Data Augmentation for Multi-domain and Multi-modal Generalised Zero-shot Learning

Rafael Felix

February 24, 2020

Thesis submitted for the degree of
Doctor of Philosophy
in
Computer Science
at The University of Adelaide
Faculty of Engineering, Computer and Mathematical Sciences
School of Computer Science

THE UNIVERSITY
of ADELAIDE

Abstract

This thesis addresses the problem of combining data augmentation with multi-domain and multi-modal training and inference for Generalised Zero-Shot Learning (GZSL). GZSL introduces an experimental setup, where the training set contains images and semantic information for a set of seen classes, and semantic information for a set of unseen classes, where there is no overlap between the seen and unseen classes. The semantic information can be represented by a group of attributes or some textual information that describes a visual class. The main goal of GZSL methods is to build a visual classifier that works for both the seen and unseen classes, even though there are no training images from the unseen classes. The key to solve this challenging problem is to explore the connection between the semantic and visual spaces by learning a model that can translate between these spaces.

The solutions proposed in the field have been focused on three directions: conventional Zero-shot Learning (ZSL), data augmentation and domain classification. Conventional ZSL comprises an optimisation procedure that learns a mapping from the visual to the semantic space using the seen classes. The inference maps the images of the unseen classes from the visual to the semantic space, where classification relies on a nearest neighbour classifier. The extension of ZSL to GZSL is not trivial since it biases the classification towards the seen classes given the lack of semantic and visual samples from the unseen classes during training. Such issue has driven GZSL to two alternative approaches: domain classification and data augmentation. Domain classification aims to learn a one-class classifier that estimates the likelihood that visual samples belong to the set of seen classes – this domain classifier is then used to select or modulate the visual classification of test images. More specifically, an input visual sample is first classified as seen or

unseen, and then forward to different classifiers (e.g., if it is classified as seen, then it goes to the visual classifier trained with the seen images; and if it is classified as unseen, then it goes to a conventional ZSL classifier). Even though relatively successful, this approach assumes that seen and unseen classes are drawn from different domains, which is unwarranted in GZSL because images from seen and unseen classes most likely come from the similar distribution. The other alternative approach, data augmentation, comprises the training of a generative model that produces visual samples conditioned on a semantic sample. Then, this generative model produces synthetic samples from the unseen classes, which are joined by the real visual samples from the seen classes to train a visual classifier. This approach introduces a multi-modal training, but there is no guarantee that the generated visual samples can represent well the visual samples from the unseen classes, and inference still relies only on the visual modality.

In this thesis, we propose several methods to address the issues mentioned above. Firstly, we introduce a novel data augmentation model based on a cycle-consistent multi-modal training to improve the generation of visual samples, particularly from the unseen classes. Secondly, we propose a novel domain classification method that no longer relies on one-class classifiers – instead, we use the visual samples from the generative model to train a binary domain classifier. Thirdly, we extend our proposed GZSL data augmentation framework to a multi-modal inference procedure, where we train a visual and a semantic classifier that are combined to classify a test image. Our final proposed model is based on a multi-modal and multi-domain data augmentation approach composed of multiple classifiers trained in three modalities (visual, semantic and joint latent space). Moreover, we proposed the use of a classification calibration technique to produce an effective multi-modal and multi-domain classification. We report extensive experiments for the proposed models, using several benchmark data sets, such as the Caltech-UCSD Birds 200 (CUB), Animal with Attributes (AWA), Scene Understanding Benchmark Suite (SUN), 102 Category Flower Dataset (FLO) and ImageNet. The experiments show that multi-modal and multi-domain optimisation can be combined with data augmentation to produce state-of-the-art GZSL results.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed: .. .. Date: 29/11/2019

Preface

This thesis was written at the School of Computer Science, The University of Adelaide. The main parts of the thesis are based on the following published/submitted papers in which I am the primary author:

1. Felix, R., Kumar, V. B., Reid, I., and Carneiro, G., Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision* (ECCV), pages 21-37, 2018. [39]

2. Felix, R., Harwood, B., Sasdelli, M., and Carneiro, G., Generalised Zero-Shot Learning with Domain Classification in a Joint Semantic and Visual Space. In *Digital Image Computing: Techniques and Applications* (DICTA), pages –. IEEE, 2019. [37]

3. Felix, R., Sasdelli, M., Reid, I. and Carneiro, G., Augmentation Network for Generalised Zero Shot Learning with Multi-Modal Inference. In *Submission*, 2019 [40].

4. Felix, R., Harwood, B., Sasdelli, M., and Carneiro, G., Multi-domain Generalised Zero-shot Learning using Visual, Semantic, and Joint Latent Spaces. In *Submission*, 2019 [38].

Dedication

I dedicate this thesis to my grandmother Mercês.

Acknowledgements

The PhD degree comprises of many scientific challenges. However, very little is spoken about the challenges related to personal growth that presented in this journey. The guidance and kindness of many people around me were vital in helping me overcome the obstacles in my degree, for those marvellous human beings I would like to express my gratitude.

Firstly, I would like to demonstrate my gratefulness to my supervisors. **Prof. Gustavo Carneiro** thank you for believing in my potential and providing me with a critical thinking, insightful and kind guidance. Our conversations helped me to develop resilience, and I became fearless to defend my ideas to the world. **Prof. Ian Reid** thank you for sharing your wisdom and kindness. Your feedback has always helped to broaden my perspective. **Dr. Michele Sasdelli** thank you for our uncountable discussions. Your advises have always helped me to focus my energy on the deliverable. I am truly honoured to have had the support of this supervisory panel.

Secondly, I am fortunate to be supported by wonderful peers and institutions. I am certainly thankful for the insightful conversations with my collaborators Dr Vijay Kumar and Dr Ben Harwood. To the Australian Centre of Excellence for Robotic Vision (ACRV), I am grateful for all the support and resources extended to myself and my research. Without the support of ACRV, I would not be able to complete this journey. The training and networking provided by ACRV were crucial for my professional development. The University of Adelaide has been my refuge in these years, as a PhD student. I am grateful for the numerous friends I have made in the School of Computer Science, and the Australian Institute for Machine Learning. I am thankful for Dr Bradley Alexander for the opportunity given as a casual lecturer. Moreover, I would like to demonstrate my gratitude

to the administrative staff which have always assisted me. Thank you Thuy Mai and Kate Aldridge (ACRV), and Hilary Brookes, Sharyn Liersch, and Lenka Hill (UoA). When I decided to undertake my studies at The University of Adelaide, I was clueless about how generously I would be received in Australia. I am thankful for the discussions and friendship developed with Kejie Le, Samya Bagchi, Hayden Faulkner, Adrian Johnston, Renato Hermoza, Gabriel Maicas, and Ming Cai. Thanks for the insightful conversations and good moments shared with Tobin South, Dr Jamie Sherrah, Prof. Fabio Faria, Violetta Shevchenko, Thomas Rowntree, Jerome Williams, Ehsan Abbasnejad, Prof. Wojciech Chojnacki, Shingfang Chng, Cheek Heng Chng, Gerard Snaawn, Xian Wang, Toan Tran, and Huangying Zhan.

Thirdly, I am grateful for my experience in Adelaide, where I have made important friends. My special thanks to Shikiko for her constant patience, love and kindness. Moreover, I would like to show my gratitude to my friends Travis Billingsley, Ben Schmidt, Oriana Alvarado, Allan Hotti, Allan Ramos, Jonas, Raquel, Jenilson and Patricia de Matos, to name a few. I am also very thankful for the 'Morning Crew' from Beyond Bouldering gym, which helped me to balance a healthy life. My friends have helped me to mitigate the burden of being away from my family.

Most of all, I am speechless to show gratitude to my family. Above all, they have always believed and loved me. The afternoons spent with my grandmother Merces, who always stimulated me to love books and science. My grandfather Pio showed me the ways to love and respect culture and arts. My mother Luciana has always supported my efforts and celebrated my achievements with me. My sister Natalia and her husband Andre, they have supported me in all my endeavours. My uncles, aunties have always challenged me to seek the best intellectual and moral skills. More than words, my family has taught me with their actions of unconditional love. My deepest desire is to make them proud.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

In recent years, the advent of technological advancements for the acquisition of visual data and for running deep neural networks have enabled outstanding visual classification results from deep learning models. The training of deep learning models generally requires large annotated training sets [26, 61, 81, 148], where such models can only recognise the visual classes available from that training set, even though in reality new visual classes can appear continuously after the training process has finished [5, 25, 66, 77, 125]. The recognition of these new visual classes is a challenge that has been studied by the computer vision and machine learning communities [5, 25, 66, 125]. For instance, Generalized Zero-shot Learning (GZSL) proposes a solution to recognise these new classes [150] by exploring the concepts of transfer learning and domain adaptation [52, 104]. The GZSL data sets are divided into seen and unseen class domains, where the seen domain corresponds to classes available for training, containing images that have been manually annotated, and the unseen domain comprises a set of classes that does not have any images during the training phase. GZSL methods are designed to model the unseen visual classes with alternative semantic descriptions [2,41,44,76], which can be represented by textual descriptions or lists of common attributes [69, 76, 102]. Fig. 1.1 illustrates two samples from the data set Animal with Attributes (AWA) [76] – these samples contain the visual information available from the image and their respective semantic features (in that case, represented by a list

Figure 1.1: Sample of two classes (a horse on the left hand side and humpback whale on the right hand side) from the GZSL benchmark data set Animal with Attribute [76], and their respective semantic features (for each semantic attribute, a score between 0 and 1 has been manually provided).

of attributes, and respective scores from 0 to 1) [69, 102]. The main motivation for using the semantic description of classes is that, while the visual annotation process is expensive and time-consuming, the semantic information is cheap to obtain, and consequently, widely available [76, 150]. For instance, one can rely on dictionary definitions of visual classes as an alternative semantic description.

Conventional GZSL methods are optimised [2, 41, 44, 77] as depicted in Fig. 1.2. In a pre-processing step, images are transformed to be represented as feature vectors in a visual space, and semantic descriptors are represented as feature vectors in a semantic space. In the first step (training), the GZSL model, represented by the attribute prediction, is trained to regress visual samples to their respective semantic samples [77]. In the second step (inference), the visual samples from the test set are mapped, by the attribute prediction, into the semantic space, which is used in a nearest neighbour classification process to predict the class label. Recent studies have shown that these conventional GZSL approaches have two important drawbacks [150]: (i) the bias issue towards the seen classes, and (ii) the single modality inference. Firstly, the pipeline suffers from a **bias towards the seen classes issue**, which consists of an imbalanced performance concerning the seen and unseen classes [19, 149, 150]. More specifically, the classification accuracy for the seen classes tends to be substantially higher than for the unseen classes [150]. Studies have pointed out that the bias issue is caused by two aspects of the GZSL training: 1) the missing visual information for the unseen classes when training the GZSL model [39], and 2) the **asymmetry of the training sets for the seen classes containing visual and semantic features, and for the unseen classes con-**

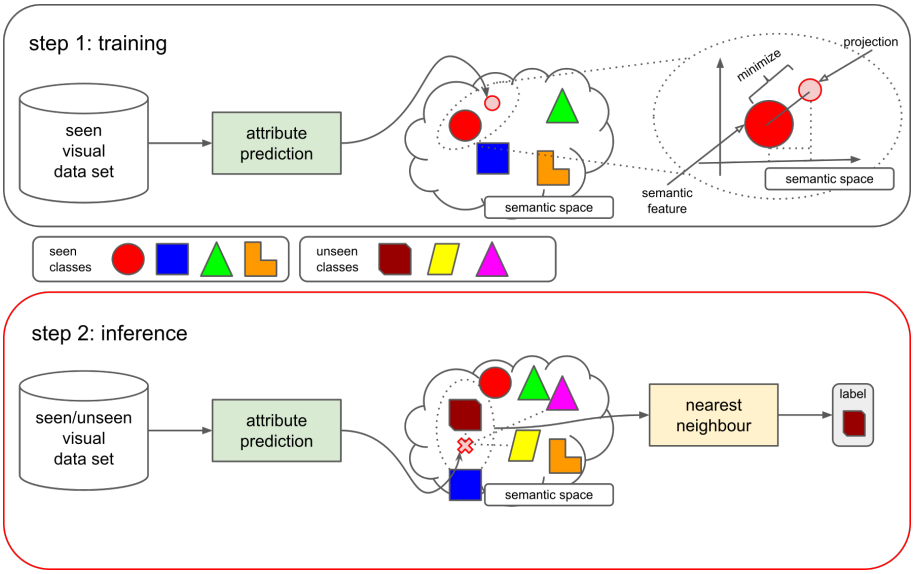


Figure 1.2: Overview of the conventional GZSL/ZSL model. During training, seen classes are used to train an attribute prediction model that transforms visual to semantic samples. During inference, visual samples from seen and unseen classes are input to the attribute prediction that outputs a semantic sample, which is then used in a nearest neighbour classification.

taining only semantic features [37]. The second drawback of conventional GZSL approaches is that the optimisation procedure for training GZSL models uses multiple modalities, but the inference procedure tends to rely on a single-modality inference. In recent literature, several studies have been conducted to address these two issues with GZSL approaches.

Frome et al. [42] suggested that deep neural networks can be used for optimising the pre-processing step of GZSL methods, explained above. In their approach, a deep neural network is used to encode the image and semantic samples [42]. This approach, based on deep learning, has shown to improve the training and inference of GZSL models [42]. Moreover, this work motivated subsequent studies that employ deep learning for solving the GZSL problem [2–4, 150, 151]. Recent studies have also proposed that the GZSL seen and unseen domains can be represented with distinct probability distributions [8, 133, 158]. In particular, these papers address the bias issue with a one-class seen domain classifier that is used to select or modulate a GZSL classifier. More specifically, the domain classifier is

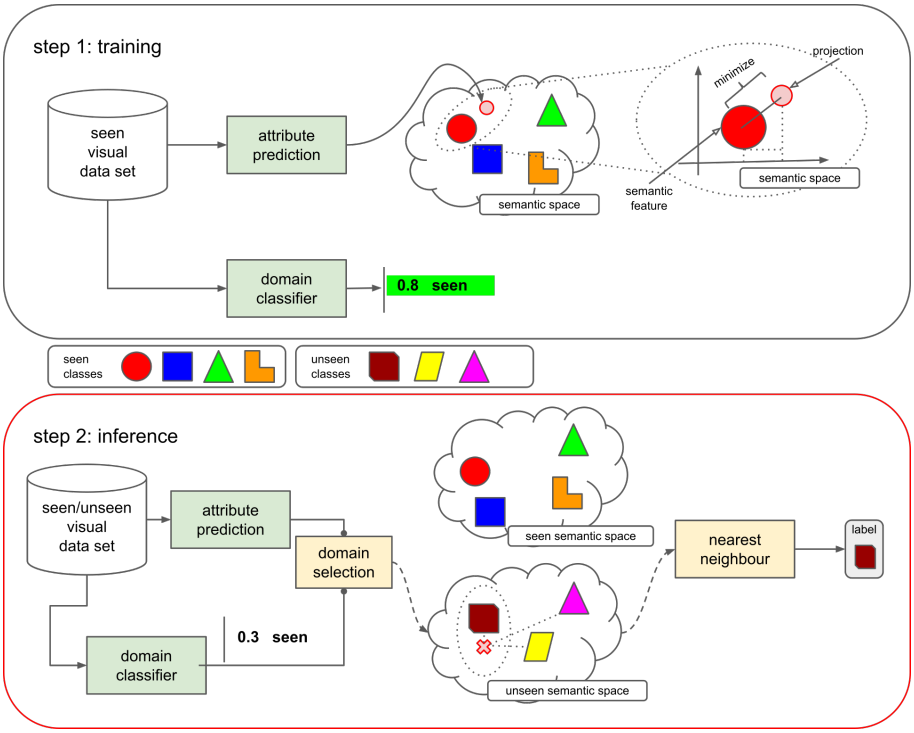


Figure 1.3: Overview of the domain classifier for GZSL. During training, the seen class samples are used to learn an attribute prediction and a domain classifier. During inference, the domain classifier estimates the probability that a test sample belongs to the seen domain. Considering a threshold value, the classification will be performed with the seen or the unseen classes, respectively.

trained with novelty detection methods (i.e., one-class classifiers) to estimate the probability that input samples belong to the seen domain. Then, either the seen or unseen classifier is selected by the domain classifier [133] for the GZSL inference process, as depicted in Fig. 1.3. The main issue with this approach is that there is not sufficient evidence to suggest that the seen and unseen classes form distinct distributions in the visual feature space [37], making the training of the one-class classifier challenging.

Another GZSL approach is based on data augmentation [21, 112, 142, 151], depicted in Fig. 1.4. This approach assumes that the seen and unseen visual samples share a similar probability distribution. The probability distribution can be learned by a generative model, which is conditioned on the joint distribution of visual and semantic samples [39, 151]. These generative models can augment the

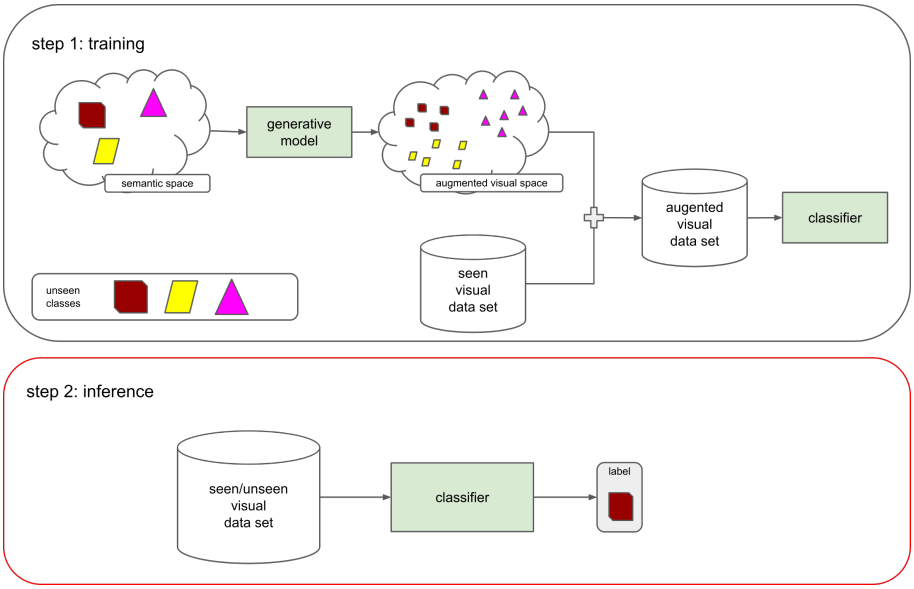


Figure 1.4: Overview of the GZSL data augmentation framework. During training, this model learns to synthesize visual samples that are used to train a visual classifier. During inference, a novel visual sample from seen or unseen classes is tested by this visual classifier.

visual data set to train a visual classifier with real samples from the seen classes and generated samples from unseen classes [151]. This approach alleviates the bias towards the seen classes because it enables training with the same number of samples from seen and unseen classes. However, the fact that these approaches do not explore multi-modal inference can be seen as a weakness given the promising results observed in other types of multi-modal inference applications [39, 144, 151].

The development and enhancement of GZSL approaches will be beneficial for the computer vision community for a large number of applications, such as visual classification [74, 151], segmentation [84, 87, 99, 119], generative models [7, 101, 122, 151], image retrieval [28, 32, 130, 155], and object detection [10, 50, 51, 113]. The progress of this field can potentially ease the burden of annotating large data sets, improve the learning for open-set applications [12, 125, 127] and contribute to a broader understanding of human-robot interaction applications [67, 79, 93, 94, 139].

1.2 Motivations

In this thesis, we focus on the design of new GZSL solutions that explore effective data augmentation, multi-modal and multi-domain training and inference processes. We aim to address the bias towards the seen classes issue, the multi-modal inference, and the asymmetry of the training sets between the seen and unseen classes.

We propose a novel method that aims to alleviate the **bias towards the seen classes issue** with a data augmentation approach that generates samples for seen and unseen classes. Our contribution involves a novel multi-modal cycle consistency loss that forces the generated visual samples to be transformed back into their respective semantic samples – the proposed method is referred to as $cycle-WGAN$ [39] (see Fig. 1.5). Similarly to previous data augmentation approaches, $cycle-WGAN$ aims to turn the GZSL problem into a "supervised" visual classification problem, using the real visual samples from the seen classes and the synthetic visual samples from the unseen classes. With a cycle consistency loss, $cycle-WGAN$ enables more efficient learning of the generative model. This leads to more effective training for the visual classifier when compared to previous approaches that do not rely on such cycle-consistency loss.

We also tackle the **asymmetry of the training sets for the seen classes containing visual and semantic samples, and for the unseen classes containing only semantic samples** with the formulation of a new GZSL model based on a domain classification strategy. In this work, we rely on a framework for learning to transform visual and semantic features into a single latent space – as depicted in Fig 1.6. The proposed binary domain classifier uses this joint latent space to estimate whether a sample belongs to the seen or unseen classes [37] and also to train the GZSL classifier. During inference, the result from the domain classifier modulates the class estimation produced by the GZSL classifier. The main novelty of our method is the use of this latent joint space for training a single GZSL classifier for both the seen and unseen classes and a binary domain classification procedure. The main advantage is that our approach no longer needs the training of a novelty detector, involving a complex modelling of one-class classifiers [37].

Furthermore, we propose two methods that tackle the **single-modality infer-**

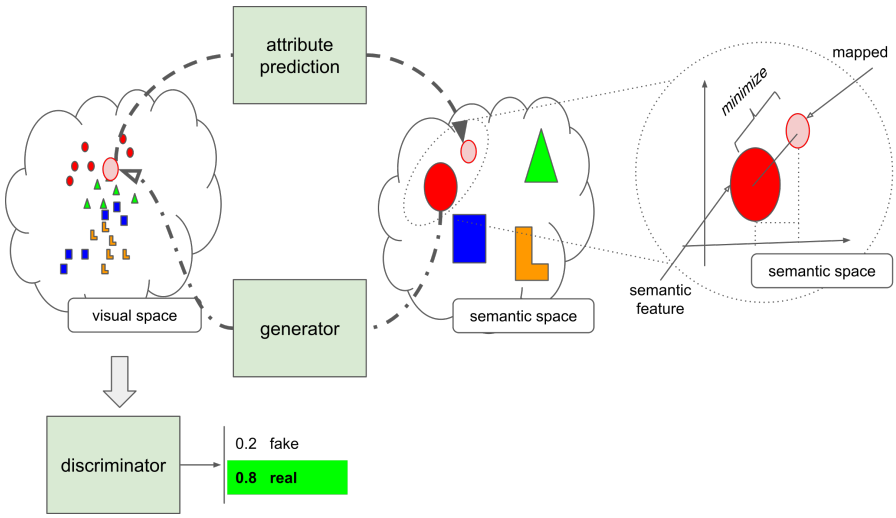


Figure 1.5: Overview of cycle-WGAN. This figure depicts the proposed cycle consistency loss that takes the generated visual samples to train a regressor that maps the visual samples back to their semantic samples. This proposed loss enables more effective data augmentation GZSL methods than previously proposed models that do not rely on such cycle-consistency loss.

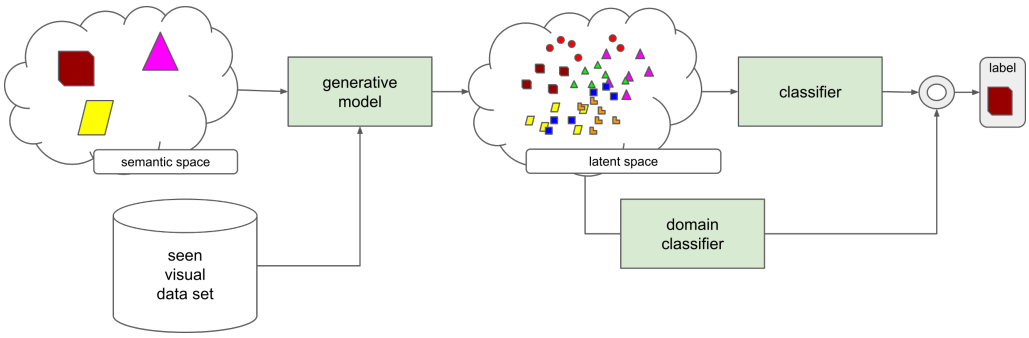


Figure 1.6: Overview of the proposed binary domain classifier method. In this approach, the generative model produces samples in a latent space from the seen and unseen classes – these samples are used to train a binary domain classifier and a GZSL classifier. During inference, a test sample is transformed into this latent space, classified by the GZSL classifier and modulated by the domain classifier.

ence problem and the **asymmetric training problem** with new **multi-modal** and **multi-domain** training and inference processes [38, 40]. One important contribution of these methods is the use of data augmentation to train not only the visual [21, 112, 142], but also the semantic and joint multi-modal classifiers. In the first method, we proposed the augmentation network for multi-domain and multi-modal GZSL (AN-GZSL) [40], where we present a novel GZSL architecture that combines the use of generative models and multiple classifiers trained using multiple modalities. Fig. 1.7 illustrates the first method. The second model consists of a novel GZSL architecture that explores the use of variational autoencoders to achieve multi-modal inference in GZSL [38]. We also propose a calibration strategy for the combination of multiple models, which can be considered a simpler and more effective solution for the multi-domain problem when compared to the domain classification approach based on novelty detection methods.

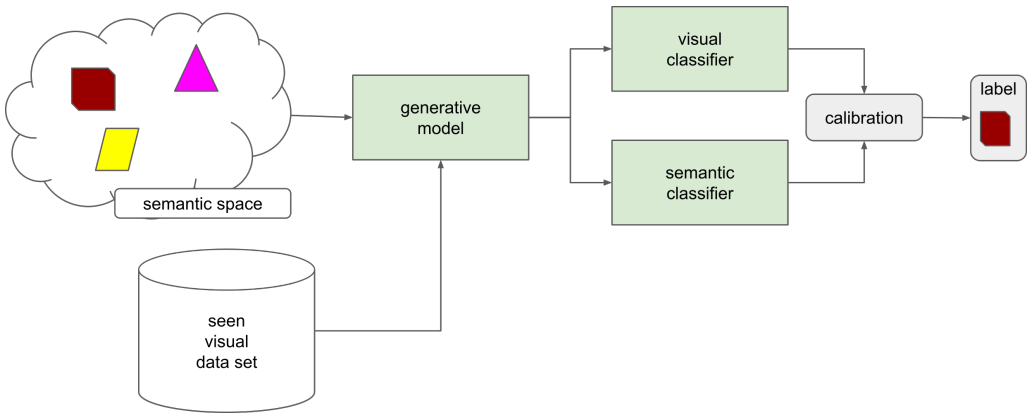


Figure 1.7: Overview of multi-modal multi-domain data augmentation GZSL model. During training, the generative model is used to synthesise samples for the training of the visual and semantic classifiers. During inference, the calibrated visual and semantic GZSL classifiers produce a multi-modal and multi-domain seen and unseen class estimation.

The main contributions proposed in this thesis are designed to address the bias towards the seen classes issue, the asymmetric training problem and the multi-modal/multi-domain training and inference. The empirical results show that the approaches proposed in this thesis established new state-of-the-art results at the time of their publication. We report the results taking into consideration several

measures commonly used in GZSL benchmarks, such as the seen and unseen classification accuracy and their harmonic mean (**H-mean**). The main benchmark data sets utilised in this thesis are: Caltech-UCSD Birds 200 (CUB) [146], Scene Understanding Benchmark (SUN) [152], ImageNet [26], Animal with Attributes (AWA1 and AWA2) [76, 150], and 102 Flower Category Database (FLO) [98].

1.3 Contributions

In this thesis, we present several contributions to the field of generalised zero-shot learning, which can be summarised as follows:

- We investigate the advantages of deep generative models in GZSL. In particular, we propose novel generative models to augment GZSL data sets by synthesising visual samples from semantic features. These synthetic visual samples are employed to train a visual classifier for the seen and unseen classes, where the seen class contains real and synthetic visual samples, while the unseen classes contain exclusively synthetic samples. We explore the use of two generative models for GZSL: Generative Adversarial Networks (GANs) [55] and Variational Autoencoders (VAE) [29]. Our main contribution in this area is the introduction of a loss function that includes a multi-modal cycle consistency objective function, which yields accuracy improvement and eases the learning process of the GAN, providing faster convergence.

- We propose a novel method that tackles domain classification trained with a dual encoder/decoder framework, which is composed of a visual and semantic variational autoencoders sharing a latent space. We show that this joint latent space can be used for training a binary domain classifier and a GZSL classifier. The estimation from the domain classifier modulates the output of the GZSL classifier to produce a balanced classification between the seen and unseen domains.

- We propose a multi-domain multi-modal classifier for GZSL. This work tackles the use of a generative model for 1) training multiple classifiers in several modalities; and 2) generating samples from multiple domains

to mitigate the bias towards the seen classes issue. This method achieves a good balance between the seen and unseen classification accuracy by a multi-domain classification calibration [56] and no longer requires a domain classifier [38, 40].

1.4 Outline

In this section we provide the thesis outline, where we briefly discuss the contents of each chapter as follows:

In Chapter 1, we present the overview, motivations and contributions of this thesis.

In Chapter 2, we provide a literature review of this thesis, where we aim to contextualize all important topics related to generalised zero-shot learning, such as deep learning, zero-shot learning, domain classification, and generative models. Furthermore, we also provide an extensive background description of previous GZSL methods.

In Chapter 3, we propose the $cycle - WGAN$ [39], which is a novel GZSL method. In this chapter, we show how generative adversarial networks (GAN) can be used to synthesise visual samples using the semantic samples of the unseen classes. We propose a cycle consistency regularisation for GANs in the GZSL problem. This cycle consistency regularisation ensures that the generated visual samples can be transformed back to their respective semantic features – this regularisation improves the training process for the GAN, and improves GZSL performance.

In Chapter 4, we propose a binary domain classification approach. This chapter introduces a novel GZSL method that learns a joint latent representation space for training two classifiers: a GZSL classifier, and a GZSL domain classifier that estimates the probability that an input visual sample belongs to the seen or unseen classes. The domain classifier modulates the GZSL classifier, alleviating the bias towards the seen classes issue.

In Chapter 5, we introduce a multi-domain multi-modal GZSL training and inference by proposing a novel method, called $AN - GZSL$. This model combines data augmentation with multi-domain and multi-modal optimisation. The $AN -$

GZSL architecture comprise a generative adversarial network, a visual network and a semantic network. The visual and semantic networks consist of calibrated classifiers to guarantee a classification that is well balanced between seen and unseen classes. Therefore, this approach does not require any gating mechanisms or external domain classifiers.

In Chapter 6, we formulate the hypothesis that reconstruction spaces from variational autoencoders can be used for multi-domain multi-modal GZSL training and inference. In particular, we propose a novel GZSL method that combines visual, semantic and joint latent space classifiers that are calibrated to promote a balanced seen and unseen classification. The proposed model replaces domain classification by a simple combination of multiple calibrated modal classifiers.

We conclude this thesis and discuss future work in Chapter 7.

CHAPTER 2

Literature Review

In the first part of this chapter, we review the literature in conventional ZSL. Then, we describe the current research in Generalised Zero-Shot Learning, where we highlight the main gaps in the literature which motivated the approaches proposed in this thesis.

2.1 Overview

In recent years, deep learning methods have achieved outstanding progress on several pattern recognition tasks [61, 81, 128]. An increasing number of studies have demonstrated the efficiency of deep learning for several applications, such as image classification [58,136], segmentation [87,99,119], object detection [50,51,113], 3D reconstruction [64,70], and many others. In fact, recent studies have reported results that surpassed human-level performance for several tasks in large-scale data sets [54]. Despite these advances, the deployment of deep learning methods in unstructured environments is not thoroughly explored [12,126,127]. One issue is that it is impractical to collect a labelled training set that contains all possible visual classes that might eventually appear for a deep learning model [126]. Therefore, there is a growing interest in the development of deep learning methods that can deal with previously unseen visual classes, either by recognising them [150] or by detecting them as a novel object [126].

The recognition of previously unseen visual classes has been formalised as zero-shot learning (ZSL) [36, 76, 157]. The lack of visual data from the unseen

classes is compensated with information from other modalities, such as shared attributes [77], semantic features [108,109,109,114,133], and contextual information [156]. More specifically, during training, the set of seen classes contain visual and non-visual information, while the unseen classes only contain non-visual data, and during testing, only visual samples of unseen classes are presented for classification [36,76,157].

In this section, we provide a general view of Zero-Shot Learning. Firstly, we introduce the ZSL problem defined by Lampert et al. [76]. Secondly, we discuss significant milestones for the field, such as the development of conventional ZSL methods and the implementation of deep learning-based ZSL (deep ZSL). Then, we present the paper by Xian et al. [150] that describes a novel experimental setup which introduces Generalised Zero-Shot Learning (GZSL). Moreover, we also discuss a widely accepted data set split proposed for GZSL benchmark data sets [150].

2.1.1 Conventional Zero-Shot Learning

Conventional ZSL is defined as a transfer learning problem [23,52,103,104,140], where the seen classes (or source domain) and the unseen classes (or target domain) are disjoint during training [76]. Lampert et al. [76] address this challenging problem by introducing a novel method based on attribute prediction [76] which consists of transferring the knowledge between the visual and attribute spaces by using the seen classes to train a regressor that transforms visual samples to their respective attribute samples [19,36,76,77,157]. A new data set to benchmark ZSL methods is introduced in that paper [76] – this data set is named 'Animal with Attributes' (AWA). AWA contains 30,000 labelled images from 50 animal classes, and a set of 85 semantic attributes that describe properties such as shape, colour, or geographic information [69,76,102] for each of those animal classes. In contrast to supervised learning, where the data sets are split into the train, validation and test sets [5,25,54,66,148], the data set AWA is divided into two domains: the seen domain, represented by 40 classes, and the unseen domain, represented by 10 classes [76]. Following this setup, ZSL methods can access the visual and attribute samples from the seen domain during training. During inference, the trained regressor transforms the visual features of a test visual sample into the semantic

space (i.e., the attribute features). The classification is estimated by using this semantic sample in a nearest neighbour search among the semantic samples of the classes in the unseen domain.

From the work proposed by Lampert et al. [76], ZSL research has grown into four main branches:

1. **Handcrafted features**: consists of hand-designing visual features derived from image processing algorithms [53]. These features are composed of edges, corners, colours or salient key-points in an image [53]. In the literature, we observed many studies investigating the impacts of handcrafted features to boost the performance of conventional ZSL, such as SIFT [43,76,85,89,91], rgSIFT [77,143], LSS [86], ORB [120], SURF [11,77], HOG [22], PHOG [16,34, 77,91], Fisher vectors [4,91,107], and Vector Quantization [36,76,123,133, 156,157].

2. **Semantic features**: another research focus for conventional ZSL has been the exploration of different semantic attributes [108,109,114]. In recent years, several studies have proposed the expansion of conventional ZSL with the attribute descriptions based on geometrical and morphological information, such as shape, colour, and texture [157], and/or hierarchical information of the classes [59,78,92,117]. Another focus has been to ease the burden of annotation with web data-mining strategies [13,91,108,109,133]. The methods in this approach rely on mining semantic information from the web to cover the gap between seen and unseen classes. Recent studies have reported the use of skip-gram text models (e.g., word2vec) [42,73,106], Wikipedia descriptions [108,109,133], textual information [34,114], and natural language processing [59,78,92,108,109].

3. **Learning strategies**: a large amount of conventional ZSL research has focused on the exploration of different learning strategies to transform samples from the visual to the semantic space. More specifically, in subsequent years from Lampert et al.'s paper [76], we have observed the proposal of several learning strategies for conventional ZSL, including: direct attribute prediction [76], indirect attribute prediction [76], learning manifolds [45,133], novelty detection [133], category-level recognition [156], compatibility in

label-embedding [3] , online incremental learning [67], sketch [32,130,155], learning unreliability of attributes [65], and co-occurrence of attributes [91]. Akata et al. [4] proposed a structured joint embedding optimisation based on a max-margin objective function for the ZSL problem. Likewise, Zhang et al. [161] extended that method by incorporating the semantic representation of the unseen classes.

4. **Large scale data sets**: part of the ZSL community has focused on the proposal of large-scale data sets, such as ImageNet [42,117], Caltech-UCSD Birds 200 (CUB) [34,91] and Scene Understanding (SUN) [65,91].

The effort to tackle large scale data sets, such as ImageNet [26], has led Frome et al. [42] to propose a novel deep visual-semantic embedding (DeViSE), which is the first model to introduce the use of convolutional neural networks (e.g., AlexNet [75]) to address the ZSL problem. Later, this work has motivated the development of many other studies that explore the use of deep learning for solving ZSL.

2.1.2 Deep Learning for Zero-Shot Learning (Deep ZSL)

In recent years, deep learning has enabled the automated learning of data representation for machine learning applications [27,54]. In Fig. 2.1, we illustrate how a deep learning model extracts discriminant features from images. Deep neural networks can be roughly divided into two parts. The first part consists of a feature extraction network composed of convolutional filters, activation functions (e.g. ReLu [90], LeakyRelu [90]) and regularisation techniques [62,71,90,122,135]. When an image is processed by this model, it is forwarded through this first part, where the initial layers are able to represent low-level visual information such as texture, lines, colour, and edges [80,141,162]. The intermediate and final layers can combine the representations from the initial layers to detect more complex visual features, which can be part of more complex visual objects, such as a car, a person, or an animal. The last layer outputs a vector of visual features, which are processed by the second part of the model, represented by a Multilayer Perceptron (MLP) [95]. The MLP is a discriminative neural network that learns the conditional output probability of a class given a visual feature [24,54].

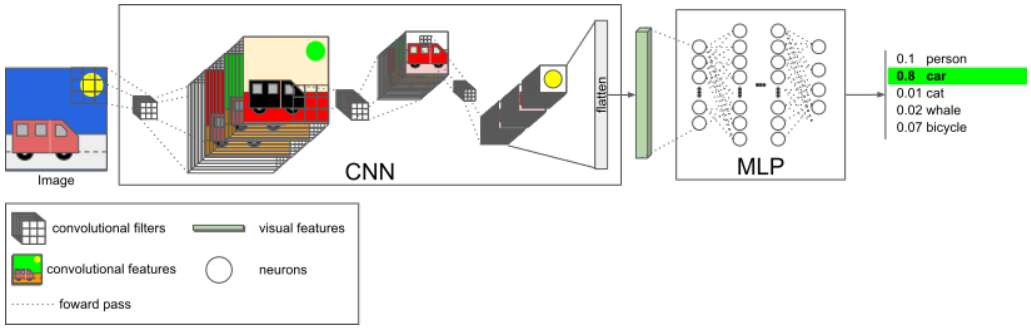


Figure 2.1: [54]. Depiction of a Convolutional Neural Network pipeline. The image is input into the convolutional neural network (CNN) that extracts discriminative features to be used by a fully-connected neural network, represented by a multi-layer perceptron (MLP), which estimates a class label.

The training of deep neural networks progresses by adjusting the network parameters. These adjustments are achieved with an optimisation algorithm, such as Stochastic Gradient Descent [116, 121]. The objective of this algorithm is to minimise a loss function, such as categorical cross-entropy [15, 54] that aims to train a classifier from the MLP output. Deep neural networks represent the currently dominating model used in GZSL problems. In recent years, several studies proposed the use of several deep learning architectures for GZSL, such as GoogLeNet [137], AlexNet [75], VGG [131] and ResNet [58].

A recent study [42] has shown that the visual features obtained from the last convolutional layer from pre-trained deep learning models can be adapted to the ZSL task. In particular, the deep learning models trained on the ImageNet challenge appear to extract effective features for ZSL benchmark data sets [42, 54]. The use of deep learning for feature extraction can facilitate ZSL training, and possibly boost the ZSL performance when compared to the use of handcrafted features. The same principle of feature extraction has been explored in ZSL for textual and hierarchical information [34, 92, 109, 114, 133]. More specifically, deep learning models have been trained on the English dictionary [92], and articles from Wikipedia [34]. Despite the outstanding results achieved by ZSL methods implemented with deep learning models, some questions still require further investigation. For instance, when deployed in unstructured environments, these methods will be required to handle samples from both seen and unseen classes [77].

Hence, the ZSL constraint that seen class samples are not present during testing is rather unrealistic for the deployment of ZSL in unstructured environments. Recent studies show that conventional ZSL models perform poorly when this constraint is violated [19, 138], which motivated the development of methods that work well in the classification of both seen and unseen classes.

2.2 Generalised Zero-Shot Learning

Generalised Zero-Shot Learning (GZSL) relaxes the ZSL constraint that seen class samples are not present during testing [19]. Chao et al. [19] argue that for a ZSL model to be truly useful, it should be able to identify classes from the seen and unseen domains at test time. Chao et al. [19] highlighted the issues with conventional ZSL setup, and propose to evaluate conventional ZSL with a novel generalised ZSL setup.

Partially motivated by the work above, researchers have switched their focus from ZSL to GZSL [19, 39, 150, 151]. This interest can be attributed to the standardisation of benchmark data sets proposed by Xian et al. [150], where they highlight two important issues in conventional ZSL [150]. First, the naive use of ImageNet [26] pre-trained CNNs, such as ResNet [58], violates the conventional ZSL conditions because of an existing overlap between the zero-shot (i.e., unseen) classes and the classes from ImageNet used to train deep learning models [26]. This issue is reported for several benchmark data sets (e.g., AWA [76], CUB [146] and SUN [152]). Xian et al. [150] then propose a solution to overcome this issue – a new data set split that takes into consideration the class overlapping with ImageNet classes. The second issue reported by Xian et al. [150] regards the application of conventional ZSL model in GZSL conditions. In particular, classification results from ZSL models show that samples from the unseen domain are prone to be mislabelled into one of the seen classes. This is defined as the bias towards the seen class (a.k.a. hubness problem [31], or class imbalance [19]) [77, 150]. Therefore, Xian et al. [150] introduced a new test set of images from the seen classes, and a new way to assess GZSL methods with a novel set of metrics that takes into consideration the bias issue, and the balancing performance between the seen and unseen samples.

Contemporary GZSL models can be categorised into three groups: **conventional attribute prediction from ZSL to GZSL**, **domain classification**, and **visual data augmentation**. In the following sections, we describe these categories.

2.3 Conventional Attribute Prediction from ZSL to GZSL

In conventional ZSL, the attribute prediction model learns a mapping function from the visual to the semantic space [3, 4, 42]. This model is optimised using the training data set composed of visual and semantic samples from the seen class domain. When tested in GZSL conditions, the inference procedure is achieved by mapping a visual sample to the semantic space, in which the classification is performed via a nearest neighbour procedure. Where, the distance between the predicted semantic feature is computed to the semantic features from the unseen classes. The label inference is achieved by attributing the class label obtained from the neighbour with minimal distance. The main compatibility functions used to compute the nearest neighbour are based on the minimum Euclidean distance [77], or the maximum cosine similarity [42].

Preliminary research in GZSL has explored several learning methods for attribute prediction. Farhadi et al. [36] and Lampert et al. [76] propose an attribute prediction model optimised with the minimization of the difference between transformed visual samples and their respective semantic representations. In this approach, they assumed that the attributes are mutually independent given the class label, which can be considered a naive approach. Socher et al. [133] propose to overcome this simplistic assumption by using a lower-dimensional manifold approach learned with an unaligned language corpora acquired from online sources. Similarly, an approach based on representing images by the co-occurrence of the visual concepts has also been proposed [91]. This co-occurrence of the training classes is used as discriminative features to estimate the classification for the unseen classes. Jayaraman et al. [65] claim that attribute prediction models are ineffective to transform visual samples into semantic samples, and propose a random forest method that measures the unreliability of the attributes predicted by regressor models. This method aims to learn statistics about the

attribute prediction errors tendencies to weight the inference procedure for the unseen classes.

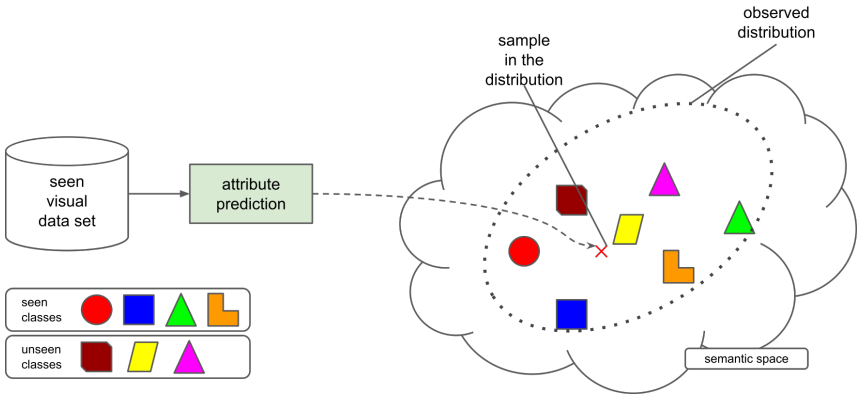
Frome et al. [42] argue that attribute prediction models transform the visual samples into a semantic space that is agnostic to the relative similarity of GZSL classes, and propose to tackle ZSL as a ranking problem [3, 4, 42, 161]. The main assumption is that the nearest neighbour inference can be solved by training the attribute prediction model with a rank optimisation, rather than a minimization optimisation. In this rank optimisation, the attribute prediction model aims to approximate the prediction to the ground truth semantic sample and at the same time to push the prediction further from the other GZSL classes. Frome et al. explore this assumption by proposing a pairwise bi-linear function with a margin restriction [42]. Akata et al. [4] extend this approach into a structured joint embedding optimisation, and later with a max-margin rank objective function [3]. Zhang et al. [161] propose a novel bi-linear objective function that also take into account the semantic features from the unseen classes during the optimisation. However, the conventional ZSL methods still suffer from the bias toward the seen classes issue, when tested under GZSL conditions.

In recent years, there has been an interest in two different strategies to mitigate the bias issue, namely: domain classification and data augmentation, which are covered in the sections below.

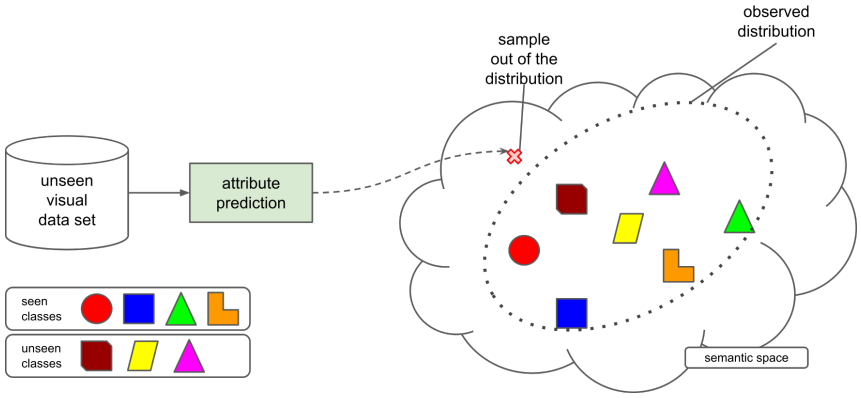
2.4 Domain Classification for GZSL

The unknown nature of the unseen classes raises the hypothesis that the GZSL problem can be handled with a domain classifier using a novelty detection strategy. In the literature, novelty detection consists of the ability of a system to estimate whether a non-observed sample belongs to the same distribution of the data used for training the model [68]. Fig. 2.4 illustrates the application of a novelty detection method to two visual samples (one from the seen and another from the unseen class) transformed into the semantic space. The main assumption in domain classification is that it is possible to robustly distinguish between samples from the seen and the unseen classes [133].

Several papers introduced methods that suggest that domain classification is



(a) novelty detection for a random seen sample



(b) novelty detection for a random unseen sample

Figure 2.2: Illustration of a novelty detection model. Diagram (a) shows how a sample from a seen class is classified by a novelty detector as belonging to the seen domain distribution, while diagram (b) illustrates that unseen samples are classified as outliers (or novelty) by the novelty detector. Although the illustration represents the novelty detection being computed in the semantic space, we highlight that the same operation can be performed in different embedding spaces [8,37,158].

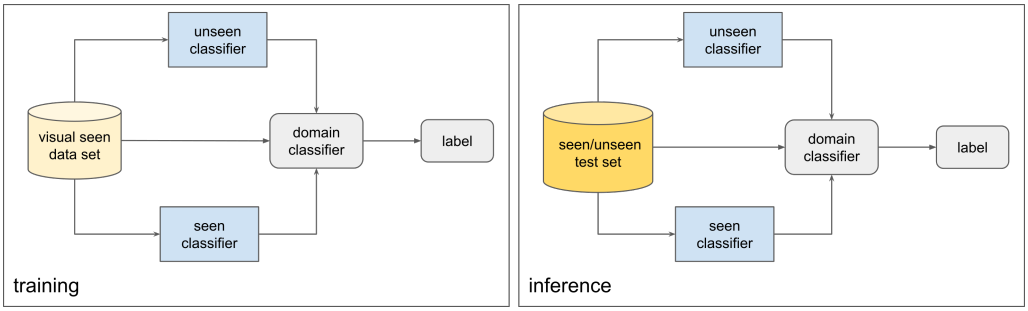


Figure 2.3: Illustration of the general pipeline for Domain Classification methods. During training, these multiple models are trained with samples from the seen visual classes 1) to perform classification for the seen and unseen class samples, and 2) to classify input samples as belonging or not to the seen class distribution (i.e., domain classification). During inference, test samples from seen and unseen classes are presented to the models, and the domain classifier selects (or modulates) the model to compute inference.

useful in GZSL problems [8, 37, 133, 158]. Domain classification for GZSL consists of an external domain classifier that modulates the classification of the seen and unseen classes, as depicted in Fig. 2.3. In these methods, the training consists of optimising a classifier that can estimate the probability that visual samples belong to the seen class distribution. The GZSL classifier is trained to estimate class labels for the seen and unseen domains, but this inference is modulated by the domain classifier.

Socher et al. [133] propose to use a one-class Gaussian Process (GP) [111, 147] as the domain classifier that selects the seen class classifier (using the visual space) or the unseen class classifier (using the semantic space). This is the first approach to explore domain classification, but it does not scale well with the number of training samples due to limitations of the GP model training [111, 147]. In a different approach, Zhang et al. [158] propose a mechanism that explores the use of multiple neural networks in domain classification. In this approach, a classifier for the seen classes and another classifier for the unseen classes are trained using a generative model and the domain classifier is trained to select between the seen and unseen classifiers with a threshold. Even though successful, the approach by Zhang et al. [158] suffers from a non-trivial asynchronous optimisation of

multiple networks. Atzmon and Chechik [8] introduce a gating network for domain classification. In this approach, the gating network comprises a model that modulates the output probability of the seen and the unseen classifiers. Even though these methods present promising results, there are still major questions that need to be answered by domain classification approaches. Firstly, these approaches consist of the disjoint training of multiple models for the seen and unseen domains, which can be considered non-optimal. Secondly, the assumption that the seen and unseen visual samples belong to two distinct distributions is too strong and may not hold in practice, limiting the applicability of domain classifiers.

In [37], we proposed a novel model that aims to address these questions. The proposed model is composed of a binary domain classifier and a GZSL classifier, which are optimised in an end-to-end training. The first network is a GZSL classifier trained with real visual samples from the seen data set, and synthetic visual samples of the unseen classes. This network is capable of classifying all GZSL classes, rather than the classes from specific domains. The second network comprises a binary domain classifier trained with the same data set – this model modulates the outputs of the GZSL classifier. This approach extends the domain classification into a binary classification problem, which is arguably simpler than currently used one-class domain classification. Another approach that we explore in this thesis is the extension of current methods to a well balanced multi-domain method without the use of any domain classifier [38, 40]. We note that such implementation is possible with the use of classification calibration, which is explained in the next section.

2.4.1 Neural Network Calibration

A recent study by Guo et al. [56] highlights that the confidence output estimated from deep learning models is generally poorly calibrated. A neural network that provides a calibrated confidence reflects the correctness of the output probability. The study by Guo et al. [56] shows that despite their outstanding performance in terms of accuracy, the miscalibrated outputs from neural network classifiers do not allow their deployment in real-life environments [56]. Guo et al. [56] show that the optimisation tools available for modern deep neural networks, such as the number

of layers, the large number of parameters, weight decay, batch normalization and dropout have largely contributed to the poor calibration of deep learning models. One of the most effective approaches that they propose for calibrating neural networks consists of the optimisation of the temperature scaling in the softmax activation function [56]. This temperature scale factor can be optimised with a simple post-processing step based on a held-out validation set to provide a calibrated confidence output.

In this thesis, we rely on deep learning classification calibration [56] to produce a multi-domain and multi-modal GZSL classifiers without the use of an external domain classifier.

2.5 Visual Data Augmentation for GZSL

Deep generative models estimate the joint probability distribution of observed and target variables [54,72,101,122]. In recent years, we have seen remarkable progress in the optimisation of deep generative models, and their capability of generating synthetic data, such as image, video, and audio [54]. In GZSL, generative models are applied to generate synthetic visual samples that can reduce the bias towards the seen classes issue. We refer to this framework as GZSL data augmentation. This model is trained with seen classes by conditioning the generation of visual samples on their respective semantic samples. Then, we can generate visual samples from the semantic samples of the unseen classes, and use them, together with the real visual samples of the seen classes, to train a GZSL visual classifier. Two types of generative models have received attention from GZSL researchers: Generative Adversarial Network (GAN) [55] and Variational Autoencoder (VAE) [29]. Below, we provide a brief explanation of GANs and VAEs and list studies that have employed them for GZSL.

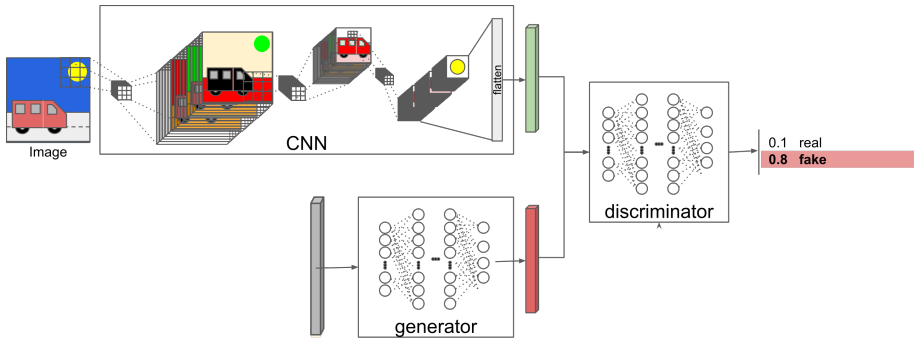
2.5.1 Generative Adversarial Networks

The term GANs refers to methods that rely on adversarial training for estimating generative models [55]. Fig. 2.4(a) illustrates the GAN framework, which is composed of the generative and the discriminative networks. The generative model aims to learn the training set distribution using a noise input randomly

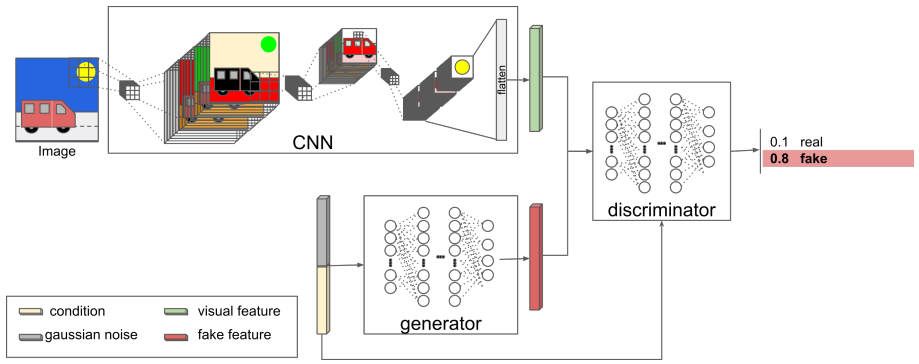
drawn from a known probability distribution (e.g. Gaussian distribution), and the output consists of a generated (i.e., fake) sample that is similar to samples from the observed distribution [55]. The discriminative model estimates the probability that a sample belongs to the training set distribution of the observable variable. During training, the discriminator receives real and fake samples and has to predict whether a sample is from the real distribution [55]. In this framework, the discriminative and generative models are trained simultaneously, by optimising an adversarial process, where the objective function can be decomposed into two terms. In the first term, the goal of the generative model is to fool the discriminative model into assigning high probabilities to the fake samples. In the second term, the main goal of the discriminative model consists of differentiating the fake and real samples [55]. The convergence of the GAN optimisation is achieved when the discriminative and generative models reach the Nash equilibrium [122].

Figure 2.4(b) depicts the conditional GAN that extends the original GAN framework [55]. The conditional GAN aims to learn the joint probability distribution of multiple modalities, where the generator receives as input a vector composed by a latent variable (sampled from a known noise function) and a conditional variable, represented by a one-hot vector [48], semantic features [151] or from other visual spaces [63, 101, 134, 154]. The condition leads the GAN to generate conditional visual samples. Likewise, the discriminator receives the same conditional variable concatenated with the respective visual sample [63, 101, 134, 151, 154] – the main goal of this discriminator is to determine whether this pair of visual and semantic samples belong to the same distribution.

Recently, there is an increasing number of studies employing GANs to GZSL models. Xian et al. [151] introduce a conditional GAN framework, where the conditional variable is represented by the semantic sample concatenated to a latent variable sampled from a Gaussian distribution. This conditional vector is forwarded through the generative model to obtain a synthetic visual sample. Besides the discriminative network, they also introduce a classifier network that evaluates the quality of the synthetic visual features generated. The discriminator and generator are regularised with (i) a Wasserstein loss [7], and (ii) a cross-entropy objective function computed from the fully-connected classifier for the seen classes [151]. The generative model is used to generate visual samples conditioned on the se-



(a) GAN



(b) conditional GAN

Figure 2.4: Illustration of a Generative Adversarial Network (GAN) framework. (a) depicts the GAN that generates visual samples from a latent noise variable. This visual sample is assessed to be real or fake by discriminator network. (b) illustrates a conditional GAN which generates visual samples conditioned on a concatenation of the respective semantic samples and a latent noise variable. The generated visual sample is again used as an input with the semantic feature to be assessed by the discriminator network.

mantic samples from the unseen classes [151] – these generated samples are used to augment the original data set and train a visual GZSL classifier [151]. Felix et al. [39] extend that method [151] by replacing the cross-entropy objective function with a cycle-consistency loss [164] that minimises the mean square error between the original semantic sample and a reconstructed semantic sample produced by a regressor trained with the generated visual sampled at the input. This approach aims to improve the domain transfer in GANs and to produce synthetic samples which yield better GZSL classification accuracy. Then, many attempts have been made to improve the regularisation of GANs for the GZSL problem [33,60,82,124]. For instance, Elhoseiny [33] investigates a framework [151] that extends the regularisation of GANs with a loss function that promotes the generation of realistic synthetic samples from the unseen classes and maximises the entropy computed by the classification output from a classifier. More recently, Li et al. [82] introduce the use of soul samples to regularise the generation of samples. These soul samples can be described as a class meta-representation, which is obtained by averaging meaningful samples from the data set distribution. Sariyildiz et al. [124] extend the optimisation in [151] with an objective function based on a gradient matching strategy that consists of the difference between the gradients of the real and fake samples, which provides smoother training and more discriminative synthetic samples. In [60], a novel approach explores the optimisation of the GAN with a novel discriminator network that aims to address the use of metric learning [60].

The main issue with the models presented above is that, although they all propose multi-modal training processes, none of them relies on a multi-modal inference, despite significant research in multi-modal machine learning [9] that shows solid classification results.

2.5.2 Variational Autoencoder

Variational Autoencoder (VAE) [29] is composed of an encoder and a decoder, as depicted in Fig. 2.5. This model is designed to learn the data distribution representation, where the encoder network represents a transformation function from an observable variable (e.g. an image) into a latent variable from a continuous space that enables random sampling and interpolation [29]. One of the VAE loss function terms enforces the probability density function (pdf) of the latent variable

to be a normal distribution. The other loss term enforces the decoder network to reconstruct the original data from the estimated latent variable [29].

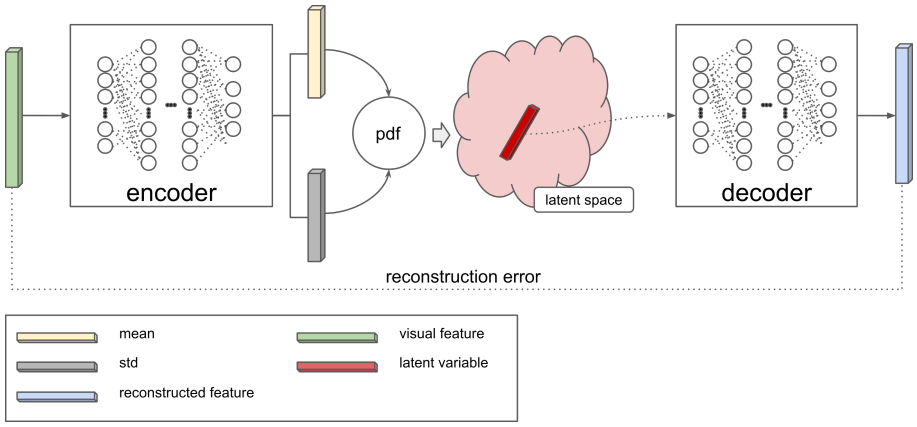


Figure 2.5: Illustration of a Variational Autoencoder model. During training, the visual sample is transformed into the latent space by the encoder. The decoder utilises samples from the latent space to reconstruct the original visual sample. The backpropagation is achieved by minimising the reconstruction error between the original and reconstructed visual samples and the divergence between the prior and observed distributions in the latent space.

The forward pass in the VAE framework consists of three steps: encoding, sampling and decoding. First, the encoder network estimates the mean and variance of the latent normal distribution that represents an input visual sample. Then, we randomly sampled from this latent distribution and forward the sample to the decoder network that aims to reconstruct the visual sample.

There have been a few GZSL approaches that employed VAEs in GZSL data augmentation [96, 144]. Mishra et al. [96] propose a VAE-based data augmented GZSL, where visual samples are concatenated to semantic samples to be used as input to the encoder, and in the reconstruction, the semantic samples are concatenated to the latent variable. This optimisation procedure maximises the joint representation likelihood of these modalities [96]. The generation of synthetic visual samples is achieved by randomly sampling a point from the latent space, and concatenating it with the semantic sample from an unseen class. This vector is then used by the decoder to generate a visual sample. Kodirov et al. [35] introduces an autoencoder that optimises the following model: $visual \rightarrow semantic \rightarrow visual$.

In contrast to [96], this approach enforces the semantic sample to match the latent space distribution. In this approach, the generation of visual samples is achieved by using the semantic samples from unseen classes to draw points in the latent space. These points are then decoded into the visual space. The inference is achieved by a nearest neighbour classifier in the visual space Later, an investigation proposed by Schonfeld et al. [129] has shown that a stack of VAEs can be used to boost GZSL performance. They introduce a VAE framework composed of two parallel encoder and decoder networks that share the same latent space. The first VAE is denoted as visual encoder/decoder (it encodes and decodes the visual samples), and the second is denoted as semantic encoder/decoder (it encodes and decodes the semantic samples). Considering the shared latent space, they propose an objective function that introduces a cross-alignment and a distribution-alignment regularisation on the latent space [129]. The cross-alignment term enforces that a pair of visual-semantic samples transformed into the join latent space can be decoded by the decoder of a different modality [129].

Similarly to the GZSL data augmentation with GANs, the VAE methods presented in this section do not explore multi-modal inference, which can be considered a major weakness of those approaches.

Chapter 3

Multi-modal Cycle-consistent Generalized Zero-Shot Learning

The work contained in this chapter has been published as the following paper:

Felix, R., Kumar, V. B., Reid, I., and Carneiro, G., Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision* (ECCV), pages 21-37, 2018 [39].

Statement of Authorship

Title of Paper	Multi-modal cycle-consistent generalized zero-shot learning
Publication Status	[X] Published [ ] Accepted for Publication [ ] Submitted for Publication [ ] Unpublished or Unsubmitted work
Publication Details	Felix, R., Kumar, V. B., Reid, I., and Carneiro, G., Multi-modal cycle-consistent generalized zero-shot learning. In Proceedings of the European Conference on Computer Vision (ECCV), pages 21-37, 2018.

Principal Author

Name of Principal Author (Candidate)	Rafael Felix Alves		
Contribution to the Paper	- Development of the main idea of the paper; - Implementing and conducting the experiments; - Writing and coordinating the revisions;		
Overall percentage (%)	60		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	10/22/2019

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Vijay Kumar K. B.		
Contribution to the Paper	- Help with the development of the idea; - Help writing, revision and discussions;		
Signature		Date	10/22/2019

Name of Co-Author	Ian Reid		
Contribution to the Paper	- Help writing, revision and discussions;		
Signature		Date	22/11/19

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	- Help with the development of the idea; - Help writing, revision and discussions;		
Signature		Date	04-11-2019

Abstract

In generalized zero shot learning (GZSL), the set of classes are split into seen and unseen classes, where training relies on the semantic features of the seen and unseen classes and the visual representations of only the seen classes, while testing uses the visual representations of the seen and unseen classes. Current methods address GZSL by learning a transformation from the visual to the semantic space, exploring the assumption that the distribution of classes in the semantic and visual spaces is relatively similar. Such methods tend to transform unseen testing visual representations into one of the seen classes' semantic features instead of the semantic features of the correct unseen class, resulting in low accuracy GZSL classification. Recently, generative adversarial networks (GAN) have been explored to synthesize visual representations of the unseen classes from their semantic features - the synthesized representations of the seen and unseen classes are then used to train the GZSL classifier. This approach has been shown to boost GZSL classification accuracy, but there is one important missing constraint: there is no guarantee that synthetic visual representations can generate back their semantic feature in a multi-modal cycle-consistent manner. This missing constraint can result in synthetic visual representations that do not represent well their semantic features, which means that the use of this constraint can improve GAN-based approaches. In this paper, we propose the use of such constraint based on a new regularization for the GAN training that forces the generated visual features to reconstruct their original semantic features. Once our model is trained with this multi-modal cycle-consistent semantic compatibility, we can then synthesize more representative visual representations for the seen and, more importantly, for the unseen classes. Our proposed approach shows the best GZSL classification results in the field in several publicly available data sets.

3.1 Introduction

Generalized Zero-shot Learning (GZSL) separates the classes of interest into a sub-set of seen classes and another sub-set of unseen classes. The training process uses the semantic features of both sub-sets and the visual representations of only

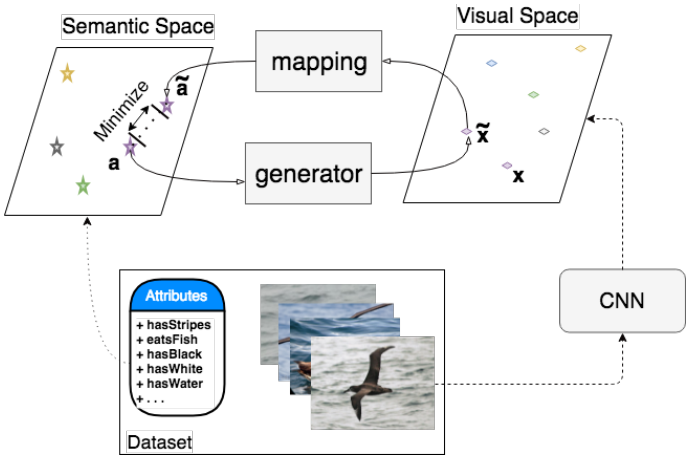


Figure 3.1: Overview of the proposed multi-modal cycle-consistent GZSL approach. Our approach extends the idea of synthesizing visual representations of seen and unseen classes in order to train a classifier for the GZSL problem [151]. The main contribution of the paper is the use of a new multi-modal cycle consistency loss in the training of the visual feature generator that minimizes the reconstruction error between the semantic feature $\mathbf{a}$, which was used to synthesize the visual feature $\widetilde{\mathbf{x}}$, and the reconstructed semantic feature $\widetilde{\mathbf{a}}$ mapped from $\widetilde{\mathbf{x}}$. This loss is shown to constrain the optimization problem more effectively in order to produce useful synthesized visual features for training the GZSL classifier.

the seen classes; while the testing process aims to classify the visual representations of both sub-sets [150, 161]. The semantic features available for both the training and testing classes are typically acquired from other domains, such as visual features [77], text [109, 133, 161], or learned classifiers [156]. The traditional approach to address this challenge [150] involves the learning of a transformation from the visual to the semantic space of the seen classes. Testing is then performed by transforming the visual representation of the seen and unseen classes into this semantic space, where classification is typically achieved with a nearest neighbor classifier that selects the closest class in the semantic space. In contrast to Zero-shot Learning (ZSL), which uses only the unseen domain for testing, GZSL approaches tend to be biased towards the seen classes, producing poor classification results, particularly for the unseen testing classes [151].

These traditional approaches rely on the assumption that the distributions observed in the semantic and visual spaces are relatively similar. Recently, this assumption has been relaxed to allow the semantic space to be optimized together with the transformation from the visual to the semantic space [88] - this alleviates the classification bias mentioned above to a certain degree. More recent approaches consist of building a generative adversarial network (GAN) that synthesizes visual representations of the seen and unseen classes directly from their semantic representation [17, 88]. These synthesized features are then used to train a multi-class classifier of seen and unseen classes. This approach has been shown to improve the GZSL classification accuracy, but an obvious weakness is that the unconstrained nature of the generation process may let the approach generate unrepresentative synthetic visual representations, particularly of the unseen classes (i.e., representations that are far from possible visual representations of the test classes).

The **main contribution** of this paper is a **new regularization of the generation of synthetic visual representations in the training of GAN-based methods that address the GZSL classification problem**. This regularization is **based on a multi-modal cycle consistency loss term that enforces good reconstruction from the synthetic visual representations back to their original semantic features** (see Fig. 3.1). This regularization is motivated by the cycle consistency loss applied in training GANs [164] that forces the generative training approach to produce

more constrained visual representations. We argue that this constraint preserves the semantic compatibility between visual features and semantic features. Once our model is trained with this multi-modal cycle consistency loss term, we can then synthesize visual representations for unseen classes in order to train a GZSL classifier [142, 151].

Using the experimental setup described by Xian et al. [150], we show that our proposed regularization provides significant improvements not only in terms of GZSL classification accuracy, but also ZSL on the following datasets: Caltech-UCSD-Birds 200-2011 (CUB) [146, 150], Oxford-Flowers (FLO) [98], Scene Categorization Benchmark (SUN) [36, 150], Animals with features (AWA) [77, 150], and *ImageNet* [26] . In fact, the experiments show that our proposed approach holds the current best ZSL and GZSL classification results in the field for these datasets.

3.2 Related Work

The starting point for our literature review is the work by Xian et al. [150, 151], who proposed new benchmarks using commonly accepted evaluation protocols on publicly available datasets. These benchmarks allow a fair comparison among recently proposed ZSL and GZSL approaches, and for this reason we explore those benchmarks to compare our results with the ones obtained from the current state of the art in the field. We provide a general summary of the methods presented in [150], and encourage the reader to study that paper in order to obtain more details on previous works. The majority of the ZSL and GZSL methods tend to compensate the lack of visual representation of the unseen classes with the learning of a mapping between visual and semantic spaces [20], [6]. For instance, a fairly successful approach is based on a bi-linear compatibility function that associates visual representation and semantic features. Examples of such approaches are ALE [3], DEVISE [42], SJE [4], ESZSL [118], and SAE [35]. Despite their simplicity, these methods tend to produce the current state-of-the-art results on benchmark datasets [150]. A straightforward extension of the methods above is the exploration of a non-linear compatibility function between visual and semantic spaces. These approaches, exemplified by LATEM [149] and CMT [133], tend not to be as competitive as their bi-linear counterpart, probably because the more

complex models need larger training sets to generalize more effectively. Seminal ZSL and GZSL methods were based on models relying on learning intermediate feature classifiers, which are combined to predict image classes (e.g., DAP and IAP) [77] – these models tend to present relatively poor classification results. Finally, hybrid models, such as SSE [161], CONSE [100], SYNC [18], rely on a mixture model of seen classes to represent images and semantic embeddings. These methods tend to be competitive for classifying the seen classes, but not for the unseen classes.

The main disadvantage of the methods above is that the lack of visual training data for the unseen classes biases the mapping between visual and semantic spaces towards the semantic features of seen classes, particularly for unseen test images. This is an issue for GZSL because it has a negative effect in the classification accuracy of the unseen classes. Recent research address this issue using GAN models that are trained to synthesize visual representations for the seen and unseen classes, which can then be used to train a classifier for both the seen and unseen classes [17,88]. However, the unconstrained generation of synthetic visual representations for the unseen classes allows the production of synthetic samples that may be too far from the actual distribution of visual representations, particularly for the unseen classes. In GAN literature, this problem is known as unpaired training [164], where not all source samples (e.g., semantic features) have corresponding target samples (e.g., visual features) for training. This creates a highly unconstrained optimization problem that has been solved by Zhu et al. [164] with a cycle consistency loss to push the representation from the target domain back to the source domain, which helped constraining the optimization problem. In this paper, we explore this idea for GZSL, which is a novelty compared to previous GAN-based methods proposed in GZSL and ZSL.

3.3 Multi-modal Cycle-consistent Generalized Zero-Shot Learning

In GZSL and ZSL [150], the dataset is denoted by $\mathcal{D} = \{(\mathbf{x}, \mathbf{a}, y)_i\}_{i=1}^{|\mathcal{D}|}$ with $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^K$ representing visual representation (e.g., image features from deep residual

nets [58]), $\mathbf{a} \in \mathcal{A} \subseteq \mathbb{R}^L$ denoting L-dimensional semantic feature (e.g., set of binary attributes [77] or a dense *word2vec* representation [92]), $y \in \mathcal{Y} = \{1, ..., C\}$ denoting the image class, and $|.|$ representing set cardinality. The set $\mathcal{Y}$ is split into seen and unseen subsets, where the seen subset is denoted by $\mathcal{Y}_S$ and the unseen subset by $\mathcal{Y}_U$, with $\mathcal{Y} = \mathcal{Y}_S \cup \mathcal{Y}_U$ and $\mathcal{Y}_S \cap \mathcal{Y}_U = \emptyset$. The dataset $\mathcal{D}$ is also divided into mutually exclusive training and testing subsets: $\mathcal{D}^{Tr}$ and $\mathcal{D}^{Te}$, respectively. Furthermore, the training and testing sets can also be divided in terms of the seen and unseen classes, so this means that $\mathcal{D}_S^{Tr}$ denotes the training samples of the seen classes, while $\mathcal{D}_U^{Tr}$ represents the training samples of the unseen classes (similarly for $\mathcal{D}_S^{Te}$ and $\mathcal{D}_U^{Te}$ for the testing set). During training, samples in $\mathcal{D}_S^{Tr}$ contain the visual representation $\mathbf{x}_i$, semantic feature $\mathbf{a}_i$ and class label y_i; while the samples in $\mathcal{D}_U^{Tr}$ comprise only the semantic feature and class label. During ZSL testing, only the samples from $\mathcal{D}_U^{Te}$ are used; while in GZSL testing, all samples from $\mathcal{D}^{Te}$ are used. Note that for ZSL and GZSL problems, only the visual representation of the testing samples is used to predict the class label.

Below, we first explain the f-CLSWGAN model [151], which is the baseline for the implementation of the main contribution of this paper: the multi-modal cycle consistency loss used in the training for the feature generator in GZSL models based on GANs. The loss, feature generator, learning and testing procedures are explained subsequently.

3.3.1 f-CLSWGAN

Our approach is an extension of the feature generation method proposed by Xian et al. [151], which consists of a classification regularized generative adversarial network (f-CLSWGAN). This network is composed of a generative model $G : \mathcal{A} \times \mathcal{Z} \rightarrow \mathcal{X}$ (parameterized by θ_G) that produces a visual representation $\widetilde{\mathbf{x}}$ given its semantic feature $\mathbf{a}$ and a noise vector $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ sampled from a multi-dimensional centered Gaussian, and a discriminative model $D : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$ (parameterized by θ_D) that tries to distinguish whether the input $\mathbf{x}$ and its semantic representation $\mathbf{a}$ represent a true or generated visual representation and respective semantic feature. Note that while the method developed by Yan et al. [154] concerns the generation of realistic images, our proposed approach, similarly to [17, 88, 151], aims to generate visual representations, such as the features from

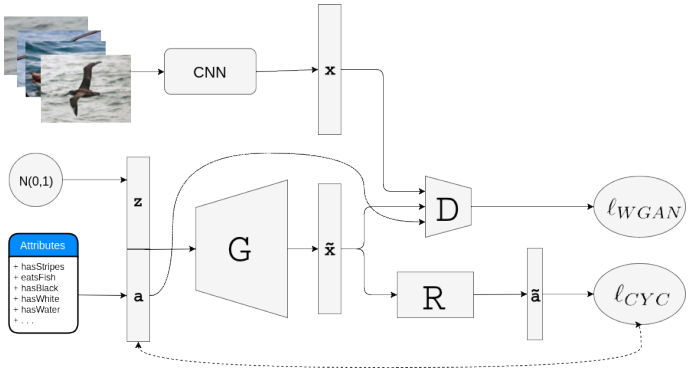


Figure 3.2: Overview of the multi-modal cycle-consistent GZSL model. The visual features, represented by $\mathbf{x}$, are extracted from a state-of-art CNN model, and the semantic features, represented by $\mathbf{a}$, are available from the training set. The generator $G(.)$ synthesizes new visual features $\widetilde{\mathbf{x}}$ using the semantic feature and a randomly sampled noise vector $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the discriminator $D(.)$ tries to distinguish between real and synthesized visual features. Our main contribution is focused on the integration of a multi-modal cycle consistency loss (at the bottom) that minimizes the error between the original semantic feature $\mathbf{a}$ and its reconstruction $\widetilde{\mathbf{a}}$, produced by the regressor $R(.)$.

a deep residual network [58] - the strategy based on visual representation has shown to produce more accurate GZSL classification results compared to the use of realistic images. The training algorithm for estimating θ_G and θ_D follows a minimax game, where $G(.)$ generates synthetic visual representations that are supposed to fool the discriminator, which in turn tries to distinguish the real from the synthetic visual representations. We rely on one of the most stable training methods for GANs, called Wasserstein GAN, which uses the following loss function [7]:

$$\theta_G^*, \theta_D^* = \arg\min_{\theta_G} \max_{\theta_D} \ell_{WGAN}(\theta_G, \theta_D), \tag{3.1}$$

with

$$\begin{aligned} \ell_{WGAN}(\theta_G, \theta_D) = \mathbb{E}_{(\mathbf{x},\mathbf{a})\sim\mathbb{P}^{x,a}}[D(\mathbf{x}, \mathbf{a}; \theta_D)] - \mathbb{E}_{(\widetilde{\mathbf{x}},\mathbf{a})\sim\mathbb{P}_G^{x,a}}[D(\widetilde{\mathbf{x}}, \mathbf{a}; \theta_D)] \\ - \lambda \mathbb{E}_{(\hat{\mathbf{x}},\mathbf{a})\sim\mathbb{P}_{\hat{\alpha}}^{x,a}}[(||\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}}, \mathbf{a}; \theta_D)||_2 - 1)^2], \end{aligned} \tag{3.2}$$

where $\mathbb{E}[.]$ represents the expected value operator, $\mathbb{P}_S^{x,a}$ is the joint distribution

of visual and semantic features from the seen classes (in practice, samples from that distribution are the ones in $\mathcal{D}_S^{Tr}$), $\mathbb{P}_G^{x,a}$ represents the joint distribution of semantic features and the visual features produced by the generative model $G(.)$, λ denotes the penalty coefficient, and $\mathbb{P}_\alpha^{x',a}$ is the joint distribution of the semantic features and the visual features produced by $\hat{\mathbf{x}} \sim \alpha\mathbf{x} + (1-\alpha)\widetilde{\mathbf{x}}$ with $\alpha \sim \mathcal{U}(0,1)$ (i.e., uniform distribution).

Finally, the f-CLSWGAN is trained with the following objective function:

$$\theta_G^*, \theta_C^*, \theta_D^* = \arg\min_{\theta_G,\theta_C}\max_{\theta_D} \ell_{WGAN}(\theta_G,\theta_D) + \beta\ell_{CLS}(\theta_C,\theta_G), \tag{3.3}$$

where $\ell_{CLS}(\theta_C,\theta_G) = -\mathbb{E}_{(\widetilde{\mathbf{x}},y)\sim\mathbb{P}_G^{x,y}}[\log P(y|\widetilde{\mathbf{x}},\theta_C)]$, with

$$P(y|\widetilde{\mathbf{x}},\theta_C) = \frac{\exp((\theta_C(y))^T\widetilde{\mathbf{x}})}{\sum_{c\in\mathcal{Y}}\exp((\theta_C(c))^T\widetilde{\mathbf{x}})} \tag{3.4}$$

representing the probability that the sample $\widetilde{\mathbf{x}}$ has been predicted with its true label y, and β is a hyper-parameter that weights the contribution of the loss function. This regularization with the classification loss was found by Xian et al. [151] to enforce $G(.)$ to generate discriminative visual representations. The model obtained from the optimization in (3.3) is referred to as **baseline** in the experiments.

3.3.2 Multi-modal Cycle Consistency Loss

The main issue present in previously proposed GZSL approaches based on generative models [17, 88, 151] is that the unconstrained nature of the generation process (from semantic to visual features) may produce image representations that are too far from the real distribution present in the training set, resulting in an ineffective multi-class classifier training, particularly for the unseen classes. The approach we propose to alleviate this problem consists of constraining the synthetic visual representations to generate back their original semantic features - this regularization has been inspired by the cycle consistency loss [164]. Figure 3.2 shows an overview of our proposal. This approach, representing the main contribution of

this paper, is represented by the following loss:

$$\begin{aligned}\ell_{CYC}(\theta_R, \theta_G) = \mathbb{E}_{\mathbf{a}\sim\mathbb{P}^a_S, \mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\mathbf{a} - R(G(\mathbf{a}, \mathbf{z}; \theta_G); \theta_R)\|_2^2\right] \\ + \mathbb{E}_{\mathbf{a}\sim\mathbb{P}^a_U, \mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\mathbf{a} - R(G(\mathbf{a}, \mathbf{z}; \theta_G); \theta_R)\|_2^2\right],\end{aligned} \tag{3.5}$$

where $\mathbb{P}^a_S$ and $\mathbb{P}^a_U$ denote the distributions of semantic features of the seen and unseen classes, respectively, and $R : \mathcal{X} \to \mathcal{A}$ represents a regressor that estimates the original semantic features from the visual representation generated by $G(.)$.

3.3.3 Feature Generation

Using the losses proposed in Sections 3.3.1 and 3.3.2, we can propose several feature generators. First, we pre-train the regressor $R(.)$ defined below in (3.6), by minimizing a loss function computed only from the seen classes, as follows:

$$\ell_{REG}(\theta_R) = \mathbb{E}_{(\mathbf{a},\mathbf{x})\sim\mathbb{P}^{a,x}_S}\left[\|\mathbf{a} - R(\mathbf{x}; \theta_R)\|_2^2\right], \tag{3.6}$$

where $\mathbb{P}^{a,x}_S$ represents the real joint distribution of image and semantic features present in the seen classes. In practice, this regressor is defined by a multi-layer perceptron, whose output activation function depends on the format of the semantic vector.

Our first strategy to build a feature generator consists of pre-training a regressor (using samples from seen classes) optimized by minimizing ℓ_{REG} in (3.6), which produces θ^*_R and training the generator and discriminator of the WGAN using the following optimization function:

$$\theta^*_G, \theta^*_D = \arg\min_{\theta_G}\max_{\theta_D} \ell_{WGAN}(\theta_G, \theta_D) + \lambda_1 \ell_{CYC}(\theta^*_R, \theta_G), \tag{3.7}$$

where ℓ_{WGAN} is defined in (3.2), ℓ_{CYC} is defined in (3.5), and λ_1 weights the importance of the second optimization term. The optimization in (3.7) can use both the seen and unseen classes, or it can rely only the seen classes, in which case the loss ℓ_{CYC} in (3.5) has to be modified so that its second term (that depends on unseen classes) is left out of the optimization. The feature generator model in (3.7) trained with seen and unseen classes is referred to as **cycle-(U)WGAN**, while the feature generator trained with only seen classes is labeled **cycle-WGAN**.

The second strategy explored in this paper to build a feature generator involves pre-training the regressor in (3.6) using samples from seen classes to produce θ_R^*, and pre-training a softmax classifier for the seen classes using ℓ_{CLS}, defined in (3.3), which results in θ_C^*. Then we train the combined loss function:

$$\theta_G^*, \theta_D^* = \arg\min_{\theta_G} \max_{\theta_D} \ell_{WGAN}(\theta_G, \theta_D) + \lambda_1 \ell_{CYC}(\theta_R^*, \theta_G) + \lambda_2 \ell_{CLS}(\theta_C^*, \theta_G). \quad (3.8)$$

The feature generator model in (3.8) trained with seen classes is referred to as **cycle-CLSWGAN**.

3.3.4 Learning and Testing

As shown in [151] the training of a classifier using a potentially unlimited number of samples from the seen and unseen classes generated with $\mathbf{x} \sim G(\mathbf{a}, \mathbf{z}; \theta_G^*)$ produces more accurate classification results compared with multi-modal embedding models [3, 4, 42, 118]. Therefore, we train a final softmax classifier $P(y|\mathbf{x}, \theta_C)$, defined in (3.4), using the generated visual features by minimizing the negative log likelihood loss $\ell_{CLS}(\theta_C, \theta_G^*)$, as defined in (3.3), where θ_G^* has been learned from one of the feature learning strategies discussed in Sec. 3.3.3 - the training of the classifier produces θ_C^*. The samples used for training the classifier are generated based on the task to be solved. For instance, for ZSL, we only use generated visual representations from the set of unseen classes; while for GZSL, we use the generated samples from seen and unseen classes.

Finally, the testing is based on the prediction of a class for an input test visual representation $\mathbf{x}$, as follows:

$$y^* = \arg\max_{y \in \widetilde{\mathcal{Y}}} P(y|\mathbf{x}, \theta_C^*), \quad (3.9)$$

where $\widetilde{\mathcal{Y}} = \mathcal{Y}$ for GZSL or $\widetilde{\mathcal{Y}} = \mathcal{Y}_U$ for ZSL.

3.4 Experiments

In this section, we first introduce the datasets and evaluation criteria used in the experiments, then we discuss the experimental set-up and finally show the results of our approach, comparing with the state-of-the-art results.

3.4.1 Datasets

We evaluate the proposed method on the following ZSL/GZSL benchmark datasets, using the experimental setup of [150], namely: CUB-200-2011 [146, 151], FLO [98], SUN [150], and AWA [76, 150] – where CUB, FLO and SUN are fine-grained datasets, and AWA coarse. Table 3.4.1 shows some basic information about these datasets in terms of number of seen and unseen classes and number of training and testing images. For CUB-200-2011 [146,151] and Oxford-Flowers [98], the semantic feature has 1024 dimensions produced by the character-based CNN-RNN [114] that encodes the textual description of an image containing fine-grained visual descriptions (10 sentences per image). The sentences from the unseen classes are not used for training the CNN-RNN and the per-class sentence is obtained by averaging the CNN-RNN semantic features that belong to the same class. For the FLO dataset [98], we used the same type of semantic feature with 1024 dimensions [114] as was used for CUB (please see description above). For the SUN dataset [150], the semantic features have 102 dimensions. Following the protocol from Xian et al. [150], visual features are represented by the activations of the 2048-dim top-layer pooling units of ResNet-101 [58], obtained from the entire image. For AWA [76,150], we use a semantic feature containing 85 dimensions denoting per-class attributes. In addition, we also test our approach on *ImageNet* [26], for a split containing 100 classes for testing [145].

The input images do not suffer any pre-processing (cropping, background subtraction, etc.) and we do not use any type of data augmentation. This ResNet-101 is pre-trained on ImageNet with 1K classes [26] and is not fine tuned. For the synthetic visual representations, we generate 2048-dim CNN features using one of the feature generation models, presented in Sec. 3.3.3.

For CUB, FLO, SUN, and AWA we use the zero-shot splits proposed by Xian et al. [150], making sure that none of the training classes are present on ImageNet [26]. Differently from these datasets (i.e., CUB, FLO, SUN, AWA), we observed that there is a lack of standardized experimental setup for GZSL on *Imagenet*. Recently, papers have used *ImageNet* for GZSL using several splits (e.g., 2-hop, 3-hop), but we noticed that some of the supposedly unseen classes can actually be seen during training (e.g., in split **2-hop**, we note that the class *American mink* is assumed

Table 3.1: Information about the datasets CUB [146], FLO [98], SUN [152], AWA [150], and ImageNet [26]. Column (1) shows the number of seen classes, denoted by $|\mathcal{Y}_S|$, split into the number of training and validation classes (train+val), (2) presents the number of unseen classes $|\mathcal{Y}_U|$, (3) displays the number of samples available for training $|\mathcal{D}^{Tr}|$ and (4) shows number of testing samples that belong to the unseen classes $|\mathcal{D}_U^{Te}|$ and number of testing samples that belong to the seen classes $|\mathcal{D}_S^{Te}|$.

Name	$\lvert\mathcal{Y}_S\rvert$ (train+val)	$\lvert\mathcal{Y}_U\rvert$	$\lvert\mathcal{D}^{Tr}\rvert$	$\lvert\mathcal{D}_U^{Te}\rvert + \lvert\mathcal{D}_S^{Te}\rvert$
CUB	150 (100+50)	50	7057	1764+2967
FLO	82 (62+20)	20	1640	1155+5394
SUN	745 (580+65)	72	14340	2580+1440
AWA	40 (27+13)	10	19832	4958+5685
ImageNet	1000 (1000 + 0)	100	1.2×10^6	5200+0

to be unseen, while class *Mink* is seen, but these two classes are arguably the same). Nevertheless, in order to demonstrate the competitiveness of our proposed **cycle-WGAN**, we compare it to the **baseline** using carefully selected 100 unseen classes [145] (i.e., no overlap with 1k training seen classes) from *ImageNet*.

3.4.2 Evaluation Protocol

We follow the evaluation protocol proposed by Xian et al. [151], where results are based on average per-class top-1 accuracy. For the ZSL evaluation, top-1 accuracy results are computed with respect to the set of unseen classes $\mathcal{Y}_U$, where the average accuracy is independently computed for each class, which is then averaged over all unseen classes. For the GZSL evaluation, we compute the average per-class top-1 accuracy on seen classes $\mathcal{Y}_S$, denoted by s, the average per-class top-1 accuracy on unseen classes $\mathcal{Y}_U$, denoted by u, and their harmonic mean, i.e. $H = 2 \times (s \times u)/(s + u)$.

3.4.3 Implementation Details

In this section, we explain the implementation details of the generator $G(.)$, the discriminator $D(.)$, the regressor $R(.)$, and the weights used for the hyper param-

Table 3.2: Summary of cross-validated hyper-parameters in our experiments.

	$R(.)$			GAN: $G(.)$ and $D(.)$				Classifier		
	$lr_{R(.)}$	**batch**	**#ep**	$lr_{G(.)}$	$lr_{D(.)}$	**batch**	**#ep**	lr	**batch**	**#ep**
CUB	$1e^{-4}$	64	100	$1e^{-4}$	$1e^{-3}$	64	926	$1e^{-4}$	4096	80
FLO	$1e^{-4}$	64	100	$1e^{-4}$	$1e^{-3}$	64	926	$1e^{-4}$	2048	100
SUN	$1e^{-4}$	64	100	$1e^{-2}$	$1e^{-2}$	64	926	$1e^{-4}$	4096	298
AWA	$1e^{-3}$	64	50	$1e^{-4}$	$1e^{-3}$	64	350	$1e^{-4}$	2048	37
ImageNet	$1e^{-4}$	2048	5	$1e^{-4}$	$1e^{-3}$	256	300	$1e^{-3}$	2048	300

eters in the loss functions in (3.2),(3.3),(3.7) and (3.8) - all these terms have been formally defined in Sec. 3.3 and depicted in Fig. 3.2. The generator consists of a multi-layer perceptron (MLP) with a single hidden layer containing 4096 nodes, where this hidden layer is activated by LeakyReLU [90], and the output layer, with 2048 nodes, has a ReLU activation [97]. The weights of $G(.)$ are initialized with a truncated normal initialization with mean 0 and standard deviation 0.01 and the biases are initialized with 0. The discriminator $D(.)$ is also an MLP consisting of a single hidden layer with 4096 nodes, which is activated by LeakyReLU, and the output layer has no activation. The initialization of $D(.)$ is the same as for $G(.)$. The regressor $R(.)$ is a linear transform from the visual space $\mathcal{X}$ to the semantic space $\mathcal{A}$. Following [151], we set $\lambda = 10$ in (3.2), $\beta = 0.01$ in (3.3) and $\lambda_1 = \lambda_2 = 0.01$ in (3.7) and (3.8). We ran an empirical evaluation with the training set and noticed that when λ_1 and λ_2 share the same value, the training becomes stable, but a more systematic evaluation to assess the relative importance of these two hyper-parameters is still needed. Table 3.2 shows the learning rates for each model (denoted by $lr_{\{R(.),G(.),D(.)\}}$), batch sizes (**batch**) and number of epochs (**#ep**) used for each dataset and model – the values for $G(.)$ and $D(.)$ have been estimated to reproduce the published results of our implementation of f-CLSWGAN (explained below), and the values for $R(.)$ have been estimated by cross validation using the training and validation sets.

Regarding the number of visual representations generated to train the classifier, we performed a few experiments and reached similar conclusions, compared to [151]. For all experiments in the paper, we generated 300 visual representations per class [151]. We reached this number after a study that shows that for a

small number of representations (below 100), the classification results were not competitive; for values superior to 200 or more, results became competitive, but unstable; and above 300, results were competitive and stable.

Table 3.3: Comparison between the reported results of **f-CLSWGAN** [151] and our implementation of it, labeled **baseline**, where we show the top-1 accuracy on the unseen test $\mathcal{Y}_U$ (GZSL), the top-1 accuracy for seen test $\mathcal{Y}_S$ (GZSL), the harmonic mean H (GZSL), and the top-1 accuracy for ZSL ($T1_Z$).

	CUB				**FLO**				**SUN**				**AWA**			
Classifier	$\mathcal{Y}_U$	$\mathcal{Y}_S$	H	$T1_Z$	$\mathcal{Y}_U$	$\mathcal{Y}_S$	H	$T1_Z$	$\mathcal{Y}_U$	$\mathcal{Y}_S$	H	$T1_Z$	$\mathcal{Y}_U$	$\mathcal{Y}_S$	H	$T1_Z$
f-CLSWGAN [151]	43.7	57.7	49.7	57.3	59.0	73.8	65.6	67.2	42.6	36.6	39.4	60.8	57.9	61.4	59.6	68.2
baseline	43.8	60.6	50.8	57.7	58.8	70.0	63.9	66.8	47.9	32.4	38.7	58.5	56.0	62.8	59.2	64.1

Since our approach is based on the f-CLSWGAN [151], we re-implemented this methodology. In the experiments, the results from our implementation of f-CLSWGAN using a softmax classifier is labeled as **baseline**. The results that we obtained from our baseline are very similar to the reported results in [150], as shown in Table 3.3. For ImageNet, note that we use a split [145] that is different from previous ones used in the literature, as explained above in Sec. 3.4.1, so it is not possible to have a direct comparison between f-CLSWGAN [151] and our **baseline**. Nevertheless, we show in Table 3.6 that the results we obtain for the split [145] are in fact similar to the reported results for f-CLSWGAN [151] for similar ImageNet splits. We developed our code [1] and perform all experiments using Tensorflow [1].

3.5 Results

In this section we show the GZSL and ZSL results using our proposed models **cycle-WGAN**, **cycle-(U)WGAN** and **cycle-CLSWGAN**, the baseline model f-CLSWGAN, denoted by **baseline**, and several other baseline methods previously used in the field for benchmarking [150]. Table 3.4 shows the **GZSL results** and Table 3.5 shows the **ZSL results** obtained from our proposed methods, and several baseline approaches on CUB, FLO, SUN and AWA datasets. The results

[1] Code is available at: https://github.com/rfelixmg/frwgan-eccv18

in Table 3.6 shows that the top-1 accuracy on ImageNet for **cycle-WGAN** and **baseline** [151].

Table 3.4: GZSL results using per-class average top-1 accuracy on the test sets of unseen classes $\mathcal{Y}_U$, seen classes $\mathcal{Y}_S$, and the harmonic mean result H – all results shown in percentage. Results from previously proposed methods in the field extracted from [150]

	CUB			**FLO**			**SUN**			**AWA**		
Classifier	$\mathcal{Y}_U$	$\mathcal{Y}_S$	H	$\mathcal{Y}_U$	$\mathcal{Y}_S$	H	$\mathcal{Y}_U$	$\mathcal{Y}_S$	H	$\mathcal{Y}_U$	$\mathcal{Y}_S$	H
DAP [76]	4.2	25.1	7.2	–	–	–	1.7	67.9	3.3	0.0	**88.7**	0.0
IAP [76]	1.0	37.8	1.8	–	–	–	0.2	**72.8**	0.4	2.1	78.2	4.1
DEVISE [42]	23.8	53.0	32.8	9.9	44.2	16.2	16.9	27.4	20.9	13.4	68.7	22.4
SJE [4]	23.5	59.2	33.6	13.9	47.6	21.5	14.7	30.5	19.8	11.3	74.6	19.6
LATEM [149]	15.2	57.3	24.0	6.6	47.6	11.5	14.7	28.8	19.5	7.3	71.7	13.3
ESZSL [118]	12.6	**63.8**	21.0	11.4	56.8	19.0	11.0	27.9	15.8	6.6	75.6	12.1
ALE [3]	23.7	62.8	34.4	13.3	61.6	21.9	21.8	33.1	26.3	16.8	76.1	27.5
SAE [35]	8.8	18.0	11.8	–	–	–	7.8	54.0	13.6	1.8	77.1	3.5
baseline [151]	43.8	60.6	50.8	58.8	70.0	63.9	47.9	32.4	38.7	56.0	62.8	59.2
`cycle-WGAN`	46.0	60.3	52.2	59.1	71.1	64.5	48.3	33.1	39.2	56.4	63.5	59.7
`cycle-CLSWGAN`	45.7	61.0	52.3	59.2	**72.5**	65.1	**49.4**	33.6	**40.0**	56.9	64.0	**60.2**
`cycle-(U)WGAN`	**47.9**	59.3	**53.0**	**61.6**	69.2	**65.2**	47.2	33.8	39.4	**59.6**	63.4	59.8

3.6 Discussion

Regarding the GZSL results in Table 3.4, we notice that there is a clear trend of all of our proposed feature generation methods (**cycle-WGAN**, **cycle-(U)WGAN**), and **cycle-CLSWGAN**) to perform better than **baseline** on the unseen test set. In particular, it seems advantageous to use the synthetic samples from unseen classes to train the **cycle-(U)WGAN** model since it achieves the best top-1 accuracy results in 3 out of the 4 datasets, with improvements from 0.7% to more than 4%. In general, the top-1 accuracy improvement achieved by our approaches in the seen test set is less remarkable, which is expected given that we prioritize to improve the results for the unseen classes. Nevertheless, our approaches achieved improvements from 0.4% to more than 2.5% for the seen classes. Finally, the harmonic

Table 3.5: ZSL results using per-class average top-1 accuracy on the test set of unseen classes $\mathcal{Y}_U$ – all results shown in percentage. Results from previously proposed methods in the field extracted from [150]

	ZSL			
Classifier	**CUB**	**FLO**	**SUN**	**AWA**
DEVISE [42]	52.0	45.9	56.5	54.2
SJE [4]	53.9	53.4	53.7	65.6
LATEM [149]	49.3	40.4	55.3	55.1
ESZSL [118]	53.9	51.0	54.5	58.2
ALE [3]	54.9	48.5	58.1	59.9
baseline [151]	57.7	66.8	58.5	64.1
`cycle-WGAN`	57.8	68.6	59.7	65.6
`cycle-CLSWGAN`	58.4	70.1	**60.0**	66.3
`cycle-(U)WGAN`	**58.6**	**70.3**	59.9	**66.8**

Table 3.6: ZSL and GZSL ImageNet results using per-class average top-1 accuracy on the test sets of unseen classes $\mathcal{Y}_U$ – all results shown in percentage.

Classifier	**ZSL**	**GZSL**
baseline [151]	7.5	0.7
`cycle-WGAN`	8.7	1.5

mean results also show that our approaches improve over the **baseline** in a range of between 1% and 2.2%. Notice that this results are remarkable considering the outstanding improvements achieved by f-CLSWGAN [151], represented here by **baseline**. In fact, our proposed methods produce the current state of the art GZSL results for these four datasets.

Analyzing the ZSL results in Table 3.5, we again notice that, similarly to the GZSL case, there is a clear advantage in using the synthetic samples from unseen classes to train the **cycle-(U)WGAN** model. For instance, top-1 accuracy results show that we can improve over the **baseline** from 0.9% to 3.5%. The results in this table show that our proposed approaches currently hold the best ZSL results for these datasets.

It is interesting to see that, compared to GZSL, the ZSL results from previous method in the literature are far more competitive, achieving results that are relatively close to ours and the **baseline**. This performance gap between ZSL and GZSL, shown by previous methods, enforces the argument in favor of using generative models to synthesize images from seen and unseen classes to train GZSL models [17, 88, 151]. As argued throughout this paper, the performance produced by generative models can be improved further with methods that help the training of GANs, such as the cycle consistency loss [164].

In fact, the experiments clearly demonstrate the advantage of using our proposed multi-modal cycle consistency loss in training GANs for GZSL and ZSL. In particular, it is interesting to see that the use of synthetic examples of unseen classes generated by **cycle-(U)WGAN** to train the GZSL classifier provides remarkable improvements over the **baseline**, represented by f-CLSWGAN [151]. The only exception is with the SUN dataset, where the best result is achieved by **cycle-CLSWGAN**. We believe that **cycle-(U)WGAN** is not the top performer on SUN due to the number of classes and the proportion of seen/unseen classes in this dataset. For CUB, FLO and AWA we notice that there is roughly a $(80\%, 20\%)$ ratio between seen and unseen classes. In contrast, SUN has a $(91\%, 9\%)$ ratio between seen and unseen classes. We also notice a sharp increase in the number of classes from 50 to 817 – GAN models tend not to work well with such a large number of classes. Given the wide variety of GZSL datasets available in the field, with different number of classes and seen/unseen proportions, we believe that

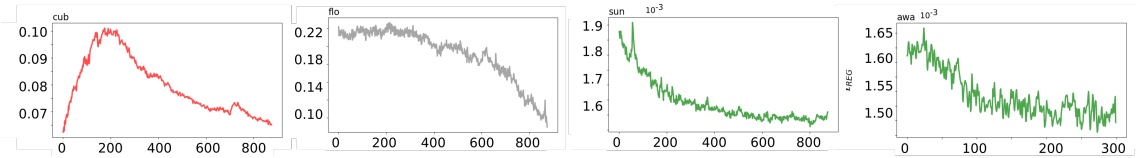


Figure 3.3: Evolution of ℓ_{REG} in terms of the number of epochs for CUB, FLO, SUN and AWA.

there is still lots of room for improvement for GZSL models.

Regarding the large-scale study on ImageNet, the results in Table 3.6 show that the top-1 accuracy classification results for **Baseline** and **cycle-WGAN** are quite low (similarly to the results observed in [151] for several ImageNet splits), but our proposed approach still shows more accurate ZSL and GZSL classification.

An important question about out approach is whether the regularisation succeeds in mapping the generated visual representations back to the semantic space. In order to answer this question, we show in Fig. 3.3 the evolution of the reconstruction loss ℓ_{REG} in (3.6) as a function of the number of epochs. In general, the reconstruction loss decreases steadily over training, showing that our model succeeds at such mapping. Another relevant question is if our proposed methods take more or less epochs to converge, compared to the **Baseline** – Fig. 3.4 shows the classification accuracy of the generated training samples from the seen classes for the proposed models **cycle-WGAN** and **cycle-CLSWGAN**, and also for the **baseline** (note that **cycle-(U)WGAN** is a fine-tuned model from the **cycle-WGAN**, so their loss functions are in fact identical for the seen classes shown in the graph). For three out of four datasets, our proposed **cycle-WGAN** converges faster. However, when the ℓ_{CLS} in included in (3.7) to form the loss in (3.8) (transforming **cycle-WGAN** into **cycle-CLSWGAN**), then the convergence of **cycle-CLSWGAN** is comparable to that of the **baseline**. Hence, **cycle-WGAN** tends to converge faster than the **baseline** and **cycle-CLSWGAN**.

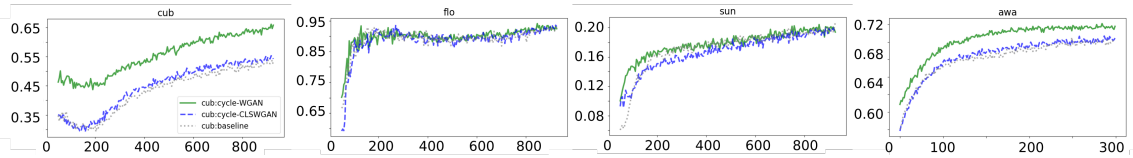


Figure 3.4: Convergence of the top-1 accuracy in terms of the number of epochs for the generated training samples from the seen classes for CUB, FLO, SUN and AWA.

3.7 Conclusions and Future Work

In this paper, we propose a new method to regularize the training of GANs in GZSL models. The main argument explored in the paper is that the use of GANs to generate seen and unseen synthetic examples for training GZSL models has shown clear advantages over previous approaches. However, the unconstrained nature of the generation of samples from unseen classes can produce models that may not work robustly for some unseen classes. Therefore, by constraining the generation of samples from unseen classes, we target to improve the GZSL classification accuracy. Our proposed constraint is motivated by the cycle consistency loss [164], where we enforce that the generated visual representations maps back to their original semantic feature – this represents the multi-modal cycle consistency loss. Experiments show that the use of such loss is clearly advantageous, providing improvements over the current state of the art f-CLSWGAN [151] both in terms of GZSL and ZSL.

As noticed in Sec. 3.6, GAN-based GZSL approaches offer indisputable advantage over previously proposed methods. However, the reliance on GANs to generate samples from unseen classes is challenging because GANs are notoriously difficult to train, particularly in unconstrained and large scale problems. Therefore, future work in this field should be focused on targeting these problems. In this paper, we provide a solution that addresses the unconstrained problem, but it is clear that other regularization approaches could also be used. In addition, the use of GANs in large scale problems (regarding the number of classes) should also be more intensively studied, particularly when dealing with real-life datasets

and scenarios. Therefore, we will focus our future research activities in solving these two issues in GZSL.

Chapter 4

Generalised Zero-Shot Learning with Domain Classification in a Joint Semantic and Visual Space

The work contained in this chapter has been published as the following paper:

Felix, R., Harwood, B., Sasdelli, M., and Carneiro, G., Generalised Zero-Shot Learning with Domain Classification in a Joint Semantic and Visual space. In *Digital Image Computing: Techniques and Applications* (DICTA), pages –. IEEE, 2019 [37].

Statement of Authorship

Title of Paper	Generalised zero-shot learning with domain classification in a joint semantic and visual space
Publication Status	☐ Published ☒ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished or Unsubmitted work
Publication Details	Felix, R., Harwood, B., Sasdelli, M., and Carneiro, G., Generalised zero-shot learning with domain classification in a joint semantic and visual space. In Digital Image Computing: Techniques and Applications (DICTA), IEEE, 2019

Principal Author

Name of Principal Author (Candidate)	Rafael Felix Alves
Contribution to the Paper	- Development of the main idea of the paper; - Implementing and conducting the experiments; - Writing and coordinating the revisions;
Overall percentage (%)	70
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 29/11/2019

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Ben Harwood
Contribution to the Paper	- Help writing, revision and discussions;

Signature		Date	06/11/2019

Name of Co-Author	Michele Sasdelli		
Contribution to the Paper	- Help with the development of the idea; - Help writing, revision and discussions;		
Signature		Date	29/11/2019

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	- Help writing, revision and discussions;		
Signature		Date	04-11-2019

Abstract

Generalised zero-shot learning (GZSL) is a classification problem where the learning stage relies on a set of seen visual classes and the inference stage aims to identify both the seen visual classes and a new set of unseen visual classes. Critically, both the learning and inference stages can leverage a semantic representation that is available for the seen and unseen classes. Most state-of-the-art GZSL approaches rely on a mapping between latent visual and semantic spaces without considering if a particular sample belongs to the set of seen or unseen classes. In this paper, we propose a novel GZSL method that learns a joint latent representation that combines both visual and semantic information. This mitigates the need for learning a mapping between the two spaces. Our method also introduces a domain classification that estimates whether a sample belongs to a seen or an unseen class. Our classifier then combines a class discriminator with this domain classifier with the goal of reducing the natural bias that GZSL approaches have toward the seen classes. Experiments show that our method achieves state-of-the-art results in terms of harmonic mean, the area under the seen and unseen curve and unseen classification accuracy on public GZSL benchmark data sets. Our code will be available upon acceptance of this paper.

4.1 Introduction

Humans have a powerful ability to learn about new visual objects without actually seeing them. This process generally involves the use of language to describe how a new visual object would look like. The textual description then allows for a new class of object to be formed in a person's mind. Our understanding of exactly how the human brain functions for this task is limited, but it is clear that humans make some sort of association between visual objects and semantic textual descriptions. Conceptually, objects with similar descriptions can naturally be viewed as being near to each other in some latent space, representing visual and semantic information. The research topic is known as generalised zero-shot learning (GZSL) aims to mimic this recognition ability of humans. In general, GZSL approaches employ an auxiliary set of semantic information that describes a

set of visual classes. This additional information, such as tags or descriptions, can be utilised to overcome missing visual information in some of the classes [150].

Traditional GZSL approaches aim to recognise the visual classes available during the training process (i.e. the **seen**, source or known classes), and also classes that are not available during training (i.e. **unseen**, target or novel classes). Due to this constraint, GZSL approaches are intrinsically divided into two main tasks: (1) the training of a model that learns a transformation from the visual to the semantic space, using the visual samples and semantic information from seen classes; and (2) the transformation of a new test image by the model above into the semantic space, followed by a search of the closest semantic sample representing a seen or unseen class. In recent years, GZSL researchers have become increasingly interested in pairwise functions for disentangling these domains [42], and deep generative models [39, 129] for learning to transform between the visual and semantic representations. In general, GZSL methods do not try to estimate if a test sample belongs to the set of seen or unseen classes – this issue inevitably biases GZSL approaches toward seen classes. Only recently this issue has been acknowledged with a method that automatically combines the classification of Zero-Shot Learning (ZSL) for unseen classes with the classification of seen classes, by automatically weighting (using the test sample) the contribution of each classifier [8]. Although that approach is in the right direction, it has the issue of relying on the training of multiple classifiers. Another issue with the methods above is that they do not consider a latent space jointly optimised for the visual and semantic representation, which we believe is a crucial part of the inference process performed by humans that should be imitated by GZSL methods. In Fig. 4.1, we illustrate the idea explored in this paper for GZSL. The visual and semantic samples are represented in a joint latent space. This space is used to learn a classifier of visual classes and a domain classifier for seen and unseen domains.

In this paper, we aim to explore two observations about the latent space for the domain classification. The first observation is that samples from unseen classes that are visually similar to one of the seen classes tend to be projected relatively close to other seen classes distributions, instead of outside of the distribution of seen classes, as proposed by Socher et al. [133]. Our second observation is that samples from unseen classes that are visually different from any of the seen classes, tend to

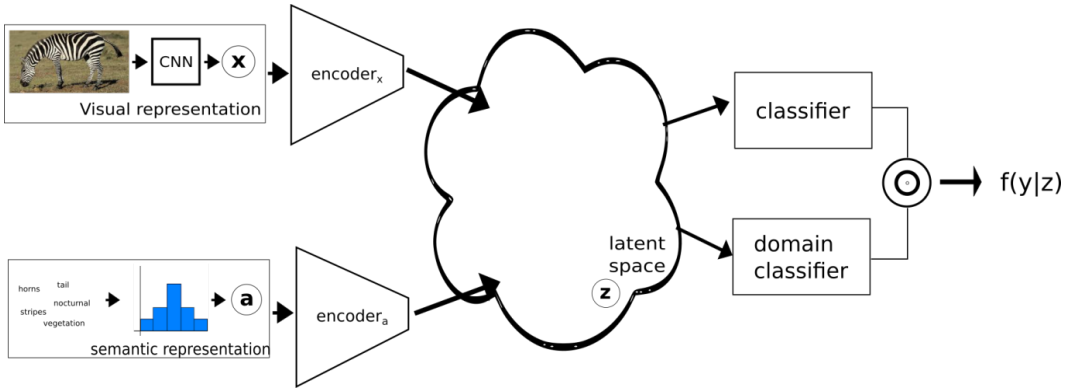


Figure 4.1: Depiction of the method proposed in this paper – our approach learns the latent space for the visual and semantic modalities. We train two classifiers using samples from this latent space: one to classify all the seen and unseen visual classes, and another to classify between the seen and unseen domains. The final classification combines the results of these two classifiers.

be projected outside the distribution of seen classes [133]. Atzmon and Chechik [8] propose a general framework that combines domain expert classifiers, such as DAP [76] for unseen classes, and LAGO for the seen classes [8]. However, this method relies on the disjoint training of both experts models, and the assumption that unseen samples are projected outside the distribution of seen classes [133]. Hence, this method can be considered to be in general sub-optimal. We propose a general framework for learning and combining the visual and domain classifiers using the latent space. More specifically, we first introduce a general framework for latent space learning from cycle-WGAN [39] and CADA-VAE [129]. Then, we propose a novel method for the seen and unseen domain classification from this latent space. Finally, we introduce a way to combine the visual and domain classifiers. The empirical results show that our proposed framework outperforms previous approaches in terms of unseen accuracy and harmonic mean (**H-mean**) on several GZSL benchmark data sets, such as CUB [146], SUN [150], AWA1 [76, 150] and AWA2 [76,150]. In terms of unseen accuracy, our method shows improvements of 4.5%, 5.6%, 2.5%, 1.5% for CUB, SUN, AWA1, and AWA2, respectively. Moreover, our method shows substantial improvements in terms of area under the curve of seen and unseen accuracy (**AUSUC**) [19]. For AUSUC we improved from 0.3698, 0.5238, 0.5216 to 0.3743, 0.5247, 0.5219, on CUB, AWA1 and AWA2, respectively.

4.2 Related Work

In this section, we discuss relevant literature that motivates and contextualises our work.

4.2.1 Traditional Zero-Shot Learning

Zero-shot learning (ZSL) is similar to GZSL, with a crucial difference: during inference, only the visual samples from the unseen classes are considered [77,150]. This difference makes ZSL a special case of GZSL. Therefore, critical problems present in GZSL are not considered in this approach, such as the natural bias of the visual classifier toward the seen classes. Unfortunately, this setup not only reduces the applicability of ZSL methods but also makes it unrealistic for real-world applications [39,151]. Also, ZSL fails to handle jointly the seen and unseen data [18,19]. Due to the simplicity and unrealistic assumptions of ZSL, the whole field moved toward the GZSL problem, which is introduced in the next section.

4.2.2 Generalised Zero-Shot Learning

In GZSL, the algorithm is trained using visual samples from the seen classes, but the inference involves the analysis of samples from the seen and unseen classes. The main issue faced by GZSL methods is the bias toward the seen classes naturally present during inference, so a great deal of research has focused on mitigating this problem [39,151]. Particularly important examples of this type of research are anomaly detection [133], domain balancing [19] and generative data augmentation for GZSL [39,129,151]. Despite the advances in GZSL with the approaches mentioned above, we note that little attention has been devoted to addressing the seen/unseen domain classification in GZSL based on a latent space that is jointly learned to represent the visual and semantic representations. Moreover, we argue that the multi-modal nature of this joint latent space carries interesting properties to perform domain classification. In this paper, we show that classifying the seen and unseen domains plays an important role in improving domain balancing in GZSL.

4.2.3 Data Augmentation for Zero-Shot Learning

A particularly successful GZSL method is based on data augmentation, where artificial visual samples of the unseen classes are generated from the semantic representation to train the visual classifier [39, 129, 144, 151]. This approach has produced the current state-of-the-art results in GZSL benchmark data sets. Overall, these studies focus on how to learn generative models conditioned on the semantic information that is used to augment the data set for the unseen classes. Among the main approaches, we observe the use of Generative Adversarial Networks (GAN) [39, 151] and Variational Autoencoders (VAE) [129, 144]. In this paper, we formalise these approaches as a framework for generative probabilistic latent space learning. Additionally, we show that these latent spaces have interesting properties that allow our approach to classifying samples into the seen or unseen domains for GZSL.

4.2.4 Domain Classification

Recent research has tackled the problem of GZSL as a novelty detection problem [133]. This approach assumes that unseen classes are projected out of the distribution of seen classes. Therefore, these unseen classes samples can be handled as an outlier of the seen classes distribution [133]. However, this approach fails to notice that samples from unseen classes can be projected relatively close to one of the seen classes. Atzmon and Chechik [8] aims to tackle this novelty detection issue by providing a framework that handles domain classification for GZSL. The gist of that approach consists of a gating method that performs domain adaptation to combine an unseen class classifier (e.g., DAP [77], DeVISE [42]), and CMT [133]), and a seen class classifier [8]. Even though this method achieves remarkable performance in GZSL, it still relies on a sub-optimal disjoint training of multiple classifiers. In this paper, we mitigate these two issues by combining a seen/unseen class discriminator with a domain classifier that uses samples from a latent space that is trained to represent both the visual and semantic spaces.

4.3 Method

In this section, we introduce the problem formulation and our proposed approach.

4.3.1 Generalised Zero-Shot Learning

In order to formulate the method of learning a classifier that can recognise visual samples from unseen visual classes, we define a visual data set $\mathcal{D} = \{(\mathbf{x}, y)_i\}_{i=1}^{N}$, where $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^K$ denotes the visual representation, and $y \in \mathcal{Y} = \{1, ..., C\}$ denotes the visual class. Recent research shows that such visual representation, $\mathbf{x}$, can be acquired from networks specialised in feature extraction. These are widely available in the literature, such as pre-trained deep residual nets [58].

In GZSL, the set of classes $\mathcal{Y}$ is split into two domains: seen domain $\mathcal{Y}^S = \{1, ..., |S|\}$, and the unseen domain $\mathcal{Y}^U = \{(|S| + 1), ..., (|S| + |U|)\}$. Hence, the total number of classes is $C = |S| + |U|$, with $\mathcal{Y} = \mathcal{Y}^S \cup \mathcal{Y}^U$, $\mathcal{Y}^S \cap \mathcal{Y}^U = \emptyset$. During training, we can only access visual samples from $\mathcal{Y}^S$, but during testing, samples can come from any class in $\mathcal{Y}$. This lack of visual samples from unseen classes during training is compensated with a semantic data set that includes semantic information for the seen and unseen classes. Therefore, we introduce the semantic data set $\mathcal{R} = \{(\mathbf{a}, y)_j\}_{j \in \mathcal{Y}}$, which associates visual classes with semantic samples, where $\mathbf{a} \in \mathcal{A} \subseteq \mathbb{R}^L$ represents a semantic feature (e.g., set of continuous features such as *word2vec* [150], or *BoW*). Note that the semantic data set only has a single element per class.

In comparison with the supervised learning paradigm, the problem of GZSL has a distinct setup. The data set $\mathcal{D}$ is divided into mutually exclusive training and testing visual subsets $\mathcal{D}^{Tr}$ and $\mathcal{D}^{Te}$, respectively. The $\mathcal{D}^{Tr}$ contains a subset of the visual samples belonging to the seen classes, and $\mathcal{D}^{Te}$ contains the visual samples from the seen classes that are held out from training and all samples from the unseen classes. The training data set is composed of the semantic data set $\mathcal{R}$ and the training visual subset $\mathcal{D}^{Tr}$, while the testing data set relies only on the testing visual subset $\mathcal{D}^{Te}$.

4.3.2 Data Augmentation Framework

In this section, we first introduce the components for the latent space learning applied to GZSL models, then we describe CADA-VAE and cycle-WGAN. Finally, we introduce the domain classification for these latent space.

In recent years, we note an increasing number of models that use data augmentation for GZSL models [35, 39, 96, 129, 151, 166]. Overall, these methods aim to learn a generative model that produces artificial samples from unseen visual classes conditioned on their semantic representation. These artificial samples lie in a latent space. In this paper, we aim to demonstrate that our proposed domain classification can be adapted to GZSL models that rely on data augmentation, such as CADA-VAE [129] and cycle-WGAN [39]. Although these two models consist of different training approaches, we observe that their components can be generally described as a framework for latent space learning. Below, we introduce three components of such models: the encoder (or generator), the decoder (or regressor), and the discriminator.

The encoder transforms samples from an input space (i.e., visual or semantic) into a latent space. We represent the encoder with

$$\mathbf{z}_x = Encoder_x(\mathbf{x}) \tag{4.1}$$

for the visual space and similarly for the semantic space with $\mathbf{z}_a = Encoder_a(\mathbf{a})$, where the vector $\mathbf{z}_{\{x,z\}} \in \mathbb{R}^Z$ lies in the latent space. The decoder transforms from the latent space into one of the input modalities. We represent the decoder with

$$\tilde{\mathbf{x}} = Decoder_x(\mathbf{z}), \tag{4.2}$$

and similarly for the semantic space with $\tilde{\mathbf{a}} = Decoder_a(\mathbf{z})$. The latent space discriminator, used to determine whether a sample $\mathbf{z}$ belongs to the latent space given the input $\mathbf{x}$, is represented by

$$p(\mathbf{z} \mid \mathbf{x}) = Discriminator(\mathbf{z}; \mathbf{x}). \tag{4.3}$$

We consider the simplified models above to describe CADA-VAE [129] and cycle-WGAN [39] as the latent space learning models.

CADA-VAE: This model is a special type of variational autoencoder (VAE) for GZSL [129]. In this approach, the VAE aims to learn the latent space with cross

alignment and distribution alignment losses, as depicted in Fig. 4.2. The overall loss by Schonfeld et al. [129] can be described with

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{VAE} + \gamma\Big(\sum_{i}^{L}\sum_{j\neq i}^{L} || \mathbf{x}^{(j)} - \tilde{\mathbf{x}}^{(i)} || \Big) \\ &+ \delta\Big(|| \mu^{(j)} - \mu^{(i)} ||_2^2 + || \Sigma_{(j)}^{\frac{1}{2}} - \Sigma_{(i)}^{\frac{1}{2}} ||_{Frobenius}^2 \Big),\end{aligned} \tag{4.4}$$

where the first term represents the VAE loss [129], the second term denotes the reconstruction error between L modalities – that is, during training, the encoder projects input samples in the latent space (e.g. $Encoder_x$ for $\mathbf{x}$), then the decoder of a different modality is used (e.g.$Decoder_a$ from $\mathbf{z}_x$ – see Fig. 4.2), which constraints the visual and semantic projections to be in the same region of the latent space represented by the mean μ and variance Σ of the samples produced by the encoder [129].

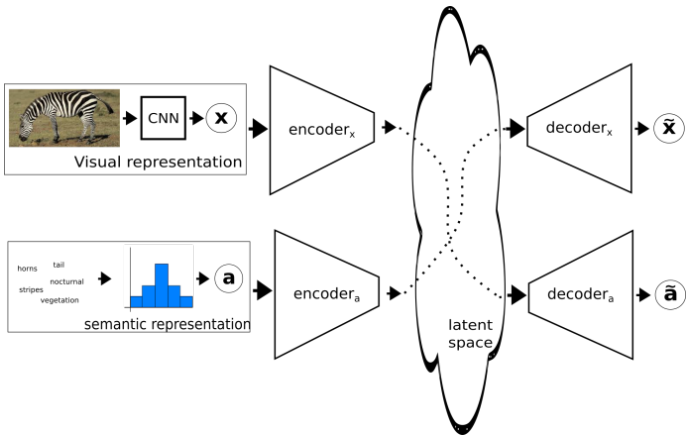


Figure 4.2: Depiction of the method CADA-VAE [129]. In this method encoders for the visual and semantic representation project samples into a shared latent space.

cycle-WGAN: Fig. 4.3 depicts the model cycle-WGAN [39]. This model is optimised as a Generative Adversarial Network (GAN), regularised by a cycle consistent term, described with

$$\mathcal{L} = \mathcal{L}_{WGAN} + \gamma\big(|| \mathbf{a} - \tilde{\mathbf{a}} ||_2^2 \big), \tag{4.5}$$

where the first term, $\mathcal{L}_{WGAN}$, represents a *Wasserstein* Generative Adversarial Loss (WGAN [39]), and the second term denotes the reconstruction loss (cycle) for

the semantic representation. Thus, the generative projection of a given semantic representation into the latent space is encouraged to be back projected near the original semantic representation.

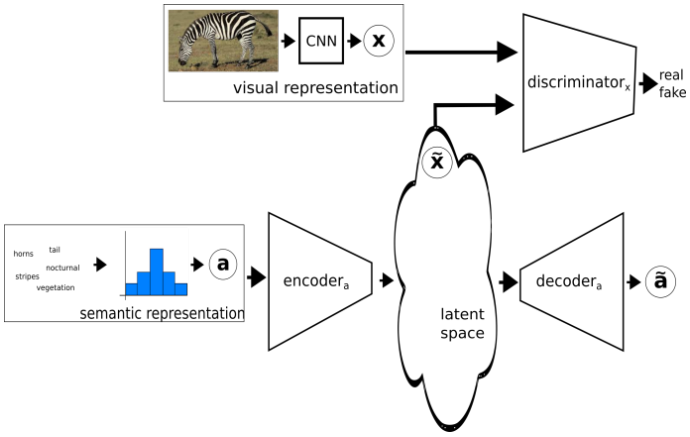


Figure 4.3: Depiction of the cycle-WGAN method [39]. This method encodes the semantic space into a latent visual space. The decoder produces semantic vectors that are used to regularise the learning process.

4.3.3 Domain Classification

From the previous section, we note that the latent space is an embedding space for visual and semantic samples. Therefore, we can use this latent space to learn a discriminative model given by

$$f(y \mid \mathbf{x}) = \int_v \int_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x}) f(y, v \mid \mathbf{z}) dv d\mathbf{z}, \tag{4.6}$$

where the function $f(.)$ represents the GZSL classifier and can be described in terms of domains, $v \in \{s, u\}$ ($s = seen$ and $u = unseen$), with

$$f(y \mid \mathbf{x}) = \sum_{v \in \{s,u\}} p(y \mid \mathbf{z}_x) f(v \mid \mathbf{z}_x, y), \tag{4.7}$$

where we assume from (4.6) that $p(\mathbf{z} \mid \mathbf{x})$ is a delta function at $\mathbf{z}_x = Encoder_x(\mathbf{x})$. The term $p(y \mid \mathbf{z}_x)$ in (4.7) is represented by a simple deep learning classifier with softmax activation. We define the function $f(.)$ in (4.7) by

$$f(v \mid \mathbf{z}_x, y) = \begin{cases} p(v \mid \mathbf{z}_x, y), \text{ if } v, y \text{ are in same domain} \\ 0, \text{ otherwise,} \end{cases} \tag{4.8}$$

where "same domain" means the domain of seen or the unseen classes, and $p(v \mid \mathbf{z}_x, y)$ is denoted by a deep learning classifier with softmax activation. The function in (4.8) represents our proposed domain classifier (DC). During the DC training, for training samples of the seen domain, we optimise $p(v = s|\mathbf{z}, y)$ with samples drawn from the latent space. These samples are acquired from visual and semantic representations projected in the latent space. For the unseen domain, $p(v = u|\mathbf{z}, y)$, we use the semantic projections in the latent space.

4.4 Experiments

In this section, we present the benchmark datasets, as well as the evaluation criteria for our experimental setup. We then show the results of our method and compare them with the current state-of-the-art. Finally, we provide ablation studies to explore our method.

4.4.1 Data Sets

We assess our method on four publicly available benchmark GZSL data sets: CUB-200-2011 [146]; SUN [150]; AWA1 [76,150], and AWA2 [76,150]. To guarantee that our experiments are reproducible, we use the GZSL experimental setup described by Xian et al. [150]. As the CUB data set is generally regarded as fine-grained, there is an intrinsic expectation that the novel unseen classes tend to have their class modes close to the seen classes. Thus, such dense visual representation space is a challenging problem for GZSL approaches. We also explore the use of coarse data sets, such as AWA1, AWA2, and SUN. Given the diversity of classes for such coarse data sets, there is an intrinsic expectation that novel classes will be projected far away from the samples of seen classes in the latent space, making the domain classification a trivial task. However, we argue that this statement does not always hold, particularly for classes that are visually similar (e.g. zebra/horse, whale/dolphin, leopard/bobcat), as depicted in Fig. 4.4. Table 4.1 contains some basic information about the data sets in terms of the number of seen and unseen classes and the number of training and testing images.

We represent the visual space by extracting image features from the activation of the 2048-dimensional top pooling layer of ResNet-101 [58]. For the semantic

Figure 4.4: Example of two classes that are visually similar from the benchmark dataset AWA1 [150]. (A) the sample leopard belongs to the seen classes, and (B) the sample bobcat belongs to the unseen classes. We speculate that samples from these two classes will lie close to each other in the latent space even though they come from different domains, challenging the view that samples from new unseen classes will lie far from samples of the seen classes in the latent space.

Table 4.1: The benchmarks for GZSL: CUB [146], SUN [152], AWA1 [150], and AWA2 [150]. Column (1) shows the number of seen classes, denoted by $|\mathcal{Y}^S|$, split into the number of training and validation classes (train+val), (2) presents the number of unseen classes $|\mathcal{Y}^U|$, (3) displays the number of samples available for training $|\mathcal{D}^{Tr}|$ and (4) shows number of testing samples that belong to the unseen classes $|\mathcal{D}_U^{Te}|$ and number of testing samples that belong to the seen classes $|\mathcal{D}_S^{Te}|$ from [39, 151]

Name	$\lvert\mathcal{Y}^S\rvert$ (train+val)	$\lvert\mathcal{Y}^U\rvert$	$\lvert\mathcal{D}^{Tr}\rvert$	$\lvert\mathcal{D}_U^{Te}\rvert + \lvert\mathcal{D}_S^{Te}\rvert$
CUB	150 (100+50)	50	7057	1764+2967
SUN	745 (580+65)	72	14340	2580+1440
AWA1 [1]	40 (27+13)	10	19832	4958+5685
AWA2	40 (27+13)	10	23527	5882+7913

representation of the data set CUB-200-2011 [150], we use the 1024-dimensional vector produced by CNN-RNN [114]. These semantic samples represent a written description of each image using 10 sentences per image. To define a unique semantic sample per-class, we average the semantic samples of all images belonging to each class [150]. We use manually annotated semantic samples containing 102 and 85 dimensions respectively, for the data sets SUN [150], AWA1 [150], and AWA2 [150]. To prevent a violation of the ZSL constraints, where the test classes should not be accessed during training, all the features were extracted according to training splits proposed in [150].

4.4.2 Evaluation Protocol

Xian et al. [150] formalised the current evaluation protocol for GZSL. We first compute the average per-class top-1 accuracy measured independently for each class, then we calculate the overall mean. We calculate the mean-class accuracy for each domain separately, i.e., the seen ($\mathcal{Y}^S$) and the unseen ($\mathcal{Y}^U$) classes. Then, we also compute the harmonic mean (H-mean) of the seen and unseen domains accuracy [150]. Furthermore, we show results by measuring the area under the seen and unseen curve (AUSUC) [19] by varying the domain expertise [19]. This domain expertise consists of a hyper-parameter to perform the trade-off between the performance in the seen and unseen classes [19].

4.4.3 Implementation Details

In this section, we describe the architecture and training procedures for learning the proposed latent space. As described in Sec. 4.3, we extend the following two models for our experimental setup: CADA-VAE [129] and cycle-WGAN [39]. The model CADA-VAE contains the following models that are parameterised as neural networks: $Encoder_x(.)$, $Encoder_a(.)$ in (4.1), $Decoder_x(.)$, and $Decoder_a(.)$ in (4.2). The training of CADA-VAE aims to produce a latent space that satisfies (4.4). In terms of the model architecture and hyper-parameters (*e.g. the number of epochs, batch size, the number of layers, learning rate, and, weight decay*), we followed the specifications provided by [129]. The encoder for visual representation is parameterised with 1560 hidden neurons, and the encoder for the semantic

representation is parameterised with 1450 hidden neurons. The decoders for the visual and semantic representation are parameterised with $1560, 660$ hidden neurons, respectively. For both modalities, the encoders project samples into the latent space, which is represented with 64-dimension vectors in the latent space. The model is optimised with Adam for 100 epochs [71]. We use an adaptive scheduling rate for the hyper-parameters γ, δ, by $(0.044, 0.0026)$, with respective epochs $(21 - 75, 0 - 90)$ [129]. We also extended cycle-WGAN [39], as explained in Sec. 4.3. The model cycle-WGAN contains the following functions that are parameterised as neural networks: $Encoder_a(.)$ in (4.1), $Decoder_a(.)$ in (4.2), and $Discriminator(.)$ in (4.3). We followed the hyper-parameters choice (e.g. *number of epochs, batch size, number of layers, learning rate, and weight decay, learning rate decay*) defined in [39]. The encoder is parameterised with a single hidden layer containing 4096 nodes with LeakyReLU activation [90], and the output layer, with 2048 nodes, has a ReLU activation [97]. The decoder is parameterised with a linear layer, and the discriminator is a network with a single hidden layer with 4096 nodes. The network has a LeakyReLU activation, and the output layer has no activation.

The domain classifier (DC)[2] is implemented as a neural network with binary output, representing the seen and unseen domains. The model is trained with Adam optimiser [71] to recognise the domains. The output probability of the domain classifier tends not to be well calibrated [8, 47]. Therefore, we calibrate the model output using the validation set [47, 150]. Then, the domain classification is performed as described in (4.7) [19].

4.4.4 Results

In this section, we present the results for our proposed approach. The first question aimed to be answered in this paper consists of whether the proposed latent space contains relevant information that enables our approach to learn the domain classifier for GZSL. Thus, we provide numerical evidence that our method outperforms both baselines (i.e., CADA-VAE and cycle-WGAN) and previous GZSL. In Table 4.3, we show the results in terms of unseen class accuracy $\mathcal{Y}^U$, seen class

[2]The code will be available at https://github.com/rfelixmg/gzsl-domain-classification.

accuracy $\mathcal{Y}^S$ and harmonic mean H, as described in Sec. 4.4.2. These results are given for the data sets CUB, SUN, AWA1 and AWA2. We compare our approach with 12 leading GZSL methods, which are divided into three groups: semantic (SJE [4], ALE [3], LATEM [149], ESZSL [118], SYNC [18], DEVISE [42]), latent space learning (SAE [35], f-CLSWGAN [151], cycle-WGAN [39] and CADA-VAE [129]) and domain classification (CMT [133] and DAZSL [8]). The semantic group contains methods that only use the seen class visual and semantic samples to learn a transformation function from the visual to the semantic space, and classification is based on nearest neighbour classification in that semantic space. The latent space learning group relies on visual samples from seen classes and semantic samples from seen and unseen classes during training, and are detailed in Sec. 4.3. The domain classification group relies on methods that weight the classification of seen and unseen classes. We discuss the numeral results in Table 4.3 in Section 4.5.

4.4.5 Ablation Studies

In Table 4.2 we report the area under the curve of seen and unseen accuracy (AUSUC) [19] for the benchmark data sets CUB, SUN, AWA1, and AWA2. We compare the results of the original CADA-VAE [129] and cycle-WGAN [39] with and without the DC. Similar to harmonic mean, the AUSUC is an evaluation metric that measures the trade-off between the seen and unseen domains.

Table 4.2: Area under the curve of seen and unseen accuracy (AUSUC). The highlighted values per column represent the best results in each data set. The notation * represents the results that we reproduced.

Classifier	**CUB**	**SUN**	**AWA1**	**AWA2**
EZSL	0.3020	0.1280	0.3980	–
DAZSL [8]	0.3570	**0.2390**	0.5320	–
f-CLSWGAN [151]	0.3550	0.2200	0.4610	–
cycle-WGAN [39]*	0.4180	0.2321	0.4730	–
CADA-VAE [129]*	0.3698	0.2362	0.5238	0.5216
cycle-WGAN + DC	**0.4262**	0.2321	0.4744	–
CADA + DC	0.3743	0.2364	**0.5247**	**0.5219**

Table 4.3: GZSL results using per-class average top-1 accuracy on the test sets of unseen classes $\mathcal{Y}^U$, seen classes $\mathcal{Y}^S$, and H-mean result H; and ZSL results on the unseen classes exclusively – all results shown in percentage. The results from previously proposed methods in the field were extracted from [150]. The highlighted values represent the best ones in each column. The methods below the double horizontal line represent the ones that use the semantic vectors from unseen classes during training. The notation * represents the results that we reproduced, and results represented with $-$ were not available in the literature, or hyper-parameters were not given.

	CUB			**SUN**			**AWA1**			**AWA2**		
Classifier	$\mathcal{Y}^S$	$\mathcal{Y}^U$	H	$\mathcal{Y}^S$	$\mathcal{Y}^U$	H	$\mathcal{Y}^S$	$\mathcal{Y}^U$	H	$\mathcal{Y}^S$	$\mathcal{Y}^U$	H
Semantic approach												
SJE [4]	59.2	23.5	33.6	30.5	14.7	19.8	74.6	11.3	19.6	73.9	8.0	14.4
ALE [3]	62.8	23.7	34.4	33.1	21.8	26.3	76.1	16.8	27.5	81.8	14.0	23.9
LATEM [149]	57.3	15.2	24.0	28.8	14.7	19.5	71.7	7.3	13.3	77.3	11.5	20.0
ESZSL [118]	63.8	12.6	21.0	27.9	11.0	15.8	75.6	6.6	12.1	77.8	5.9	11.0
SYNC [18]	70.9	11.5	19.8	43.3	7.9	13.4	87.3	8.9	16.2	90.5	10.0	18.0
DEVISE [42]	53.0	23.8	32.8	27.4	16.9	20.9	68.7	13.4	22.4	74.7	17.1	27.8
Generative approach												
SAE [35]	18.0	8.8	11.8	54.0	7.8	13.6	77.1	1.8	3.5	82.2	1.1	2.2
f-CLSWGAN [151]	57.7	43.7	49.7	36.6	42.6	39.4	61.4	57.9	59.6	68.9	52.1	59.4
cycle-WGAN [39]	60.3	46.0	52.2	33.1	48.3	39.2	63.5	56.4	59.7	–	–	–
CADA-VAE [129]	53.5	51.6	52.4	35.7	47.2	40.6	72.8	57.3	64.1	75.0	55.8	63.9
CADA-VAE [129]*	57.2	48.4	52.4	36.8	45.1	40.6	76.6	55.0	64.1	75.3	55.5	63.9
Domain Classification												
CMT [133]	49.8	7.2	12.6	21.8	8.1	11.8	87.6	0.9	1.8	90.0	0.5	1.0
DAZSL [8]	56.9	47.6	51.8	37.2	45.6	**41.4**	76.9	54.7	63.9	–	–	–
cycle-WGAN + DC (ours)	61.9	45.9	**52.7**	39.3	41.3	40.3	68.6	53.4	60.0	–	–	–
CADA-VAE + DC (ours)	52.4	**52.9**	52.6	34.0	**50.7**	40.7	72.6	**57.5**	**64.2**	74.9	**57.0**	**64.3**

4.5 Discussions

In this section, we discuss the main contributions presented by our approach. We performed our experiments by combining previous GZSL approaches (such as CADA-VAE [129]) and cycle-WGAN [39]) with our Domain Classification in order to enhance the balancing of the seen and unseen domains for GZSL.

Firstly, in Table 4.3 we provide quantitative information that shows that our method outperforms existing methods in terms of unseen accuracy, $\mathcal{Y}^u$. This demonstrates that by learning to classify the domain for each sample, our method improves the classification of the unseen classes. Specifically, for CUB, SUN, AWA1 and AWA2 data sets, the baseline unseen classification results of 48.4%, 45.1%, 55.0%, and 55.5% have become 52.9%, 50.7%, 57.5%, and 57.0%. This improvement was achieved given a minor trade-off with the seen classes.

Secondly, despite the trade-off mentioned above, our approach is still able to achieve minor improvements in terms of **H-mean**. Table 4.3 shows an improvement of 0.2%, 0.1%, 0.1% and 0.4%, when compared to the baseline CADA-VAE. Although these results can be considered minor, we argue that our model does not directly optimise the H-mean. Thus, this improvement indicates that our approach has a more balanced performance than previous models.

We note similar behaviour for the cycle-WGAN model [39], where the proposed method achieves improvement for **H-mean** from 52.2% to 52.7% for CUB, from 39.2% to 40.3% for SUN, and from 59.7% to 60.0% for AWA1. However, such improvement is achieved due to the positive trade-off towards the seen domain. We argue that this difference, when compared to CADA-VAE, is due to the inherent differences in the latent space learning of each of the approaches. In fact, the approach CADA-VAE is directly optimised by a variational autoencoder, where the control on the latent space is guided by a divergence measure for the visual and semantic representation jointly. On the other hand, the cycle-WGAN model is directly optimised by an adversarial loss from a generative adversarial network conditioned mainly on the semantic representation.

In terms of AUSUC, the proposed approach achieves improvements for both cycle-WGAN [39] and CADA-VAE [129]. For CADA-VAE, the domain classification yielded improvements from 0.3698,0.2362, 0.5238,0.5216 to 0.3743, 0.2364,

0.5247, 0.5219, for CUB, SUN, AWA1 and AWA2, respectively. Likewise, for cycle-WGAN [39], the DC provided improvements from 0.4180, 0.4730 to 0.4268, 0.4744 for CUB and AWA1, respectively.

4.6 Conclusion and Future Work

In this paper, we introduce a principled method to classify the seen and unseen domains in GZSL. In particular, we presented our domain classifier that learns directly from the latent space of visual and semantic information. We have demonstrated that our proposed approach can be combined with previous latent space learning models, such as CADA-VAE and cycle-WGAN. Our approach yielded improvements for each one of those models by automatically balancing the seen and unseen domains in benchmark experiments on four available data sets: CUB, SUN, AWA1, and AWA2.

Our experimental results show that our proposed approach has achieved state-of-the-art H-mean results for CUB, AWA1 and AWA2, and unseen accuracy for CUB, SUN, AWA1, and AWA2. In particular, our results are substantially better than the state of the art on CUB and SUN, which contain a large number of classes. On AWA1, AWA2, which are smaller data sets, our results are marginally better. Furthermore, our model produces substantial improvements in terms of AUSUC results for CUB, AWA1 and marginally better on AWA2.

As stated previously, our domain classification learns to discriminate between samples from the seen and unseen domains. We observe that the improvement of CADA-VAE and cycle-WGAN are different. The CADA-VAE model tends to improve in terms of the unseen domain when the DC is applied. Whereas cycle-WGAN tends to improve in terms of the seen domain. On one hand, we note that the training strategy for both models follows different guidelines, VAE and GAN. On the other hand, our model does not impose direct constraints in order to optimise GZSL metrics, such as accuracy or H-mean. In fact, we believe that these aspects are the main factors for the contrasting outcomes for CADA-VAE and cycle-WGAN models. With that in mind, we believe that the differences between these two data augmentation approaches should be studied in future generalised zero-shot learning research.

In the future, we intend to further study the reasons behind the performance difference observed between the data sets. Moreover, we also plan to develop a more extensive framework that can incorporate domain classification for approaches that do not rely on latent space learning.

Chapter 5

Augmentation Network for Generalised Zero Shot Learning with Multi-Modal Inference

The work contained in this chapter is in submission as the following paper:

Felix, R., Sasdelli, M., Reid, I. and Carneiro, G., Augmentation Network for Multi-modal and Multi-domain Generalised Zero Shot Learning. In *Submission*, 2019 [40].

Statement of Authorship

Title of Paper	Augmentation Network for Generalized Zero Shot Learning with Multi-Modal Inference
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished or Unsubmitted work
Publication Details	Felix, R., Sasdelli, M., Reid, I., and Carneiro, G., Augmentation Network for Generalized Zero Shot Learning with Multi-Modal Inference. In submission (2019)

Principal Author

Name of Principal Author (Candidate)	Rafael Felix Alves
Contribution to the Paper	- Development of the main idea of the paper; - Implementing and conducting the experiments; - Writing and coordinating the revisions;
Overall percentage (%)	60
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date: 29/11/2019

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Michele Sasdelli
Contribution to the Paper	- Help with the development of the idea; - Help writing, revision and discussions;

Signature		Date	29/11/2019

Name of Co-Author	Ian Reid
Contribution to the Paper	- Help writing, revision and discussions;
Signature	Date 22/11/2019

Name of Co-Author	Gustavo Carneiro
Contribution to the Paper	- Help with the development of the idea; - Help writing, revision and discussions;
Signature	Date 04-11-2019

Abstract

Generalised zero-shot learning (GZSL) is defined by a training process containing a set of visual samples from seen classes and a set of semantic samples from seen and unseen classes, while the testing process consists of the classification of visual samples from the seen and the unseen classes. Current approaches are based on inference processes that rely on the result of a single classifier running on only one modality (visual, semantic, or latent joint space) that balances the classification between the seen and unseen classes using gating mechanisms. There are a couple of problems with such approaches: 1) multi-modal classifiers are known to generally be more accurate than single modality classifiers, and 2) the gating mechanisms rely on a complex one-class training of an external domain classifier that modulates the seen and unseen classifiers. In this paper, we mitigate these issues by proposing a new GZSL method – augmentation network that tackles multi-modal and multi-domain inference for generalised zero-shot learning (AN-GZSL). Our approach consists of a multi-modal inference that combines visual and semantic classification and automatically balances the seen and unseen classification using temperature calibration, without requiring any gating mechanisms or external domain classifiers. Experiments show that our method produces the new state-of-the-art GZSL results for fine-grained benchmark data sets CUB and FLO and for the large-scale data set ImageNet. We also obtain competitive results for coarse-grained data sets SUN and AWA1. We show an ablation study that justifies each stage of the proposed AN-GZSL.

5.1 Introduction

As computer vision systems start to be deployed in unstructured environments, they must have the ability to recognise not only the visual classes used during the training process (i.e., the seen classes) but also classes that are not available during training (i.e., unseen classes). The importance of such ability lies in the impracticality of collecting visual samples from all possible classes that will be shown to the system. In this context, approaches categorised as Generalised Zero-Shot Learning (GZSL) [19, 39, 150] play an important role due to their ability to classify visual

samples from seen and unseen classes. In general, the training of GZSL methods involves the use of visual samples from seen classes and semantic samples (e.g., textual definition) from seen and unseen classes. The rationale behind the use of semantic samples is that they are readily available from various sources, such as Wikipedia, English dictionary [92], or manually annotated attributes [76]. Such training setup can potentially mitigate the issue of collecting visual samples from all possible unseen classes, and the success of GZSL lies in the effective transferring of knowledge between the semantic and visual modalities.

In recent years, we note three different approaches for solving GZSL. One type of GZSL approach has focused on training a mapping function that transforms samples from the visual to the semantic space [77], and inference is then based on a classification process that works exclusively in the semantic space. Another type of GZSL method is based on training a conditional generative model that generates visual samples from their respective semantic samples. The generated visual samples of the unseen classes and the true visual samples from the seen classes are then used for training a visual classifier [17,39,60,82,105,124,129,144,151] – inference is based on a visual classification process. Another type of GZSL method relies on an external domain classifier (trained with the visual samples from the seen classes via a one-class learning problem) that modulates the classification between the seen and unseen classes [8,14,37,133,158], where the classification in each domain typically uses a single modality.

There are a couple of issues with the GZSL methods above: 1) even though the training process involves some sort of interaction between the visual and semantic modalities, the inference usually does not rely on a truly multi-modal classification (i.e., where both modalities are jointly used in the process) [8,17,39, 76,77,133,144,151], which can be considered a weakness given the strong evidence that multi-modal inference can improve classification accuracy [8,158,163]; and 2) the one-class training of external domain classifiers that modulate the seen and unseen classification [8,14,37,133,158] is not a trivial process given the similarity between the seen and unseen class domains. In fact, it can be argued that samples from these domains are drawn from the same distribution, making it challenging to distinguish between them.

In this paper, we introduce the Augmentation Network for multi-modal and

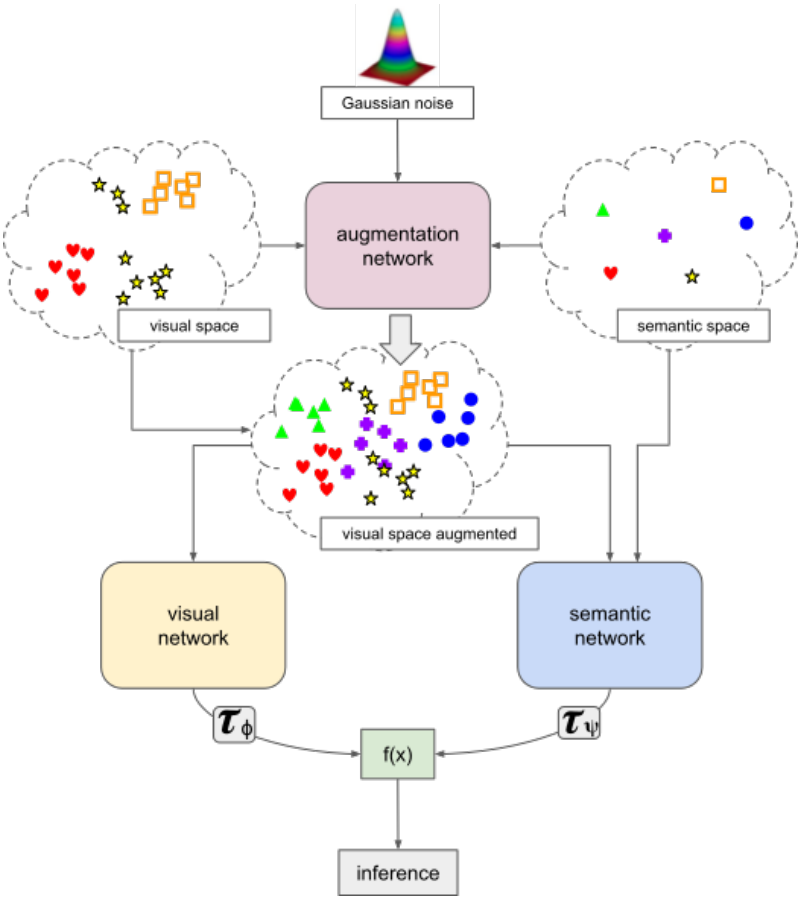


Figure 5.1: Depiction of our proposed model Augmentation Network for multi-modal and multi-domain Generalised Zero-Shot Learning (AN-GZSL). AN-GZSL is composed of the augmentation network (that generates visual samples for training the visual and the semantic networks), the visual and semantic networks, a classification calibration (represented by τ_ψ and τ_ϕ in (5.2)) that enables multi-domain classification, and the multi-modal classification that combines the visual and semantic modules.

multi-domain Generalised Zero-Shot Learning (AN-GZSL) designed to address the two problems listed above – see Fig. 5.1. The approach consists of an augmentation network, a visual network, a semantic network, a classification calibration and a multi-modal classifier. The augmentation network is a generative model that produces visual samples conditioned on the semantic data, where these generated visual samples are used by the visual network to learn a visual classifier and by the semantic network to model a semantic classifier. The visual and semantic classifiers are then calibrated to enable an effective modulation-free multi-modal classification. Then, the two calibrated classifiers are combined in a multi-modal classification. We show that the proposed approach produces state-of-the-art GZSL results on the fine-grained benchmark data sets CUB [146,150] and FLO [98] and on the large-scale data set ImageNet [26,145]. We also achieve competitive results for the coarse-grained data sets SUN [150] and AWA1 [76]. The experiments also show an ablation study that tests the importance of each component of the proposed model.

5.2 Literature Review

In this section we describe relevant literature that contextualises and motivates the proposed approach.

Generalised Zero-Shot Learning (GZSL). In recent years, we have observed a growing interest in GZSL. A catalyst for such interest was the paper by Xian et al. [150] that formalises the GZSL problem. Their work introduces a solid experimental setup and a robust evaluation metric based on the harmonic mean between the classification accuracy results of the seen and the unseen visual classes.

Recently proposed GZSL methods can be roughly divided into three categories: **semantic attribute prediction**, **visual data augmentation**, and **domain balancing**. **Semantic attribute prediction** methods [3,57,76] tackle GZSL by training a regressor that maps visual samples from seen classes to their respective semantic samples. Hence, given a test visual sample (from a seen or unseen class), the regressor maps it into the semantic space, which is then used in a nearest neighbour semantic classification process. The main assumption of this approach is that

the mapping from visual to semantic spaces learned from the seen class domain can be transferred to the unseen class domain. Unfortunately, such assumption is unwarranted, and a typical issue of this approach is that test visual samples from seen classes are classified correctly and samples from unseen classes are often incorrectly classified into one of the seen classes – this is referred to as a bias toward the seen classes [150]. Recent research exploring matching functions between visual and semantic samples can address the issue mentioned above, but they still show biased classification toward the seen classes [3].

Visual data augmentation relies on a generative model trained to produce visual samples from the corresponding semantic samples [17, 39, 60, 82, 105, 124, 129, 144, 151]. Such model allows the generation of visual samples for the unseen classes, which are then used in the modelling of a visual classifier that is trained with real visual samples from seen classes and generated visual samples from unseen classes. Methods based on this approach are effective because they solve, to a certain extent, the bias toward the seen classes. Recently, the training process of this approach has been extended, forcing generated visual samples to regress to the corresponding semantic samples, in a multi-modal cycle consistent training [39, 144]. This extension represents the first attempt at a multi-modal training, which allowed further improvements in GZSL results. However, none of the methods above relies on a multi-modal inference process. It is interesting to note that the inference process of **semantic attribute prediction** focuses exclusively on the semantic space, while **visual data augmentation** works solely on the visual space. A multi-modal inference process that effectively merges the two spaces has yet to be proposed.

Domain balancing methods solve the bias toward the seen classes issue with a gating mechanism that modulates the classification of seen and unseen classes [8, 14, 37, 133, 158]. In particular, these methods consist of a (generally visual) classifier trained for the seen classes, a (usually semantic) classifier trained for the unseen classes, and a domain classifier for the modulation process [8, 37, 158]. Even though domain balancing approaches hold outstanding results [8, 158], they have the following challenges: 1) the training of multiple domain-specific classifiers, and 2) the non-trivial training of a gating mechanism that needs to classify between seen and unseen classes using a one-class classification process, which is a hard

task considering that these classes arguably come from the same data distribution. In this paper, we also rely on visual data augmentation and domain balancing, but differently from the approaches above, our multi-modal classification relies on visual and semantic classifiers trained on all seen and unseen classes (i.e., they are not domain-specific). Furthermore, the balancing between seen and unseen domains is achieved with a classification calibration approach that does not need any gating mechanism.

5.3 Method

In the next sub-sections, we first formulate the GZSL problem. Then, we introduce our proposed augmentation network for multi-modal and multi-domain generalised zero-shot learning (AN-GZSL), with the explanation of the inference, architecture and training processes.

5.3.1 Problem Formulation

To formulate the GZSL problem [19,150], we first define the visual data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^K$ denotes the visual representation (acquired from the second to last layer of a pre-trained deep residual nets [58]), and $y \in \mathcal{Y} = \{1, ..., C\}$ denotes the visual class, which can also be described with a one-hot vector $\mathbf{h} \in \{0,1\}^C$, where the y-th position in $\mathbf{h}$ is assigned to 1, and all the others 0. The visual data set has N samples, denoting the number of images. We also need to define the semantic data set $\mathcal{R} = \{\mathbf{a}_y\}_{y \in \mathcal{Y}}$, which associates visual classes with semantic samples, where $\mathbf{a}_y \in \mathcal{A} \subseteq \mathbb{R}^L$ represents a semantic feature (e.g., *word2vec* features [150]). The semantic data set has as many elements as the number of classes. The set $\mathcal{Y}$ is split into the seen subset $\mathcal{Y}^S = \{1, ..., S\}$, and the unseen subset $\mathcal{Y}^U = \{(S+1), ..., (S+U)\}$. Therefore, $C = S + U$, with $\mathcal{Y} = \mathcal{Y}^S \cup \mathcal{Y}^U$, $\mathcal{Y}^S \cap \mathcal{Y}^U = \emptyset$. Furthermore, $\mathcal{D}$ is also divided into mutually exclusive training and testing visual subsets $\mathcal{D}^{Tr}$ and $\mathcal{D}^{Te}$, respectively, where $\mathcal{D}^{Tr}$ contains a subset of the visual samples belonging to the seen classes, and $\mathcal{D}^{Te}$ has the visual samples from the seen classes held out from training and all samples from the unseen classes. The training data set comprises the semantic data set $\mathcal{R}$ and the training

visual subset $\mathcal{D}^{Tr}$, while the testing data set consists of the testing visual subset $\mathcal{D}^{Te}$ and the same semantic data set $\mathcal{R}$.

5.3.2 AN-GZSL Calibrated Inference

The inference procedure consists of estimating the class label of a test visual sample $\mathbf{x}$ that optimises

$$f(y|\mathbf{x},\mathcal{R}) = \sigma_{\phi}(\phi(y|\mathbf{x}),\tau_{\phi}) + \sigma_{\psi}(\psi(y|\mathbf{x},\mathcal{R}),\tau_{\psi}), \tag{5.1}$$

where $f(.)$ denotes the classification function, $\phi(.)$ and $\psi(.)$ represent the visual network (defined in Sec. 5.3.5) and the semantic network (Sec. 5.3.4) that return a logit, and $\sigma_{\phi}(.)$ and $\sigma_{\psi}(.)$ represent the softmax activation function with temperature calibration [56], defined by

$$\sigma(l_y,\tau) = \frac{\mathrm{e}^{(l_y/\tau)}}{\sum_{c=1}^{C}\mathrm{e}^{(l_c/\tau)}}, \tag{5.2}$$

where the logit $l_y \in \mathbb{R}$ represents the y^{th} output of a network (i.e., the visual or the semantic), and the temperature scaling τ represents a calibrating factor. The multi-modal inference in (5.1) consists of a sum of the results from the visual and semantic classifiers, where the final classification is achieved by

$$y^{*} = \underset{y\in\mathcal{Y}}{\operatorname{argmax}} f(y|\mathbf{x},\mathcal{R}). \tag{5.3}$$

The GZSL inference in (5.3) balances the seen and unseen classification with a confidence calibrated by the temperature scaling, which can be considered to be a much simpler strategy compared to previous gating mechanisms [8, 133, 158] that had to deal with complicated one-class domain classification problems. Furthermore, (5.3) shows a simple multi-modal inference without any hyper-parameter to combine the contributions of each classifier.

5.3.3 Augmentation Network

The augmentation network relies on a generative model [151] trained to produce visual samples conditioned on their semantic samples. After training this generative model, it is then possible to generate visual samples from the unseen classes

to train a classifier using real visual samples from the seen classes and generated visual samples from the unseen classes [17,39,60,82,105,124,129,144,151]. This approach has been recently extended with a cycle consistency loss that regularises the training process [39]. The augmentation network is optimised with a Wasserstein generative adversarial network (WGAN) [7] loss and cycle-consistent loss [39], defined by

$$\ell_{AN} = \ell_{WGAN} + \ell_{CYC}, \tag{5.4}$$

where ℓ_{WGAN} represents the WGAN loss [7] that optimises a conditional generator network $g(.)$ and discriminator network $d(.)$. The loss ℓ_{WGAN} is defined by

$$\begin{aligned} \ell_{WGAN} = \mathbb{E}_{(\mathbf{x},\mathbf{a})\sim\mathbb{P}_s^{x,a}}[d(\mathbf{x},\mathbf{a};\theta_d)] - \mathbb{E}_{(\tilde{\mathbf{x}},\mathbf{a})\sim\mathbb{P}_g^{x,a}}[d(\tilde{\mathbf{x}},\mathbf{a};\theta_d)] \\ - \kappa\mathbb{E}_{(\tilde{\mathbf{x}},\mathbf{a})\sim\mathbb{P}_\alpha^{x,a}}[(||\nabla_{\tilde{\mathbf{x}}} d(\tilde{\mathbf{x}},\mathbf{a};\theta_D)||_2 - 1)^2], \end{aligned} \tag{5.5}$$

where $\mathbb{E}[.]$ represents the expected value operator. The joint distribution of visual and semantic samples from the seen classes is given by $\mathbb{P}_s^{x,a}$, and $\mathbb{P}_g^{x,a}$ represents the joint distribution of semantic and visual samples produced by the augmented network with the generator network, as follows: $\tilde{\mathbf{x}} \sim g(\mathbf{a}, \mathcal{N}(0,\mathbf{I}); \theta_g)$. The coefficient κ in (5.5) weights the contribution of the third term of the loss, and the joint distribution of the semantic and visual samples produced by $\hat{\mathbf{x}} \sim \alpha\mathbf{x} + (1-\alpha)\tilde{\mathbf{x}}$ with $\alpha \sim \mathcal{U}(0,1)$ (i.e., uniform distribution) is given by $\mathbb{P}_\alpha^{x,a}$. In this network, the generator receives the semantic samples and a noise vector to generate visual samples. Then, the discriminator network aims to differentiate the generated from the real visual samples [39]. Then, the loss ℓ_{CYC} provides a cycle-consistent training regularisation which guarantees that the generated visual samples can reconstruct their respective semantic samples. The loss ℓ_{CYC} is defined by

$$\begin{aligned} \ell_{CYC} = \mathbb{E}_{\mathbf{a}\sim\mathbb{P}_s^a, \mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\mathbf{a} - r(g(\mathbf{a},\mathbf{z};\theta_g);\theta_r)\|_2^2\right] \\ + \mathbb{E}_{\mathbf{a}\sim\mathbb{P}_u^a, \mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\mathbf{a} - r(g(\mathbf{a},\mathbf{z};\theta_g);\theta_r)\|_2^2\right], \end{aligned} \tag{5.6}$$

where the function $r(.)$ represents a regressor network parameterised by θ_r that estimates the original semantic samples from the visual samples generated by $g(.)$, the latent variable $\mathbf{z}$ represents a Gaussian noise, and the distributions of the semantic samples of both seen and unseen domains are represented by $\mathbb{P}_s^a$

and $\mathbb{P}_u^a$. In contrast to previous approaches [17, 39, 60, 82, 105, 124, 129, 144, 151], our proposed augmentation network feeds the visual *and* semantic networks with generated visual samples from both the seen and unseen domains – this allows the visual and semantic classifiers to jointly learn an effective discriminating space for *all seen and unseen classes*.

5.3.4 Semantic Network

The semantic network extends the ranking loss proposed by Akata et al. [3]. We define the semantic network as $\psi(y|\mathbf{x}, \mathcal{R}) = \mathbf{x}^T \theta_\psi \mathbf{a}_y$, represented by a bi-linear model parameterised by $\theta_\psi \in \mathbb{R}^{K \times L}$, with $\mathbf{a}_y \in \mathcal{R}$ and $\mathbf{x}$ being either a real sample from a seen class or a generated sample from an unseen class – note that *the use of such data augmentation represents the main difference between our proposed semantic classifier and the one in [3].* This semantic network learns the joint relationship between the visual sample $\mathbf{x}$ and semantic sample $\mathbf{a}_y$ for a particular class y. For training this network, we extend the loss from [3], defined by

$$\ell_{SN} = \sum_{i=1}^{M} \sum_{c=1}^{C} \lambda(\mathbf{x}_i, \theta_\psi, \mathbf{a}_{y_i}, \mathbf{h}_i) \Big[\beta(\tilde{\mathbf{x}}_i, \theta_\psi, \mathbf{a}_c, \mathbf{a}_{y_i}, h_{i,c}) \Big]_+, \tag{5.7}$$

where $[.]_+$ represents the hinge loss, $\mathbf{a}_{y_i}$ denotes the semantic vector associated with the class y_i of the i^{th} training sample, $h_{i,c}$ represents the c^{th} position of the one-hot vector for the i^{th} training sample $\mathbf{h}_i$, and M is the size of the training set, which includes the generated visual features. In (5.7), the term $\beta(.)$ consists of a compatibility bi-linear loss, defined by

$$\beta(\mathbf{x}_i, \theta_\psi, \mathbf{a}_c, \mathbf{a}_{y_i}, h_{i,c}) = h_{i,c} + \mathbf{x}_i^T \theta_\psi \mathbf{a}_{y_i} - \mathbf{x}_i^T \theta_\psi \mathbf{a}_c, \tag{5.8}$$

and $\lambda(.)$ represents a ranking regularization, defined by

$$\lambda(\mathbf{x}_i, \theta_\psi, \mathbf{a}_{y_i}, \mathbf{h}_i) = \frac{1}{\sum_{c=1}^{C} \mathbb{1}\Big(\beta(\mathbf{x}_i, \theta_\psi, \mathbf{a}_c, \mathbf{a}_{y_i}, h_{i,c})\Big)}, \tag{5.9}$$

where $\mathbb{1}(.)$ represents a Heaviside step function, with the divisor computing the rank of the transformation according to the semantic data set.

The optimisation of (5.7) forces $\psi(y|\mathbf{x}, \mathcal{R})$ to be higher when $\mathbf{x}$ and $\mathbf{a}_y$ match correctly. This result is then calibrated by (5.2) to enable an effective multi-domain classification.

5.3.5 Visual Network

The visual network is a fully connected neural network represented by $\phi(y|\mathbf{x})$, parameterised by θ_ϕ, where $\mathbf{x}$ can be a real sample from a seen class or a generated sample from an unseen class. This visual network is trained with the usual cross-entropy loss defined by ℓ_{VN}. Similarly to the semantic classifier, this visual classifier is also calibrated with (5.2) for the multi-domain classification.

5.3.6 AN-GZSL Training

The loss function for our proposed AN-GZSL model is defined by

$$\ell_{AN-GZSL} = \ell_{AN} + \ell_{VN} + \ell_{SN}, \tag{5.10}$$

which is minimised to estimate the parameters $\theta_g, \theta_d, \theta_r, \theta_\phi, \theta_\psi$. For training, we use the visual samples produced by the augmentation network as input to the proposed visual and semantic networks. This approach not only augments the number of samples from the seen classes, but it also generates samples from the unseen classes. In practice, we perform an alternating training where we first optimise θ_g, θ_d and θ_r, then we optimise θ_ψ and θ_ϕ. Empirically, we have observed that the augmentation network tend to generate random samples at early stages of training [55]. Hence, the alternating strategy provides stronger gradients signal for the optimisation of θ_ψ and θ_ϕ, at late stages. After the optimisation of (5.10) stabilises, the temperature for each network (τ_ϕ and τ_ψ in Eq. 5.1) is learned using a validation set held out from training [56].

5.4 Experiments

In this section, we describe the benchmark data sets, evaluation criteria and the setup adopted for the experiments. Then, we present a set of ablation studies and the results of the proposed method, which are compared with the state of the art (SOTA).

Table 5.1: Information about CUB [146], FLO [98], SUN [152], AWA1 [150], and ImageNet [26,145]. Column (1) shows the number of seen classes, denoted by $|\mathcal{Y}^S|$, split into the number of training and validation classes (train+val), (2) presents the number of unseen classes $|\mathcal{Y}^U|$, (3) displays the number of samples available for training $|\mathcal{D}^{Tr}|$ and (4) shows number of testing samples that belong to the unseen classes $|\mathcal{D}_U^{Te}|$ and number of testing samples that belong to the seen classes $|\mathcal{D}_S^{Te}|$ from [39].

Name	$\lvert\mathcal{Y}^S\rvert$ (train+val)	$\lvert\mathcal{Y}^U\rvert$	$\lvert\mathcal{D}^{Tr}\rvert$	$\lvert\mathcal{D}_U^{Te}\rvert + \lvert\mathcal{D}_S^{Te}\rvert$
CUB	150 (100+50)	50	7057	1764+2967
FLO	82 (62+20)	20	1640	1155+5394
SUN	745 (580+65)	72	14340	2580+1440
AWA1 [1]	40 (27+13)	10	19832	4958+5685
ImageNet	1000 (1000 + 0)	100	1.2kk	5200+50k

5.4.1 Data Sets

We assess the proposed method on publicly available benchmark GZSL data sets. More specifically, we perform experiments on CUB-200-2011 [146,150], FLO [98], SUN [150], and AWA1 [76,150] with the GZSL experimental setup described by Xian et al. [150]. We also perform GZSL experiments on ImageNet [26,145]. The data sets CUB and FLO are generally regarded as fine-grained, while AWA1 and SUN are coarse-grained, and ImageNet is large-scale. In Table 5.1 we show some details about the data sets regarding the number of seen and unseen classes and the number of training and testing images.

For the semantic features, we use the 1024-dimensional vector produced by CNN-RNN [114] for CUB-200-2011 [150] and FLO [98]. These semantic features are extracted from a set of textual description of 10 sentences per image. To define a unique semantic sample per-class, the semantic features of all images belonging to each class were averaged [150]. For the SUN and AWA1 data sets, we use manually annotated semantic features (attributes) containing 102 and 85 dimensions, respectively [150]. For the visual samples, we follow the protocol by Xian et al. [150], where the features are represented by the activation of the 2048-dimensional top pooling layer of ResNet-101 [58], obtained for the image.

All the semantic and visual features were obtained from [150]. To guarantee reproducible and consistent results, we follow the data set split proposed by Xian et al. [150], which guarantees that there is no overlap between the unseen classes and the ImageNet training classes. Moreover, we followed the same testing setup from [150].

For the ImageNet experiment [26], there can be several testing splits for GZSL (e.g., 2-hop, 3-hop), which rely on the training set of 1K classes and testing set on 22K classes. However, recent studies reported that such splits show overlap between seen and unseen classes for GZSL [39]. We argue that although these splits may be suitable for open-set recognition approaches, further studies are required to ensure their applicability for GZSL [39]. Nevertheless, to demonstrate the robustness of the proposed approach to large data sets, we experiment with ImageNet [26] for a split containing 100 classes for testing [145] and the standard 1K classes for training [145], without any overlap between seen and unseen classes. For ImageNet, we used 500-dimensional semantic samples [145] and 2048-dimensional ResNet-features, where images are resized to 256×256 pixels, cropped to 224×224 pixels, normalised with means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$ per RGB channel.

5.4.2 Evaluation Protocol

The evaluation protocol is based on computing the average per-class top-1 accuracy measured independently for each class before dividing their cumulative sum by the number of classes [150]. For GZSL, after computing the average per-class top-1 accuracy on seen classes $\mathcal{Y}^S$ and unseen classes $\mathcal{Y}^U$, we compute the harmonic mean of the seen and the unseen classification accuracy [150]. We also show results using the receiver operating characteristics (ROC) curve that measures the seen and the unseen classification accuracy over many operating points of the classifier [19]. Using such curve, we can measure the area under the seen unseen accuracy curve (AUSUC) [19] that represents an unbiased performance of the GZSL method.

5.4.3 Implementation Details

In this section, we describe the implementation details for the augmentation network, visual network and semantic networks that compose the model AN-GZSL, in terms of the model architecture and hyper-parameters (e.g. *number of epochs, batch size, number of layers, learning rate, weight decay, and learning rate decay*)[2]. Firstly, the augmentation network (composed of a generator, a discriminator, and a regressor) is defined in terms of a generative adversarial network (GAN) with cycle-consistency loss [39]. The generator consists of a single hidden layer with 4096 nodes and LeakyReLU activation [90] with an output layer of 2048 nodes (same dimension as ResNet [58] feature layer). The discriminator consists of a single hidden layer with 4096 nodes, which has a LeakyReLU activation function, and the output layer has no activation. Secondly, the visual network consists of a model parameterised with one fully connected layer from the 2048-dimensional visual space into the label space $\mathcal{Y}$. Thirdly, the semantic network is defined as a bi-linear model [3] that matches the 2048-dimensional visual space with the semantic space. Moreover, we introduce a dropout layer for the visual and the semantic networks for regularisation during training with dropout rate equal to 0.2. We performed cross-validated experiments that showed the number of generated features that turns the model competitive and stable [151]. The augmented network generates 300 visual samples per class for training the visual and the semantic networks for all the benchmark data sets [151]. The temperature calibration in (5.2) is achieved by optimising the parameters τ_ψ and τ_ϕ with a grid search minimization of the losses for the visual and semantic networks. Finally, we perform a Bayesian inference using Monte-Carlo dropout [46] because recent results suggest that such Bayesian inference can improve classification calibration and accuracy [47]. All hyper-parameters of the proposed AN-GZSL model were estimated with standard model selection methods using the validation sets proposed by Xian et al. [150].

Table 5.2: GZSL results using per-class average top-1 accuracy on the test sets of unseen classes $\mathcal{Y}^U$, seen classes $\mathcal{Y}^S$, and H-mean result H – all results shown in percentage. The highlighted values represent the best ones for each column.

	CUB			**FLO**			**SUN**			**AWA**		
Classifier	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H
$AN-GZSL^{\phi}$	46.2	**61.5**	52.8	60.0	70.8	65.0	48.7	33.1	39.4	55.4	64.8	59.7
$AN-GZSL^{\psi}$	**77.7**	41.8	54.4	**84.9**	36.6	51.2	47.2	21.6	29.6	46.4	**67.3**	54.9
$AN-GZSL^{\tau=1}$	46.2	**61.5**	52.8	60.1	**70.9**	65.1	**53.3**	32.8	40.6	55.6	65.0	60.0
$AN-GZSL$	60.5	56.6	**58.5**	80.7	69.3	**74.5**	41.7	**37.1**	**41.7**	**58.2**	66.1	**61.9**

5.4.4 Ablation Study

In Table 5.2, we report the ablation study for the proposed method AN-GZSL. First, we report the results for inference computed by the visual network ($AN-GZSL^{\phi}$). Second, $AN-GZSL^{\psi}$ reports the results for our semantic network. Then, $AN-GZSL^{\tau=1}$ shows the combination of the visual and semantic network without the temperature calibration. The last row shows the results $AN-GZSL$ with the calibrated (i.e., multi-domain) multi-modal networks.

5.4.5 Results

In Table 5.3, we compare the GZSL results on CUB, FLO, SUN and AWA1, produced by the proposed model AN-GZSL and several other methods previously proposed in the field. These methods are split into three groups: semantic approach, generative approach and domain balancing. The semantic approach models are DAP [76], IAP [77], DEVISE [42], SJE [4], LATEM [149], ESZSL [118], ALE [3], PQZSL [83], AREN [153], and MLSE [30]. The generative approaches are SAE [35], f-CLSWGAN [150], cycle-WGAN [39], CADA-VAE [129], GDAN [60], GMN [124], Zhu et al. [165], and LisGAN [82]. The domain balancing models are: CMT [133] and DAZSL [8]. Moreover, we report the following metrics in Table 5.3: the accuracy for the unseen domain ($\mathcal{Y}^U$), the seen domain ($\mathcal{Y}^S$) and the harmonic-mean (H) between these two accuracy measures.

Table 5.4 shows the top-1 accuracy on ImageNet for the proposed AN-GZSL and the results reported by previous methods on the same ImageNet experimental

[2]Please see more details in the code, https://github.com/rfelixmg/an-gzsl

setup.

In Fig. 5.2, we show the ROC results of the proposed method AN-GZSL, and the cycle-WGAN [39], which has code available online and represents one of the SOTA methods for that measure, to the best of our knowledge. Furthermore, Figure 5.2 shows the seen and unseen classification graphs for previously published GZSL methods (please refer to Tab. 5.3 for a reference to each method). We represent previous methods [151] by single (diamond-shaped) points denoting the results for the seen and unseen classification accuracy. We can only use single points for these methods because they are the results available from the literature (i.e., previous methods only report a single operating point for the classification of seen and unseen classes).

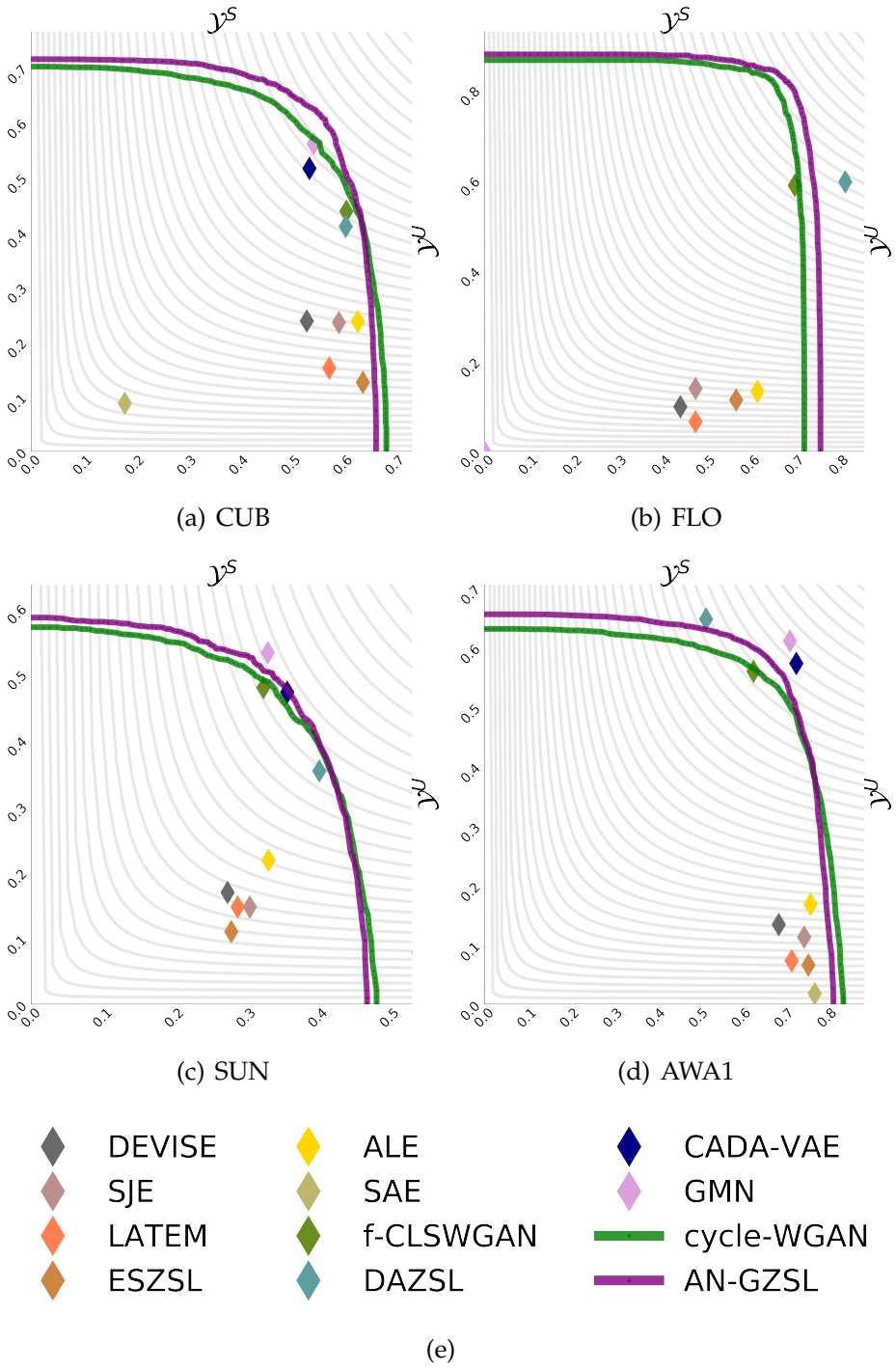
Using the graph in Fig. 5.2, we compute the AUSUC on each data set for AN-GZSL – results are shown in Table 5.5. We also added the results reported by the previous methods EZSL [118], fCLSWGAN [151], cycle-WGAN [39] and DAZSL [8] in Table 5.5.

5.5 Discussions

Ablation study. Table 5.2 shows the importance of each component of AN-GZSL, where the H-mean tends to be higher for the multi-modal approach, compared to each individual modality, and the multi-domain multi-modal method at the last row shows the best performance in all data sets. The high similarity between the results of $AN-GZSL^{\tau=1}$ and $AN-GZSL^{\phi}$ suggests that the un-calibrated multi-modal classifiers rely entirely on the visual classifiers. This is explained by the fact that the classification results produced by the un-calibrated semantic classifier shows classification probabilities close to a uniform distribution, in contrast to the un-calibrated visual classifier that shows more non-uniform distributions. However, when calibration is applied, the classification probabilities produced by both classifiers are pushed further away from the uniform distribution, which means that the sum of calibrated classifiers can produce results that are different from the original visual and semantic classifiers. In fact, Table 5.2 shows that the multi-modal calibrated classification accuracy is always higher than single-modality classification results. This multi-modal calibrated classifier also produces

Table 5.3: GZSL results using per-class average top-1 accuracy on the test sets of unseen classes $\mathcal{Y}^u$, seen classes $\mathcal{Y}^s$, and H-mean result H; – all results shown in percentage. The highlighted values represent the best for each column.

	CUB			**FLO**			**SUN**			**AWA**		
Classifier	$\mathcal{Y}^u$	$\mathcal{Y}^s$	H	$\mathcal{Y}^u$	$\mathcal{Y}^s$	H	$\mathcal{Y}^u$	$\mathcal{Y}^s$	H	$\mathcal{Y}^u$	$\mathcal{Y}^s$	H
Semantic approach												
DAP [76]	4.2	25.1	7.2	–	–	–	1.7	67.9	3.3	0.0	**88.7**	0.0
IAP [76]	1.0	37.8	1.8	–	–	–	0.2	72.8	0.4	2.1	78.2	4.1
DEVISE [42]	23.8	53.0	32.8	9.9	44.2	16.2	16.9	27.4	20.9	13.4	68.7	22.4
SJE [4]	23.5	59.2	33.6	13.9	47.6	21.5	14.7	30.5	19.8	11.3	74.6	19.6
LATEM [149]	15.2	57.3	24.0	6.6	47.6	11.5	14.7	28.8	19.5	7.3	71.7	13.3
ESZSL [118]	12.6	63.8	21.0	11.4	56.8	19.0	11.0	27.9	15.8	6.6	75.6	12.1
ALE [3]	23.7	62.8	34.4	13.3	61.6	21.9	21.8	33.1	26.3	16.8	76.1	27.5
PQZSL [83]	43.2	51.4	46.9	–	–	–	35.1	35.3	35.2	31.7	70.9	43.8
AREN [153]	38.9	78.7	52.1	–	–	–	19.0	38.8	25.5	–	–	–
MLSE [30]	22.3	71.6	34.0	–	–	–	20.7	36.4	26.4	–	–	–
Generative approach												
SAE [35]	8.8	18.0	11.8	–	–	–	7.8	54.0	13.6	1.8	77.1	3.5
f-CLSWGAN [151]	43.8	60.6	50.8	58.8	70.0	63.9	47.9	32.4	38.7	56.0	62.8	59.2
cycle-WGAN [39]	46.0	60.3	52.2	59.1	71.1	64.5	48.3	33.1	39.2	56.4	63.5	59.7
CADA-VAE [129]	51.6	53.5	52.4	–	–	–	47.2	35.7	40.6	57.3	72.8	64.1
GDAN [60]	39.3	66.7	49.5	–	–	–	38.1	**89.9**	**53.4**	–	–	–
GMN [124]	56.1	54.3	55.2	–	–	–	**53.2**	33.0	40.7	61.1	71.3	**65.8**
Zhu et al. [165]	33.4	**87.5**	48.4	–	–	–	–	–	–	–	–	–
LisGAN [82]	46.5	57.9	51.6	57.7	**83.8**	68.3	42.9	37.8	40.2	52.6	76.3	62.3
Domain balancing												
CMT [133]	7.2	49.8	12.6	–	–	–	8.1	21.8	11.8	0.9	87.6	1.8
DAZSL [8]	41.0	60.5	48.9	59.6	81.4	68.8	35.3	40.2	37.6	**64.8**	51.7	57.5
Ours												
$AN-GZSL$	**60.5**	56.6	**58.5**	**80.7**	69.3	**74.5**	41.7	37.1	41.7	58.2	66.1	61.9



(a) CUB (b) FLO

(c) SUN (d) AWA1

(e)

Figure 5.2: ROC curves for the proposed method AN-GZSL, and several baseline and state-of-the-art methods (please see text and Table 5.3 for details about the methods). Note that these graphs are used to compute the AUSUC in Table 5.5.

Table 5.4: GZSL ImageNet results – all results shown in percentage. Please see caption of Table 5.3 for details on each measure. The highlighted values represent the best ones in each column.

Classifier	$\mathcal{Y}^U$	$\mathcal{Y}^S$	H
f-CLSWGAN [151]	0.7	–	–
cycle-WGAN [39]	1.5	**66.5**	2.8
$AN-GZSL$	**2.5**	47.4	**4.8**

Table 5.5: Area under the curve of seen and unseen accuracy (AUSUC). The highlighted values per column represent the best results in each dataset.

Classifier	**CUB**	**FLO**	**SUN**	**AWA**
ESZSL [118]	30.2	25.7	12.8	39.8
fCLSWGAN [151]	34.5	53.1	22.0	45
cycle-WGAN [39]	42.6	60.8	23.2	47.4
DAZSL [8]	35.6	58.1	21.0	**55.9**
$AN-GZSL$	**43.7**	**64.6**	**23.6**	47.9

the most balanced classification results between the seen and unseen domains for all data sets. These results suggest that our proposed multi-modal calibrated classifiers provide a way to correct the mistakes made by each modality classifier. For example, this can happen when the classification probabilities of the correct class are relatively high for both modalities, but not the highest in any modality, and when summed, the correct class receives the highest confidence.

Another important point to notice from Table 5.2 is that our proposed AN-GZSL seems to be more advantageous in fine-grained (i.e., CUB and FLO) than in coarse-grained (i.e., SUN and AWA1) data sets. In coarse-grained data sets, the results from the calibrated visual classifier are almost binary, with the highest classification probability close to one and all other probability values close to zero. The calibrated semantic classifier shows a more uniform distribution, which when combined with the almost binary results of the visual classifier is less effective (than in fine-grained problems) to change a possibly incorrect visual classifier result for the multi-domain multi-modal model.

Comparison with SOTA. In Table 5.3, we notice a clear tendency of the pro-

posed model AN-GZSL to perform substantially better than the SOTA in terms of H-mean and classification accuracy on unseen classes for fine-grained (CUB and FLO) data sets, and competitively for coarse-grained data sets (SUN and AWA1).

This result shows that the more challenging classification problem offered by the fine-grained data sets represents an ideal situation for exploring multi-modal and multi-domain classification. We discuss in the ablation study paragraph above, the reasons behind the superior performance in fine-grained data sets of our proposed AN-GZSL method. Another interesting point to observe from Table 5.3 is that none of the competing methods stand out as a clear SOTA approach for all data sets since one method can be better in one data set, but worse in all others. In fact, out of the four data sets studied, AN-GZSL is better in two, GDAN is better in one and GMN is better in another. It is also worth comparing the performance of previous semantic approaches in Table 5.4, and our proposed semantic network, represented by $AN - GZSL^{\psi}$ in Table 5.2. This comparison is important because our proposed semantic network introduces one significant novelty, which is the use of visual data augmentation for training the semantic classifier. Our proposed $AN - GZSL^{\psi}$ produces substantially better results in terms of H-mean and classification accuracy on unseen classes for CUB, FLO and AWA1.

In terms of the large-scale data set ImageNet, we show in Table 5.4 that the proposed method establishes a new SOTA in terms of the H-mean result. More specifically, the proposed method achieves around 80% of relative H-mean improvement. We speculate that these results can be explained by the similar challenges present in fine-grained and large-scale data sets. Also, the proposed approach scales as well as f-CLSWGAN [151] and cycle-WGAN [39] with respect to the number of classes and samples.

Seen and unseen classification graphs. Figure 5.2 shows the trade-off between the classification of seen and unseen classes for GZSL methods. In particular, it is interesting to notice a fact that is prevalent in GZSL methods, which is the classification imbalance that usually favours the seen classes – the figure illustrates that the majority of the previous methods (represented by diamonds) lie at the bottom-right part of the graphs, indicating the preference for seen classes. In terms of seen and unseen curves, the more balanced methods (see Table 5.3) usually lies

close to the elbow of the curve, located at the top-right part of the graph. The results suggest that our method is more robust to fine-grained data sets, such as CUB and FLO. We argue that this is achieved due to the proximity of the distributions of the classes in the feature spaces. This proximity increases the effectiveness of the Bayesian inference because a large number of samples of the parameter values can show a more calibrated classification result for each class. On the other hand, coarse data sets have class distributions that are likely to be far from each other, resulting in an ineffective Bayesian inference that keeps showing over-confident classification results.

AUSUC. Table 5.5 shows that the proposed approach, AN-GZSL, outperforms previous methods on data sets CUB, SUN and FLO. For AWA1, we achieve competitive performance, where the proposed method is the second best. It is worth emphasising that the AUSUC measure provides a more complete assessment of GZSL methods, where it is no longer necessary to commit to a particular operating point of the classification of seen and unseen classes.

5.6 Conclusions and Future Work

In this paper, we introduce a new approach to perform GZSL using a multi-modal multi-domain augmentation network. The proposed approach is the first to explore visual data augmentation for training visual *and* semantic classifiers, enabling a truly multi-modal training and inference. In addition, we show that the calibration of those visual and semantic classifiers provide an effective multi-domain classification, where the classification of seen and unseen classes are accurate and well balanced. The experimental results show that the proposed approach has established new state-of-the-art GZSL harmonic mean results for three benchmark data sets (CUB, FLO, and Imagenet). In particular, we report results that are substantially better than the previous methods on CUB and FLO, which are fine-grained data sets, and competitive on SUN and AWA1, which are coarse-grained data sets. Moreover, the results of the proposed approach outperform previous methods on Imagenet data set by a large margin. Also, our proposed AN-GZSL achieves the best performance in terms of AUSUC for three benchmark data sets.

In the future, we intend to study more thoroughly the reason behind the performance difference observed between fine-grained and coarse-grained data sets. We will also investigate why it is challenging to obtain high classification accuracy on the unseen classes of the large scale ImageNet data set.

Chapter 6

Multi-domain Generalised Zero-shot Learning using Visual, Semantic, and Joint Latent Spaces

The work contained in this chapter is in submission as the following paper:

Felix, R., Harwood, B., Sasdelli, M., and Carneiro, G., Multi-domain Generalised Zero-shot Learning using Visual, Semantic, and Joint Latent Spaces. In *Submission*, 2019 [38].

Statement of Authorship

Title of Paper	Multi-domain Generalised Zero-shot Learning using Visual, Semantic, and Joint Latent Spaces
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished or Unsubmitted work
Publication Details	Felix, R., Harwood, B., Sasdelli, M., and Carneiro, G., Multi-domain Generalised Zero-shot Learning using Visual, Semantic, and Joint Latent Spaces. In submission (2019)

Principal Author

Name of Principal Author (Candidate)	Rafael Felix Alves		
Contribution to the Paper	- Development of the main idea of the paper; - Implementing and conducting the experiments; - Writing and coordinating the revisions;		
Overall percentage (%)	60		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	29/11/2019

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

x. the candidate's stated contribution to the publication is accurate (as detailed above);

xi. permission is granted for the candidate in include the publication in the thesis; and

xii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Ben Harwood
Contribution to the Paper	- Help with the development of the idea; - Help writing, revision and discussions;

Signature		Date	06/11/2019

Name of Co-Author	Michele Sasdelli		
Contribution to the Paper	- Help with the development of the idea; - Help writing, revision and discussions;		
Signature		Date	29/11/2019

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	- Help writing, revision and discussions;		
Signature		Date	04-11-2019

Abstract

Generalised zero-shot learning (GZSL) methods aim to classify previously seen and unseen visual classes by leveraging the semantic information of those classes. In the context of GZSL, semantic information is non-visual data such as a text description of the seen and unseen classes. Previous GZSL methods have explored transformations between visual and semantic spaces, as well as the learning of a latent joint visual and semantic space. In these methods, even though learning has explored a combination of spaces (i.e., visual, semantic or joint latent space), inference tended to focus on using just one of the spaces. By hypothesising that inference must explore all three spaces, we propose a new GZSL method based on a multi-modal classification over visual, semantic and joint latent spaces. Another issue affecting current GZSL methods is the intrinsic bias toward the classification of seen classes – a problem that is usually mitigated by a domain classifier which modulates seen and unseen classification. Our proposed approach replaces the modulated classification by a computationally simpler multi-domain classification based on averaging the multi-modal calibrated classifiers from the seen and unseen domains. Experiments on GZSL benchmarks show that our proposed GZSL approach achieves competitive results compared with the state-of-the-art.

6.1 Introduction

In the usual visual classification setup, training comprises a set of visual classes, each of which containing a large set of visual samples to model the classifier [54]. The inference process consists of classifying new visual samples into one of the classes used for training. Although useful, this setup bears little resemblance with real-world visual classification problems (e.g., self-driving cars or robotic personal assistant), where previously unseen visual classes must be handled in a reasonable manner. One possible way to address such real-world problems is with the generalised zero-shot learning (GZSL) setup [150] that contains a set of seen and another set of unseen classes – seen classes contain visual samples for training, while unseen classes do not have any visual samples for training. In the

GZSL setup, the recognition of unseen classes depends on semantic information collected from different modalities, such as textual descriptions [114] or a list of attributes [77] for the seen and unseen classes. One of the GZSL challenges lies in how to handle the multi-modal information contained in the visual samples from the seen classes and the semantic samples from the seen and unseen classes. Another GZSL challenge is how to properly balance the classification of new samples from seen and unseen classes because the classification model will be naturally biased toward the classification of seen classes given the availability of visual samples from those classes during training [37, 133].

Traditional GZSL methods aim to build a function that transforms samples from the visual to the semantic space so that the classification of seen and unseen classes are performed exclusively in the semantic space [150]. More recent approaches rely on a generative model to produce visual samples from their respective semantic samples [39, 60, 82, 105, 124, 129, 144, 151]. The generated visual samples from unseen classes and the original visual samples from the seen classes are then used to train a visual classifier that is used during testing in a single modality (i.e., visual) classification. Note that these generative methods are the first GZSL approaches to train a visual classifier with visual samples from both seen and unseen domains. Alternative approaches encode the semantic and the visual data into a joint latent embedding space [129] or with pairwise compatibility functions [165], which are then used to train a classifier that works exclusively in just one of the modalities. It is worth noting that the previous methods presented above explore the multi-modality aspect of GZSL during training, but they always rely on a single modality classifier for testing. We hypothesise that a multi-modal inference has the potential to improve current GZSL results because of a more effective use of the visual and semantic information available.

Another major issue affecting GZSL methods is the imbalance in the classification results for the seen and unseen classes [150]. One of the first GZSL methods [133] noticed that and proposed the use of a domain classification that classifies input visual samples into the set of seen or unseen classes, where in the former case, the sample would go to a visual classifier, and in the latter case, the sample would be transformed into a semantic sample to be classified by a semantic classifier. Therefore, this method [133] not only addressed multi-modal training

and inference, but it also tried to balance the seen and unseen classification. However, its classification accuracy is underwhelming, particularly compared with recent methods. More recent methods also proposed the use of an external domain classifier [8,37], but they always rely on a single modality classification. The major drawback of the approaches above lies in the need to train a domain classifier using visual samples from the seen classes, which is a hard classification problem given that there is no guarantee that the divergence within the seen classes is smaller than the divergence between seen and unseen classes.

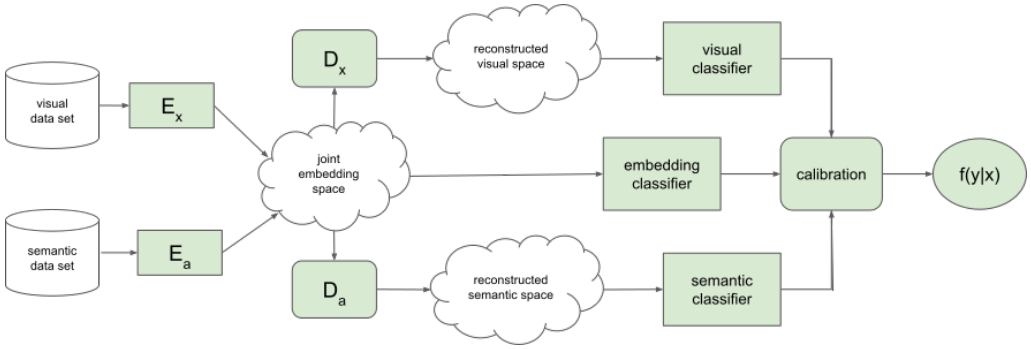


Figure 6.1: Our model consists of encoders from visual and semantic spaces to a latent joint embedding space. Samples from this joint space are used to train decoders that reconstruct the original samples from visual and semantic spaces. Samples from these visual, semantic and joint spaces are then used to train and calibrate classifiers for each space. The final multi-domain classification confidence, represented by $f(y|\mathbf{x})$, is obtained from averaging the results of the multi-modal calibrated classifiers.

In this paper, we introduce a new GZSL approach that relies on multi-modal training and inference, where the multi-domain classification is based on calibrating the classifier from each modality, without the use of any external domain classifier – see Fig. 6.1. More specifically, our model consists of a visual and a semantic encoder that transforms samples from these two domains into samples in a joint latent space. The proposed model also contains decoders from the joint space back to the visual and semantic spaces. The samples from the visual, semantic, and joint latent spaces are used to train the visual, semantic and joint classifiers. By calibrating [56] those multi-modal classifiers, we obtain good balancing between the classification of seen and unseen classes without an external domain classifier.

Experiments include an ablation study that highlights the importance of each modality and the classification calibration. Using public GZSL benchmarks, we show that our method has results that are competitive with the state-of-the-art.

6.2 Literature Review

In this section, we review the recent literature in zero-shot learning (ZSL), GZSL, and domain balancing for GZSL.

6.2.1 Zero-Shot Learning

ZSL is defined as a classification problem, where the set of seen visual classes used for training does not overlap with the set of unseen visual classes used for testing [77, 150]. The main solution explored by ZSL methods is based on the use of an auxiliary semantic space, where each visual class has a particular semantic representation. With the learning of a transformation function that projects samples from visual to semantic spaces, it is then possible to transform samples from unseen visual classes to the semantic space. This approach is motivated by the assumption that the unseen visual clusters can be transferred with same structure into the semantic space for computing inference. However, a recent review of the literature in this field shows that the ZSL set-up limits the applicability of ZSL methods [39, 151] because the testing procedure completely ignores the seen classes [18, 19]. Although limited, ZSL methods can be seen as an expert model for the unseen visual classes [8].

6.2.2 Generalized Zero-Shot Learning

GZSL extends the ZSL framework with the recognition of the seen and unseen visual classes during testing. This extension is challenging due to the bias toward the seen classes issue reported in [19, 133, 150], which has motivated the development of several GZSL approaches [151]. Previously, studies in GZSL have been based on an ensemble of classifiers that combines semantic classifiers [30], approaches that learn transformations between the visual and semantic spaces [159, 165], methods that combine seen and unseen classifications [8, 19, 133], and algorithms that

generate synthetic unseen visual samples [39, 60, 82, 105, 124, 129, 144, 151].

The most successful GZSL approaches are based on methods that generate synthetic visual samples for the unseen classes, given their semantic representation [39, 60, 82, 105, 124, 129, 144, 151]. These synthetic unseen visual samples, together with the real seen visual samples, are used to train a visual classifier of seen and unseen classes. The generative models explored by these methods are the Generative Adversarial Networks (GAN) [39, 60, 82, 105, 124, 151] and Variational Autoencoders (VAE) [129, 144]. The approaches above do not have a testing stage that can handle multi-modal (i.e., visual and semantic) classification. In fact, during the testing stage, these approaches only deal with samples either in the visual space or in a joint visual and semantic latent space. We hypothesise that the use of all spaces (i.e., visual, semantic and joint latent spaces) can improve recognition accuracy.

The first method to address the bias toward the seen classes was proposed by Socher et al. [133]. Their paper realised that GZSL classifiers were biased towards the seen classes because of the availability of visual samples from seen classes and the lack of unseen visual samples during training. This issue is usually handled with a domain classifier that classifies test samples into the seen or unseen classes, and use different classifiers for each domain [37, 133, 158]. More recently, the approach developed by Atzmon and Chechik [8] tackles the bias issue toward seen classes in a similar manner. Their solution involves a classifier that combines the result of a ZSL classifier for the unseen classes and a seen class classifier, where this combination is achieved with a (seen/unseen) gating network. Even though this approach achieves outstanding results, it can be criticised for not exploring more effectively the multi-modality nature of the problem and for relying on a computationally complex domain classifier that is challenging to be trained given the assumption that samples from unseen classes come from a distribution that has a high divergence with respect to the seen class distribution, which is hard to guarantee.

6.3 Methods

In this section, we first present the GZSL problem. Then we introduce our proposed model that consists of a calibrated classifiers over the visual, semantic and joint latent spaces.

6.3.1 Generalised Zero-Shot Learning

GZSL methods rely on visual and semantic data modalities. The data set for the visual modality is represented by $\mathcal{D} = \{(\mathbf{x}, y)_i\}_{i=1}^{N}$, where $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^X$ denotes the visual representation, and $y \in \mathcal{Y} = \{1, ..., C\}$ denotes the visual class. The visual representation consists of visual features extracted by pre-trained deep neural networks, such as ResNet [58], and VGG [131]. In GZSL problems, $\mathcal{D}$ is split into two disjoint domains: the seen domain $\mathcal{Y}^S = \{1, ..., |S|\}$, and the unseen domain $\mathcal{Y}^U = \{(|S|+1), ..., (|S|+|U|)\}$, where $\mathcal{Y} = \mathcal{Y}^S \cup \mathcal{Y}^U$, and $\mathcal{Y}^S \cap \mathcal{Y}^U = \emptyset$. Visual samples from $\mathcal{Y}^S$ can be accessed during training time, but samples from the unseen domain $\mathcal{Y}^U$ are only available during test time. Therefore the main challenge in GZSL consists of classifying samples that are drawn from $\mathcal{Y}$, independently if they come from the seen or unseen domain [150]. The data set for the semantic modality is defined as $\mathcal{R} = \{\mathbf{a}_y\}_{y \in \mathcal{Y}}$, where each $\mathbf{a}_y \in \mathcal{A} \subseteq \mathbb{R}^A$ is associated to a visual class from $\mathcal{Y}$. The semantic representation consists of a semantic information (e.g., textual description, or a set of attributes) available for the visual classes. This information can be transformed into an embedding space by feature representation methods (e.g., set of continuous features such as *word2vec* [150]). The semantic data set has only one representation per visual class.

GZSL has a particular set up for the training and testing stages. The data set $\mathcal{D}$ is divided into two subsets: $\mathcal{D}^{tr}$ for training, and $\mathcal{D}^{ts}$ for testing. The training set contains visual samples drawn from the seen classes $\mathcal{Y}^S$ and the testing set contains visual samples from both the seen and unseen domains. The semantic data set, $\mathcal{R}$, is available during training and testing.

6.3.2 GZSL with Calibrated Classifiers over Visual, Semantic and Joint Latent Spaces

The inference of our proposed model estimates the visual class y of a test image $\mathbf{x}$, as follows:

$$y^* = \arg\max_{y \in \mathcal{Y}} f(y|\mathbf{x}), \tag{6.1}$$

with

$$f(y|\mathbf{x}) = \sigma_x(y|\tilde{\mathbf{x}}, \tau_x, \theta_x) + \sigma_a(y|\tilde{\mathbf{a}}, \tau_a, \theta_a) + \sigma_z(y|\tilde{\mathbf{z}}, \tau_z, \theta_z), \tag{6.2}$$

where $\tilde{\mathbf{x}} \in \mathcal{X}$ represents a generated visual sample, $\tilde{\mathbf{a}} \in \mathcal{A}$ denotes a generated semantic sample, $\tilde{\mathbf{z}} \in \mathcal{Z} \subseteq \mathbb{R}^Z$ is a generated joint latent sample, and $\sigma_x(.), \sigma_a(.), \sigma_z(.)$ represent the softmax classifiers for the visual, semantic and joint latent spaces – these classifiers are parameterised by $\theta_x, \theta_a, \theta_z$, and calibrated by τ_x, τ_a, τ_z, respectively. Note that the inference defined in Eq. 6.1 and Eq. 6.2 shows the main contributions of this paper: 1) the multi-modal inference, and 2) the domain balancing by classifier calibration without any external domain classifier to distinguish samples from seen and unseen classes.

The whole model depicted in Fig. 6.1 shows other components that are defined below. The visual and semantic encoders are defined by

$$\begin{aligned} \tilde{\mathbf{z}} &\sim p_x^{(E)}(\mathbf{z}|\mathbf{x}, \theta_x^{(E)}), \\ \tilde{\mathbf{z}} &\sim p_a^{(E)}(\mathbf{z}|\mathbf{a}, \theta_a^{(E)}), \end{aligned} \tag{6.3}$$

where $p_x^{(E)}(.)$ and $p_a^{(E)}(.)$ denote the visual and semantic encoding models. The visual and semantic decoders are defined by

$$\begin{aligned} \tilde{\mathbf{x}} &\sim p_x^{(D)}(\mathbf{x}|\mathbf{z}, \theta_x^{(D)}), \\ \tilde{\mathbf{a}} &\sim p_a^{(D)}(\mathbf{a}|\mathbf{z}, \theta_a^{(D)}), \end{aligned} \tag{6.4}$$

where $p_x^{(D)}(.)$ and $p_a^{(D)}(.)$ represent the visual and semantic decoding models.

There have been many GZSL methods that rely on the generation of synthetic visual samples, given their semantic representation [39, 60, 82, 105, 124, 129, 144, 151], as described in Sec. 6.2.2. In this paper, we extend the model proposed by Schonfeld et al. [129]. In particular, the training of the model defined in Eq. 6.1-

Eq. 6.4 is an end-to-end process that minimises the following loss function:

$$\ell(\mathcal{D}^{tr}, \mathcal{R}) = \gamma_{PD}\ell_{PD} + \ell_{VAE} + \gamma_{CM}\ell_{CM} + \gamma_{DA}\ell_{DA}. \tag{6.5}$$

The first term in Eq. 6.5 enables the training of a GZSL model taking into consideration the joint domain optimisation (with the seen and unseen domain) and the multi-modal inference (visual, semantic and latent spaces). The sample-wise loss ℓ_{PD} is defined as the cross-entropy loss for the classifiers in Eq. 6.2, as follows:

$$\begin{aligned} \ell_{PD} = &- \mathbf{h}_y \log(\sigma_x(y|\tilde{\mathbf{x}}, \tau_x, \theta_x)) - \mathbf{h}_y \log(\sigma_a(y|\tilde{\mathbf{a}}, \tau_a, \theta_a)) \\ &- \mathbf{h}_y \log(\sigma_z(y|\tilde{\mathbf{z}}, \tau_z, \theta_z)), \end{aligned} \tag{6.6}$$

where $\mathbf{h}_y$ represents the y^{th} dimension of a one-hot representation of the label y, the sample $\tilde{\mathbf{z}}$ is generated according to Eq. 6.3 using the encoders from the semantic and visual spaces, and the samples $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{a}}$ are generated with the decoders in Eq. 6.4. It is important to notice in Eq. 6.6 that there is no hyper-parameter or external domain classifier that weights the classification for each modality, as is the case in previous GZSL methods [8,133]. Instead, we rely entirely on calibrating the classifiers using temperature scaling [56], which, for the case of the softmax classifier, is defined by

$$\sigma_x(y|\mathbf{x}, \tau_x, \theta_x) = \frac{\mathrm{e}^{(\pi_x(y|\mathbf{x}, \theta_x)/\tau_x)}}{\sum_{c=1}^{C} \mathrm{e}^{(\pi_x(c|\mathbf{x}, \theta_x)/\tau_x)}}, \tag{6.7}$$

where $\pi_x(y|\mathbf{x}, \theta_x)$ represents the logit for the visual classification (and similarly for $\sigma_a(y|\mathbf{a}, \tau_a, \theta_a)$ and $\sigma_z(y|\mathbf{z}, \tau_z, \theta_z)$ in Eq. 6.2). In traditional supervised learning, the temperature scaling factor τ is assumed to be equal to one. However, recent research shows that this parameter can be used for calibrating the classification confidence [56]. After calibrating each classifier, the ensemble consists of summing the three classification results from Eq. 6.2. The calibration parameters are learned based on the validation set held out from training, as proposed in [150].

The second term in Eq. 6.5 represents the variational auto-encoder (VAE) error [29], defined by [129]. The sample-wise loss for that second term is denoted

by

$$\begin{aligned}\ell_{VAE} =& \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\lambda)}\big[\log(p_x^{(D)}(\mathbf{x}|\mathbf{z},\theta_x^{(D)}))\big] \\ &+ \mathbb{E}_{q(\mathbf{z}|\mathbf{a},\lambda)}\big[\log(p_a^{(D)}(\mathbf{a}|\mathbf{z},\theta_x^{(D)}))\big] \\ &- \mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x},\lambda_x)||p_\phi(\mathbf{z})\big) \\ &- \mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{a},\lambda_a)||p_\phi(\mathbf{z})\big),\end{aligned} \tag{6.8}$$

which represents the variational loss, where the first term aims to minimize the reconstruction error for the visual features, the second term minimises the reconstruction error for the semantic features, and the last two terms represent the Kullback-Leibler divergence between the prior distribution $p_\phi(\mathbf{z})$ (assumed to be Gaussian) and the variational distributions $q_\phi(\mathbf{z} \mid \mathbf{x}, \lambda_x)$ and $q_\phi(\mathbf{z} \mid \mathbf{x}, \lambda_a)$, also assumed to be Gaussian.

The third term in Eq. 6.5 denotes the cross-modality alignment loss that calculates the reconstruction error between the visual and semantic modalities [129]. The sample-wise loss for that third term is defined by:

$$\ell_{CM} = \|\mathbf{x} - \tilde{\mathbf{x}}\| + \|\mathbf{a} - \tilde{\mathbf{a}}\|, \tag{6.9}$$

where $\tilde{\mathbf{x}}$ is sampled from the decoder $p_x^{(D)}(\mathbf{x}|\tilde{\mathbf{z}},\theta_x^{(D)})$ in Eq. 6.4, with $\tilde{\mathbf{z}}$ being sampled from $p_a^{(E)}(\mathbf{z}|\mathbf{a},\theta_a^{(E)})$ in Eq. 6.3 and $\mathbf{x}$ and $\mathbf{a}$ belonging to the same class. Similarly in Eq. 6.9, $\tilde{\mathbf{a}}$ is sampled from the decoder $p_a^{(D)}(\mathbf{a}|\tilde{\mathbf{z}},\theta_a^{(D)})$ in Eq. 6.4, with $\tilde{\mathbf{z}}$ being sampled from $p_x^{(E)}(\mathbf{z}|\mathbf{x},\theta_x^{(E)})$ in Eq. 6.3 and $\mathbf{x}$ and $\mathbf{a}$ belonging to the same class.

The fourth term in Eq. 6.5 consists of the distribution-alignment loss of samples belonging to the same class. The loss is defined by [129]:

$$\ell_{DA} = || \mu_x - \mu_a ||_2^2 + || \Sigma_x^{\frac{1}{2}} - \Sigma_a^{\frac{1}{2}} ||_F^2, \tag{6.10}$$

where $\mu_x \in \mathcal{Z}$ and $\Sigma_x \in \mathcal{Z} \times \mathcal{Z}$ are the mean vector and co-variance matrix of the latent samples from a particular class produced by the encoder $p_x^{(E)}(\mathbf{z}|\mathbf{x},\theta_x^{(E)})$ (similarly for μ_a and Σ_a for $p_a^{(E)}(\mathbf{z}|\mathbf{a},\theta_x^{(E)})$), and $\|.\|_F$ represents the Frobenius norm. This loss assumes a uni-modal Gaussian distribution of the latent vectors of a particular class, and approximates the distributions produced by the visual and semantic classes. The training is achieved by minimising the loss in Eq. 6.10 with the average of the sample-wise losses defined in Equations 6.6, 6.8, 6.9, where the hyper-parameters are estimated with grid search using the validation set.

6.4 Experiments

In this section, we introduce the experimental setup to demonstrate the performance of the proposed method. First, we present the benchmark data sets, then we describe the evaluation criteria for the experimental setup. We then show the results of the proposed method compared with previous models from the literature. Finally, we provide ablation studies to explore the functionality of the proposed method.

Table 6.1: The benchmarks for GZSL: AWA1 [150], AWA2 [150], CUB [146], and SUN [152]. Column (1) shows the number of seen classes, denoted by $|\mathcal{Y}^S|$, split into the number of training and validation classes (train+val), (2) presents the number of unseen classes $|\mathcal{Y}^U|$, (3) displays the number of samples available for training $|\mathcal{D}^{Tr}|$ and (4) shows number of testing samples that belong to the unseen classes $|\mathcal{D}_U^{Te}|$ and number of testing samples that belong to the seen classes $|\mathcal{D}_S^{Te}|$ from [39, 151]

Name	$\lvert\mathcal{Y}^S\rvert$ (train+val)	$\lvert\mathcal{Y}^U\rvert$	$\lvert\mathcal{D}^{Tr}\rvert$	$\lvert\mathcal{D}_U^{Te}\rvert + \lvert\mathcal{D}_S^{Te}\rvert$
AWA1	40 (27+13)	10	19832	4958+5685
AWA2	40 (27+13)	10	23527	5882+7913
CUB	150 (100+50)	50	7057	1764+2967
SUN	745 (580+65)	72	14340	2580+1440

6.4.1 Data Sets

We evaluate the proposed method on four publicly available[1] benchmark GZSL data sets: AWA1 [76, 150], AWA2 [76, 150], CUB [146], and SUN [150]. Recent research argues that GZSL approaches that use pre-trained models must take into consideration the overlap between unseen classes and the ImageNet classes [150]. Therefore, we use the GZSL experimental setup described by Xian et al. [150], which prevents that the GZSL unseen classes overlap with the ImageNet classes [26, 150]. These data sets can be either fine or coarse-grained. The CUB data set [146] is fine-grained, where the visual classes are similar to each other, and

[1] Data sets from https://cvml.ist.ac.at/AwA2/.

the semantic representation contains discriminative details. The data sets SUN, AWA1 and AWA2 are coarse-grained, where visual classes are better separated. In particular, SUN represents a challenging GZSL problem due to the number and diversity of classes [150]. Table 6.4 contains basic information about the data sets in terms of the number of seen and unseen classes and the number of training and testing images.

6.4.2 Feature Representation

The visual representation for all the benchmark data sets is extracted from the activation of the 2048-dimensional top pooling layer of ResNet-101 [58]. The semantic representation of CUB [150] consists of the 1024-dimensional vector produced by CNN-RNN [114]. These semantic samples represent a written description of each image using 10 sentences per image. To define a unique semantic sample per-class, we average the semantic samples of all images belonging to each class [150]. For AWA1, AWA2 and SUN we used the semantic features proposed by Xian et al. [150], where we use the 102-dimensional feature for SUN [150], and the 85-dimensional feature for AWA1 [150] and AWA2 [150].

6.4.3 Evaluation Protocol

We evaluate the proposed model with Xian et al.'s [150, 151] protocol, which has been widely used for GZSL evaluation. This protocol relies on three measures: top-1 accuracy for the seen samples, top-1 accuracy for the unseen samples, and the harmonic mean. The top-1 accuracy is computed by the average per-class, then we calculate the overall mean over all classes. We calculate the mean-class accuracy for each domain separately, i.e., the seen ($\mathcal{Y}^S$) and the unseen ($\mathcal{Y}^U$) classes. The harmonic mean (H-mean) is a measure that combines the accuracy for the seen and unseen domains [150]. We also present experiments using the area under the seen and unseen curve (AUSUC) [19]. The AUSUC is achieved by varying a balancing factor between the seen and the unseen contributions for the harmonic-mean [19]. The AUSUC is a more general assessment of GZSL methods, compared with the measures above, because it does not commit to any operating point of the seen and unseen classification. In fact, AUSUC shows the overall performance of the

GZSL method, where several operating points are considered, with each point representing different classification biases for the unseen and seen classes. The evaluation protocol follows the guidelines reported by [150].

6.4.4 Implementation Details

In this section, we describe the architecture for the proposed model. We first describe the variational auto-encoder network, where the visual encoder is a network comprising one hidden layer with 1560 nodes, and the semantic encoder is a network consisting of one hidden layer with 1450 nodes. The visual decoder and the semantic decoder are represented by networks with one hidden layer containing 1560 and 660 nodes, respectively. The latent space $\mathcal{Z}$ contains 64 dimensions. The whole model is optimised with Adam for 100 epochs [71]. The hyper-parameters γ_{PD}, γ_{CM} and γ_{DA} are estimated with cross-validation. The multi-modal classifiers in Eq. 6.1 are represented by a neural network with one linear layer transformation and an output layer of size $|\mathcal{Y}| = C$. As proposed in Eq. 6.7, all these classifier networks have a softmax activation function after the linear layer. The training of these classifiers relies on multi-class cross-entropy loss and Adam optimiser [71], with a learning rate of 0.001. To alleviate the lack of unseen samples, we generated artificial samples from the semantic representation for all benchmark data sets during the training of the classifiers. The training uses only seen visual samples. We propose the optimisation of the loss function in Eq. 6.5, by alternating the training of each component. Furthermore, we calibrate the predictions with temperature scaling for GZSL models, as described in Eq. 6.7, where this optimisation process depends on the validation set provided by Xian et al [150], and each classifier has a singular temperature scale.[2]

6.4.5 Results

Table 6.2 shows an ablation study of the proposed model. The ablation results show the accuracy of the classifiers trained for each modality: the joint visual/semantic embedding space $classifier(\tilde{\mathbf{z}})$ (similarly to Schonfeld et al. [129]); the reconstructed visual space $classifier(\tilde{\mathbf{x}})$; and the reconstructed semantic space

[2]Code available at https://github.com/rfelixmg/multi-spaces-gzsl.

$classifier(\tilde{\mathbf{a}})$. We also show the results with our multi-modal approach trained without temperature calibration, denoted by 'ours ($\tau = 1$)'. The last row in Table 6.2 shows the result of our proposed multi-modal approach with calibration.

Table 6.2: Ablation study of our GZSL approach, using per-class average top-1 accuracy on the test sets of unseen classes $\mathcal{Y}^U$, seen classes $\mathcal{Y}^S$, and H-mean result H – all results shown in percentage. We report the results for each of the embedding spaces used for classification, the simple average combination without classification calibration (denoted as $\tau = 1$ in Eq. 5.2), and the proposed temperature calibrated method. The best result per column is highlighted.

	AWA1			**AWA2**			**CUB**			**SUN**		
Classifier	$\mathcal{Y}^S$	$\mathcal{Y}^U$	H	$\mathcal{Y}^S$	$\mathcal{Y}^U$	H	$\mathcal{Y}^S$	$\mathcal{Y}^U$	H	$\mathcal{Y}^S$	$\mathcal{Y}^U$	H
$classifier(\tilde{\mathbf{x}})$	76.5	44.1	56.0	81.4	43.8	57.0	65.0	28.0	39.1	28.9	48.7	36.3
$classifier(\tilde{\mathbf{a}})$	77.0	42.1	54.4	81.9	47.9	60.4	61.5	25.0	35.6	24.7	36.7	29.5
$classifier(\tilde{\mathbf{z}})$	76.6	55.0	64.1	75.3	55.5	63.9	57.2	48.4	52.4	**36.8**	45.1	40.6
ours ($\tau = 1$)	**80.0**	51.3	62.5	**84.4**	52.0	64.4	**66.7**	30.1	41.5	32.8	**49.2**	39.3
ours	75.2	**57.3**	**65.0**	73.2	**58.5**	**65.0**	55.2	**52.7**	**54.0**	35.6	47.4	**40.7**

In Table 6.3, we evaluate the performance of the proposed approach, referred to as 'ours', and compare it to several models in the literature. More specifically, we show the results for the data sets AWA1, AWA2, CUB, and SUN and compare the proposed model to recently proposed and baseline GZSL methods. We define three distinct groups of GZSL approaches: semantic approach, generative approach and models that combine domain classifiers. In the semantic approach we compare the results from the proposed approach to SJE [4], ALE [3], LATEM [149], ESZSL [118], SYNC [18], DEVISE [42], AREN [153], PQZSL [83], and MLSE [30]. This group focuses on learning a transformation from visual to semantic representation, then the classification is based on nearest neighbour classification in the semantic space. For the generative approach we compare the proposed model to SAE [35], f-CLSWGAN [151], cycle-WGAN [39], CADA-VAE [129], Zhu et al. [165], LisGAN [82], GMN [124], and GDAN [60]. This group of GZSL approaches rely on generative models to produce synthetic visual features for the unseen classes. We also compare the proposed model to approaches that combine the seen and unseen domain classifiers: CMT [133], DAZSL [8], and SABR [105].

In Table 6.4 and Fig. 6.2, we show the area under the curve of seen and unseen

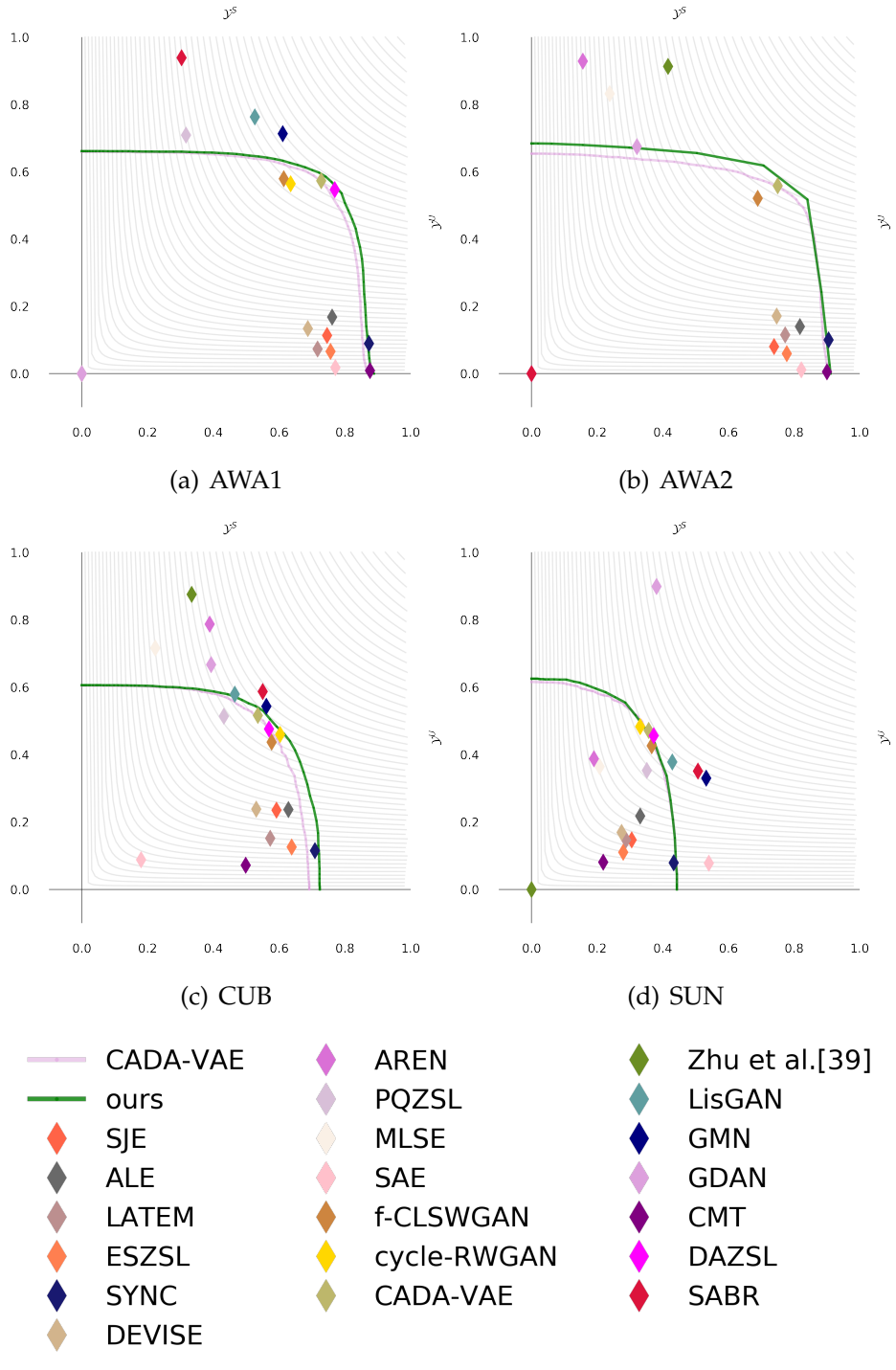


Figure 6.2: The area for seen and unseen accuracy curve for the proposed method (green) and CADA-VAE [129] (pink), which is the closest model to ours (please see text and Table 6.3 for details about the methods). Note that these graphs are used to compute the AUSUC in Table 6.4

accuracy (AUSUC) results [19]. We evaluate the proposed model in terms of AUSUC for the benchmark data sets AWA1, AWA2, CUB, and SUN; and compare the results with the following GZSL models: ESZSL [118], DAZSL [8], f-CLSWGAN [151], cycle-WGAN [39] and CADA-VAE [129]. We only show AUSUC results for methods that published those results or for methods that have code available online, which allowed us to run and obtain the results.

Table 6.3: GZSL results using per-class average top-1 accuracy on the test sets of unseen classes $\mathcal{Y}^U$, seen classes $\mathcal{Y}^S$, and H-mean result H – all results shown in percentage. The results from previously proposed methods in the field were extracted from [150]. The highlighted values represent the best ones in each column.

	AWA1			**AWA2**			**CUB**			**SUN**		
Classifier	$\mathcal{Y}^S$	$\mathcal{Y}^U$	H	$\mathcal{Y}^S$	$\mathcal{Y}^U$	H	$\mathcal{Y}^S$	$\mathcal{Y}^U$	H	$\mathcal{Y}^S$	$\mathcal{Y}^U$	H
Semantic approach												
SJE [4]	74.6	11.3	19.6	73.9	8.0	14.4	59.2	23.5	33.6	30.5	14.7	19.8
ALE [3]	76.1	16.8	27.5	81.8	14.0	23.9	62.8	23.7	34.4	33.1	21.8	26.3
LATEM [149]	71.7	7.3	13.3	77.3	11.5	20.0	57.3	15.2	24.0	28.8	14.7	19.5
ESZSL [118]	75.6	6.6	12.1	77.8	5.9	11.0	63.8	12.6	21.0	27.9	11.0	15.8
SYNC [18]	87.3	8.9	16.2	90.5	10.0	18.0	**70.9**	11.5	19.8	43.3	7.9	13.4
DEVISE [42]	68.7	13.4	22.4	74.7	17.1	27.8	53.0	23.8	32.8	27.4	16.9	20.9
AREN [153]	–	–	–	15.6	92.9	26.7	38.9	78.7	52.1	19.0	38.8	25.5
PQZSL [83]	31.7	70.9	43.8	–	–	–	43.2	51.4	46.9	35.1	35.3	35.2
MLSE [30]	–	–	–	23.8	83.2	37.0	22.3	71.6	34.0	20.7	36.4	26.4
Generative approach												
SAE [35]	77.1	1.8	3.5	82.2	1.1	2.2	18.0	8.8	11.8	**54.0**	7.8	13.6
f-CLSWGAN [151]	61.4	57.9	59.6	68.9	52.1	59.4	57.7	43.7	49.7	36.6	42.6	39.4
cycle-WGAN [39]	63.5	56.4	59.7	–	–	–	60.3	46.0	52.2	33.1	48.3	39.2
CADA-VAE [129]	72.8	57.3	64.1	75.0	55.8	63.9	53.5	51.6	52.4	35.7	47.2	40.6
Zhu et al. [165]	–	–	–	41.6	**91.3**	57.2	33.4	**87.5**	48.4	–	–	–
LisGAN [82]	52.6	76.3	62.3	–	–	–	46.5	57.9	51.6	42.9	37.8	40.2
GMN [124]	61.1	71.3	**65.8**	–	–	–	56.1	54.3	55.2	53.2	33.0	40.7
GDAN [60]	–	–	–	32.1	67.5	43.5	39.3	66.7	49.5	38.1	**89.9**	**53.4**
Combining classifiers												
CMT [133]	**87.6**	0.9	1.8	**90.0**	0.5	1.0	49.8	7.2	12.6	21.8	8.1	11.8
DAZSL [8]	80.0	52.8	63.6	–	–	–	57.8	44.4	50.2	37.7	44.9	41.0
SABR [105]	30.3	**93.9**	46.9	–	–	–	55.0	58.7	**56.8**	50.7	35.1	41.5
ours	75.2	57.3	65.0	73.2	58.5	**65.0**	55.2	52.7	54.0	35.6	47.4	40.7

Table 6.4: Area under the curve of seen and unseen accuracy (AUSUC). The highlighted values per column represent the best results in each data set. The notation * represents the results that we reproduced. The best result per column is highlighted.

Classifier	**AWA1**	**AWA2**	**CUB**	**SUN**
EZSL [118]	39.8	–	30.2	12.8
DAZSL [8]	53.2	–	35.7	23.9
f-CLSWGAN [151]	46.1	–	35.5	22.0
cycle-WGAN [39]*	47.3	–	**41.8**	23.2
CADA-VAE [129]*	52.4	52.2	37.0	23.6
ours	**53.2**	**54.9**	39.3	**24.0**

6.5 Discussions

The ablation results in Table 6.2 shows that the proposed approach is more accurate than each one of the single modality classifiers (joint semantic/visual space, reconstructed visual and reconstructed semantic spaces). We also show in Table 6.2 that the calibration of all classifiers provides a substantial improvement in terms of H-mean for all data sets, compared with a simple combination of un-calibrated classifiers. This suggests that the proposed combination of multi-modal calibrated classifiers enables an accurate multi-domain classification with a good balance between seen and unseen classification.

Table 6.3 shows that there is not a dominant method in the current GZSL literature for top-1 accuracy measures. For instance, for AWA1, we notice that GMN [124] and our approach are the top performing methods, with similar H-mean results. For AWA2, our method is the best, with CADA-VAE [129] being slightly worse, but comparable. For CUB, we notice that SABR [105], GMN [124] and our approach are the top performing methods, with comparable H-mean results. For SUN, GDAN [60] is significantly better than all other approaches. Therefore, these results suggest that the top performing GZSL methods in the field are GMN and ours, with other methods being superior on one data set and inferior on other data sets (e.g., GDAN [60] and SABR [105]). It is also important to notice that our approach produces better H-mean results than CADA-VAE [129], which

is the most influential method for our proposed approach. Also, on the SUN data set, our approach is in fact competitive with all other methods in the field, except for the recently proposed GDAN [60] that is more than 10% better than any other approach in the field.

Furthermore, Table 6.4 and Fig. 6.2 show that the proposed approach achieved solid improvement in terms of AUSUC compared to the previous state of the art. More specifically, the proposed approach produces the highest AUSUC in three out of the four data sets (SUN, AWA1, and AWA2), and also improves over CADA-VAE [129] on all four data sets. For CUB, our AUSUC result is the second best among the methods in Table 6.4.

6.6 Conclusions

In this paper, we introduced an approach that explores multi-modal (i.e., visual, semantic and joint latent modalities) and multi-domain (seen and unseen classes) GZSL classifiers. The multi-modal aspect of our proposal is based on a dual encoder-decoder method that uses a joint latent space to transform samples between the visual and semantic spaces. This mechanism allows us to generate samples for seen and unseen classes for each of the visual, semantic, and latent joint modalities, forming a multi-modal GZSL classification. By calibrating each modality classifier, we show that we can achieve a good balance between the classification of seen and unseen classes, producing an accurate multi-domain classification method. The experimental results provide evidence for these contributions and demonstrate that the proposed approach achieves competitive results in common GZSL benchmarks. Specifically, the proposed proposed method achieved state-of-the-art H-mean results for AWA1, AWA2, and CUB. Moreover, the proposed model achieves state-of-the-art results in terms of AUSUC for SUN, AWA1 and AWA2.

In Sec. 5.5, we discussed how the proposed method can combine complementary information from multiple modalities and domains. We believe that our result can motivate further study in GZSL on how to combine other modalities and domains. We also believe that we can extend the proposed model to work with different generative models, which can potentially produce better synthetic

samples to train the GZSL models.

Chapter 7

Conclusion and Future Directions

In this thesis, we investigated the impact of multi-domain and multi-modal optimisation, combined with data augmentation, for Generalised Zero-Shot Learning methods. In this chapter, we discuss the key contributions of this thesis and highlight the future work in this field.

7.1 Summary of the Contributions

In Chapter 3, we introduced a *Multi-modal Cycle-consistent Generalised Zero-Shot Learning* (cycle-WGAN) [39]. The proposed model contains a multi-modal cycle-consistent loss term, which regularises the optimisation of a GAN model for GZSL. The proposed cycle-consistent loss aims to guarantee that the synthetic visual samples can be transformed back to the semantic samples utilised in the generative process. The model cycle-WGAN is the first GZSL data augmentation approach to introduce a cycle-consistency loss [164]. This new constraint promotes faster convergence and better generalisation when compared to the previous data augmentation GZSL methods. We validate the work with extensive experiments on five benchmark data sets, where the proposed model established a new state-of-the-art for the data sets CUB, AWA, SUN, FLO and ImageNet (at the time of its publication).

In Chapter 4, we presented a *Generalised Zero-Shot Learning with Domain Classification in a Joint Semantic and Visual Space*. This work represents a novel domain classifier that estimates the source domain (i.e., seen or unseen) of a sample. We

design this domain classifier using a joint latent space that contains samples from both domains and, therefore, no longer relies on a challenging one-class classifier, as previous GZSL domain classifier approaches [133]. We demonstrate that the proposed approach is effective when combined with GZSL data augmentation methods based on GANs and VAEs. We report results with experiments on the following benchmark data sets: CUB, SUN, AWA1 and AWA2. The proposed method outperforms previous approaches in terms of harmonic-mean for the benchmark data sets CUB, AWA1 and AWA2 (at the time of its publication).

In Chapter 5, we proposed the *Augmentation Network for Multi-modal and Multi-domain Generalised Zero-Shot Learning* (AN-GZSL). The proposed AN-GZSL model is trained by an optimisation procedure that uses multi-modal and multi-domain information. The augmentation network uses semantic samples from the seen and unseen domains, and visual samples from the seen domain to promote a multi-domain optimisation. Furthermore, the synthetic samples from the augmentation network feed a multi-modal classifier. We calibrate the output of the multi-modal models with temperature scaling. For this investigation, we report experiments in five benchmark data sets: CUB, FLO, SUN, AWA and ImageNet. The proposed method achieves state-of-the-art results for CUB, FLO and ImageNet, and it shows competitive results for SUN and AWA.

In Chapter 6, we proposed a *Multi-domain Generalised Zero-Shot Learning using Visual, Semantic, and Joint Latent Spaces*. This work extends the GZSL data augmentation approach based on a dual VAE. The main contribution comprises the use of reconstruction spaces to train three multi-modal classifiers, namely the visual, semantic and joint latent classifiers that are calibrated to enable a balanced multi-domain classification. In this paper, we report results for the benchmark data sets CUB, SUN, AWA1 and AWA2, where the proposed model achieves competitive performance to previous methods in the GZSL problem.

7.2 Limitations and Future Directions

GZSL data augmentation methods represent most of the leading approaches with outstanding performance [144, 151]. However, the 2-step scheduled training, where the generator is trained first, followed by the training of the GZSL classifier,

can be considered a weakness of our method. Therefore, we plan to design an end-to-end optimisation for GZSL data augmentation methods that can train the generator and the GZSL classifier at the same time.

Moreover, recent studies have measured the qualitative performance of deep generative models by their capacity to fool humans with synthetic samples of images, audios and textual information. In this test, a human operator is asked to choose whether a sample is real or synthetic [55,115]. Despite the outstanding quantitative progress achieved with GZSL data augmentation models, less effort has been devoted to the design of models that can produce realistic synthetic images for the unseen visual classes. In the future, we aim to investigate a deep generative model framework that can synthesise realistic images from the semantic descriptions of unseen classes. The capacity of generating images from unseen classes based on descriptions can be potentially useful in forensics [160] and machine creativity [49].

Furthermore, this thesis focuses on the two main modalities available for GZSL benchmark data sets, which are the visual and semantic modalities. Current findings reveal that multi-modal training and inference improves the performance of machine learning models [38–40]. Several benchmark data sets are composed of multiple modalities which have not been explored by GZSL approaches [146,152]. Future work should concentrate on the extension of multi-modal concepts to additional modalities, such as audio, sketches, and videos [10,28,51,110,130,132, 160]. It remains to be investigated whether different modalities other than text can be utilised for improving GZSL applications.

Finally, this thesis focuses on the training and inference steps from the GZSL pipeline that do not include a pre-processing step for the visual and semantic data sets. This is due to the standardization of GZSL benchmark data sets with an experimental setup [150], which takes into consideration the overlap between the classes used to pre-train deep learning models and GZSL classes. In the future, we plan to investigate the data representation for GZSL models, where we will replace the GZSL models based on pre-trained backbones with an end-to-end training pipeline that includes the learning of the pre-processing step.

Despite the limitations listed in this chapter, this thesis proposed substantial contributions to the field of Generalised Zero-Shot Learning. We systematically

addressed some of the gaps in this field by proposing methods that use data augmentation, multiple modalities and multiple domains in the training/inference of Generalised Zero-Shot Learning models. At the time of their publications, the empirical results show that the approaches proposed in this thesis established new state-of-the-art., showing their practical importance to the field.

Acknowledgement

The author gratefully acknowledge the support of the Australian Research Council through the Centre of Excellence for Robotic Vision (project number CE140100016), Laureate Fellowship FL130100102 to IR and Discover Project DP180103232.

Bibliography

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] 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*, pages 819–826, 2013.

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

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

[5] E. Alpaydin. *Introduction to machine learning*. MIT press, 2009.

[6] Y. Annadani and S. Biswas. Preserving semantic relations for zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223, 2017.

[8] Y. Atzmon and G. Chechik. Adaptive confidence smoothing for generalized zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11671–11680, 2019.

[9] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[10] 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)*, pages 384–400, 2018.

[11] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[12] A. Bendale and T. E. Boult. Towards open set deep networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010.

[14] S. Bhattacharjee, D. Mandal, and S. Biswas. Autoencoder based novelty detection for generalized zero shot learning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3646–3650. IEEE, 2019.

[15] C. M. Bishop. *Pattern recognition and machine learning*. Springer Science+ Business Media, 2006.

[16] 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*, pages 401–408. ACM, 2007.

[17] M. Bucher, S. Herbin, and F. Jurie. Generating visual representations for zero-shot classification. In *International Conference on Computer Vision (ICCV) Workshops: TASK-CV: Transferring and Adapting Source Knowledge in Computer Vision*, 2017.

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

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

[20] L. Chen, H. Zhang, J. Xiao, W. Liu, and S.-F. Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks.

In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[21] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation policies from data. *Conference on Computer Vision and Pattern Recognition*, 2019.

[22] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2005.

[23] H. Daume III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

[24] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

[25] L. N. De Castro. *Fundamentals of natural computing: basic concepts, algorithms, and applications*. Chapman and Hall/CRC, 2006.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[27] L. Deng, D. Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[28] S. Dey, P. Riba, A. Dutta, J. Llados, and Y.-Z. Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[29] P. K. Diederik, M. Welling, et al. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

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

[31] G. Dinu, A. Lazaridou, and M. Baroni. Improving zero-shot learning by mitigating the hubness problem. *International Conference on Learning Representations (ICLR) - Workshop track*, 2015.

[32] 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*, pages 5089–5098, 2019.

[33] M. Elhoseiny and M. Elfeki. Creativity inspired zero-shot learning. *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[34] 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*, pages 2584–2591, 2013.

[35] T. X. Elyor Kodirov and S. Gong. Semantic autoencoder for zero-shot learning. *IEEE CVPR 2017*, July 2017.

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

[37] R. Felix, B. Harwood, M. Sasdelli, and G. Carneiro. Generalised zero-shot learning with domain classification in a joint semantic and visual space. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages –. IEEE, 2019.

[38] R. Felix, B. Harwood, M. Sasdelli, and G. Carneiro. Generalised zero-shot learning with multi-modal embedding spaces. *In Submission*, 2019.

[39] R. Felix, B. V. Kumar, I. Reid, and G. Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *European Conference on Computer Vision*, pages 21–37. Springer, 2018.

[40] R. Felix, M. Sasdelli, I. Reid, and G. Carneiro. Augmentation network for multi-modal and multi-domain generalised zero shot learning. *In Submission*, 2019.

[41] V. Ferrari and A. Zisserman. Learning visual attributes. In *Advances in neural information processing systems*, pages 433–440, 2008.

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

[43] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multimodal latent attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):303–316, 2013.

[44] Y. Fu and L. Sigal. Semi-supervised vocabulary-informed learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5337–5346, 2016.

[45] Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2635–2644, 2015.

[46] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Insights and applications. In *Deep Learning Workshop, ICML*, 2015.

[47] Y. Gal, J. Hron, and A. Kendall. Concrete dropout. In *Advances in Neural Information Processing Systems*, pages 3581–3590, 2017.

[48] J. Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2, 2014.

[49] P. Gemeinboeck and R. Saunders. Creative machine performance: Computational creativity and robotic art. In *ICCC*, pages 215–219, 2013.

[50] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[51] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[52] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520, 2011.

[53] R. C. Gonzalez and P. Wintz. Digital image processing(book). *Reading, Mass., Addison-Wesley Publishing Co., Inc.(Applied Mathematics and Computation*, 13:451, 1977.

[54] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[55] 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*, pages 2672–2680, 2014.

[56] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

[57] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

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

[59] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016.

[60] 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*, pages 801–810, 2019.

[61] K. Huang, A. Hussain, Q.-F. Wang, and R. Zhang. *Deep Learning: Fundamentals, Theory and Applications*, volume 2. Springer, 2019.

[62] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[63] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[64] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings*

of the 24th annual ACM symposium on User interface software and technology, pages 559–568. ACM, 2011.

[65] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *Advances in neural information processing systems*, pages 3464–3472, 2014.

[66] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[67] 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*, pages 3657–3664. IEEE, 2012.

[68] M. Kemmler, E. Rodner, and J. Denzler. One-class classification with gaussian processes. In *Asian Conference on Computer Vision*, pages 489–500. Springer, 2010.

[69] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5, 2006.

[70] K. Khoshelham and S. O. Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454, 2012.

[71] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[72] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

[73] S. Kottur, R. Vedantam, J. M. Moura, and D. Parikh. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4985–4994, 2016.

[74] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[75] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

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

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

[78] A. Lazaridou, N. T. Pham, and M. Baroni. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[79] M. Lazaro-Gredilla, D. Lin, J. S. Guntupalli, and D. George. Beyond imitation: Zero-shot task transfer on robots by learning concepts as cognitive programs. *arXiv preprint arXiv:1812.02788*, 2018.

[80] Y. LeCun, Y. Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[81] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. nature, 521 (7553). *DOI= http://dx. doi. org/10.1038/nature14539*, page 436, 2015.

[82] J. Li, M. Jin, K. Lu, Z. Ding, L. Zhu, and Z. Huang. Leveraging the invariant side of generative zero-shot learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[83] J. Li, X. Lan, Y. Liu, L. Wang, and N. Zheng. Compressing unknown images with product quantizer for efficient zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5463–5472, 2019.

[84] Z. Li, E. Gavves, T. Mensink, and C. G. M. Snoek. Attributes make sense on segmented objects. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars,

editors, *Computer Vision – ECCV 2014*, pages 350–365, Cham, 2014. Springer International Publishing.

[85] T. Lindeberg. Feature detection with automatic scale selection. *International journal of computer vision*, 30(2):79–116, 1998.

[86] J. Liu, G. Zeng, and J. Fan. Fast local self-similarity for describing interest regions. *Pattern Recognition Letters*, 33(9):1224–1235, 2012.

[87] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[88] Y. Long, L. Liu, F. Shen, L. Shao, and X. Li. Zero-shot learning using synthesised unseen visual data with diffusion regularisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[89] D. G. Lowe et al. Object recognition from local scale-invariant features. In *iccv*, volume 99, pages 1150–1157, 1999.

[90] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[91] T. Mensink, E. Gavves, and C. G. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2441–2448, 2014.

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

[93] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2018.

[94] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2018.

[95] M. Minsky and S. A. Papert. *Perceptrons: An introduction to computational geometry*. MIT press, 2017.

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

[97] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[98] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on*, pages 722–729. IEEE, 2008.

[99] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[100] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014.

[101] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org, 2017.

[102] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith. Default probability. *Cognitive Science*, 15(2):251–269, 1991.

[103] S. J. Pan, J. T. Kwok, Q. Yang, et al. Transfer learning via dimensionality reduction. In *AAAI*, volume 8, pages 677–682, 2008.

[104] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[105] A. Paul, N. C. Krishnan, and P. Munjal. Semantically aligned bias reducing zero shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7056–7065, 2019.

[106] 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)*, pages 1532–1543, 2014.

[107] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

[108] R. Qiao, L. Liu, C. Shen, and A. v. d. Hengel. Visually aligned word embeddings for improving zero-shot learning. *British Machine Vision Conference (BMVC)*, 2017.

[109] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 2016.

[110] S. Rahman, S. Khan, and N. Barnes. Polarity loss for zero-shot object detection. *arXiv preprint arXiv:1811.08982*, 2018.

[111] C. E. Rasmussen and C. K. Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

[112] A. J. Ratner, H. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré. Learning to compose domain-specific transformations for data augmentation. In *Advances in neural information processing systems*, pages 3236–3246, 2017.

[113] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

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

[115] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, pages 1060–1069, 2016.

[116] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[117] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR 2011*, pages 1641–1648. IEEE, 2011.

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

[119] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[120] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, volume 11, page 2. Citeseer, 2011.

[121] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[122] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[123] J. Sanchez and F. Perronnin. Image classification employing image vectors compressed using vector quantization, May 20 2014. US Patent 8,731,317.

[124] 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*, pages 2168–2178, 2019.

[125] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2013.

[126] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, July 2013.

[127] W. J. Scheirer, L. P. Jain, and T. E. Boult. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2317–2324, 2014.

[128] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

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

[130] 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*, pages 3598–3607, 2018.

[131] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2014.

[132] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[133] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, pages 935–943, 2013.

[134] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

[135] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

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

[137] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[138] K. D. Tang, M. F. Tappen, R. Sukthankar, and C. H. Lampert. Optimizing one-shot recognition with micro-set learning. In *2010 IEEE Computer Society*

Conference on Computer Vision and Pattern Recognition, pages 3027–3034. IEEE, 2010.

[139] W. Thomason and R. A. Knepper. Recognizing unfamiliar gestures for human-robot interaction through zero-shot learning. In *International Symposium on Experimental Robotics*, pages 841–852. Springer, 2016.

[140] L. Torrey and J. Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2010.

[141] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[142] T. Tran, T. Pham, G. Carneiro, L. Palmer, and I. Reid. A bayesian data augmentation approach for learning deep models. In *Advances in Neural Information Processing Systems*, pages 2794–2803, 2017.

[143] K. Van De Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1582–1596, 2009.

[144] V. K. Verma, G. Arora, A. Mishra, and P. Rai. Generalized zero-shot learning via synthesized examples. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[145] P. Wang, L. Liu, C. Shen, Z. Huang, A. van den Hengel, and H. T. Shen. Multi-attention network for one shot learning. In *2017 IEEE conference on computer vision and pattern recognition, CVPR*, pages 22–25, 2017.

[146] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[147] C. K. Williams and C. E. Rasmussen. Gaussian processes for regression. In *Advances in neural information processing systems*, pages 514–520, 1996.

[148] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

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

[150] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *CoRR*, abs/1707.00600, 2017.

[151] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[152] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.

[153] 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*, pages 9384–9393, 2019.

[154] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

[155] S. K. Yelamarthi, S. K. Reddy, A. Mishra, and A. Mittal. A zero-shot framework for sketch based image retrieval. In *European Conference on Computer Vision*, pages 316–333. Springer, 2018.

[156] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 771–778, 2013.

[157] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *European conference on computer vision*, pages 127–140. Springer, 2010.

[158] H. Zhang and P. Koniusz. Model selection for generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[159] H. Zhang, Y. Long, Y. Guan, and L. Shao. Triple verification network for generalized zero-shot learning. *IEEE Transactions on Image Processing*, 28(1):506–517, 2018.

[160] Y. Zhang, C. McCullough, J. R. Sullins, and C. R. Ross. Human and computer evaluations of face sketches with implications for forensic investigations. In *2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems*, pages 1–7. IEEE, 2008.

[161] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4166–4174, Washington, DC, USA, 2015. IEEE Computer Society.

[162] L. Zheng, Y. Yang, and Q. Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1224–1244, 2017.

[163] Z.-H. Zhou. *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC, 2012.

[164] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[165] 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*, pages 2995–3003, 2019.

[166] 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*, pages 1004–1013, 2018.