Generalised Zero-shot Learning with Multi-modal Embedding Spaces

Rafael Felix*†‡, Michele Sasdelli*†‡, Ben Harwood*§, Gustavo Carneiro*†‡

*Australian Centre of Excellence for Robotic Vision
†Australian Institute for Machine Learning
‡The University of Adelaide
{rafael.felixalves, michele.sasdelli, gustavo.carneiro}@adelaide.edu.au
§ Monash University
ben.harwood@monash.edu

***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.**

I. 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 [1]. 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 [1]. One possible way to address such real-world problems is with the generalised zero-shot learning (GZSL) setup [2] 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 [3] or a list of attributes [4] 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 [5], [6].

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 [2]. More recent approaches rely on a generative model to produce visual samples from their respective semantic samples [7]. 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 [8] or with pairwise compatibility functions [9], 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 [10].

Another major issue affecting GZSL methods is the imbalance in the classification results for the seen and unseen classes [2]. One of the first GZSL methods [6] 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 [6] 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 [5], [11], but they always rely on a single modality classification. The major drawback of the approaches

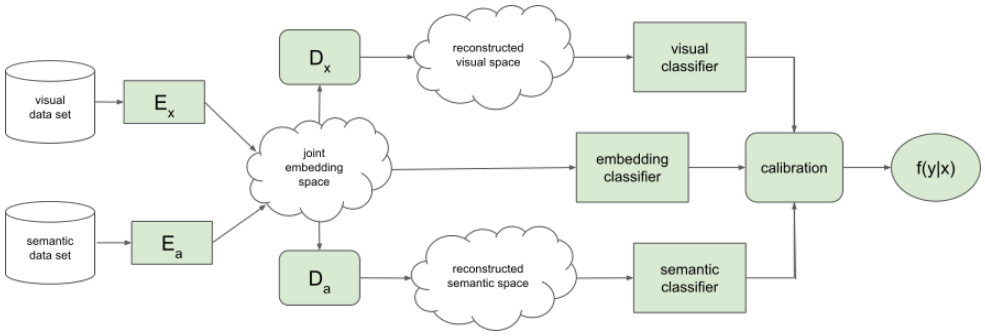

Fig. 1. Our model consists of encoders from visual and semantic spaces to a latent joint embedding space. Samples from the joint space are used to train decoders that reconstruct the original samples from visual and semantic spaces. Samples from the visual, semantic and joint spaces are then used to train and calibrate classifiers for each space. The final multi-domain classifier averages the results of the multi-modal calibrated classifiers.

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.

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. 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 [12] 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.

II. Literature Review

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

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 [4], [13]. 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 [14]. 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 [7], [15] because the testing procedure completely ignores the seen classes [16], [17]. Although limited, ZSL methods can be seen as an expert model for the unseen visual classes [11].

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 [2], [6], [16], which has motivated the development of several GZSL approaches [15]. Previously, studies in GZSL have been based on an ensemble of classifiers that combines transformations between the visual and semantic spaces [9], [18], methods that combine seen and unseen classifications [6], [11], [16], and algorithms that generate synthetic unseen visual samples [7], [8].

The most successful GZSL approaches are based on methods that generate synthetic visual samples for the unseen classes, given their semantic representation [15]. 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) [7], [15] and Variational Autoencoders (VAE) [8], [19]. 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 [20]. The first method to address the bias toward the seen classes was proposed by Socher et al. [6]. 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 [5], [6], [21]. More recently, the approach developed by Atzmon and Chechik [11] 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.

III. METHOD

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.

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 $\mathbf{x} \in \mathcal{X} \subseteq R^X$ 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 [22], and VGG [23]. 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 [2]. The data set for the semantic modality is defined as $\mathcal{R} = \{\mathbf{a}_y\}_{y \in \mathcal{Y}}$, where each $\mathbf{a}_y \in \mathcal{A} \subseteq R^A$ 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* [2]). 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.

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:

$$y^* = \arg\max_{y \in \mathcal{Y}} f(y|\mathbf{x}), \tag{1}$$

with

$$f(y|\mathbf{x}) = \sigma_x(y|\tilde{\mathbf{x}}, \tau_x, \theta_x) + \sigma_a(y|\tilde{\mathbf{a}}, \tau_a, \theta_a) + \sigma_z(y|\tilde{\mathbf{z}}, \tau_z, \theta_z), \tag{2}$$

where $\tilde{\mathbf{x}} \in \mathcal{X}$ represents a generated visual sample, $\tilde{\mathbf{a}} \in \mathcal{A}$ denotes a generated semantic sample, $\tilde{\mathbf{z}} \in \mathcal{Z} \subseteq R^Z$ 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. 1 and Eq. 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. 1 shows other components that are defined below. The visual and semantic encoders are defined by

$$\begin{aligned} \tilde{\mathbf{z}} &\sim p_x^{(E)}(\mathbf{z}|\mathbf{x}, \theta_x^{(E)}), \\ \tilde{\mathbf{z}} &\sim p_a^{(E)}(\mathbf{z}|\mathbf{a}, \theta_a^{(E)}), \end{aligned} \tag{3}$$

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

$$\begin{aligned} \tilde{\mathbf{x}} &\sim p_x^{(D)}(\mathbf{x}|\mathbf{z}, \theta_x^{(D)}), \\ \tilde{\mathbf{a}} &\sim p_a^{(D)}(\mathbf{a}|\mathbf{z}, \theta_a^{(D)}), \end{aligned} \tag{4}$$

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 [7], [8], [15], as described in Sec. II. In this paper, we extend the model proposed by Schonfeld et al. [8]. In particular, the training of the model defined in Eq. 1- Eq. 4 is an end-to-end process that minimises the following loss function:

$$\ell(\mathcal{D}^{tr}, \mathcal{R}) = \gamma_{PD}\ell_{PD} + \ell_{VAE} + \gamma_{CM}\ell_{CM} + \gamma_{DA}\ell_{DA}. \tag{5}$$

The first term in Eq. 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. 2, as follows:

$$\begin{aligned} \ell_{PD} = &-\mathbf{h}_y \log(\sigma_x(y|\tilde{\mathbf{x}}, \tau_x, \theta_x)) \\ &-\mathbf{h}_y \log(\sigma_a(y|\tilde{\mathbf{a}}, \tau_a, \theta_a)) \\ &-\mathbf{h}_y \log(\sigma_z(y|\tilde{\mathbf{z}}, \tau_z, \theta_z)), \end{aligned} \tag{6}$$

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. 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. 4. It is important to notice in Eq. 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 [6], [11]. Instead, we rely entirely on calibrating the classifiers using temperature scaling [12], which, for the case of the softmax classifier, is defined by

$$\sigma_x(y|\mathbf{x}, \tau_x, \theta_x) = \frac{\mathrm{e}^{(\pi_x(y|\mathbf{x}, \theta_x)/\tau_x)}}{\sum_{c=1}^{C} \mathrm{e}^{(\pi_x(c|\mathbf{x}, \theta_x)/\tau_x)}}, \tag{7}$$

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. 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 [12]. After calibrating each classifier, the ensemble consists of summing the three classification results from Eq. 2. The calibration parameters are learned based on the validation set held out from training, as proposed in [2].

The second term in Eq. 5 represents the variational auto-encoder (VAE) error [24], defined by [8]. The sample-wise loss for that second term is denoted by

$$\begin{aligned} \ell_{VAE} = &E_{q(\mathbf{z}|\mathbf{x}, \lambda)}\big[\log(p_x^{(D)}(\mathbf{x}|\mathbf{z}, \theta_x^{(D)}))\big] \\ &+E_{q(\mathbf{z}|\mathbf{a}, \lambda)}\big[\log(p_a^{(D)}(\mathbf{a}|\mathbf{z}, \theta_x^{(D)}))\big] \\ &-\mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}, \lambda_x)||p_\phi(\mathbf{z})\big) \\ &-\mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{a}, \lambda_a)||p_\phi(\mathbf{z})\big), \end{aligned} \tag{8}$$

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. 5 denotes the cross-modality alignment loss that calculates the reconstruction error between the visual and semantic modalities [8]. The sample-wise loss for that third term is defined by:

$$\ell_{CM} = \|\mathbf{x} - \tilde{\mathbf{x}}\| + \|\mathbf{a} - \tilde{\mathbf{a}}\|, \tag{9}$$

where $\tilde{\mathbf{x}}$ is sampled from the decoder $p_x^{(D)}(\mathbf{x}|\tilde{\mathbf{z}}, \theta_x^{(D)})$ in Eq. 4, with $\tilde{\mathbf{z}}$ being sampled from $p_a^{(E)}(\mathbf{z}|\mathbf{a}, \theta_a^{(E)})$ in Eq. 3 and $\mathbf{x}$ and $\mathbf{a}$ belonging to the same class. Similarly in Eq. 9, $\tilde{\mathbf{a}}$ is sampled from the decoder $p_a^{(D)}(\mathbf{a}|\tilde{\mathbf{z}}, \theta_a^{(D)})$ in Eq. 4, with $\tilde{\mathbf{z}}$ being sampled from $p_x^{(E)}(\mathbf{z}|\mathbf{x}, \theta_x^{(E)})$ in Eq. 3 and $\mathbf{x}$ and $\mathbf{a}$ belonging to the same class.

The fourth term in Eq. 5 consists of the distribution-alignment loss of samples belonging to the same class. The loss is defined by [8]:

$$\ell_{DA} = || \mu_x - \mu_a ||_2^2 + || \Sigma_x^{\frac{1}{2}} - \Sigma_a^{\frac{1}{2}} ||_F^2, \tag{10}$$

where $\mu_x \in \mathcal{Z}$ and $\Sigma_x \in \mathcal{Z} \times \mathcal{Z}$ are the mean vector and covariance 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. 10 with the average of the sample-wise losses defined in Equations 6, 8, 9, where the hyper-parameters are estimated with grid search using the validation set.

IV. 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.

Data Sets - We evaluate the proposed method on four publicly available[1] benchmark GZSL data sets: AWA1 [2], [25], AWA2 [2], [25], CUB [26], and SUN [2]. Recent research argues that GZSL approaches that use pre-trained models must take into consideration the overlap between unseen classes and the ImageNet classes [2]. Therefore, we use the GZSL experimental setup described by Xian et al. [2], which prevents that the GZSL unseen classes overlap with the ImageNet classes [2], [27]. These data sets can be either fine or coarse-grained. The CUB data set [26] is fine-grained, where the visual classes are similar to each other, and 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 [2]. Table I 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.

The visual representation for all the benchmark data sets is extracted from the activation of the 2048-dimensional top pooling layer of ResNet-101 [22]. The semantic representation of CUB [2] consists of the 1024-dimensional vector produced by CNN-RNN [3]. 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 [2]. For AWA1, AWA2 and SUN we used the semantic features proposed by Xian et al. [2], where we use the 102-dimensional feature for SUN [2], and the 85-dimensional feature for AWA1 [2] and AWA2 [2].

TABLE I. The benchmarks for GZSL: AWA1 [2], AWA2 [2], CUB [26], and SUN [28]. The number of seen classes, denoted by $|\mathcal{Y}^S|$, split into training and validation classes (train+val), the number of unseen classes $|\mathcal{Y}^U|$, the number of samples available for training $|\mathcal{D}^{Tr}|$ and testing samples that belong to the unseen classes $|\mathcal{D}_U^{Te}|$ and testing samples from the seen classes $|\mathcal{D}_S^{Te}|$ [7], [15].

Name	$\|\mathcal{Y}^S\|$ (train+val)	$\|\mathcal{Y}^U\|$	$\|\mathcal{D}^{Tr}\|$	$\|\mathcal{D}_U^{Te}\| + \|\mathcal{D}_S^{Te}\|$
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

Evaluation Protocol - We evaluate the proposed model with Xian et al.'s [2], [15] 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 [2]. We also present experiments using the area under the seen and unseen curve (AUSUC) [16].

Implementation Details - 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

TABLE II. 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 [29]	39.8	–	30.2	12.8
DAZSL [11]	53.2	–	35.7	23.9
f-CLSWGAN [15]	46.1	–	35.5	22.0
cycle-WGAN [7]*	47.3	–	**41.8**	23.2
CADA-VAE [8]*	52.4	52.2	37.0	23.6
ours	**53.2**	**54.9**	39.3	**24.0**

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

TABLE III. 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 [2]. THE HIGHLIGHTED VALUES REPRESENT THE BEST ONES IN EACH COLUMN WITHIN A CONFIDENCE OF $\pm 1\%$.

	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 [30]	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 [31]	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 [32]	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 [29]	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 [17]	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 [33]	68.7	13.4	22.4	74.7	17.1	27.8	53.0	23.8	32.8	27.4	16.9	20.9
PQZSL [34]	31.7	70.9	43.8	–	–	–	43.2	51.4	46.9	35.1	35.3	35.2
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 [15]	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 [7]	63.5	56.4	59.7	–	–	–	60.3	46.0	52.2	33.1	48.3	39.2
CADA-VAE [8]	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. [9]	–	–	–	41.6	**91.3**	57.2	33.4	**87.5**	48.4	–	–	–
LisGAN [36]	52.6	76.3	62.3	–	–	–	46.5	57.9	51.6	42.9	37.8	40.2
GMN [37]	61.1	71.3	**65.8**	–	–	–	56.1	54.3	55.2	53.2	33.0	40.7
GDAN [38]	–	–	–	32.1	67.5	43.5	39.3	66.7	49.5	38.1	**89.9**	**53.4**
Combining classifiers												
CMT [6]	**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 [11]	76.9	54.7	63.9	–	–	–	56.9	47.6	51.8	37.2	45.6	41.4
SABR [39]	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

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 [40]. The hyper-parameters γ_{PD}, γ_{CM} and γ_{DA} are estimated with cross-validation. The multi-modal classifiers in Eq. 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. 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 [40], 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. We propose the optimisation of the loss function in Eq. 5, by alternating the training of each component. Furthermore, we calibrate the predictions with temperature scaling for GZSL models, as described in Eq. 7, where this optimisation process depends on the validation set provided by Xian et al [2], and each classifier has a singular temperature scale.

V. RESULTS AND DISCUSSIONS

AUSUC – In Table II, we show the area under the curve of seen and unseen accuracy (AUSUC) results [16]. We evaluate the proposed model in terms of AUSUC for the benchmark data sets AWA1, AWA2, CUB, and SUN; and compare to several literature methods [7], [8], [11], [15], [29]. 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 [8] on all four data sets. For CUB, our AUSUC result is the second best among the methods in Table II. The AUSUC is achieved by varying a balancing factor between the seen and the unseen contributions for the harmonic-mean [16]. 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, the AUSUC shows in Fig. 2 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.

GZSL – In Table III, 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 in the table: semantic approach, generative approach and models that combine domain classifiers. The semantic group focuses on learning a transformation from visual to semantic representation, then the classification is based on nearest neighbour classification in the semantic space [17], [29]–[34]. The generative group of GZSL approaches rely on generative models to produce synthetic visual features for the unseen classes [7]–[9], [15], [35]–[38]. We also compare the proposed model to approaches that combine the seen and unseen domain classifiers [6], [11], [39]. Table III 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 [37] and our approach are the top performing methods, with similar H-mean results. For AWA2, our method is the best, with CADA-VAE [8] being slightly worse, but comparable. For CUB, we notice that SABR [39], GMN [37] and our approach are the top performing methods, with comparable H-mean results. For SUN, GDAN [38] 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 [38] and SABR [39]). It is also important to notice that our approach produces better H-mean results than CADA-VAE [8], 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 [38] that is more than 10% better than any other approach in the field.

Ablation studies – Table IV shows an ablation study of the

TABLE IV. 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. 7), 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**

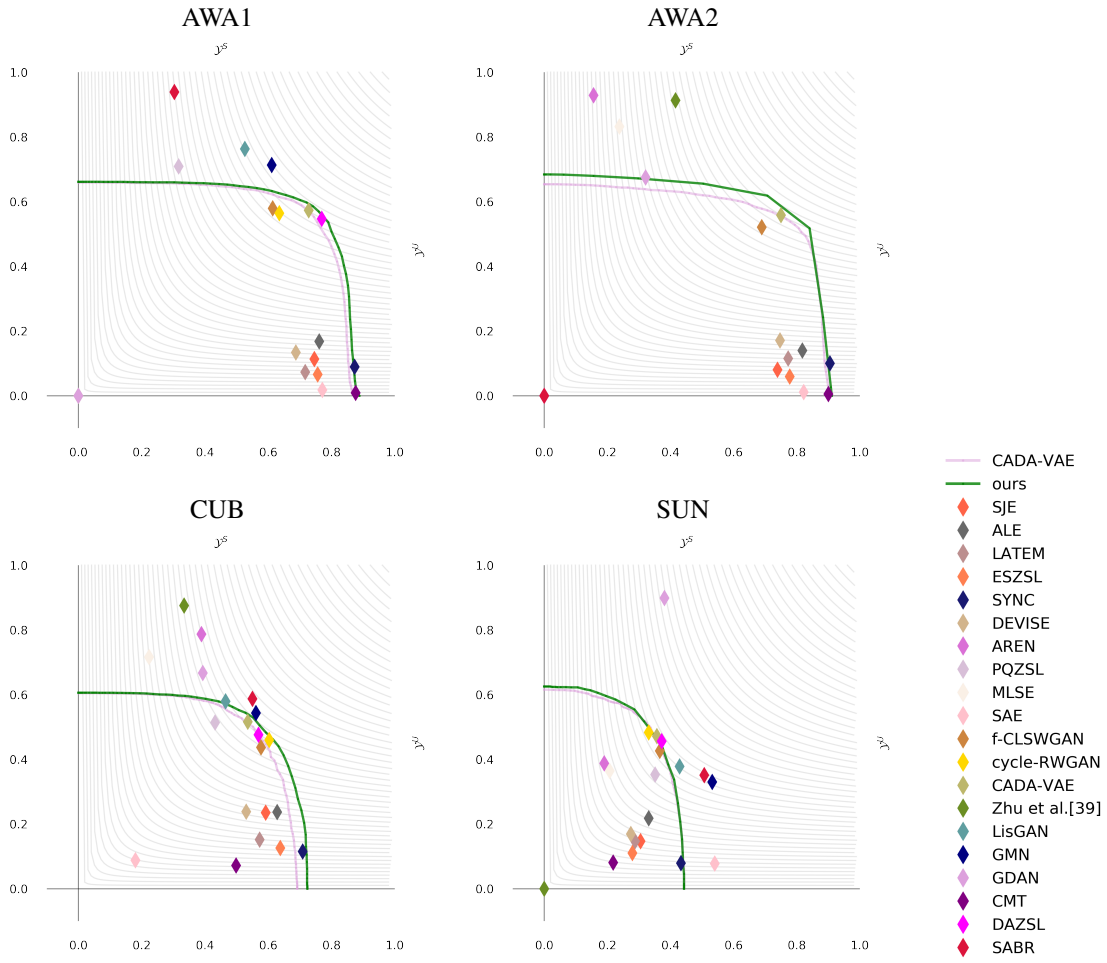

Fig. 2. The area for seen and unseen accuracy curve for the proposed method (green) and CADA-VAE [8] (pink), which is the closest model to ours (please see text and Table III for details about the methods). Note that these graphs are used to compute the AUSUC in Table II

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. [8]); the reconstructed visual space $classifier(\tilde{\mathbf{x}})$; and the reconstructed semantic space $classifier(\tilde{\mathbf{a}})$. We also show the results with our multi-modal approach trained without temperature calibration, denoted by 'ours ($\tau = 1$)'. The last row in Table IV shows the result of our proposed multi-modal approach with calibration. This study 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 IV 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.

The proposed method proposes a novel appoach for solving GZSL, which demonstrates by Tables I, II and III outstanding performance. In particularly, the proposed method improves over the baseline [8]. In fact, as shown on Tables II and III, the combination of multiple classifiers improves the accuracy of the baseline [8] for all datasets. Furthermore, our method is competitive in all datasets, achieving competitive AUSUC results in AWA1, AWA2 and SUN. Regarding H-mean, our approach is competitive in all datasets, particularly in AWA1

and AWA2.

VI. Conclusions and Future Work

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. V, 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.

VII. Acknowledge

This work was partially supported by Australian Research Council grants (FT190100525 and CE140100016).

References

[1] Yu Liu, Li Liu, Yanming Guo, and Michael S Lew. Learning visual and textual representations for multimodal matching and classification. *Pattern Recognition*, 84:51–67, 2018.

[2] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *CoRR*, abs/1707.00600, 2017.

[3] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt 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.

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

[5] Rafael Felix, Ben Harwood, Michele Sasdelli, and Gustavo 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.

[6] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, pages 935–943, 2013.

[7] Rafael Felix, BG Vijay Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *European Conference on Computer Vision*, pages 21–37. Springer, 2018.

[8] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep 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.

[9] Pengkai Zhu, Hanxiao Wang, and Venkatesh 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.

[10] Titir Dutta and Soma Biswas. Cross-modal retrieval in challenging scenarios using attributes. *Pattern Recognition Letters*, 2019.

[11] Yuval Atzmon and Gal 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.

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian 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.

[13] Teng Long, Xing Xu, Fumin Shen, Li Liu, Ning Xie, and Yang Yang. Zero-shot learning via discriminative representation extraction. *Pattern Recognition Letters*, 109:27–34, 2018.

[14] Ziad Al-Halah, Lukas Rybok, and Rainer Stiefelhagen. Transfer metric learning for action similarity using high-level semantics. *Pattern Recognition Letters*, 72:82–90, 2016.

[15] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei 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.

[17] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5327–5336, 2016.

[18] Haofeng Zhang, Yang Long, Yu Guan, and Ling Shao. Triple verification network for generalized zero-shot learning. *IEEE Transactions on Image Processing*, 28(1):506–517, 2018.

[19] V Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] Gaurav Bhatt, Piyush Jha, and Balasubramanian Raman. Representation learning using step-based deep multi-modal autoencoders. *Pattern Recognition*, 2019.

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

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2013.

[25] 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.

[26] 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.

[27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li 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.

[28] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio 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.

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

[30] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt 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.

[31] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2016.

[32] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016.

[33] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[34] Jin Li, Xuguang Lan, Yang Liu, Le Wang, and Nanning 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.

[35] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3174–3183, 2017.

[36] Jingjing Li, Mengmeng Jin, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. *arXiv preprint arXiv:1904.04092*, 2019.

[37] Mert Bulent Sariyildiz and Ramazan Gokberk 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.

[38] He Huang, Changhu Wang, Philip S Yu, and Chang-Dong 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.

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

[40] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedingsof 3rd International Conference on Learning Representations*, 2014.