PC-GZSL: Prior Correction for Generalized Zero Shot Learning

S Divakar Bhat[*] Amit More[*] Mudit Soni Bhuvan Aggarwal
Honda R&D Co., Ltd.
Tokyo, Japan

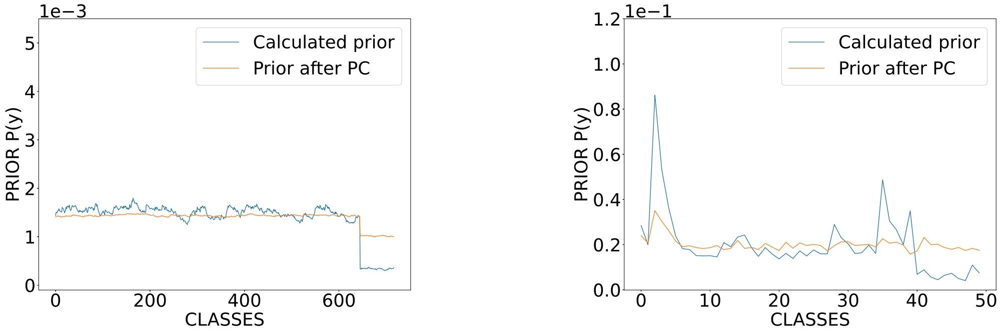

Figure 1. The bias in the trained model [33] for seen and unseen classes is calculated using proposed approach for SUN [39] and AWA2 [51] datasets. We can notice higher average probability score for seen classes. The proposed approach, Prior Correction, not only removes the model bias for seen classes, it does so between individual seen classes and unseen classes as well, evident by the smoothness of the curve.

Abstract

Generalized Zero Shot Learning (GZSL) aims at achieving a good accuracy on both seen and unseen classes by relying on the information acquired from auxiliary attributes. Existing approaches have devised many frameworks to make this knowledge transfer more efficient and informative. Despite their effectiveness on boosting the overall performance, there has always been a strong bias in the model towards the seen classes which makes GZSL problem more challenging. The effect of this bias on the model performance has never been properly explored. We observe that GZSL algorithms in literature have an evident bias towards the seen classes. Further we also show that techniques like calibrated stacking [7] fall short of resolving this conflict between the seen and unseen classes effectively. In this work we analyze and develop a logit-adjustment approach in GZSL setting and propose a simple, yet effective method to remove the bias from trained models in a post-hoc manner. Moreover, as a consequence of the post-hoc nature of the proposed approach, there is no additional training cost. We exhaustively compare the proposed method on both embedding-based and generative-based GZSL frameworks surpassing the SOTA results by 3.1%, 4.6% and 3.1% on CUB, SUN and AwA2 datasets. We also present theoretical analysis showing effectiveness of proposed approach.

1. Introduction

In real world datasets, some classes maybe very rare to encounter and some other categories may just emerge in everyday life with no labelled samples available. Developing DNN models which can extend the knowledge gained during training to a completely novel set of classes while testing has been one of the most challenging direction of research. Zero-shot learning, as it has come to be known in recent years, primarily aims to perform well on a test set which exclusively consists of novel unseen classes. Whereas in Generalized Zero Shot Learning (GZSL) [38] performance on both seen and unseen classes are evaluated.

GZSL primarily have two broad approaches, first is the embedding-based method [1, 2, 50, 57] which rely on the semantic attributes and image data to learn a global alignment between the visual representation and the corresponding semantic embedding. Second is the generative approach [43, 52] where the insufficiency in the feature expression of unseen classes is compensated via a two stage approach consisting of the generator learning and classifier learning. For any GZSL framework, the training data consist only of seen classes while the test data is made up of both seen and unseen classes. Auxiliary attributes are used

[*] Equal contribution

Data	AWA2				CUB			
Method	$\mathcal{A}_{U\to U}$	$\mathcal{A}_{S\to S}$	$\mathcal{A}_{U\to T}$	$\mathcal{A}_{S\to T}$	$\mathcal{A}_{U\to U}$	$\mathcal{A}_{S\to S}$	$\mathcal{A}_{U\to T}$	$\mathcal{A}_{S\to T}$
PSVMA [33]	77.57	95.55	34.90 (-42.8)	95.28 (-0.3)	77.65	88.60	36.05 (-41.6)	87.46 (-1.1)
f-CLSWGAN [52]	68.20	78.63	52.78 (-15.4)	68.98 (-9.7)	75.72	68.36	61.09 (-14.6)	59.3 (-9.1)

Table 1. Accuracy in % when the seen and unseen prior is relaxed during evaluation. $\mathcal{A}_{U\to U}$ is the unseen to unseen (ZSL) evaluation and $\mathcal{A}_{S\to S}$ indicates seen to seen evaluation. $\mathcal{A}_{U\to T}$ and $\mathcal{A}_{S\to T}$ denotes the unseen and seen class accuracy in the GZSL setting. T denotes $\{U \cup S\}$. One may notice a drop in accuracy, especially for unseen classes, indicating a bias toward the seen classes.

during training along with the seen class data to help reduce the information gap between the seen classes and the novel unseen classes encountered at testing. Although the attributes serve as a good means for knowledge transfer between the seen and unseen classes, they cannot compensate for the skewed model prior, a result of over representation of seen classes.

The poor accuracy on the unseen classes in any GZSL framework can be attributed to the inherent imbalance in the distribution of training data where only seen class samples are available for training. Further, the over-fitting nature of DNNs also has a strong influence in creating the bias towards the seen classes where the classifier tends to classify unseen classes into seen categories. To mitigate the effect of this bias on model performance, the generative methods use a large sample size while sampling from the generated unseen distribution to enhance the influence of synthetic unseen samples on the trained model. However, these synthetic samples have been found to have significant deviation from the distribution of the real unseen samples [8]. Even though many of these methods have contributed significantly in improving the performance of GZSL algorithms, some residual bias still remains. In the present work we focus on addressing this issue and propose to remove the seen-unseen class bias from a trained model.

One of the early works which explored the seen-unseen class bias in GZSL proposed a very simple method called calibrated stacking [7]. It is a heuristic approach which suppresses the scores of all the seen classes in comparison to the unseen classes. However, it completely ignores the varying degree of biases within the individual seen classes while making this correction. Further [8] proposed a logit adjustment based approach for generative GZSL frameworks. Inspired from the logit-adjustment approaches [32, 35] in long tailed learning, they modify their loss function by using a prior computed from class frequencies. Prior computation using class frequencies inherently comes with a strong assumption that the model accurately learns the posterior distribution. However, this depends on a lot of factors like network complexity, amount of training data, etc [28, 42].

In this work we make two observations. First, the learned prior (also referred to as bias) on class labels would differ significantly from a fixed priors used in literature such as calibrated stacking [7] or sample frequencies for different classes [8] and hence creates a strong bias toward the seen classes. Second, the DNN algorithms implicitly assume that the test set and train set have similar distributions for class labels. Outside this assumption, performance of DNN models starts to deteriorate, long tailed learning [18, 21, 26] is a good example for this phenomenon. The GZSL problem can be viewed as an extreme case of class imbalance with zero samples from test classes. And thus the generalized zero-shot learning scenario also suffers from the issue of distribution mismatch.

In Figure 1 we show the effective prior of a trained model, computed as an average model response over test dataset. The figure evidently implies a strong bias towards the seen classes. We also show the corrected prior using our prior correction (PC) method clearly indicating the reduction in this bias between the seen and unseen classes. Note that the corrected prior is smoother in nature indicating a reduced inter class bias. Further, in Table 1 we report the results on embedding based and generative GZSL framework by restricting the target class set. We can clearly observe that for the unseen classes the performance of ZSL ($\mathcal{A}_{U\to U}$) task degrades significantly in GZSL ($\mathcal{A}_{U\to T}$) setting, while the degradation for seen classes is relatively lower. This dip in the performance on unseen classes is due to the bias towards the seen classes where unseen class samples are incorrectly classified into seen class categories.

These observations demands the need of a more statistically grounded non-heuristic mechanism for bias removal. We propose a simple but effective prior correction method to remove the bias in GZSL algorithms. First, we present an approach to calculate the bias, effective prior, of a trained model from the data. We then show how to adjust the predictions by the model to remove bias and show theoretical optimality of proposed approach. We also present analysis from the perspective of improving the harmonic mean between seen and unseen class accuracies and provide further modifications to model predictions leading to improved performance. A summary of our contributions is as follows:

- We show that existing GZSL approaches have strong bias for seen classes and present an approach to capture this bias in the trained model.

- We present theoretically motivated method for adjusting model predictions to remove the model bias for seen classes.
- Proposed approach is exhaustively validated on three benchmark datasets and outperforms the SOTA for both embedding and generative based approaches.
- We show that proposed method can be used as a plugin approach with existing frameworks and can improve the performance further without any training.

2. Related Work

2.1. Generalized Zero Shot Learning

Existing algorithms in the literature can be broadly classified into embedding-based and generation-based methods. Embedding based methods [1,2,50,57] is a family of GZSL methods which focus on projecting the semantic and visual domain features into a common space and align the information. These early works faced difficulties in capturing a sufficiently discriminative global visual information. More recent methods started to employ part based learning strategies to leverage the distinct information from the local visual regions. [31, 62] obtain coordinate positions using attention mechanisms to zoom-in and crop on distinctive local regions. Graph networks [24, 55] or attention guidance [34, 36, 54] are also employed for highlighting the significant visual features. Semantic guided methods [10, 11, 23, 25, 47] introduced the sharing attribute prototypes for localizing the attribute-related regions.

Generative methods rely on the synthesis of unseen samples using either generative adversarial networks [19, 20, 29, 52] or variational auto-encoders [27, 53]. Therefore the quality of visual semantic alignment of the synthesized unseen domain features play a crucial role in determining the model performance. [37] designed a recurrent structure and enforced semantic alignment in every stage while [12] relied on fine-tuning the visual features. Further [20] proposed to classify using projected visual features in the latent space. Although generative methods handled data imbalance using synthetic samples and showed improved performance, some residual bias still remains.

In [8] authors presented logit-adjustment approach for GZSL problem similar to us. They derived a lower bound on the harmonic mean and proposed a class frequency based logit-adjustment loss and showed improvements. Different from [8] our framework is motivated from a probabilistic perspective where we adjust the logits in a post-hoc manner using learned bias of the model. Further we also use the harmonic mean bound and propose additional correction factor. Our approach is orthogonal to [8] and many other methods in the literature in a sense that due to its post-hoc nature, it can be combined with other methods as a plug in, and further improve the overall performance.

2.2. Data-Imbalanced Learning

As GZSL is a special case of a data imbalanced learning, we review the relevant literature here. Supervised contrastive learning has been proposed to tackle class imbalance [17, 30, 61] where data scarcity is mitigated by data augmentation where additional contrastive loss is applied on augmented samples. Other frameworks, such as ensemble of multiple experts and gradient manipulation [3,44,48], have been shown to be effective to improve model performance.

Related to our work, logit-adjustment is a popular technique used for mitigating effects of data imbalance. In [4] authors posed the class-imbalance problem as a distribution misalignment problem and proposed to adjust the decision boundaries using class dependent margin by logit-adjustment. Further work in bridging the gap between the learned posterior and the target distribution was carried out by [22,35,41] where data imbalance is modelled using class frequencies and the bias is corrected during training or post training. [49, 58] proposed to learn a logit calibration layer to reduce the effect of bias by rescaling the output logits.

In principle, the logit-adjustment framework can be combined with generative approach for solving GZSL problem as already shown in [8], however their extension to embedding based approaches is non-trivial. In contrast, present approach is flexible and can be combined with both embedding based and generative methods as we have shown.

2.3. Vision Language Models

Recently multi-modal vision-language models (VLM) trained with internet scale data, such as CLIP [40], have shown impressive performance on many datasets in zero-shot setting. Prompt learning has been proposed to improve few-shot accuracies of such VLMs in [59, 60] however it has resulted in decrease in zero-shot performance of the model. In [45, 45] authors proposed to augment class descriptor prompts with visual discriptors using Large Language Models showing improved zero-shot performance.

3. Problem Formulation

Consider a classification problem where data samples are taken from distribution $P(x)$ and belong to different unique classes from distribution $P(y)$, where x and y represent the input data and ground truth class, respectively. Let $P_s(y)$ represent the distribution over a set $\mathcal{Y}_s$ of seen classes whose samples are available during training. Further, let $P_u(y)$ represent the distribution over the set of unseen classes $\mathcal{Y}_u$. In general, input data x is represented by features of a trained DNN model and the goal is to train

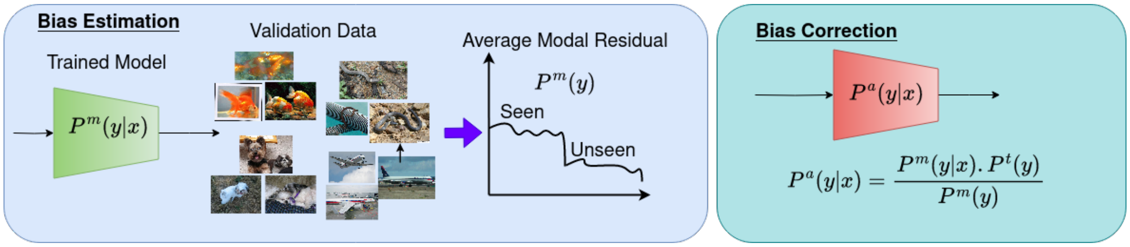

Figure 2. We summarize proposed approach in the figure. Effective prior, model bias, of a trained model is computed first using validation data. The adjusted, debiased, model is then estimated as shown.

a classification model to map these samples to the target classes. In GZSL problem, samples from $\mathcal{Y}_s$ are available during training and the goal is to model *a posteriori* distribution $P(y|x)$ using available trained data in such a way that it can be extended to the unseen classes $\mathcal{Y}_u$ as well.

Let us denote the posterior distribution on the test data by $P^t(y|x)$. It is evident that the trained model can only approximate the target test distribution as no samples from the unseen classes are available during training. If we denote the distribution over samples from the test set classes $\mathcal{Y}_t = \{\mathcal{Y}_s \cup \mathcal{Y}_u\}$ as $P^t(y)$ then we have $P^t(y) \neq P(y)$. In particular, these distributions differ for both seen and unseen class. For seen classes, the distribution $P_s(y)$ is only approximated by given models such as DNN [42]. For the unseen classes, the distribution $P_u(y)$ is not available during training and it is learned by using auxiliary information and hence may deviate from the actual distribution.

3.1. Logit Adjustment for bias removal

Using Baye's theorem we can write the relation between different distributions for test dataset as below,

$$P^t(y|x) = \frac{P^t(x|y)P^t(y)}{P^t(x)} \tag{1}$$

If we denote by $P^m(y|x)$ the posterior distribution of a trained model, then we can write,

$$P^m(y|x)P^t(x) = P^m(x|y)P^m(y) \tag{2}$$

We note that, $P^t(x)$ is the target test distribution on both seen and unseen classes. The right hand side of the equation shows the class priors $P^m(y)$ and the class conditional data distribution $P^m(x|y)$ of a learned model on the test dataset. Thus we may write,

$$P^m(y|x) = \frac{P^m(x|y)P^m(y)}{P^t(x)} \tag{3}$$

We assume that the class conditional data distributions $P^t(x|y)$ and $P^m(x|y)$ are the same. Thus taking the ratio of above equations we have

$$P^t(y|x) = P^m(y|x)\frac{P^t(y)}{P^m(y)} \tag{4}$$

The above equation defines the relationship between posterior distributions $P^t(y|x)$ and $P^m(y|x)$ over the test set $\mathcal{Y}_t$. The Eq. 4 represents classical logit-adjustment framework popular in long-tailed literature where model output is adjusted by the ratios of class priors for train and test datasets. In case of GZSL problem, modelling $P^m(y)$ by training data distribution $P(y)$ is challenging as no samples from $\mathcal{Y}_u$ are available. Thus we propose to estimate it from the validation samples assuming validation data and test data both follow the same distribution.

In general, the ideal posterior distribution and the prior on the test data satisfy following property,

$$\int P^t(y|x)P^t(x)dx = P^t(y) \tag{5}$$

Thus we propose to estimate the prior $P^m(y)$ such that the adjusted distribution satisfies Eq. 5.

Theorem 1.

Let $P^a(y|x)$ represent the adjusted probabilities $P^m(y|x)$ of the trained model as below

$$P^a(y|x) = P^m(y|x)\frac{P^t(y)}{P^m(y)} \tag{6}$$

then the marginal distribution $P^a(y)$ will satisfy Eq. 5 if

$$P^m(y) = \int P^m(y|x)P^t(x)dx \tag{7}$$

Proof. It is easy to see that,

$$\begin{aligned}
P^a(y) &= \int P^a(y|x)P^t(x)dx && (8)\\
&= \int P^m(y|x)\frac{P^t(y)}{P^m(y)}P^t(x)dx && (9)\\
&= \frac{P^t(y)}{P^m(y)}\int P^m(y|x)P^t(x)dx && (10)\\
&= \frac{P^t(y)}{P^m(y)}P^m(y) && (11)\\
&= P^t(y) && (12)
\end{aligned}$$

The above theorem show that the prior of a trained model can be estimated using test data as shown in Eq. 7. The resultant adjusted model has the same prior as the test dataset and removes the bias between seen and unseen classes.

3.2. Optimizing harmonic mean

The GZSL problem is focused on improving accuracy on novel unseen classes while maintaining a good accuracy on the seen classes. The harmonic mean between seen and unseen class accuracies represents such a criteria. It is maximum only when both seen and unseen class accuracies are high and well balanced. However, typical training framework focuses on improving the arithmatic mean of the accuracy over training samples. This does not necessarily result in optimal performance in terms of harmonic mean accuracy. In this section we analyse the relation between model performance and the harmonic mean. We use the lower bound on harmonic mean shown in [8] and derive additional posterior adjustment required. Following [8, 16] the model accuracy for class y can be written in terms of posterior probabilities as

$$A(y) = E_x\big[P^t(y|x)P^m(y|x)\big] \tag{13}$$

where, $E_x[.]$ represents the expectation over $P^t(x)$. Typically, individual classes may have different number of samples, thus class balanced accuracy is used to avoid most frequent classes from dominating the evaluation metric. Thus class balanced accuracies are used as shown below

$$A(y) = E_x\big[\frac{P^t(y|x)P^m(y|x)}{P^t(y)}\big] \tag{14}$$

The average accuracy on seen classes is then given by,

$$As = \frac{1}{|\mathcal{Y}_s|}\sum_{y\in\mathcal{Y}_s} E_x\big[\frac{P^t(y|x)P^m(y|x)}{P^t(y)}\big] \tag{15}$$

Using Jensen-Shannon inequality we have,

$$\frac{1}{As} \leq \frac{1}{|\mathcal{Y}_s|}\sum_{y\in\mathcal{Y}_s} E_x\big[\frac{P^t(y)}{P^t(y|x)P^m(y|x)}\big] \tag{16}$$

The accuracy A_u on the unseen classes can also be written in a similar way. The harmonic mean is defined as

$$\frac{2}{A_h} = \frac{1}{A_s} + \frac{1}{A_u} \tag{17}$$

Thus we can define an upper bound on the harmonic mean using the bounds on seen and unseen class accuracies.

$$\begin{aligned}\frac{2}{A_h} \leq\ & \frac{1}{|\mathcal{Y}_s|}\sum_{y\in\mathcal{Y}_s} E_x\big[\frac{P^t(y)}{P^t(y|x)P^m(y|x)}\big]\\ &+ \frac{1}{|\mathcal{Y}_u|}\sum_{y\in\mathcal{Y}_u} E_x\big[\frac{P^t(y)}{P^t(y|x)P^m(y|x)}\big] \qquad (18)\\ =\ & \frac{1}{|\mathcal{Y}_t|}\sum_{y\in\mathcal{Y}_t} \frac{|\mathcal{Y}_t|}{\mathbb{1}_{\mathcal{Y}s}(y)|\mathcal{Y}_s| + \mathbb{1}_{\mathcal{Y}u}(y)|\mathcal{Y}_u|}\\ & E_x\big[\frac{P^t(y)}{P^t(y|x)P^m(y|x)}\big] \qquad (19)\end{aligned}$$

where $\mathcal{Y}_t = \mathcal{Y}_s \cup \mathcal{Y}_u$ as defined before and we have,

$$\mathbb{1}_{\mathcal{Y}s}(y) = \begin{cases} 1 & \text{if } y \in \mathcal{Y}_s \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

and $\mathbb{1}_{\mathcal{Y}u}(y)$ is defined accordingly. Thus we may write,

$$\frac{2}{A_h} \leq \frac{1}{|\mathcal{Y}_t|}\sum_{y\in\mathcal{Y}_t} E_x\big[\frac{P^t(y)}{P(\mathcal{Y})P^t(y|x)P^m(y|x)}\big] \tag{21}$$

where, $P(\mathcal{Y})$ is defined as $\frac{|\mathcal{Y}_s|}{|\mathcal{Y}_t|}$ for seen classes and $\frac{|\mathcal{Y}_u|}{|\mathcal{Y}_t|}$ for unseen classes representing the empirical seen and unseen class probabilities, $P_s(y)$ and $P_u(y)$, when y is a member of the respective set. Thus harmonic mean can be optimized by maximizing the product $P(\mathcal{Y})P^t(y|x)P^m(y|x)$ in Eq. 21. One may note that the term $P(\mathcal{Y})P^t(y|x)$ in the denominator represents the adjusted ground truth required to maximize the harmonic mean. Thus the optimal adjustment to the predicted probabilities to maximize the denominator term and hence maximize the harmonic mean, can be given by the adjusted model posterior as $P(\mathcal{Y})P^m(y|x)$.

One may interpret these findings in terms of cost-sensitive learning where seen vs unseen class prior is taken into account during evaluation. Nonetheless, we validate our findings with extensive experiments and show that post-hoc adjustment using empirical seen vs unseen class priors indeed boosts model performance.

It should be noted here that our interpretation of the bound on the harmonic mean in Eq. 21 is different from [8]. We only consider $P(\mathcal{Y})$ for the adjustment as it is enough to lower the bound. However, in [8] authors use the term $P(\mathcal{Y})/P^t(y)$ using class frequencies along with other hyper-parameters for the adjustment. Nonetheless, our experiments show that our formulation is effective and leads to better results.

3.3. Overall Prior Correction (PC) Adjustment

In this section we combine both proposed logit-adjustments, one for bias removal and other for improv-

Method	Arch.	CUB			SUN			AwA2		
		$\mathcal{U}$	$\mathcal{S}$	$\mathcal{H}$	$\mathcal{U}$	$\mathcal{S}$	$\mathcal{H}$	$\mathcal{U}$	$\mathcal{S}$	$\mathcal{H}$
TF-VAEGAN* [37] ECCV'20	R101	54.75	62.68	58.45	39.72	35.70	37.60	58.97	74.73	65.92
GCM-CF [56] CVPR'21	R101	61.00	59.70	60.30	47.90	37.80	42.20	60.40	75.10	67.00
HSVA [13] NeurIPS'21	R101	52.70	58.30	55.30	48.60	39.00	43.30	56.70	**79.80**	66.30
SDGZSL [15] ICCV'21	R101	59.90	66.40	63.00	-	-	-	64.60	73.60	68.80
FREE+ESZSL [6] ICLR22	R101	51.60	60.40	55.70	48.20	36.50	41.50	51.30	78.00	61.80
CDL+OSCO [5] TPAMI23	R101	29.00	**69.00**	40.60	32.00	**65.00**	42.90	48.00	71.00	57.10
CE-GZSL* [20] CVPR'21	R101	67.36	67.51	67.43	**51.53**	35.04	41.71	63.99	75.11	69.11
CE-GZSL + PC (Our)	R101	**67.73**	67.91	**67.82**	50.14	38.14	**43.32**	**67.32**	72.53	**69.83**
Transzero* [9] AAAI'22	R101	69.18	68.31	68.74	52.29	33.35	40.66	61.13	82.23	70.13
MSDN* [10] CVPR'22	R101	68.58	67.34	67.95	51.87	33.95	41.04	61.94	74.37	67.59
DUET [14] AAAI'23	ViT-Base	62.9	72.8	67.5	45.7	45.8	45.8	63.7	84.7	72.7
VADS [23] CVPR'24	ViT-Base	74.1	74.6	74.3	64.6	49.0	55.7	**75.4**	83.6	79.3
ZSLViT [11] CVPR'24	ViT-Base	69.4	78.2	73.6	45.9	48.4	47.3	66.1	84.6	74.2
PSVMA [33] CVPR'23	ViT-Base	70.1	77.8	73.8	61.7	45.3	52.3	73.6	77.3	75.4
PSVMA*	ViT-Base	73.1	74.9	74.0	62.4	44.9	52.2	69.9	84.3	76.4
PSVMA + PC (Our)	ViT-Base	**75.68**	**78.63**	**77.13**	**65.9**	49.96	**56.83**	72.69	87.61	**79.46**

Table 2. Evaluation on other GZSL frameworks. We highlight the **best results** in bolt and underline the second best results. One may note that proposed logit-adjustment achieves best harmonic mean accuracy amongst all methods for both generative and embeding based methods. (* indicates reproduced results)

ZSLA	PC	AwA2			CUB		
		$\mathcal{U}$	$\mathcal{S}$	$\mathcal{H}$	$\mathcal{U}$	$\mathcal{S}$	$\mathcal{H}$
X	X	57.3	71.9	63.8	63.1	60.9	62.0
X	✓	60.4	70.3	65.0	64.7	60.4	62.5
✓	X	59.9	73.1	65.8	64.6	60.4	62.4
✓	✓	60.1	74.3	**66.5**	66.3	59.7	**62.9**

Table 3. We compare proposed approach with ZSLA [8]. The table show our reproduced results using vanilla softmax classifier with ZSLA framework. Baseline results are obtained by removing logit adjustment during training from ZSLA framework.

ing the harmonic mean. If we represent with Z_y the unnormalized probability scores for class y, then the overall adjustment is written as

$$Z_y^a = Z_y + \alpha \log \frac{P^t(y)}{P^m(y)} + \beta \log P(\mathcal{Y}) \tag{22}$$

where, we have included hyper-parameters α and β to control the amount of adjustment. We compute the model bias $P^m(y)$ by empirical estimate of the Eq. 7 as

$$P^m(y) = \frac{1}{N} \sum_{\forall x} P^m(y|x) \tag{23}$$

i.e. the bias is computed by averaging the model response over validation dataset with N samples. The term $P^t(y)$ represents the average number of samples for individual classes and $P(\mathcal{Y})$ represents seen and unseen class probabilities estimated using sample numbers.

We note here that, proposed approach does not involve any model training and uses a trained model and validation data to estimate and tune the adjustment terms. We only tune two scalar hyper-parameters, α and β. We summarize the overall approach in Figure. 2 and give algorithmic flow in Supplementary Material.

4. Experiments

4.1. Experimental setup

We use three benchmark datasets to evaluate the proposed method, Caltech-UCSD Birds-200-2011 (CUB) [46], SUN Attribute [39] and Animals with Attributes2 (AwA2) [51]. We follow the standard split proposed in [51]. We denote the average of per class top-1 accuracy of the seen and unseen splits as $\mathcal{S}$ and $\mathcal{U}$ respectively. Further following [51], harmonic mean denoted using $\mathcal{H}$ and arithmetic mean A, is used to evaluate the overall performance. There have been many approaches from both the generative and embedding based GZSL in literature used to improve the performance of GZSL task. We have compared the proposed method with many recent methods from both these families. Embedding based approaches are represented by many strong baselines with high performance such as VADS [23], ZSLViT [11], PSVMA [33],

Method	CUB			SUN			AwA2		
	$\mathcal{U}$	$\mathcal{S}$	$\mathcal{H}$	$\mathcal{U}$	$\mathcal{S}$	$\mathcal{H}$	$\mathcal{U}$	$\mathcal{S}$	$\mathcal{H}$
TF-VAEGAN + CS	54.68	63.00	58.55(+0.10)	39.72	35.78	37.65(+0.05)	63.18	71.36	67.02(+1.10)
TF-VAEGAN + PC	58.30	64.30	**61.15**(+2.70)	41.74	36.74	**39.08**(+1.48)	64.03	71.34	**67.49**(+1.57)
CE-GZSL + CS	67.61	67.22	67.41(-0.02)	51.60	34.92	41.65(-0.06)	65.03	74.14	69.29(+0.18)
CE-GZSL + PC	67.73	67.91	**67.82**(+0.39)	50.14	38.14	**43.32**(+1.61)	67.32	72.53	**69.83**(+0.72)
TransZero + CS	69.51	68.03	68.77(+0.03)	52.71	33.18	40.72(+0.06)	61.30	81.96	70.14(+0.01)
TransZero + PC	70.06	71.90	**70.97**(+2.23)	52.29	35.35	**42.18**(+1.52)	63.14	81.93	**71.32**(+1.19)
PSVMA + CS	73.10	74.90	74.00(+22.94)	62.40	44.90	52.20(+16.45)	69.90	84.30	76.40(+25.31)
PSVMA + PC	75.68	78.63	**77.13**(+26.07)	65.9	49.96	**56.83**(+21.08)	72.69	87.61	**79.46**(+28.37)

Table 4. Results of our method compared with Calibrated stacking (CS) [7] approach. CS and PC indicate results obtained when using calibrated stacking and proposed prior correction. We highlight the improvement over baseline in red.

DUET [14], MSDN [10], TransZero [9], etc some of which are ViT based methods. Further, we also compare with some of the prominent generative methods like CE-GZSL [20], FREE [12] and HSVA [13]. We exclude large-scale models such as CLIP [40] and its extensions from the comparative analysis.

For both embedding-based and generation-based frameworks we employ our approach as a simple plugin method during testing. Note that given the post-hoc nature of the proposed approach, there is no additional training required. Although a simple hyper-parameter tuning of α and β is needed to obtain the best performance.

5. Results

5.1. Comparison with SOTA

We show our results on embedding-based and generation-based methods in Table 2. For embedding-based methods we apply our post-hoc prior correction on the PSVMA [33] method. We reproduce the results of the original paper, and use the improved result as the baseline. Further our method is applied on PSVMA by removing any existing calibrated stacking technique employed in the framework. As this further deteriorates the baseline, thus demanding a higher margin of improvement to surpass the state-of-the-art results. Despite this our method is able to surpass the results of all other existing works as shown in the table. Further we also employ our method over the generation-based methods taking the already strong CE-GZSL [20] as the baseline. Note that our approach boosts the accuracy of the CE-GZSL framework despite the generative methods already trying to minimize the imbalance by oversampling the synthetic unseen class samples. It improves the performance of [20] to a harmonic mean of 67.82%, 43.32% and 69.83% on CUB, SUN and AwA2 datasets respectively. Further our method becomes the new SOTA by boosting [33] to achieve a harmonic mean of 77.13% on the CUB, 56.83% on SUN and 79.46% on AwA2 datasets improving already strong baseline by 3.1%, 4.6% and 3.1%, respectively. Note that our method demonstrates a significant boost on the fine-grained SUN dataset surpassing the SOTA by a margin of 4.6%.

5.2. Comparison with Zero-Shot Logit-Adjustment

ZSLA [8] presents a strong baseline with similar logit-adjustment framework as ours. In particular, authors adjust model predictions during training using empirical class frequencies. We compare proposed approach as a plugin method with ZSLA in Table 3. We note that, proposed approach (2^{nd} row) achieves comparable accuracy to that of ZSLA (3^{rd} row) on both datasets. Further when combined with ZSLA (4^{th} row), proposed approach improves the performance further showing its effectiveness.

5.3. Results on pre-trained models

Given a trained model our proposed approach allows easy calculation of the prior and thereby remove the bias. We show the result of using our method as a plug-in over other existing GZSL frameworks in Table 4. For embedding-based methods we consider Transzero [9]and PSVMA [33] and generation-based methods are represented by TF-VAEGAN [37] and CE-GZSL [20]. From the table it can be observed that our method consistently provides a non-trivial performance boost for all the baseline frameworks. The margin of improvement can be as high as 4.63% on SUN dataset for embedding-based method PSVMA and 2.6% on CUB dataset for the generation-based approach TF-VAEGAN. The higher margin of improvement seen on embedding-based methods can be attributed to the higher imbalance between the seen and unseen classes.

5.4. Comparison with Calibrated Stacking

Calibrated stacking (CS) [7] presents a simple approach to remove seen class bias from the model by suppressing

$\log \frac{P^t(y)}{P^m(y)}$	$\log P(\mathcal{Y})$	$\mathcal{U}$	$\mathcal{S}$	$\mathcal{H}$	A
		CUB			
X	X	36.1	87.5	51.1	55.2
✓	X	74.7	79.1	76.8	76.7
X	✓	63.3	81.3	71.2	70.0
✓	✓	75.78	78.6	**77.1**	**77.3**
		SUN			
X	X	26.9	53.4	35.8	43.9
✓	X	63.6	50.2	56.1	55.0
X	✓	55.6	48.1	51.6	50.8
✓	✓	65.9	50.0	**56.8**	**55.7**

Table 5. We show contribution of individual adjustment terms in removing seen unseen class bias. Overall best improvement is observed when both components are used.

all seen class probabilities by a scale factor. We compare CS with propose method in Table 4. While the improvement provided by CS is minimal, proposed approach clearly provides a significant gain in performance across all the datasets and baselines. This validates the efficiency of proposed method in removing the seen unseen bias and show that trivially suppressing all seen class accuracies by same scale factor is not optimal.

5.5. Ablation Experiments:

Effectiveness of different components: We validate the effectiveness of different adjustment terms in Table 5. We note that each component improves model performance in terms of Harmonic mean as well as Arithmetic mean. We further validate effectiveness of Theorem 1 in Table 6. We show results when adjustment is performed using either $\log P^t(y)$ or $\log P^m(y)$ term only. When compared with baseline (no adjustment), adjustment due to $\log P^t(y)$ significantly boosts the performance showing that effective prior $P(y)$ learned by the model was quite different from the required prior $P^t(y)$. Further, removing incorrect prior $P(y)$ of a trained model estimated using $P^m(y)$ also lead to improvements.

Validating debiasing performance In Figure 3 we show accuracy when seen class examples are correctly classified (+ve) vs when they are mis-classified into other seen classes (seen -ve) and unseen classes (unseen -ve). Similar analysis is shown for unseen classes. We note that, for baseline model, most of the examples are misclassified into seen categories, showing a strong bias. After PC correction, the overall confusion is reduced and uniformly spread.

6. Limitations

We have seen that proposed PC-GZSL approach is quite robust and improves accuracy of many existing methods. However, as it only corrects the already trained model predictions, its performance is bounded by the baseline model's inherent accuracy. In particular, a poorly trained model will not improve much even after PC correction as original probabilities would be noisy leading to noisy bias estimation and incorrect adjustment. Further, proposed approach shares similar limitation as existing methods i.e. requirement of validation data for bias estimation, which can be challenging for ZSL settings.

Adjustment	$\mathcal{U}$	$\mathcal{S}$	$\mathcal{H}$	A
		CUB		
-	36.1	87.5	51.1	55.2
$\log P^t(y)$	72.0	74.9	73.38	73.5
$\log P^m(y)$	56.4	**85.2**	67.9	67.2
$\log \frac{P^t(y)}{P^m(y)}$	**74.7**	79.1	**76.8**	**76.7**
		SUN		
-	26.9	53.4	35.8	43.9
$\log P^t(y)$	62.5	44.8	52.2	51.2
$\log P^m(y)$	29.4	**58.0**	39.0	47.7
$\log \frac{P^t(y)}{P^m(y)}$	**63.6**	50.2	**56.1**	**55.0**

Table 6. We validate the effectiveness of bias removal using $\log P^t(y)$ and $\log P^m(y)$ terms. We use both terms separately and notice some improvements compared to baseline.

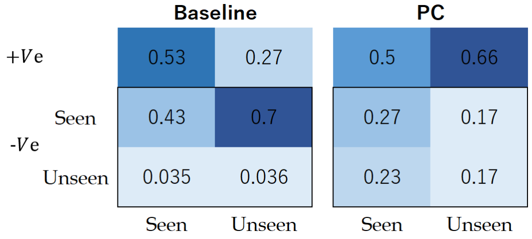

Figure 3. We show the confusion between seen and unseen classes for incorrectly classified examples for the SUN dataset.

7. Conclusions

We presented a framework to calculate the effective prior of a trained model in GZSL showing that significant bias towards the seen classes lead to sub-optimal performance. We show theoretically optimal way to remove model bias by adjusting the posterior leading to improved performance. We further use the lower bound on the harmonic mean to calculate additional correction to posterior probabilities leading to further improvement. We show the effectiveness of our approach as a simple plugin method which does not require any model training and can boost performance of many existing methods, both in embedding and generative classes.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 819–826, 2013. 1, 3

[2] 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. 1, 3

[3] S Divakar Bhat, Amit More, Mudit Soni, and Yuji Yasui. Robust loss function for class imbalanced semantic segmentation and image classification. *IFAC-PapersOnLine*, 56(2):7934–7939, 2023. 3

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 3

[5] Jacopo Cavazza, Vittorio Murino, and Alessio Del Bue. No adversaries to zero-shot learning: Distilling an ensemble of gaussian feature generators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12167–12178, 2023. 6

[6] Samet Cetin, Orhun Buğra Baran, and Ramazan Gökberk Cinbiş. Closed-form sample probing for learning generative models in zero-shot learning. 2022. 6

[7] 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 *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 52–68. Springer, 2016. 1, 2, 7

[8] Dubing Chen, Yuming Shen, Haofeng Zhang, and Philip HS Torr. Zero-shot logit adjustment. *arXiv preprint arXiv:2204.11822*, 2022. 2, 3, 5, 6, 7

[9] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. Transzero: Attribute-guided transformer for zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 330–338, 2022. 6, 7

[10] Shiming Chen, Ziming Hong, Guo-Sen Xie, Wenhan Yang, Qinmu Peng, Kai Wang, Jian Zhao, and Xinge You. Msdn: Mutually semantic distillation network for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7612–7621, 2022. 3, 6, 7

[11] Shiming Chen, Wenjin Hou, Salman Khan, and Fahad Shahbaz Khan. Progressive semantic-guided vision transformer for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23964–23974, 2024. 3, 6

[12] Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. Free: Feature refinement for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 122–131, 2021. 3, 7

[13] Shiming Chen, Guosen Xie, Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, and Ling Shao. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Advances in Neural Information Processing Systems*, 34:16622–16634, 2021. 6, 7

[14] Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Wen Zhang, Yin Fang, Jeff Z Pan, and Huajun Chen. Duet: Cross-modal semantic grounding for contrastive zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 405–413, 2023. 6, 7

[15] Zhi Chen, Yadan Luo, Ruihong Qiu, Sen Wang, Zi Huang, Jingjing Li, and Zheng Zhang. Semantics disentangling for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8712–8720, 2021. 6

[16] Guillem Collell, Drazen Prelec, and Kaustubh Patil. Reviving threshold-moving: a simple plug-in bagging ensemble for binary and multiclass imbalanced data. *arXiv preprint arXiv:1606.08698*, 2016. 5

[17] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724, 2021. 3

[18] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 2

[19] Rafael Felix, Ian Reid, Gustavo Carneiro, et al. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 21–37, 2018. 3

[20] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2371–2381, 2021. 3, 6, 7

[21] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 2

[22] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6626–6636, 2021. 3

[23] Wenjin Hou, Shiming Chen, Shuhuang Chen, Ziming Hong, Yan Wang, Xuetao Feng, Salman Khan, Fahad Shahbaz Khan, and Xinge You. Visual-augmented dynamic semantic prototype for generative zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23627–23637, 2024. 3, 6

[24] Yang Hu, Guihua Wen, Adriane Chapman, Pei Yang, Mingnan Luo, Yingxue Xu, Dan Dai, and Wendy Hall. Graph-based visual-semantic entanglement network for zero-shot image recognition. *IEEE Transactions on Multimedia*, 24:2473–2487, 2021. 3

[25] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4483–4493, 2020. 3

[26] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002. 2

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[28] Steve Lawrence, Ian Burns, Andrew Back, Ah Chung Tsoi, and C Lee Giles. Neural network classification and prior class probabilities. In *Neural networks: tricks of the trade*, pages 299–313. Springer, 2002. 2

[29] Jingjing Li, Mengmeng Jing, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7402–7411, 2019. 3

[30] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6928, 2022. 3

[31] Yan Li, Junge Zhang, Jianguo Zhang, and Kaiqi Huang. Discriminative learning of latent features for zero-shot recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7463–7471, 2018. 3

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[33] Man Liu, Feng Li, Chunjie Zhang, Yunchao Wei, Huihui Bai, and Yao Zhao. Progressive semantic-visual mutual adaption for generalized zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15337–15346, 2023. 1, 2, 6, 7

[34] Yang Liu, Jishun Guo, Deng Cai, and Xiaofei He. Attribute attention for semantic disambiguation in zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6698–6707, 2019. 3

[35] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 2, 3

[36] Shaobo Min, Hantao Yao, Hongtao Xie, Chaoqun Wang, Zheng-Jun Zha, and Yongdong Zhang. Domain-aware visual bias eliminating for generalized zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12664–12673, 2020. 3

[37] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 479–495. Springer, 2020. 3, 6, 7

[38] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. *Advances in neural information processing systems*, 22, 2009. 1

[39] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2751–2758. IEEE, 2012. 1, 6

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 7

[41] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020. 3

[42] Michael D Richard and Richard P Lippmann. Neural network classifiers estimate bayesian a posteriori probabilities. *Neural computation*, 3(4):461–483, 1991. 2, 4

[43] Yuming Shen, Jie Qin, Lei Huang, Li Liu, Fan Zhu, and Ling Shao. Invertible zero-shot recognition flows. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 614–631. Springer, 2020. 1

[44] Jingru Tan, Xin Lu, Gang Zhang, Changqing Yin, and Quanquan Li. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1685–1694, 2021. 3

[45] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Argue: Attribute-guided prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28578–28587, 2024. 3

[46] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[47] Chaoqun Wang, Shaobo Min, Xuejin Chen, Xiaoyan Sun, and Houqiang Li. Dual progressive prototype network for generalized zero-shot learning. *Advances in Neural Information Processing Systems*, 34:2936–2948, 2021. 3

[48] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020. 3

[49] Yidong Wang, Bowen Zhang, Wenxin Hou, Zhen Wu, Jindong Wang, and Takahiro Shinozaki. Margin calibration for long-tailed visual recognition. In *Asian Conference on Machine Learning*, pages 1101–1116. PMLR, 2023. 3

[50] 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. 1, 3

[51] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 1, 6

[52] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018. 1, 2, 3

[53] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10275–10284, 2019. 3

[54] Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. Attentive region embedding network for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9384–9393, 2019. 3

[55] Guo-Sen Xie, Li Liu, Fan Zhu, Fang Zhao, Zheng Zhang, Yazhou Yao, Jie Qin, and Ling Shao. Region graph embedding network for zero-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 562–580. Springer, 2020. 3

[56] Zhongqi Yue, Tan Wang, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, 2021. 6

[57] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2021–2030, 2017. 1, 3

[58] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021. 3

[59] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 3

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3

[61] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2022. 3

[62] Yizhe Zhu, Jianwen Xie, Zhiqiang Tang, Xi Peng, and Ahmed Elgammal. Semantic-guided multi-attention localization for zero-shot learning. *Advances in Neural Information Processing Systems*, 32, 2019. 3