine-grained. A more balanced version has 1k classes with 1000 images in each category.

4.2 Dataset Splits

Here we discuss the original splits of the datasets as well as the other splits proposed for the Zero-shot problem.

Standard Splits (SS). In ZSL problems, unseen classes should be disjoint to seen classes and test time samples limited to unseen classes, thus the original splits aim to follow this setting. SUN [4] proposed to use 645 classes for training among which 580 of the classes are used for training, 65 classes are for validation and the remaining 72 classes will be used for testing. For CUB, [23] introduces the split of 150 training classes (including 50 validation classes) and 50 test classes. As for AWA1, [4] introduced the standard split of 40 classes for training (13 validation classes) and 10 classes for testing. The same splits are used for AWA2. In aPY, 20 classes of Pascal are used for training (15 classes for training and 5 for validation), while the 12 classes of Yahoo are used for testing.

Proposed Splits (PS). The standard split images from

Table 1: Common ZSL and GZSL methods categorised based on their embedding space model, with further divisions in a top-down manner.

Models	Categories	Main Features	Description
Semantic Embedding	Two-Step Learning	Attributes classifiers	DAP-Based [41], [44], [45], [4], [63], [64] IAP-Based [41], [57], [57], [4], [11] Bayesian network (BN)[58], Random Forest Model [60], HEX Graph [61]
	Direct Learning	Implicit knowledge representation	Linear [65], [10], [23], [5], [12], [50], [69], [64], [70], [7], [29], [18] or Non-Linear [21], [53], [48], [74], [26], [18] compatibility Functions
		Explicit knowledge representation	Graph Convolutional Networks (GCN) [33], Knowledge Graphs [34], [20], [32], [35], 3-Node Chains [36], Matrix Tri-Factorization with Manifold Regularization [31]
Cross-Modal Latent Embedding	Fusion-based Models	Fusion of seen class data	Combination of seen classes properties [14], [15], [75], Combination of seen class scores [25]
	Common Representation Space Models	Mapping of the visual and semantic spaces in a joint intermediate space	Parametric [22], [47], [49], [78], [24], [79], [80], [81], Non-parametric [82], [84], [87], or Both [88]
Visual Embedding	Visual Space Embedding	Learning of the semantic to visual projection	Linear [52], [16], [13] or Non-linear [89], [91], [92], [94], [27] projection functions
	Data Augmentation	Image generation	Gaussian distribution [30], [17], GAN [96], VAE [97]
		Visual feature generation	GAN [101], [55], [103], WGAN [37], [28], CVAE [38], [108], VAE+GAN [19], GMMN [99], Similar feature combination [106]
	Leveraging Web Data	Web images crawling	Dictionary learning [54]
Hybrid	Visual + Semantic Embedding	Reconstruction of the semantic features	Autoencoder [109], Adversarial Autoencoder [110], GAN with two reconstructing regressors [112], GAN an inverse GAN [113]
	Visual+Cross Modal Embedding	Feature generation with aligned semantic features	Semantic to visual mapping [40], VAE [39]
	All	The use of generator and discriminator together with the regressor	GAN + Dual Learning [115]

SUN, CUB, AWA1 and aPY overlap with some images of pre-trained ResNet-101 ImageNet model. To solve the problem, proposed splits (PS) is introduced by [121] where no test images are contained in the ImageNet 1K dataset. **ImageNet.** [121] proposes 9 ZSL splits for the ImageNet dataset; two of which evaluate the semantic hierarchy in distance-wise scales of 2-hops (1509 classes) and 3-hops (7678 classes) from the 1k training classes. The remaining six splits consider the imbalanced size of classes with increasing granularity splits starting from 500, 1K and 5K

least-populated classes to 500, 1K and 5K most-populated classes, or All which denotes a subset of 20k other classes for testing.

Seen-Unseen relatedness. To measure the relatedness of seen samples to unseen classes, [7] introduces two splits Super-Category-Shared (SCS) and Super-Category-Exclusive (SCE). SCS is the easy split since it considers the relatedness to the parent category while SCE is harder and measures the closeness of an unseen sample to that particular child node.

Table 2: Statics of the attribute datasets accounting for the number of attributes, classes plus their splits and their total number of images.

Attribute Datasets	#attributes	y	y^U	y^S	#images
SUN [4]	102	717	580+65	72	14,340
CUB [117]	312	200	100+50	50	11,788
NAB [118]	-	1,011	-	-	48,562
AWA1 [4]	85	50	27+13	10	30,475
AWA2 [6]	85	50	27+13	10	37,322
aPY [42]	64	32	15+5	12	15,339

4.3 Class Embeddings

There exist several class embeddings, each suitable for a specific scenario. Class embeddings are in forms of vectors of real numbers which can further be used to make predictions based on the similarity between them and can be obtained through three categories: attributes, word embeddings, and hierarchical ontology. The last two are done in an unsupervised manner thus do not require human labour.

4.3.1 Supervised Attribute-Embeddings

Human annotated attributes are done under the supervision of experts with a great amount of effort. Binary, relative and real-valued attributes are three types of attributes embeddings. Binary attributes depict the presence of an attribute in an image thus value is either 0 or 1. They are the easiest type and are provided in benchmark attribute datasets AWA1, AWA2, NAB, CUB, SUN, aPY. Relative attributes [122] on the other hand, show the strength of an attribute in a given image comparing to the other images. The real-valued attributes are in continuous form thus they have the most quality [5]. In attribute datasets, they have achieved confidence through averaging the binary labels from multiple annotators [116].

4.3.2 Unsupervised Word-Embeddings

Textual corpora embedding. Bag of Words (BOW) [123] is a one-shot encoding approach. It simply shows the number of occurrences of the words in a representation called bag and is negligent of word orders and grammar. One-shot encoding approaches had a drawback of giving the stopwords (like "a", "the" and "of") high relevancy counts. Later Term Frequency-Inverse Document Frequency (TF-IDF): [124] used term weighting to alleviate this problem by filtering the stopwords and to keep meaningful words. Word2Vec [125], a widely used two-layered neural embedding model and is divided into two variants CBOW and skip-gram. CBOW predicts a target word in the centre of a context using its surroundings while the skip-gram model predicts surrounding words using a target word. CBOW is

faster in train and usually results in better accuracy for frequent words while Skip-gram is preferred for rare words and it works well with sparse training data. Global Vectors (GloVe) [126] is trained by Wikipedia. It combines local context window methods and global matrix factorization. Glove learns to consider global word-word co-occurrence matrix statistics to build the word embeddings.

Word hierarchy embedding. WordNet [120] is a large-scale lexical database of semantical synsets (grouped synonym) of English words that are organized using the hierarchy distances. Approaches based on knowledge graphs often follow the WordNet.

In this article, we report the results of ZSL and GZSL using the same class embeddings as [121] that is Word2Vec trained on Wikipedia for ImageNet and per-class attributes for the attribute datasets, and for the seen-unseen relatedness task we follow [7] and consider TF-IDF for the CUB and NAB datasets.

4.4 Image Embeddings

Existing models use either shallow or deep feature representation. Examples of shallow features are SIFT [127], PHOG [128], SURF [129] and local self-similarity histograms [130]. Among the mentioned features, SIFT is the commonly used features in ZSL models like [23], [15] and [22].

Deep features are obtained from deep CNN architectures [73] and contain higher-level features. Extracted features are one of the followings: 4,096-dim top-layer hidden unit activations (fc7) of the AlexNet [131], 1000-dim last fully connected layer (fc8) of VGG-16 [132], 4,096-dim of the 6th layer (fc6) and 4,096-dim of the last layer (fc7) features of the VGG-19 [132]. 1,024-dim top-layer pooling units of the GoogleNet [133]. and 2048-dim last layer pooling units of the ResNet-101 [134].

In this paper, we consider the ResNet-101 network which is pre-trained on ImageNet-1K without any fine-tuning. That is the same image embedding used in [121]. Features are extracted from whole images of SUN, CUB, AWA1, AWA2, and ImageNet and the cropped bounding boxes of aPY. For the seen-unseen relatedness task, VGG-16 is used for CUB and NAB as proposed in [7].

4.5 Evaluation Metrics

Common evaluation criteria used for ZSL challenge are:

Classification accuracy. One of the simplest metrics is classification accuracy in which the ratio of the number of the correct predictions to samples in class c is measured.

However, it results in a bias towards the populated classes.

Average per-class accuracy. To reduce the bias problem for the populated classes, average per-class accuracies computed by multiplying the division of the classification accuracy to division of their cumulative sum.

$$acc_y = \frac{1}{\|y\|} \sum_{y=1}^{\|y\|} \frac{\# \text{correct predictions in class } y}{\# \text{samples in class } y} [6] \quad (28)$$

Harmonic mean. For performance evaluation on both seen and unseen classes (i.e. the GZSL setting), the Top-1 accuracies for the seen and unseen classes are used to compute the harmonic mean:

$$H = \frac{2 * acc_{y^S} * acc_{y^U}}{acc_{y^S} + acc_{y^U}} [6] \quad (29)$$

In this paper, we designate the Top-1 accuracies and the harmonic mean as the evaluation protocols.

5 EXPERIMENTS

In this section, first, we provide the results for ZSL, GZSL and seen-unseen relatedness on attribute datasets, then we present the experimental results on the ImageNet dataset. A minor part of the results is reported from [6] for a more comprehensive comparison.

5.1 Zero-Shot Learning Results

For the original ZSL task where only unseen classes are being estimated during the test time, we compare 21 state-of-the-art models in Table 3, among which, DAP [4], IAP [4] and CONSE [11] belong to attribute classifiers. CMT [71], LATEM [48], ALE [23], DEVISE [10], SJE [5], ESZSL [12], GFZSL [18] and DSRL [29] are from compatibility learning approaches, SSE [14] and SYNC [15] are representative models of cross-modal embedding, DEM [91], GAZSL [55], f-CLSWGAN [37], CVAE-ZSL [38], SE-ZSL [97] are visual embedding models. From the hybrid or combination category, we compare the results of SAE [109]. Three transductive approaches ALE-tran [23], GFZSL-tran [18] and DSRL [29] are also presented among the selected models. Due to the intrinsic nature of the transductive setting, the results are competitive and in some cases better than the inductive methods, i.e. for GFZSL-tran [18] the accuracy is 9.9% higher than CVAE-ZSL [38] for PS split of AWA1 dataset. However, in comparison with the inductive form of the same model, there are cases where the inductive model has better accuracies. i.e. in PS split of the aPY dataset, the performance is 38.4% vs 37.1% or for ALE-tran [23] model in PS split of SUN it's 58.1% vs

55.7%, also for PS split of CUB it is 54.9% vs 54.5% with its inductive type. GFZSL [18], a compatibility-based approach, has the best scores compared to other models of the same category in every dataset except for the CUB where SJE [5] tops the results in both splits. This superiority could be due to the generative nature of the model. GFZSL [18] performs the best on AWA1 both in inductive and transductive settings. Out of cross-modal methods, SYNC [15] performs better than SSE [14] in SUN and CUB datasets, while for AWA1, AWA2 and aPY in SS split it has lower performance than SSE [14] in the proposed split. Visual generative methods have proved to perform better as they make the problem into the traditional supervised form, among which, SE-ZSL [97] has the most outstanding performance. For the proposed split in one case on CUB dataset, SE-ZSL [97] performs better than ALE-tran [23] which is its transductive counterpart where the accuracies are 59.6% vs 54.5%. In PS split of AWA1, CVAE-ZSL [38] stays at the top, with 1.9% higher accuracy than the second-best performing model. The accuracies for SS splits are higher than PS in most cases and the reason could be the test images included in training samples, especially for AWA1 and AWA2, as reported in [121].

5.2 Generalized Zero-Shot Learning Results

A more real-world scenario where previously learned concepts are estimated alongside new ones is necessary to experiment. 21 state-of-the-art models, same as with ZSL challenge, include: DAP [4], IAP [4], CONSE [11], CMT [71], SSE [14], LATEM [48], ALE [23], DEVISE [10], SJE [5], ESZSL [12], SYNC [15], SAE [109], GFZSL [18], DEM [91], GAZSL [55], f-CLSWGAN [37], CVAE-ZSL [38], SE-GZSL [97], ALE-tran [23], GFZSL-tran [18], DSRL [29]. CADA-VAE [39] is added to the comparison as a model combining the visual feature augmentation approach with the cross-modal alignment. CMT* [71] has a novelty detection and is included in the report as an alternative version to CMT [71]. The reports in Table 4 are in PS splits. As shown in the table, the results on y^S are dramatically higher than y^U since in GZSL, the test search space includes seen classes as well as unseen classes, this gap is the most conspicuous in attribute classifiers like DAP [4] that performs poorly on AWA1 and AWA2, hybrid approaches and in GFZSL [18] where it results in 0% accuracy on SUN and CUB when training classes are estimated at test time. However for three models f-CLSWGAN [37], SE-GZSL [97] and CADA-VAE [39] in SUN dataset, the accuracy for y^U is higher than y^S , i.e. for SE-GZSL [97] it is 10.4% higher. For a fair comparison, the weighted average of training and test classes is also reported. According to harmonic means, the best model on all evaluated datasets is SE-ZSL [97], although the results haven't been reported

Table 3: Zero-shot learning results for the Standard Split (SS) and Proposed Split (PS) on SUN, CUB, AWA1, AWA2, and aPY datasets. We measure Top-1 accuracy in % for the results. † and ‡ denote inductive and transductive settings respectively.

Methods	SUN		CUB		AWA1		AWA2		aPY	
	SS	PS	SS	PS	SS	PS	SS	PS	SS	PS
DAP [4]	38.9	39.9	37.5	40.0	57.1	44.1	58.7	46.1	35.2	33.8
IAP [4]	17.4	19.4	27.1	24.0	48.1	35.9	46.9	35.9	22.4	36.6
CONSE [11]	44.2	38.8	36.7	34.3	63.6	45.6	67.9	44.5	25.9	26.9
CMT [71]	41.9	39.9	37.3	34.6	58.9	39.5	66.3	37.9	26.9	28.0
SSE [14]	54.5	51.5	43.7	43.9	68.8	60.1	67.5	61.0	31.1	34.0
LATEM [48]	56.9	55.3	49.4	49.3	74.8	55.1	68.7	55.8	34.5	35.2
ALE [23]	59.1	58.1	53.2	54.9	78.6	59.9	80.3	62.5	30.9	39.7
DEVISE [10]	57.5	56.5	53.2	52.0	72.9	54.2	68.6	59.7	35.4	39.8
† SJE [5]	57.1	53.7	55.3	53.9	76.7	65.6	69.5	61.9	32.0	32.9
ESZSL [12]	57.3	54.5	55.1	53.9	74.7	58.2	75.6	58.6	34.4	38.3
SYNC [15]	59.1	56.3	54.1	55.6	72.2	54.0	71.2	46.6	39.7	23.9
SAE [109]	42.4	40.3	33.4	33.3	80.6	53.0	80.7	54.1	8.3	8.3
GFZSL [18]	62.9	60.6	53.0	49.3	80.5	68.3	79.3	63.8	51.3	38.4
DEM [91]	-	61.9	-	51.7	-	68.4	-	67.1	-	35.0
GAZSL [55]	-	61.3	-	55.8	-	68.2	-	68.4	-	41.1
f-CLSWGAN [37]	-	60.8	-	57.3	-	68.8	-	68.2	-	40.5
CVAE-ZSL [38]	-	61.7	-	52.1	-	71.4	-	65.8	-	-
SE-ZSL [97]	64.5	63.4	60.3	59.6	83.8	69.5	80.8	69.2	-	-
‡ ALE-tran [23]	-	55.7	-	54.5	-	65.6	-	70.7	-	46.7
‡ GFZSL-tran [18]	-	64.0	-	49.3	-	81.3	-	78.6	-	37.1
‡ DSRL [29]	-	56.8	-	48.7	-	74.7	-	72.8	-	45.5

for aPY. In some cases, the attribute classifier achieves the best results on y^S . Transductive models have fluctuating results in comparison with their inductive types. CADA-VAE [39] achieves the best performance in all of the harmonic means cases (results for aPY are not reported) and shows the best results, higher than all of the transductive methods.

5.3 Seen-Unseen Relatedness Results

For fine-grained problems, sometimes it is important to measure the closeness of previously known concepts to novel unknown ones. For this purpose, a total of 11 models are compared in Table 5. MCZS [49], WAC-Linear [52], WAC-Kernel [89], ESZSL [12], SJE [5], ZSLNS [50], SynC_{fast} [15], SynC_{OVO} [15], ZSLPP [7], GAZSL [55] and CANZSL [113]. SCE is the hard split thus has lower results compared to the SCS splits. The two variations reported for SYNC [15] model, SynC_{fast} denotes the setting in which the standard Crammer-Singer loss is used, and SynC_{fast} [15] depicts setting with one-versus-other classifiers. The first setting has better accuracies on CUB. CANZSL [113] outperforms all other models in both datasets and splits and improves the accuracy by 4% from

10.3% to 14.3% on SCE split of the CUB dataset and 35.6% vs 38.1% in SCS splits of NAB compared to the next best performing model is GAZSL [55]. Similar to previous experiments, in the seen-unseen relatedness challenge, models that contain feature generating steps have the highest results.

5.4 Zero-Shot Learning Results on ImageNet

ImageNet is a large-scale single-labelled dataset with an imbalanced number of data that possesses WordNet hierarchy instead of human-annotated attributes, thus is useful mean to measure the performance of various methods in recognition-in-the-wild scenarios. The performances of 12 state-of-the-art models are reported here. They are CONSE [11], CMT [71], LATEM [48], ALE [23], DEVISE [10], SJE [5], ESZSL [12], SYNC [15], SAE [109], f-CLSWGAN [37], CADA-VAE [39] and f-VAEGAN-D2 [19]. All of the Top-1 accuracies, except for the data generating models are reported from [121] experiments. As can be understood from Figure 2a, Feature generating methods have outstanding performance compared to other approaches. Although the results of f-VAEGAN-D2 [19] are available only for 2H, 3H and all splits, it still has the high-

Table 4: Generalized Zero-Shot Learning results for the Proposed Split (PS) on SUN, CUB, AWA1, AWA2, and aPY datasets. We measure the Top-1 accuracy in % for seen (S), unseen (U) and their harmonic mean (H). † and ‡ denote inductive and transductive settings, respectively.

Methods	SUN			CUB			AWA1			AWA2			aPY		
	y^U	y^S	H	y^U	y^S	H	y^U	y^S	H	y^U	y^S	H	y^U	y^S	H
DAP [4]	4.2	25.1	7.2	1.7	67.9	3.3	0.0	88.7	0.0	0.0	84.7	0.0	4.8	78.3	9.0
IAP [4]	1.0	37.8	1.8	0.2	72.8	0.4	2.1	78.2	4.1	0.9	87.6	1.8	5.7	65.6	10.4
CONSE [11]	6.8	39.9	11.6	1.6	72.2	3.1	0.4	88.6	0.8	0.5	90.6	1.0	0.0	91.2	0.0
CMT [71]	8.1	21.8	11.8	7.2	49.8	12.6	0.9	87.6	1.8	0.5	90.0	1.0	1.4	85.2	2.8
CMT* [71]	8.7	28.0	13.3	4.7	60.1	8.7	8.4	86.9	15.3	8.7	89.0	15.9	10.9	74.2	19.0
SSE [14]	2.1	36.4	4.0	8.5	46.9	14.4	7.0	80.5	12.9	8.1	82.5	14.8	0.2	78.9	0.4
LATEM [48]	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	11.5	77.3	20.0	0.1	73.0	0.2
ALE [23]	21.8	33.1	26.3	23.7	62.8	34.4	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7
DEVISE [10]	16.9	27.4	20.9	23.8	53.0	32.8	13.4	68.7	22.4	17.1	74.7	27.8	4.9	76.9	9.2
† SJE [5]	14.7	30.5	19.8	23.5	59.2	33.6	11.3	74.6	19.6	8.0	73.9	14.4	3.7	55.7	6.9
ESZSL [12]	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
SYNC [15]	7.9	43.3	13.4	11.5	70.9	19.8	8.9	87.3	16.2	10.0	90.5	18.0	7.4	66.3	13.3
SAE [109]	8.8	18.0	11.8	7.8	54.0	13.6	1.8	77.1	3.5	1.1	82.2	2.2	0.4	80.9	0.9
GFZSL [18]	0.0	39.6	0.0	0.0	45.7	0.0	1.8	80.3	3.5	2.5	80.1	4.8	0.0	83.3	0.0
DEM [91]	20.5	34.3	25.6	19.6	57.9	29.2	32.8	84.7	47.3	30.5	86.4	45.1	11.1	75.1	19.4
GAZSL [55]	21.7	34.5	26.7	23.9	60.6	34.3	25.7	82.0	39.2	19.2	86.5	31.4	14.2	78.6	24.1
f-CLSWGAN [37]	42.6	36.6	39.4	43.7	57.7	49.7	57.9	61.4	59.6	52.1	68.9	59.4	32.9	61.7	42.9
CVAE-ZSL [38]	-	-	26.7	-	-	34.5	-	-	47.2	-	-	51.2	-	-	-
SE-GZSL [97]	40.9	30.5	34.9	41.5	53.3	46.7	56.3	67.8	61.5	58.3	68.1	62.8	-	-	-
CADA-VAE [39]	47.2	35.7	40.6	51.6	53.5	52.4	57.3	72.8	64.1	55.8	75.0	63.9	-	-	-
ALE-tran [23]	19.9	22.6	21.2	23.5	45.1	30.9	25.9	-	-	12.6	73.0	21.5	8.1	-	-
‡ GFZSL-tran [18]	0	41.6	0	24.9	45.8	32.2	48.1	-	-	31.7	67.2	43.1	0.0	-	-
DSRL [29]	17.7	25.0	20.7	17.3	39.0	24.0	22.3	-	-	20.8	74.7	32.6	11.9	-	-

est accuracies among other models.

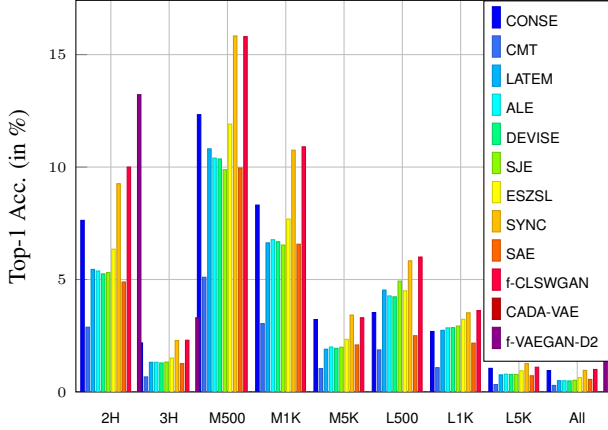
Table 5: Seen-Unseen relatedness results on CUB and NAB datasets with easy (SCS) and hard (SCE) splits. Top-1 accuracy is reported in %

Methods	CUB		NAB	
	SCS	SCE	SCS	SCE
MCZSL [49]	34.7	-	-	-
WAC-Linear [52]	27.0	5.0	-	-
WAC-Kernel [89]	33.5	7.7	11.4	6.0
ESZSL [12]	28.5	7.4	24.3	6.3
SJE [5]	29.9	-	-	-
ZSLNS [50]	29.1	7.3	24.5	6.8
SynC _{fast} [15]	28.0	8.6	18.4	3.8
SynC _{OVO} [15]	12.5	5.9	-	-
ZSLPP [7]	37.2	9.7	30.3	8.1
GAZSL [55]	43.7	10.3	35.6	8.6
CANZSL [113]	45.8	14.3	38.1	8.9

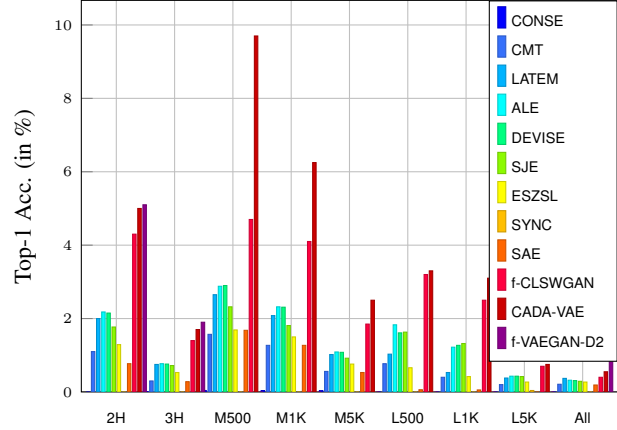
SYNC [15] and f-CLSWGAN [37] are the next best performing models with approximately the same accuracies. CONSE [11] is a representative model from attribute-classifier based models, as it is also superior to direct com-

patibility approaches. ESZSL [12], a model with linear compatibility function outperforms the other model within its category. However, in one case, SJE [5] has slightly better accuracy in L500 split setting. It can be interpreted from the figures that on coarse-grained classes, the results are conspicuously better, while fine-grained classes with few images per class have more challenges. However, if the test search space is too big then the accuracies decrease. i.e. M5K has lower accuracies compared to L500 splits, and on 20K split, it is the lowest.

The GZSL results are important in the way that they depict the models' ability to recognize both seen and unseen classes at the test time. The results for the SYNC [15] model is only reported in the L5K setting. As shown in Figure 2b, the trend is Similar to ZSL where populated classes have better results than the least populated classes, yet have poor results if the search spaces become too big like the decreasing trends in most and least populated classes. Moreover, data-generating approaches dominate other strategies. CADA-VAE [39] that has the advantages of both cross-modal alignment and data feature synthesizing methods, evidently outperforms other models. In one case, i.e M500, it nearly has double the accuracy of f-CLSWGAN [37]. For the semantic embedding category, although ESZSL [12] had better results on ZSL, it falls behind approaches like ALE [23], DEVISE [10] and SJE [5].



(a) Zero-Shot Learning



(b) Generalized Zero-Shot Learning

Figure 2: ImageNet results measured with Top-1 accuracy in % for the 9 splits including 2 and 3 hops away from ImageNet-1K training classes (2H and 3H) and 500, 1K and 5K most (M) and least (L) populated classes, and All the remaining ImageNet-20K classes.

6 DISCUSSION

A typical zero-shot learning problem is usually faced with three popular issues that need to be solved in order to enhance the performance of the model. These issues are Bias, Hubness and domain shift; and every model revolves around solving one or more of the issues mentioned. In this section, we discuss efforts done by different approaches to alleviate bias, hubness and domain-shift and infer the logic each approach owns to learn its model.

Bias. The problem with ZSL and GZSL tasks is that the imbalanced data between training and test classes cause a bias towards seen classes at prediction time. Other reasons for bias could be high-dimensionality and the devoid of manifold structure of features. Several data generating approaches have worked on alleviating bias by synthesizing visual data for unseen classes. [37] generates semantically rich CNN features of the unseen classes to make unseen embedding space more known. [38] generates pseudo seen and unseen class features, and then it trains an SVM classifier to mitigate bias. [28] improves the quality of the synthesized examples by using gradient matching loss. Models combining data generation or reconstruction along with other techniques have proved to be effective in alleviating bias. [40] uses an intermediate space to help discover the geometric structure of the features that previously didn't with the regression-based projections. [110] used calibrated stacking rule. [39] generates latent feature sizes of 64 with the idea that low-dimensional representations tend to mitigate bias. [112] uses two regressors to calculate reconstruction to diminish the bias. Transductive-based approaches like [28] are also used to solve the bias

issue. In [26], it forces the unseen classes to be projected into fixed pre-defined points to avoid results with bias.

Hubness [135]. In large-dimensional mapping spaces, samples (hubs) might end up falsely as the nearest neighbours of several other points in the semantic space and result in an incorrect prediction. To avoid hubness, [88] proposes a stage-wise bidirectional latent embedding framework. When a mapping is done from high-dimensional feature space to a low-dimensional semantic space using regressors, the distinctive features will partially fade while in the visual feature space, the structures are better preserved. Hence, the visual embedding space is well-known for mitigating hubness problem. [27] and [91] use the output of the visual space of the CNN as the embedding space.

Domain-shift. Zero-shot learning challenge can be considered as a domain adaptation problem. This is because the source labelled data is disjoint with the target unlabelled domain data. This is called project domain-shift. Domain adaptation techniques are used to learn the intrinsic relationships among these domains and transfer knowledge between the two. A considerable amount of works has been done through a transductive setting which has been successful to overcome the domain-shift issue. [22] a multi-view embedding framework, performs label propagation on graph a heuristic one-stage self-learning approach to assign points to their nearest data points. [21] introduces a regularized sparse coding based unsupervised domain adaptation framework that solves the domain shift problem. [136] uses a structured prediction method to solve the problem by visually clustering the unseen data. [27] uses a visual

constraint on the centre of each class when the mapping is being learned. Since the pure definition of the ZSL challenge is the inaccessibility of unseen data during training, several inductive approaches tried to solve the problem as well. [109] proposes to reconstruct the visual features to alleviate this issue. [29] performs sparse non-negative matrix factorization for both domains in a common semantic dictionary. MFMR [31] exploits the manifold structure of test data with a joint prediction scheme to avoid domain shift. [81] uses entropy minimization in optimization. [106] preserves the semantic similarity structure in seen and unseen classes to avoid the domain-shift occurrence. [103] mitigates projection domain-shift by generating soul samples that are related to the semantic descriptions.

These three common issues together with inferiorities each category of methods will be a motivation to decide on a particular approach when solving the ZSL problem. Attribute classifiers are considered customized since human-annotations are used; however, this makes the problem a laborious task that has strong supervision. Compatibility learning approaches have the ability to learn directly by eliminating the intermediate step but often face with the bias and hubness problem. Manifold learning solves this weakness of the semantic learning approaches by preserving the geometrical structure of the features. Cross-modal latent embedding approaches take on a different point of view and leverage both visual and semantic features and the similarity and differences between them. They often propose methods for aligning the structures between the two modes of features. This category of methods also suffers from the hubness problem for the problems dealing with high-dimensional data. Visual space embedding approaches have the advantage of turning the problem into a supervised one by generating or aggregating visual instances for the unseen classes. Plus are a favourable approach for solving hubness problem due to the high-dimensionality of the visual space that can preserve information structure better and also bias problem by alleviating the imbalanced data by generating unseen class samples. Here a challenge would be generating more realistic looking data. Another different setting is transductive learning that present solutions to bias problem, by creating balance in data by gathering unseen data, yet not applicable to many of the real-world problems since the original definition of ZSL limits the use of unseen data during the training phase.

Depending on the real-world scenarios, each way of solving the problem might be the most appropriate choice. Some approaches improve the solution by combining two or more methods to benefit from each one's strengths.

7 ZERO- TO FEW-SHOT LEARNING

Few-shot learning is the challenge of training a previously learned model on a large annotated dataset, for novel

classes where there are only one or few labelled images per class. This task falls into the categories of transductive learning and supervised learning and its main challenge is to improve the generalization ability as it often faces the overfitting problem. Like ZSL, the FSL can also be trained in the generalized model to detect both known and novel classes at the test time. ZSL can be extended to one-shot or few-shot learning by either updating the training data with one or few generated samples from augmentation techniques or by having access to a few of the unseen images during the training time. Many ZSL approaches propose to broaden the FSL related techniques and application such as [137], [138], [46], [60], [69], [18] [74], [82], [19], [39]. For the first time, the idea of using additional information (attributes) in FSL, was introduced in [139].

8 APPLICATIONS

During the years, zero-shot Learning has proved to be a necessary challenge to-be-solved for many different scenarios. The number of applications for the task of learning without access to the unseen target concepts is increasing with each year.

A very recent and global challenge of COVID-19 diagnose and recognition is a perfect real-world application of Zero-shot learning, where we do not have millions of annotated datasets available; and the symptoms of the disease and the chest x-ray of infected people may also vary from person to person. This is considered as a novel unseen target. We only know that symptoms of the infected people with COVID-19 and their chest X-ray images have partial similarities with other lung inflammatory diseases, such as asthma. So, we have to seek for a semantic relationship between training and the new unseen classes. Therefore, ZSL can help us significantly to cope with this new challenge.

Zero-shot learning is widely discussed in the computer vision area. Object recognition in general and for any applications, such as in [140] and [141] aims to locate the objects besides recognising them; several ZSL models are proposed for this purpose; i.e. [142], [143] and [144]. Zero-shot emotion recognition [145] has the task of recognizing unseen emotions while zero-shot semantic segmentation aims to segment the unseen object categories [146] and [147]. Moreover, on the task of retrieving images from large scale set of data, Zero-shot has a growing number of research [148] [149] along with sketch-based image retrieval systems [150], [151] and [152]. Zero-shot learning has an application on visual imitation learning to reduce human supervision by automating the exploration of the agent [153], [154]. Action recognition is the task of recognizing the sequence of actions from the frames of a video. However, if the new actions are not available when training, Zero-shot learning can be a solution, such as in [155], [156], [157] and [158]. Zero-shot Style Transfer in

an image is the problem of transferring the texture of source image to target image while the style is not pre-determined and it is arbitrary [159]. Zero-shot resolution enhancement problem aims at enhancing the resolution of an image without pre-defined high-resolution images for training examples [160]. Zero-shot scene classification for HSR images [161] and scene-sketch classification has been studied in [162] as other applications of ZSL in computer vision. Zero-shot learning has also left its footprint in the area of NLP. Zero-Shot Entity Linking, links entity mentions in the text using a knowledge base [163]. Many research works focus on the task of translating languages to another without pre-determined translation between pairs of samples [164], [165], [166], [167]. In sentence embedding [168] and in Style transfer of text, a common technique is to convert the source to another style via arbitrary styles like the artistic technique discussed in [169]. In the audio processing field, zero-shot based voice conversion to another speaker's voice [170] is an applicable scenario of ZSL.

Based on the recent successful applications, we can infer that in any scenarios that the goal is to reduce supervision, and the target of the problem can be learned through side information, and its relation to the seen data, the Zero-shot learning can be conducted as the learning technique.

9 CONCLUSION

In this article, we had a comprehensive review on the challenge of ZSL, its fundamentals and variants for different scenarios. We divided the recent state-of-the-arts methods into the existing space-wise embedding groups. We also reviewed the side information and went through the popular datasets and their corresponding splits for the problem of ZSL. The paper also contributed in performing the experiment results for some of the common baselines, and elaborated on the advantages and disadvantages of each group, as well as the ideas behind different areas of solutions to improve in each group. We briefly discussed the extension of the problem into few-shot learning, and finally, we reviewed the current and potential real-world applications of ZSL in the near future.

References

- [1] E. G. Miller, N. E. Matsakis, and P. A. Viola, "Learning from one example through shared densities on transforms," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, vol. 1, pp. 464–471, IEEE, 2000.
- [2] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," in *Proceedings of the annual meeting of the cognitive science society*, vol. 33:33, 2011.
- [3] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, Lille, 2015.
- [4] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2013.
- [5] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2927–2936, 2015.
- [6] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning-the good, the bad and the ugly," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4582–4591, 2017.
- [7] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6288–6297, IEEE, 2017.
- [8] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–37, 2019.
- [9] M. Rezaei and R. Klette, "Look at the driver, look at the road: No distraction! no accident!," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pp. 129–136, 2014.
- [10] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in neural information processing systems*, pp. 2121–2129, 2013.
- [11] 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.
- [12] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *International Conference on Machine Learning*, pp. 2152–2161, 2015.
- [13] Y. Guo, G. Ding, J. Han, and Y. Gao, "Zero-shot recognition via direct classifier learning with transferred samples and pseudo labels," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [14] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proceedings of the IEEE international conference on computer vision*, pp. 4166–4174, 2015.
- [15] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5327–5336, 2016.
- [16] B. Zhao, B. Wu, T. Wu, and Y. Wang, "Zero-shot learning posed as a missing data problem," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2616–2622, 2017.

- [17] Y. Li and D. Wang, “Zero-shot learning with generative latent prototype model,” *arXiv preprint arXiv:1705.09474*, 2017.
- [18] V. K. Verma and P. Rai, “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.
- [19] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, “f-vaegand2: A feature generating framework for any-shot learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10275–10284, 2019.
- [20] M. Rohrbach, S. Ebert, and B. Schiele, “Transfer learning in a transductive setting,” in *Advances in neural information processing systems*, pp. 46–54, 2013.
- [21] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, “Unsupervised domain adaptation for zero-shot learning,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2452–2460, 2015.
- [22] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, “Transductive multi-view zero-shot learning,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 11, pp. 2332–2345, 2015.
- [23] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, “Label-embedding for image classification,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1425–1438, 2015.
- [24] Z. Zhang and V. Saligrama, “Zero-shot learning via joint latent similarity embedding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6034–6042, 2016.
- [25] X. Xu, F. Shen, Y. Yang, J. Shao, and Z. Huang, “Transductive visual-semantic embedding for zero-shot learning,” in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pp. 41–49, ACM, 2017.
- [26] J. Song, C. Shen, Y. Yang, Y. Liu, and M. Song, “Transductive unbiased embedding for zero-shot learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1024–1033, 2018.
- [27] Z. Wan, D. Chen, Y. Li, X. Yan, J. Zhang, Y. Yu, and J. Liao, “Transductive zero-shot learning with visual structure constraint,” *arXiv preprint arXiv:1901.01570*, 2019.
- [28] M. B. Sariyildiz and R. G. Cinbis, “Gradient matching generative networks for zero-shot learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2168–2178, 2019.
- [29] M. Ye and Y. Guo, “Zero-shot classification with discriminative semantic representation learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7140–7148, 2017.
- [30] D. Wang, Y. Li, Y. Lin, and Y. Zhuang, “Relational knowledge transfer for zero-shot learning,” in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [31] X. Xu, F. Shen, Y. Yang, D. Zhang, H. Tao Shen, and J. Song, “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.
- [32] Y. Li, D. Wang, H. Hu, Y. Lin, and Y. Zhuang, “Zero-shot recognition using dual visual-semantic mapping paths,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3279–3287, 2017.
- [33] X. Wang, Y. Ye, and A. Gupta, “Zero-shot recognition via semantic embeddings and knowledge graphs,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6857–6866, 2018.
- [34] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C. Frank Wang, “Multi-label zero-shot learning with structured knowledge graphs,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1576–1585, 2018.
- [35] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, and E. P. Xing, “Rethinking knowledge graph propagation for zero-shot learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11487–11496, 2019.
- [36] P. Zhu, H. Wang, and V. Saligrama, “Generalized zero-shot recognition based on visually semantic embedding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2995–3003, 2019.
- [37] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, “Feature generating networks for zero-shot learning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5542–5551, 2018.
- [38] A. Mishra, S. Krishna Reddy, A. Mittal, and H. A. Murthy, “A generative model for zero shot learning using conditional variational autoencoders,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2188–2196, 2018.
- [39] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, “Generalized zero-and few-shot learning via aligned variational autoencoders,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8247–8255, 2019.
- [40] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1627–1636, 2017.
- [41] C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 951–958, IEEE, 2009.
- [42] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, “Describing objects by their attributes,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1785, IEEE, 2009.

- [43] N. Kaessli, Z. Akata, B. Schiele, and A. Bulling, "Gaze embeddings for zero-shot image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4525–4534, 2017.
- [44] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, "What helps where—and why? semantic relatedness for knowledge transfer," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 910–917, IEEE, 2010.
- [45] M. Rohrbach, M. Stark, and B. Schiele, "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," in *CVPR 2011*, pp. 1641–1648, IEEE, 2011.
- [46] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 819–826, 2013.
- [47] Y. Lu, "Unsupervised learning on neural network outputs: with application in zero-shot learning," *arXiv preprint arXiv:1506.00990*, 2015.
- [48] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 69–77, 2016.
- [49] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele, "Multicue zero-shot learning with strong supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 59–68, 2016.
- [50] R. Qiao, L. Liu, C. Shen, and A. Van Den Hengel, "Less is more: zero-shot learning from online textual documents with noise suppression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2249–2257, 2016.
- [51] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," *arXiv preprint arXiv:1612.04844*, 2016.
- [52] M. Elhoseiny, B. Saleh, and A. Elgammal, "Write a classifier: Zero-shot learning using purely textual descriptions," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2584–2591, 2013.
- [53] J. Lei Ba, K. Swersky, S. Fidler, *et al.*, "Predicting deep zero-shot convolutional neural networks using textual descriptions," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4247–4255, 2015.
- [54] L. Niu, A. Veeraraghavan, and A. Sabharwal, "Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7171–7180, 2018.
- [55] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal, "A generative adversarial approach for zero-shot learning from noisy texts," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1004–1013, 2018.
- [56] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 49–58, 2016.
- [57] P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa, "Online incremental attribute-based zero-shot learning," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3657–3664, IEEE, 2012.
- [58] X. Wang and Q. Ji, "A unified probabilistic approach modeling relationships between attributes and objects," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2120–2127, 2013.
- [59] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [60] D. Jayaraman and K. Grauman, "Zero-shot recognition with unreliable attributes," in *Advances in neural information processing systems*, pp. 3464–3472, 2014.
- [61] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam, "Large-scale object classification using label relation graphs," in *European conference on computer vision*, pp. 48–64, Springer, 2014.
- [62] M. Rezaei, "Creating a cascade of haar-like classifiers: Step by step," in *the University of Auckland*, 2013.
- [63] Z. Al-Halah, M. Tapaswi, and R. Stiefelhagen, "Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5975–5984, 2016.
- [64] Y. Atzmon and G. Chechik, "Probabilistic and-or attribute grouping for zero-shot learning," *arXiv preprint arXiv:1806.02664*, 2018.
- [65] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Advances in neural information processing systems*, pp. 1410–1418, 2009.
- [66] J. Weston, S. Bengio, and N. Usunier, "Large scale image annotation: learning to rank with joint word-image embeddings," *Machine learning*, vol. 81, no. 1, pp. 21–35, 2010.
- [67] N. Usunier, D. Buffoni, and P. Gallinari, "Ranking with ordered weighted pairwise classification," in *Proceedings of the 26th annual international conference on machine learning*, pp. 1057–1064, ACM, 2009.
- [68] I. Tschantz, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of machine learning research*, vol. 6, no. Sep, pp. 1453–1484, 2005.
- [69] M. Bucher, S. Herbin, and F. Jurie, "Improving semantic embedding consistency by metric learning for zero-shot classification," in *European Conference on Computer Vision*, pp. 730–746, Springer, 2016.
- [70] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, and L. Shao, "Attentive region embedding network for zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9384–9393, 2019.

- [71] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, “Zero-shot learning through cross-modal transfer,” in *Advances in neural information processing systems*, pp. 935–943, 2013.
- [72] M. Rezaei and M. Isehaghi, “An efficient method for license plate localization using multiple statistical features in a multilayer perceptron neural network,” in *9th Conference on Artificial Intelligence and Robotics and 2nd Asia-Pacific International Symposium*, 2018.
- [73] M. Teimouri, M. H. Delavaran, and M. Rezaei, “A real-time ball detection approach using convolutional neural networks,” in *The 23rd Annual RoboCup International Symposium*, 2019.
- [74] S. Changpinyo, W.-L. Chao, and F. Sha, “Predicting visual exemplars of unseen classes for zero-shot learning,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3476–3485, 2017.
- [75] Y. Li, J. Zhang, J. Zhang, and K. Huang, “Discriminative learning of latent features for zero-shot recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7463–7471, 2018.
- [76] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, “A multi-view embedding space for modeling internet images, tags, and their semantics,” *International journal of computer vision*, vol. 106, no. 2, pp. 210–233, 2014.
- [77] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.
- [78] T. Mukherjee and T. Hospedales, “Gaussian visual-linguistic embedding for zero-shot recognition,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 912–918, 2016.
- [79] H. Jiang, R. Wang, S. Shan, and X. Chen, “Learning class prototypes via structure alignment for zero-shot recognition,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 118–134, 2018.
- [80] S. Kolouri, M. Rostami, Y. Owechko, and K. Kim, “Joint dictionaries for zero-shot learning,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [81] M. Rostami, S. Kolouri, Z. Murez, Y. Owechko, E. Eaton, and K. Kim, “Zero-shot image classification using coupled dictionary embedding,” *arXiv preprint arXiv:1906.10509*, 2019.
- [82] Y.-H. H. Tsai, L.-K. Huang, and R. Salakhutdinov, “Learning robust visual-semantic embeddings,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3591–3600, IEEE, 2017.
- [83] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, “A kernel method for the two-sample-problem,” in *Advances in neural information processing systems*, pp. 513–520, 2007.
- [84] T. Mukherjee, M. Yamada, and T. M. Hospedales, “Deep matching autoencoders,” *arXiv preprint arXiv:1711.06047*, 2017.
- [85] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola, “On kernel-target alignment,” in *Advances in neural information processing systems*, pp. 367–373, 2002.
- [86] M. Yamada, L. Sigal, M. Raptis, M. Toyoda, Y. Chang, and M. Sugiyama, “Cross-domain matching with squared-loss mutual information,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1764–1776, 2015.
- [87] S. Liu, M. Long, J. Wang, and M. I. Jordan, “Generalized zero-shot learning with deep calibration network,” in *Advances in Neural Information Processing Systems*, pp. 2005–2015, 2018.
- [88] Q. Wang and K. Chen, “Zero-shot visual recognition via bidirectional latent embedding,” *International Journal of Computer Vision*, vol. 124, no. 3, pp. 356–383, 2017.
- [89] M. Elhoseiny, A. Elgammal, and B. Saleh, “Write a classifier: Predicting visual classifiers from unstructured text,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2539–2553, 2016.
- [90] B. Schölkopf, R. Herbrich, and A. J. Smola, “A generalized representer theorem,” in *International conference on computational learning theory*, pp. 416–426, Springer, 2001.
- [91] L. Zhang, T. Xiang, and S. Gong, “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.
- [92] Y. Long, L. Liu, and L. Shao, “Towards fine-grained open zero-shot learning: Inferring unseen visual features from attributes,” in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 944–952, IEEE, 2017.
- [93] M. Arvanaghi Jadid and M. Rezaei, “Facial age estimation using hybrid haar wavelet and color features with support vector regression,” in *Artificial Intelligence and Robotics*, 2017.
- [94] Z. Ji, Y. Fu, J. Guo, Y. Pang, Z. M. Zhang, *et al.*, “Stacked semantics-guided attention model for fine-grained zero-shot learning,” in *Advances in Neural Information Processing Systems*, pp. 5995–6004, 2018.
- [95] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, pp. 2672–2680, 2014.
- [96] B. Tong, M. Klinkigt, J. Chen, X. Cui, Q. Kong, T. Murakami, and Y. Kobayashi, “Adversarial zero-shot learning with semantic augmentation,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [97] V. Kumar Verma, G. Arora, A. Mishra, and P. Rai, “Generalized zero-shot learning via synthesized examples,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4281–4289, 2018.
- [98] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

- [99] M. Bucher, S. Herbin, and F. Jurie, “Generating visual representations for zero-shot classification,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2666–2673, 2017.
- [100] Y. Li, K. Swersky, and R. Zemel, “Generative moment matching networks,” in *International Conference on Machine Learning*, pp. 1718–1727, 2015.
- [101] R. Felix, V. B. Kumar, I. Reid, and G. Carneiro, “Multi-modal cycle-consistent generalized zero-shot learning,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 21–37, 2018.
- [102] K. Sohn, H. Lee, and X. Yan, “Learning structured output representation using deep conditional generative models,” in *Advances in neural information processing systems*, pp. 3483–3491, 2015.
- [103] J. Li, M. Jin, K. Lu, Z. Ding, L. Zhu, and Z. Huang, “Leveraging the invariant side of generative zero-shot learning,” *arXiv preprint arXiv:1904.04092*, 2019.
- [104] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *International conference on machine learning*, pp. 214–223, 2017.
- [105] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” in *Advances in neural information processing systems*, pp. 5767–5777, 2017.
- [106] C. Li, X. Ye, H. Yang, Y. Han, X. Li, and Y. Jia, “Generalized zero shot learning via synthesis pseudo features,” *IEEE Access*, vol. 7, pp. 87827–87836, 2019.
- [107] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *European conference on computer vision*, pp. 499–515, Springer, 2016.
- [108] P. Zhu, H. Wang, and V. Saligrama, “Dont even look once: Synthesizing features for zero-shot detection,” *arXiv preprint arXiv:1911.07933*, 2019.
- [109] E. Kodirov, T. Xiang, and S. Gong, “Semantic autoencoder for zero-shot learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3174–3183, 2017.
- [110] L. Chen, H. Zhang, J. Xiao, W. Liu, and S.-F. Chang, “Zero-shot visual recognition using semantics-preserving adversarial embedding networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1043–1052, 2018.
- [111] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, “Adversarial autoencoders,” *arXiv preprint arXiv:1511.05644*, 2015.
- [112] X. Shibing and G. Zishu, “Bi-semantic reconstructing generative network for zero-shot learning,” *arXiv preprint arXiv:1912.03877*, 2019.
- [113] Z. Chen, J. Li, Y. Luo, Z. Huang, and Y. Yang, “Canzsl: Cycle-consistent adversarial networks for zero-shot learning from natural language,” *arXiv preprint arXiv:1909.09822*, 2019.
- [114] C. R. Givens, R. M. Shortt, *et al.*, “A class of wasserstein metrics for probability distributions,” *The Michigan Mathematical Journal*, vol. 31, no. 2, pp. 231–240, 1984.
- [115] H. Huang, C. Wang, P. S. Yu, and C.-D. Wang, “Generative dual adversarial network for generalized zero-shot learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 801–810, 2019.
- [116] G. Patterson and J. Hays, “Sun attribute database: Discovering, annotating, and recognizing scene attributes,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2751–2758, IEEE, 2012.
- [117] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” *California Institute of Technology*, 2011.
- [118] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604, 2015.
- [119] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.
- [120] G. A. Miller, “Wordnet: a lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [121] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, “Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly,” *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [122] D. Parikh and K. Grauman, “Relative attributes,” in *2011 International Conference on Computer Vision*, pp. 503–510, IEEE, 2011.
- [123] Z. Harris, “Distributional structure. word, 10 (2-3): 146–162. reprinted in fodor, j. a and katz, jj (eds.), readings in the philosophy of language,” 1954.
- [124] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.
- [125] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, pp. 3111–3119, 2013.
- [126] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.
- [127] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [128] A. Bosch, A. Zisserman, and X. Munoz, “Representing shape with a spatial pyramid kernel,” in *Proceedings of the 6th ACM international conference on Image and video retrieval*, pp. 401–408, 2007.

- [129] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features (surf),” *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.
- [130] E. Shechtman and M. Irani, “Matching local self-similarities across images and videos,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2007.
- [131] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- [132] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [133] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.
- [134] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [135] M. Radovanović, A. Nanopoulos, and M. Ivanović, “Hubs in space: Popular nearest neighbors in high-dimensional data,” *Journal of Machine Learning Research*, vol. 11, no. Sep, pp. 2487–2531, 2010.
- [136] Z. Zhang and V. Saligrama, “Zero-shot recognition via structured prediction,” in *European conference on computer vision*, pp. 533–548, Springer, 2016.
- [137] X. Yu and Y. Aloimonos, “Attribute-based transfer learning for object categorization with zero/one training example,” in *European conference on computer vision*, pp. 127–140, Springer, 2010.
- [138] V. Sharmanska, N. Quadrianto, and C. H. Lampert, “Augmented attribute representations,” in *European Conference on Computer Vision*, pp. 242–255, Springer, 2012.
- [139] Y.-H. H. Tsai and R. Salakhutdinov, “Improving one-shot learning through fusing side information,” *arXiv preprint arXiv:1710.08347*, 2017.
- [140] M. Rezaei, M. Terauchi, and R. Klette, “Robust vehicle detection and distance estimation under challenging lighting conditions,” *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, 2015.
- [141] R. Sabzevari, A. Shahri, A. Fasih, S. Masoumzadeh, and M. Rezaei Ghahroudi, “Object detection and localization system based on neural networks for robo-pong,” in *Mechatronics and Its Applications, 5th International Symposium on*, 2008.
- [142] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, “Zero-shot object detection,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 384–400, 2018.
- [143] S. Rahman, S. Khan, and F. Porikli, “Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts,” in *Asian Conference on Computer Vision*, pp. 547–563, Springer, 2018.
- [144] B. Demirel, R. G. Cinbis, and N. Ikizler-Cinbis, “Zero-shot object detection by hybrid region embedding,” *arXiv preprint arXiv:1805.06157*, 2018.
- [145] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang, “Zero-shot emotion recognition via affective structural embedding,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1151–1160, 2019.
- [146] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, “Zero-shot semantic segmentation,” *arXiv preprint arXiv:1906.00817*, 2019.
- [147] W. Wang, X. Lu, J. Shen, D. J. Crandall, and L. Shao, “Zero-shot video object segmentation via attentive graph neural networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9236–9245, 2019.
- [148] Y. Long, L. Liu, Y. Shen, and L. Shao, “Towards affordable semantic searching: Zero-shot retrieval via dominant attributes,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [149] Y. Xu, Y. Yang, F. Shen, X. Xu, Y. Zhou, and H. T. Shen, “Attribute hashing for zero-shot image retrieval,” in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 133–138, IEEE, 2017.
- [150] A. Dutta and Z. Akata, “Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5089–5098, 2019.
- [151] S. Dey, P. Riba, A. Dutta, J. Lladós, and Y.-Z. Song, “Doodle to search: Practical zero-shot sketch-based image retrieval,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2179–2188, 2019.
- [152] Y. Shen, L. Liu, F. Shen, and L. Shao, “Zero-shot sketch-image hashing,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3598–3607, 2018.
- [153] D. Pathak, P. Mahmoudieh, G. Luo, P. Agrawal, D. Chen, Y. Shentu, E. Shelhamer, J. Malik, A. A. Efros, and T. Darrell, “Zero-shot visual imitation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2050–2053, 2018.
- [154] M. Lázaro-Gredilla, D. Lin, J. S. Guntupalli, and D. George, “Beyond imitation: Zero-shot task transfer on robots by learning concepts as cognitive programs,” *Science Robotics*, vol. 4, no. 26, p. eaav3150, 2019.
- [155] J. Gao, T. Zhang, and C. Xu, “I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8303–8311, 2019.

- [156] J. Qin, L. Liu, L. Shao, F. Shen, B. Ni, J. Chen, and Y. Wang, “Zero-shot action recognition with error-correcting output codes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2833–2842, 2017.
- [157] A. Mishra, V. K. Verma, M. S. K. Reddy, S. Arulkumar, P. Rai, and A. Mittal, “A generative approach to zero-shot and few-shot action recognition,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 372–380, IEEE, 2018.
- [158] L. Shen, S. Yeung, J. Hoffman, G. Mori, and L. Fei-Fei, “Scaling human-object interaction recognition through zero-shot learning,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1568–1576, IEEE, 2018.
- [159] L. Sheng, Z. Lin, J. Shao, and X. Wang, “Avatar-net: Multi-scale zero-shot style transfer by feature decoration,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8242–8250, 2018.
- [160] A. Shocher, N. Cohen, and M. Irani, ““zero-shot” super-resolution using deep internal learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3118–3126, 2018.
- [161] A. Li, Z. Lu, L. Wang, T. Xiang, and J.-R. Wen, “Zero-shot scene classification for high spatial resolution remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 4157–4167, 2017.
- [162] Y. Xie, P. Xu, and Z. Ma, “Deep zero-shot learning for scene sketch,” in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3661–3665, IEEE, 2019.
- [163] L. Logeswaran, M.-W. Chang, K. Lee, K. Toutanova, J. Devlin, and H. Lee, “Zero-shot entity linking by reading entity descriptions,” *arXiv preprint arXiv:1906.07348*, 2019.
- [164] J. Gu, Y. Wang, K. Cho, and V. O. Li, “Improved zero-shot neural machine translation via ignoring spurious correlations,” *arXiv preprint arXiv:1906.01181*, 2019.
- [165] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, *et al.*, “Google’s multilingual neural machine translation system: Enabling zero-shot translation,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.
- [166] T.-L. Ha, J. Niehues, and A. Waibel, “Effective strategies in zero-shot neural machine translation,” *arXiv preprint arXiv:1711.07893*, 2017.
- [167] S. M. Lakew, Q. F. Lotito, M. Negri, M. Turchi, and M. Federico, “Improving zero-shot translation of low-resource languages,” *arXiv preprint arXiv:1811.01389*, 2018.
- [168] M. Artetxe and H. Schwenk, “Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond,” *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 597–610, 2019.
- [169] K. Carlson, A. Riddell, and D. Rockmore, “Zero-shot style transfer in text using recurrent neural networks,” *arXiv preprint arXiv:1711.04731*, 2017.
- [170] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, “Autovc: Zero-shot voice style transfer with only autoencoder loss,” in *International Conference on Machine Learning*, pp. 5210–5219, 2019.