Towards Adversarial Robustness and Reducing Uncertainty Bias through Expert Regularized Pseudo-Bidirectional Alignment in Transductive Zero Shot Learning

Abhishek Kumar Sinha[1], Deepak Mishra[2], and S. Manthira Moorthi[1]

[1] Signal and Image Processing Area, Space Applications Center
{aks,smmoorthi}@sac.isro.gov.in
[2] Indian Institute of Space Science and Technology
deepak.mishra@iist.ac.in

Abstract. Transductive zero-shot learning (TZSL) aims to minimize the domain shift between the learned and true distribution of the unseen classes by allowing access to the unpaired samples from unseen classes. While many distribution alignment based methods attempt to align both visual and semantic spaces to train the classifier, their performance is still limited by confirmation bias. Additionally, bidirectional alignment approaches are based on the strong assumption that the intrinsic dimensions of visual and semantic spaces are the same, which is rarely true. In this work, we first highlight the limitations of bidirectional alignment in terms of intrinsic dimensionality. We then present a pseudo-bidirectional approach that, without any underlying assumptions on these spaces, utilizes the learned visual-to-attribute mapping to minimize the distribution shift between learned and true unseen visual feature distributions. We further utilize an entangled loss between semantic and visual space to minimize the confirmation or uncertainty bias and improve the adversarial robustness. We, theoretically and empirically, show the performance gain in addition to the adversarial robustness under the proposed setting.

Keywords: Zero-shot learning · Bias · Adversarial robustness.

1 Introduction

The aim of zero-shot learning is to recognize and classify objects or concepts for which they have not been explicitly trained. In most of the practical scenarios, the computer vision models are required to be trained on a set of large number of training examples paired with their corresponding labels, known as seen classes. The trained model are then used to infer the labels for which there are no available training examples, referred to as target labels. In inductive zero-shot learning, samples from the target (or unseen) classes are not provided for training. However, a sufficient number of paired examples are provided for the seen categories. This approach requires the classifier to learn the relation between visual and semantic spaces using the seen classes and transfers this knowledge to the unseen classes assuming that such relevant knowledge exists. This knowledge sharing requires annotated data such as vector embedding of labels, attribute features and so on. However, transfer learning without access to unseen labels can be quite challenging due to domain shift problem [5]. To simplify the problem, Transductive zero-shot learning (TZSL) [16,15,6] utilizes the unlabeled examples of the targeted classes for training. This allows access to the collective target data distribution without correspondences to ease off the burden of distribution shift.

Most of the approaches are influenced by generative modelling that intend to align the distribution of real examples and generated examples followed by training the classifier on the generated examples. Depending on the discrepancy in the learned data distribution, the classifier may suffer from confirmation bias, which means that the classifier is trained on the generated samples assuming that they are correctly paired to their target labels. Additionally, many previous graph based approaches model the attribute relation using Word2Vec or GloVe embeddings but rely on corpora training which may not provide necessary characteristics to distinguish between the classes [21]. Moreover, knowledge graph based methods [16] also has its own challenges. For example, knowledge graphs may struggle with handling ambiguous concepts or entities with multiple senses. Different classes or concepts can share similar or overlapping features, making it challenging to disambiguate them solely based on the information in the knowledge graph.

To address these limitations, we propose pseudo-bidirectional alignment that utilizes expert information to learn bidirectional-like mapping. The contribution of the proposed work is as follows:

1. We introduce pseudo-bidirectional alignment using Expert guided VAEGAN that, unlike bidirectional adversarial learning, learns the semantic-to-visual mapping based on the additional knowledge from an expert model, which is visual-to-attribute mapping in our case.
2. The proposed model improves the semantic-to-visual mapping by incorporating knowledge from an expert model to learn distribution shift in a low intrinsic dimensional space, contributing to a more robust and effective learning process.
3. A new entangled loss function is introduced for classifier training, by integrating generated visual features and pseudo-labels. This leads to reduction of confirmation bias and shows its effectiveness in terms of adversarial robustness and, providing a novel and impactful contribution to the training process.
4. Both theoretical and experimental evidence were presented, showcasing the remarkable performance of pseudo-bidirectional alignment. Finally, implicit robustness is achieved through the proposed approach, contributing to the model's resilience in the face of various challenges. The method's ability to overcome uncertainty is highlighted, making it a notable and impactful contribution to the field of Transductive zero-shot learning.

2 Related Works

2.1 Zero-Shot Learning

Zero-shot learning has garnered a lot of interest in the past few years due to its practical applications in many vision and language-related problems. Inductive zero-shot entity recognition has previously been addressed in which most of them tend to learn semantic to visual space mapping using projection mapping [34]. This approach of transferring knowledge from seen to unseen classes suffers from domain shift due to non-overlapping distributions of seen and unseen classes. Subsequently, some works utilize two networks to align the distributions in both semantic and visual spaces by using generative modelling such as VAE and GANs. For example, Cycle-WGAN [4] uses a new multi-modal cycle consistency loss which constrains the optimization problem to generate useful visual features for the training of classifier. Another method is to exploit the expert knowledge for domain alignment by leveraging the expertise of a domain expert to constrain the learning process, and it closely resembles our idea in this work. For example, Norouzi et al. [17] proposes a convex combination mapping approach for zero-shot learning. It incorporates expert knowledge in the form of semantic attributes and enforces a regularization term to constrain the model's predictions to be a convex combination of attribute vectors. Since the model is regularized to align its predictions with the provided attributes, noisy or misleading attributes might negatively impact the learning process and lead to erroneous predictions. Field-Guide-Inspired Zero-Shot Learning ([14]) is another interesting approach which directly involves a human expert to interact with the learner. In this approach, the learner is first trained on a set of base classes followed by interaction with an expert annotator to seek minimal guidance on the attributes to classify the unseen classes. The method may suffer from the knowledge gap between attribute understandings of humans and the neural net, and it is relatively difficult to align their knowledge due to variations in the human experts. Nevertheless, the absence of knowledge of unseen classes serves as the performance bottleneck and restricts the performance of inductive ZSL.

Contrary to an inductive setting, transductive ZSL allows the learner to utilize the knowledge of examples from unseen classes without correspondence. Generative models have been adopted by most of the state-of-the-art as adversarial training allows to align the distribution. Marmoreo et al. [15] proposed the idea of decoupled feature generation by encapsulating the visual patterns into structured prior to boost the performance of conditional visual feature synthesis. It uses DecGAN to capture the distribution of visual features and generate realistic descriptors. The pioneering work in zero-VAE-GAN [6] is the first to attempt the coupled Variational Autoencoder (VAE) and generative model for this task. It uses generative methods to synthesize visual features conditioned on semantic side information and learn a conventional supervised classifier from generated sample. However, when generative models are trained with seen classes, there are inherently biased when it comes to generalization to unseen classes.

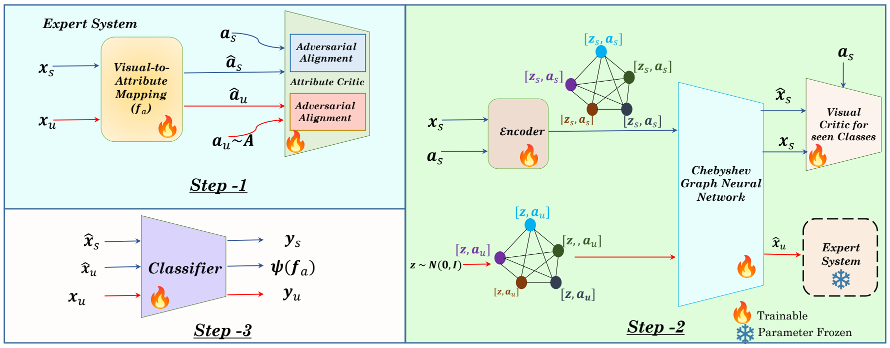


Fig. 1: Overall training methodology for proposed Transductive ZSL. Step 1 involves adversarial training of visual-to-attribute mapper (f_a) for both seen and unseen classes. Step 2 utilizes f_a from previous step for aligning distributions of true and generated visual features of unseen labels. In step 3, we utilize the paired generated feature-label and original feature-pseudolabel to train the classifier.

2.2 Graph Neural Networks in ZSL

Unlike multi-layer perceptron that has fully connected layers, graph neural nets learn the node embeddings based on the node strength and its connections to other nodes of the graph. Similar to self-attention [24,22,23], it is capable of learning global representations based on the node connections and strengths. The recent studies [31,9,16] have demonstrated the effectiveness of utilizing the graph structure in zero shot learning. Xiel et al. [31] proposed a region graph embedding network to capture the relationships between various parts of the image using graph convolutions. The graph nodes consist of local regions of the image and are connected by the edges depending on the pairwise nodes' similarity. Since the regionwise features of the image may fail to capture the extent of the relation, it translates to the edges' strength resulting in misleading interaction between patches. Similarly, the Visual-Semantic Entanglement network in [9] learns the graph embeddings of visual features and maps it to the semantic attributes using the knowledge graph. Additionally, it uses a multi-path entangled path network which feeds the visual features from CNN to GCN to learn the semantic relations resulting in self-consistent regression for graph modelling. Liu et al. [16] also exploits the knowledge graph through a transformer to learn class representations by embedding nodes in the knowledge graph. [12] exploits graph relation for attribute propagation to refine the features in semantic space based on the information aggregated from the neighbouring nodes. This approach does not add any constraint to align the learned attribute features and therefore makes a strong assumption that attribute propagation does not affect the associated attribute labels, and thus the label for the propagated attributes is the same as the original associated label before propagation. While the knowledge graph embeddings are useful in natural language-related problems, the semantic vectors may not be directly applicable to computer vision problems. The limited semantic coverage of knowledge graph vectors introduces a bias towards certain classes restricting its ability to generalize for unseen classes. To avoid this limitation, we directly utilize the attribute vectors associated with class labels to represent the connection between nodes.

3 Limitations of Bidirectional Alignment

We use the idea of Intrinsic dimensionality and Wassertein distance to highlight the limitation of bidirectional alignment [27].

Theorem 1 (Invariance of domain). *If U is an open subset of R^n and $f : U \rightarrow R^n$ is an injective continuous map, then $V = f(U)$ is an open and f is a homeomorphism between U and V.*

Bi-VAEGAN [27] learns adversarial mapping between visual space and attribute space. However, visual features comprise of additional details beyond the attribute features and therefore, the Intrinsic dimension of visual feature space (ID_v) is relatively larger than that of attribute space (ID_a). This limitation may negatively impact the diversity of learned samples in visual space.

Proposition 1. *Let $x \sim \mathcal{P}$ and $x' \sim \mathcal{P}^{'}$ be the samples from true and learned distributions, respectively such that diffrence between the intrinsic dimension of $\mathcal{P}^{'}$ and $\mathcal{P}$ is δ. If D^* is the intrinsic dimension of $\mathcal{P}$, then normalized Wasserstein distance, conditioned on $[0, w]$, is given by,*

$$\mathcal{W}_2^2(F, G, w) = \frac{2\delta^2}{(D^* + 2)(D^* + \delta + 2)(2D^* + \delta + D^{*2} + D^*\delta)} \tag{1}$$

The proof of Proposition 1 is provided in supplementary material. It shows that the Wasserstein distance is less sensitive to the distribution shift if the underlying intrinsic dimension is large. This means that, for a given shift, the Wasserstein loss in visual space remains relatively lower than that in attribute space, and therefore, in the case of Bi-VAEGAN [27], adversarial learning in visual space is not as effective as attribute space. We, therefore, propose to use only attribute space to learn pseudo-bidirectional alignment.

4 Methodology

4.1 Problem Formulation

Transductive ZSL aims to classify the unseen classes by accessing the unpaired examples from the domain of unseen classes $\mathcal{D}^u$. We denote by $\mathcal{D}^s = \{(x, y, a_y) | x \in \mathcal{X}^s, y \in \mathcal{Y}^s, a_y \in \mathcal{A}^s\}$ the domain of seen classes, where x is the visual feature, y is the corresponding label and a_y is the attribute of that category. Similarly,

$\mathcal{D}^u = \{(x, u, a_u) | x \in \mathcal{X}^u, u \in \mathcal{Y}^u, a_u \in \mathcal{A}^u\}$ is the set of unseen labels, and $\mathcal{A} = \mathcal{A}^s \bigcup \mathcal{A}^u$. The model won't have access to the labels for $x \in \mathcal{X}^u$. For generalized setting, it is assumed that $\mathcal{Y}^s \cap \mathcal{Y}^u = \Phi$. Furthermore, f_a and D_a are used to denote the visual-to-attribute mapper and the attribute critic, respectively. Additionally, $\psi \circ f_a$ assigns the label based on the cosine similarity between the generated attribute vector and attribute from unseen domain. We denote by x^{ug} the unseen generated visual feature vector. $\mathcal{E}$ and $\mathcal{G}$ are the encoders and decoders of the variational autoencoder, where $\mathcal{G}$ is a graph neural network. Let D_v^s denote the visual critic for seen classes only. The goal is to develop a framework to classify the examples from unseen classes correctly in both conventional and generalized setting.

For theoretical analysis, we denote by $\epsilon(h, f) = \mathbb{E}_{(x,a)\in D}[\mathbf{1}_{f(x)\neq h(x,a)}]$ the actual risk and $\hat{\epsilon}(h, f) = \frac{1}{|D|}\sum_{(x,a)\in D}[\mathbf{1}_{f(x)\neq h(x,a)}]$ the empirical risk. $d_{h^*}(\mathcal{D}_1, \mathcal{D}_2)$ and $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_1, \mathcal{D}_2)$ is the generative distance for optimal hypothesis h^* and $\mathcal{H}\Delta\mathcal{H}$ distance [3], respectively. The details are discussed in supplementary material.

4.2 Overall Outline

Figure 1 shows the overall training pipeline of pseudo-bidirectional alignment. The first stage involves training the visual-to-attribute mapper in adversarial fashion for both seen and unseen classes. Subsequently, it is utilized for transferring knowledge to train the VAEGAN architecture to reduce domain shift between true and generated visual features. Furthermore, it also exploits the semantic relationship among the classes through graphical structure to adversarially learn the visual features of unseen classes.

Visual-to-Attribute Mapping The first stage of training involves learning attribute feature from the given visual feature for both seen and unseen classes. It uses a simple multi-layer perceptron architecture which is trained in the supervised fashion with adversarial regularization for seen classes. For the samples from unseen classes, the model is trained only in adversarial fashion. Unlike Bi-VAEGAN [27], the adversarial learning on both seen and unseen classes helps the attribute critic to learn the interaction between their distributions. The optimization objective minimizes the L_1 norm for the examples from $\mathcal{D}^s$ given by,

$$L^s_{f_a} = \min_{f_a} \max_{D_a} ||f_a(x) - a_y||_1 + \lambda_1 L^s_{adv}, \tag{2}$$

where $L^s_{adv}(\mathcal{A}^s, \mathcal{V}^s) = \mathbb{E}[D_a(a^s)] - \mathbb{E}[D_a(f_a(x^s))] + (||\nabla_{\hat{a}^s}\mathbb{E}[D_a(\hat{a}^s)]||_2 - 1)^2$ with $\hat{a} = \alpha a^s + (1-\alpha) f_a(x^s)$. The objective for the unseen classes is similarly defined as,

$$\begin{aligned} L^u_{f_a} = L^u_{adv} = &\mathbb{E}[D_a(a^u)] - \mathbb{E}[D_a(f_a(x^u))] \\ &+ (||\nabla_{\hat{a}^u}\mathbb{E}[D_a(\hat{a}^u)]||_2 - 1)^2, \end{aligned} \tag{3}$$

where $a^u \sim \mathcal{A}^u$. The critic training includes the gradient penalty term [8] to induce better Lipschitz stability.

Pseudo-Bidirectional Alignment for Attribute to Visual Mapping The aim of training Expert-VAEGAN is to align the distributions of synthetic and true visual features. It uses the visual critic D_v^s specifically to align the generated seen visual features conditioned on their corresponding attribute features. Additionally, we use the visual-to-attribute map from the preceding stage as an expert system to transfer the knowledge about the previously learned relationship between visual and attribute space for unseen categories. We call this approach pseudo-bidirectional alignment because it tends to diversify only those components of visual features that are necessary to discriminate them in attribute space. To alleviate this issue, we re-utilize the visual-to-attribute mapper and attribute critic together to minimize the distribution shift between generated and true visual features. Since they have already been trained in preceding step, they together serve as an expert through their learned mapping. Firstly for seen categories, an encoder is explicitly used to learn the latent representations for the visual features conditioned on their respective attribute space. For unseen categories, we randomly sample the latent vector from a standard Gaussian distribution and stack them with a sampled attribute vector. The concatenated attribute and latent vectors serve as the node embeddings for the graph with the node connections defined by cosine similarity between the node attributes. For two nodes with attributes a_i and a_j, the the weight of connecting edge is defined as $e_{ij} = \frac{<a_i, a_j>}{||a_i||_2 ||a_j||_2}$. The constructed graph is then passed to the first order Chebyshev graph net (ChebNet) that computes the visual features corresponding to each node. Since it leverages the Laplacian eigenbasis of the graph to perform convolutions in the spectral domain, ChebNet captures both local and global structural dependency effectively. This, in turn, allows to learn the visual features based on their attribute similarity more effectively. Furthermore, some of elements in the visual feature vector is randomly masked with zero while training so to enhance the model's ability to learn the intra-feature connections.

For training, we apply the VAE objective on the latent space vectors of the seen classes as it is known to prevent the mode-collapse in GAN training. Furthermore, we add L1 loss to minimize the reconstruction error of visual features, and an adversarial regularization to align the synthetic and true visual features distributions. Since the visual features can be paired with its attribute features, the adversarial training aligns the learnt and true feature distribution conditioned on the attribute space. The overall training objective for the seen classes is given by,

$$\begin{aligned}\mathcal{L}^s = & \min_{\mathcal{E},\mathcal{G}} \max_{D_v^s} \mathbb{E}_{z_s \sim \mathcal{E}(x^s, a^s)}[KL(z^s || \mathcal{N}(0, I))] + \\ & \mathbb{E}_{z_s \sim \mathcal{E}(x^s, a^s)}[||\mathcal{G}(z^s, a^s) - x^s||_1] + \lambda_2 \mathcal{L}^s_{D_v^s},\end{aligned} \tag{4}$$

where $D_v^s = \mathbb{E}[D_v^s(x^s, a^s)] - \mathbb{E}_{\bar{x}^s \sim \mathcal{G}}[D_v^s(\bar{x}^s, a^s)] + (||\nabla_{\hat{x}^s} \mathbb{E}[D_v^s(\hat{x}^s, a^s)]||_2 - 1)^2$.

For unseen examples, we directly utilize the attribute critic for adversarial learning. In this case, the training can be formulated as,

$$\mathcal{L}^u = \min_{\mathcal{G}} \max_{D_a} \mathbb{E}_{\hat{x}^u \sim \mathcal{G}}[||f_a(\hat{x}^u) - a^u||_1] + \lambda_3 \mathcal{L}^u_{adv}(\mathcal{G}, \mathcal{A}^u), \tag{5}$$

where $L^u_{adv}(\mathcal{G},\mathcal{A}^u) = \mathbb{E}[D_a(a^u)] - \mathbb{E}_{\hat{x}^u\sim\mathcal{G}}[D_a(f_a(\hat{x}^u))] + (||\nabla_{\bar{a}^u}\mathbb{E}[D_a(\bar{a}^u)]||_2 - 1)^2$ and $\bar{a} = \alpha a^u + (1-\alpha)\hat{a}^u$.

Here, λ_2 and λ_3 are the hyper-parameters. Equation 5 allows to align the knowledge of the graph net with that of visual-to-attribute mapping. Instead of strictly aligning visual features' distribution for unseen labels, we utilize the expert knowledge to learn the visual features discriminative enough to classify them into correct categories.

Algorithm 1 Algorithm for pseudo-bidirectional alignment

```
$\mathcal{X}^s$,$\mathcal{Y}^s$, $\mathcal{X}^u$, ($A^u$,$A^s$), $T_1$, $T_2$ Trained $\mathcal{G}$, $\mathcal{E}$, $f_a$, $D^s_v$, $D_a$
for i in range ($T_1$) do
    Train the visual-to-attribute mapping transductively using equations 2 and 3.
end
for i in range ($T_2$) do
    Generate synthetic visual features $\hat{x}^s$ for a sampled $\{x^s$,$a^s\}$.
    Train $\mathcal{E}$ and $\mathcal{G}$ for seen classes using equation 4.
    Uniformly sample a batch of attributes $a^u \sim \mathcal{A}^u$ and $z \sim \mathcal{N}(0, I)$.
    Estimate the edge weights $E = \{e_{ij}\}$ for $a_i, a_j \in a^u$.
    Generate the corresponding synthetic visual feature $\hat{x}^u \sim \mathcal{G}(z, a^u, E)$ and get $\{\hat{x}^u$,$a^u\}$.
    Train $\mathcal{E}$ and $\mathcal{G}$ for unseen classes using equation 5.
    For training of classifier, generate a pair of true visual feature and its pseudo-label $\{x^u$,$\psi \circ f_a\}$.
    Also, generate a pair of synthetic visual feature and original attribute vector $\{\hat{x}^u$,$a^u\}$.
    Train the classifier using the loss function 6.
end
```

4.3 Training the classifier

Since transductive setting allows access to unseen classes, it adds another degree of freedom that we exploit in the loss function. It is to be noted that training Expert-VAEGAN involves two modules that can assign a label to the visual feature, one is visual-to-attribute mapping and the other one is the classifier itself. To strengthen the alignment of their predictions, we apply entanglement between true and learnt distributions in both visual and semantic space. For this, we generate pseudo-labels from f_a for the given true visual feature in addition to the paired synthetic visual feature and the attribute vector. The combined training objective for the classifier is given by,

$$\mathcal{L}_{cls} = \beta[-\mathbb{E}_{x\sim\mathcal{X}^u} P(\psi \circ f_a|x;\theta)] + (1-\beta)[-\mathbb{E}_{\hat{x}\sim\mathcal{G}} P(y|\hat{x};\theta)], \tag{6}$$

where $P(y|\hat{x};\theta)$ denotes the probability of assigning label y to the synthetic feature $\hat{x}$. Similarly, $P(\psi \circ f_a|x;\theta)$ is the probability of assigning the pseudo-label $\psi \circ f_a$ to the true features. The overall procedure for training is described in the Algorithm 1.

4.4 Theoretical Perspective

In this section, we provide theoretical arguments to support our claims on adversarial robustness and confirmation bias.

Confirmation (or Uncertainty) Bias

Theorem 2. *Let $\mathcal{R}$ and $\mathcal{H}$ denote the hypothesis space of classifier h and visual-to-attribute regressor R, respectively. Without the loss of generality, lets assume that for regressor R, $\psi \circ R$ assigns the label based on the similarity measure between predicted attribute and unseen classes' attributes. If optimal classifier h^* satisfies the condition: $h^* = \underset{h'}{\operatorname{argmin}}\ \hat{\epsilon}_s(h', f) + \hat{\epsilon}_{ug}(h', f) + \hat{\epsilon}_u(h', \psi \circ R)$, Then with probability $1 - \delta$, following inequality holds for N number of samples,*

$$\begin{aligned}\epsilon_u(h, f) \leq \hat{\epsilon}_s(h, f) + d_{h^*}(\mathcal{X}^{ug}, \mathcal{X}^u) + d_{R^*}(\mathcal{X}^u, \mathcal{X}^s) + \\ d_{h^*}(\mathcal{X}^u, \mathcal{X}^s) + \frac{1}{2} d_{H \Delta H}(\mathcal{X}^u, \mathcal{X}^s) + \epsilon_u(h^*, f) \\ + \lambda + \sqrt{\frac{1}{2N} log \frac{2}{\delta}},\end{aligned}$$

where $\lambda = \epsilon_s(h^*, f) + \epsilon_u(h^*, \psi \circ R^*) + \epsilon_{ug}(h^*, f)$

We attempt to theoretically show that uncertainty or confirmation bias has detrimental impact on the overall performance. In transductive zero shot learning, confirmation bias arises in two ways. Firstly, we assume that the synthetic visual features belong to a particular class even though the generated feature may be perturbed enough to change its category. Second, the confirmation bias may get injected into the model through the conflict between the pseudo-label assigned by visual-to-attribute mapping to the true visual features and label predicted by the classifier. If the label predicted by the classifier and the pseudo-label assigned by the mapper do not match with each other, the overall training may converge to a sub-optimal solution. Theorem 2 shows that the loss function in 6 implicitly adds a constraint of on R^* and h^* through $\epsilon_u(h^*, \psi \circ R^*)$. For prediction error to reduce on the unseen labels, the labels assigned by both of them on a given feature must agree. Additionally, $d_{h^*}(\mathcal{X}^{ug}, \mathcal{X}^u)$ constrains the distance between the distributions of $\mathcal{X}^{ug}$ and $\mathcal{X}^u$ to reduce, and second, third and fourth terms are constant for a given problem due to fixed domain shift between seen and unseen classes.

Implicit Adversarial Robustness In self-training based methods, it is quite common to apply label interpolation to improve adversarial robustness. However, [18] provides detailed analysis to contradict this assumption by showing that interpolation in noisy labels is as large an adversarial risk as the poisoning with similar noise rate.We, therefore, refrain from applying such interpolation in our loss function. Instead, we now show that loss function 6 implicitly acts as a weak upper bound on the robustness of the learned classifier. Our analysis for

	Method	Conventional			Generalized								
		AWA2	CUB	SUN	AWA2			CUB			SUN		
					S	U	H	S	U	H	S	U	H
$\mathcal{I}$	RGEN [31]	73.6	76.1	63.8	67.1	76.5	71.5	60.0	73.5	66.1	44.0	31.7	36.8
$\mathcal{I}$	APNet [12]	68.0	57.7	62.3	83.9	54.8	66.4	55.9	48.1	51.7	40.6	35.4	37.8
$\mathcal{I}$	FG [14]	-	-	-	65.0	65.8	65.4	59.6	52.8	55.8	**61.3**	41.3	49.3
$\mathcal{I}$	LSG [32]	61.1	52.9	53.4	84.9	60.4	70.6	50.4	49.6	50.0	23.1	52.8	32.2
$\mathcal{I}$	Assym. Net [26]	-	55.9	57.6	-	-	-	19.4	56.5	28.9	18.5	28.6	22.5
$\mathcal{T}$	DSRL [33]	72.8	56.8	48.7	-	-	-	25.0	17.7	20.7	39.0	17.3	24.0
$\mathcal{T}$	F-VAEGAN-D2 [30]	-	71.1	70.1	-	-	-	65.1	61.4	63.2	41.9	60.6	49.6
$\mathcal{T}$	Zero-VAEGAN [6]	89.0	69.1	68.4	87.0	70.2	77.6	57.9	64.1	60.8	35.8	53.1	42.8
$\mathcal{T}$	ZSL-KG [16]	78.1	-	-	84.4	66.8	74.6	-	-	-	-	-	-
$\mathcal{T}$	DecGAN [15]	-	-	-	-	-	-	44.3	57.2	49.9	**68.4**	60.9	**63.4**
$\mathcal{T}$	Bi-VAEGAN [27]	95.8	**76.8**	**74.2**	**91.0**	**76.1**	**90.4**	**71.7**	**71.2**	**71.5**	45.4	**66.8**	54.1
$\mathcal{T}$	Ours	**96.4**	**77.2**	**75.2**	**92.6**	**89.6**	**91.1**	**70.3**	**73.9**	**72.1**	58.7	**66.2**	**62.2**

Table 1: Performance comparison with state-of-the-art in both conventional and generalized ZSL. $\mathcal{I}$ and $\mathcal{T}$ refer to inductive and transductive settings, respectively. In generalized ZSL, U and S indicate accuracies for unseen and seen labels, respectively, and H is their harmonic mean. The best and second best results are shown in red and blue, respectively.

robustness is based on two assumptions: (1) the generated samples already contain adversarial noise, and therefore serves as adversarial examples [11], (2) We follow [20,28] to use population consistency loss as the measure of robustness on the unlabeled features from unseen classes. Based on this, we assume $x^{ug} \in \mathcal{B}_\rho$, where $\mathcal{B}_\rho(x) = \{x' : ||x' - x|| \leq \rho\}$. The population consistency loss is defined as $\mathcal{R}_B(h, x) = \mathbb{E}_{x \sim \mathcal{D}}[\mathbf{1}(\exists\ x' \in \mathcal{B}_\rho(x)\, such\, that\, h(x) \neq h(x'))]$.

Theorem 3. *Let $x \sim \mathcal{D}^u$ and x' be the true and the corresponding adversarial features, respectively. Let $R(x)$ maps the given visual feature x to its corresponding semantic feature, and $\psi \circ R(x)$ produces corresponding label based on semantic feature similarity. The population consistency loss $\mathcal{R}_B$ is weakly bounded by,*

$$\begin{aligned} \mathcal{R}_B \leq & \mathbb{E}_{x' \sim \mathcal{B}_\rho(x), x \sim \mathcal{D}^u}[\mathbf{1}(h(x') \neq f(x))] + \mathbb{E}_{x \sim \mathcal{D}^u}[\mathbf{1}(h(x) \neq \psi \circ R(x))] + \\ & \mathbb{E}_{x \sim \mathcal{D}^u}[\mathbf{1}(\psi \circ R(x) \neq f(x))], \end{aligned} \tag{7}$$

In Theorem 3, the first two terms represent the loss function 6 provided we treat the generated features as the adversarial examples. The bound is apparently weaker in the initial phase of training since we have no prior information about the labels of x for unseen categories, and therefore the third term cannot be explicitly controlled. However, as the model begins to converge, the third term approaches to zero and the bound eventually becomes tighter. In other words, the proposed loss function provides weak guarantee of adversarial robustness.

5 Experiments

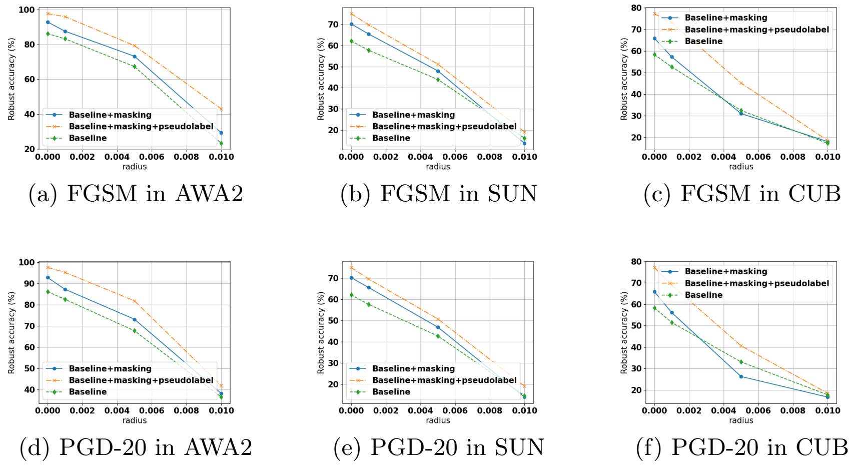
In this section, we compare the performance with other state-of-the-art methods using benchmark datasets. Additionally, we provide empirical evidences to support our theoretical analysis. We conduct our experiments on three datasets, including AWA2 [29], CUB [25] and SUN [19]. The visual features of the images are extracted using ResNet-101 pre-trained network. We analyse the performance in both conventional and generalized setting by measuring overall accuracy for all the unseen classes. The dataset and training details along with hyper-parameter settings are provided in the supplementary material. In generalized setting, we measure the accuracy for both seen (ACC^s) and unseen classes (ACC^u) and express them using Harmonic mean given by $H = \frac{2ACC^s \times ACC^u}{ACC^s + ACC^u}$. Additionally, We directly report the results from the published papers.

5.1 Performance Comparison

Since we utilize the graphical structure in VAE-GAN setup, We compare the performance with VAE-GAN setups, including F-VAEGAN-D2 [30], Zero-VAEGAN [6], DecGAN [15] and Field-Guided CADA-VAE [14], and graph based approaches such as RGEN [31], APNet [12], LSG [32], Asymmetric Graph Network [26] and ZSL-KG [16]. Table 1 presents the comprehensive comparison to the aforementioned state-of-the-art models. In Conventional setting, our method achieves the best performance in in all three datasets. In generalized setting, our approach achieves best performance in AWA2 and CUB, and second best in SUN in terms of harmonic mean. We argue that our idea is still competitive for two reasons. Firstly, most of the generative methods apply strong discriminability on both visual and semantic spaces, whereas our model learns the features in visual space by solely transferring knowledge to semantic space and leveraged it to learn the discriminant visual features. This approach reduces the dependency on additional discriminator and aids to training stability. Secondly, it can be observed that the model does not overfit on the seen categories and maintains the decent balance between the observed and unobserved samples. It is to be noted that Field-Guided CADA-VAE [14] achieved second best accuracy on the seen classes of SUN dataset at the cost of unseen ones, and therefore overfits on the seen labels. Overall, the method achieved competitive performance by simply exploiting the knowledge of expert to learn the distribution of visual features.

5.2 Ablation studies

In this section, we analyse several aspects of the method to study their impacts on the overall accuracy of the classifier. We discuss the impact of visual feature masking and how the loss function based on pseudo-labels contributes to reduce the confirmation bias in the network, and we empirically verify the Theorem 2. Additionally, we evaluate its robustness against commonly used adversarial attacks to validate Theorem 3.



(a) FGSM in AWA2 (b) FGSM in SUN (c) FGSM in CUB

(d) PGD-20 in AWA2 (e) PGD-20 in SUN (f) PGD-20 in CUB

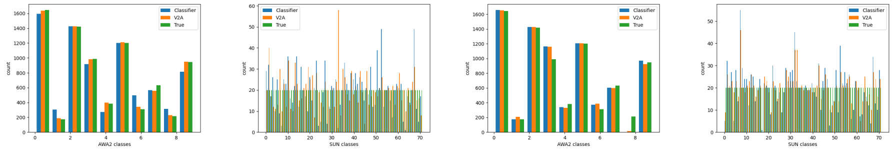
Fig. 2: Graphical comparison of performance under FGSM and PGD-20 attacks on visual features for AWA2, SUN and CUB datasets.

Method	Zero Shot Learning			Generalized Zero Shot Learning					
	AWA2	CUB	SUN	AWA2		CUB		SUN	
				S	U	S	U	S	U
Baseline	85.3	58.7	61.8	92.5	76.8	50.2	49.2	48.8	38.3
Baseline+masking-10	91.9	65.9	69.9	92.8	79.5	69.6	55.4	47.8	52.1
Baseline+masking-20	91.2	65.2	70.7	**93.5**	78.5	**72.8**	61.0	48.4	51.9
Baseline+masking-10+pseudolabel	**96.4**	**77.2**	**75.2**	92.6	**89.6**	70.3	**73.9**	**58.7**	**62.2**

Table 2: Performance comparison with Top-1 accuracy under various settings. Masking-10/20 means that 10/20 random elements in the visual feature vector are masked during training. Methods without pseudolabel setting only uses the first term of loss function for training. Best results are shown in **bold**.

Feature Masking and Pseudo-labels The motivation to learn intra-class feature dependencies is from Kong et. al. [10], which learns a new embedding to enhance the separability between seen and unseen classes. Instead of adding learning overhead, we mask some of the elements in feature elements randomly and train the model over the masked feature vectors. Since the visual features extracted using ResNet-101 may also contain some redundant information, masking lets the network explore the intra-feature relationships to extract the maximum information and predict the correct attribute. Additionally, The graphical structure allows to share the knowledge and exploit the inter-class relationships. Table 2 shows a significant performance gain when a small fraction of visual features are randomly masked. Moreover, there is no significant advantage when we increase the number of masked features from 10 to 20. Therefore, all the experiments are conducted with 10 masked elements in visual feature vector. The performance is further supplemented by reduced confirmation bias when trained with proposed loss function. The empirical studies on the confirmation bias follows in the next section.

Analysing the robustness We study the robust accuracy of our approach under two adversarial attacks, including FGSM [7] and PGD-20 [13] attacks. For this, the visual features are subjected to these attacks for different perturbation budgets, and the accuracy of classifier is then observed to evaluate the robustness. Figures 2 (a)-(c) and Figures 2 (d)-(f) compare the performances under FGSM and PGD-20 attacks, respectively. It can be observed that the model trained with loss function 6 outperforms by a relatively large margin supporting our claims in Theorem 3.



(a) AWA2@$\beta = 0$ (b) SUN@$\beta = 0$ (c) AWA2@$\beta = 0.5$ (d) SUN@$\beta = 0.5$

Fig. 3: Comparison of categorical distribution learnt by classifier and visual-to-attribute mapper when trained for $\beta = \{0.5, 0\}$

β	0	0.1	0.2	0.4	0.5	0.6	0.8	1
AwA2	52.7	94.8	94.0	95.1	96.4	94.9	94.3	91.1
SUN	71.3	71.1	71.7	72.2	75.2	69.2	68.8	64.9
CUB	65.8	70.2	70.6	71.2	77.2	72.7	71.9	62.2

Table 3: Top-1 accuracies on three datasets for different values of β. The best results are obtained when both terms in the loss function are given equal weightage.

Mitigating Confirmation Bias To further support our claims on Theorem 2, we ablate the value of β to show that minimizing both $-\mathbb{E}_{x\sim\mathcal{X}^u}P(\psi \circ f_a|x;\theta)]$ and $-\mathbb{E}_{\hat{x}\sim\mathcal{G}}P(y|\hat{x};\theta)$ enhances the overall accuracy of the trained classifier. Table 3 shows the accuracies of the classifier when trained with different values of classifier. It is clearly evident that the overall performance increases significantly when both losses are minimized with equal weightage. An intuitive explanation is that when these losses are assigned unequal weights, one of the losses reduces faster than other resulting in imbalance in their accuracy. Additionally, Figure 3 illustrates the histogram distribution of the labels predicted by the classifier and visual-to-attribute mapper (f_a). It can be observed that there is large disagreement between the classes assigned by f_a and the classifier for a given visual feature. Such conditions result in higher uncertainty due to which classifier converges to sub-optimal solution. Furthermore, the histogram distribution are much better aligned for $\beta = 0.5$ showing consistency in the behaviour of both modules.

6 Conclusions

In this work, we proposed an approach for transductive ZSL to tackle the mismatch in intrinsic dimensionality during bidirectional domain alignment. In addition, we highlighted the confirmation or uncertainty bias that were prevalent while training the classifier and compensated it through entangled loss function, and then theoretically and empirically demonstrated the advantage of the proposed loss function in terms of adversarial robustness. This approach results in notable improvement in overall performance as compared to Bi-VAEGAN.

References

1. Abu-Mostafa, Y.S., Magdon-Ismail, M., Lin, H.T.: Learning From Data. AMLBook (2012)
2. Bailey, J., Houle, M.E., Ma, X.: Local intrinsic dimensionality, entropy and statistical divergences. Entropy **24**(9) (2022)
3. David, S.B., Blitzer, S., Crammer, K., Kulesza, A., F., P., Vaughan, J.W.: A theory of learning from different domains. Machine Learning **79(1-2)**, 151–175 (2010)
4. Felix, R., Vijay Kumar, B., Reid, I., Carneiro, G.: Multi-modal cycle-consistent generalized zero-shot learning. In: Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science (2018)
5. Fu, Y., Hospedales, T.M., Xiang, T., Gong, S.: Transductive multi-view zero-shot learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(11), 2332–2345 (2015)
6. Gao, R., Hou, X., Qin, J., Chen, J., Liu, L., Zhu, F., Zhang, Z., Shao, L.: Zero-vae-gan: Generating unseen features for generalized and transductive zero-shot learning. IEEE Transactions on Image Processing **29**, 3665–3680 (2020)
7. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations, ICLR (2015)
8. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30 (2017)
9. Hu, Y., Wen, G., Chapman, A., Yang, P., Luo, M., Xu, Y., Dai, D., Hall, W.: Graph-based visual-semantic entanglement network for zero-shot image recognition. IEEE Transactions on Multimedia **24**, 2473–2487 (2022)
10. Kong, X., Gao, Z., Li, X., Hong, M., Liu, J., Wang, C., Xie, Y., Qu, Y.: En-compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9296–9305 (2022)
11. Liu, F., Xu, M., Li, G., Pei, J., Shi, L., Zhao, R.: Adversarial symmetric gans: Bridging adversarial samples and adversarial networks. Neural Networks **133**, 148–156 (2021)

12. Liu, L., Zhou, T., Long, G., Jiang, J., Zhang, C.: Attribute propagation network for graph zero-shot learning. In: AAAI (2020)
13. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations, ICLR (2018)
14. Mall, U., Hariharan, B., Bala, K.: Field-guide-inspired zero-shot learning. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9526–9535 (2021)
15. Marmoreo, F., Cavazza, J., Murino, V.: Transductive zero-shot learning by decoupled feature generation. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 3108–3117 (2021)
16. Nayak, N.V., Bach, S.: Zero-shot learning with common sense knowledge graphs. Transactions on Machine Learning Research (2022)
17. Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G.S., Dean, J.: Zero-shot learning by convex combination of semantic embeddings. In: arXiv:1312.5650 (2014)
18. Paleka, D., Sanyal, A.: A law of adversarial risk, interpolation, and label noise. In: The Eleventh International Conference on Learning Representations (2023)
19. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2751–2758 (2012)
20. Raghunathan, A., Xie, S.M., Yang, F., Duchi, J., Liang, P.: Understanding and mitigating the tradeoff between robustness and accuracy. Proceedings of Machine Learning Research (2020)
21. Senel, L.K., Utlu, I., Yucesoy, V., Koc, A., Cukur, T.: Semantic structure and interpretability of word embeddings. IEEE/ACM Transactions on Audio, Speech, and Language Processing **26**(10), 1769–1779 (oct 2018)
22. Sinha, A.K., Manthira Moorthi, S., Dhar, D.: Nl-ffc: Non-local fast fourier convolution for image super resolution. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 466–475 (2022)
23. Sinha, A.K., S, M.M.: Lips-specformer: Non-linear interpolable transformer for spectral reconstruction using adjacent channel coupling. In: 34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023. BMVA (2023)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023)
25. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: Caltech-ucsd bird. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
26. Wang, Y., Zhang, H., Zhang, Z., Long, Y.: Asymmetric graph based zero shot learning. Multimedia Tools and Applications (2019)
27. Wang, Z., Hao, Y., Mu, T., Li, O., Wang, S., He, X.: Bi-directional distribution alignment for transductive zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19893–19902 (June 2023)
28. Wei, C., Shen, K., Chen, Y., Ma, T.: Theoretical analysis of self-training with deep networks on unlabeled data. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=rC8sJ4i6kaH
29. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**, 225–2265 (2018)

30. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: F-vaegan-d2: A feature generating framework for any-shot learning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10267–10276 (2019)
31. Xie1, G.S., Liu1, L., Zhu1, F., Zhao1, F., Zhang, Z., Yao, Y., Qin, J., Shao, L.: Region graph embedding network for zero-shot learning. In: Europena Conference on Computer Vision (2020)
32. Xu, B., Zeng, Z., Lian, C., Ding, Z.: Semi-supervised low-rank semantics grouping for zero-shot learning. IEEE Transactions on Image Processing **30**, 2207–2219 (2021)
33. Ye, M., Guo, Y.: Zero-shot classification with discriminative semantic representation learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5103–5111 (2017)
34. Zhao, A., Ding, M., Guan, J., Lu, Z., Xiang, T., Wen, J.R.: Domain-invariant projection learning for zero-shot recognition. In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)

7 Proof of Proposition 1

Definition 1. *Let $F : \mathbb{R}^+ \cup 0 \to \mathbb{R}^+ \cup 0$ be a function that is positive except at $F(0) = 0$. We say that F is a smooth growth function if:*

- *There exists a value $r > 0$ such that F is monotonically increasing over $(0, r)$.*
- *F is continuous over $[0, r)$.*
- *F is differentiable over $(0, r)$.*
- *The local intrinsic dimensionality D^*_F exists and is positive.*

Define $F_w(t) = \frac{F(t)}{F(w)} = P(X \leq t | X \leq w)$ for a random variable X and $w > 0$.

Definition 2. *The Normalized 2-Wasserstein distance between F and G, conditioned on $[0, w]$, is defined as,*

$$\mathcal{W}_2(F, G, w) = \frac{1}{w}\Big(\int_0^1 |F_w^{-1}(u) - G_w^{-1}|^2 du\Big)^{\frac{1}{2}} \tag{8}$$

Referring to [2],

$$\mathcal{W}_2^2(F, G, w) = \frac{1}{\frac{2}{D^*_F} + 1} + \frac{1}{\frac{2}{D^*_G} + 1} - \frac{2}{\frac{1}{D^*_F} + \frac{1}{D^*_G} + 1} \tag{9}$$

Assume that D^* is the intrinsic dimension of the true distribution $\mathcal{P}$. Furthermore, we can assume that the intrinsic dimension of the learned distribution $\mathcal{P}^{'}$ is $D^* + \delta$, where δ is the extent of error. Substituting them in 9, we get

$$\mathcal{W}_2^2(F, G, w) = \frac{2\delta^2}{(D^* + 2)(D^* + \delta + 2)(2D^* + \delta + {D^*}^2 + D^*\delta)} \tag{10}$$

8 Proof of Theorem 2

Definition 3 ($\mathcal{H}\Delta\mathcal{H}$*distance* [3]). *For two feature distributions $\mathcal{D}_g$ and $\mathcal{D}_r$, and the hypothesis class $\mathcal{H}$, the $\mathcal{H}\Delta\mathcal{H}$ distance is defined as,*

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_g, \mathcal{D}_r) = 2 \sup_{h,h' \in \mathcal{H}} |\mathbb{P}_{x\sim\mathcal{D}_g}[h(x) \neq h'(x)] - \mathbb{P}_{x\sim\mathcal{D}_r}[h(x) \neq h'(x)]|$$

It can be rewritten in terms of actual risks as,

$$\begin{aligned} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_g, \mathcal{D}_r) &= 2 \sup_{h,h' \in \mathcal{H}} |\mathbb{P}_{x\sim\mathcal{D}_g}[h(x) \neq h'(x)] - \mathbb{P}_{x\sim\mathcal{D}_r}[h(x) \neq h'(x)]| \\ &\leq 2|\mathbb{E}_{x\sim\mathcal{D}_g}[\mathbf{1}(h(x) \neq h^*(x))] - \mathbb{E}_{x\sim\mathcal{D}_r}[\mathbf{1}(h(x) \neq h^*(x))]| \\ &= 2|\epsilon_g(h, h^*) - \epsilon_r(h, h^*)| \end{aligned} \tag{11}$$

Furthermore, we refer $h^*\Delta f$-distance between $\mathcal{D}_g$ and $\mathcal{D}_r$ as $d_{h^*}(\mathcal{D}_g, \mathcal{D}_r)$.

Lemma 1 (Mostafa et. al. [1]). *For a fixed hypothesis, the actual risk $\epsilon(h, f)$ can be estimated from the empirical one $\hat{\epsilon}(h, f)$ for m samples with probability $1 - \delta$,*

$$\epsilon(h, f) \leq \hat{\epsilon}(h, f) + \sqrt{\frac{1}{2m} log \frac{1}{2\delta}} \tag{12}$$

We use a following inequality to derive the proof.

$$\begin{aligned} |\epsilon_{\mathcal{D}}(h, f) - \epsilon_{\mathcal{D}}(h, h^*)| &= |\mathbb{E}_{x\sim\mathcal{D}}\mathbf{1}(f \neq h) - \mathbb{E}_{x\sim\mathcal{D}}\mathbf{1}(h^* \neq h)| \\ &= |\mathbb{E}_{x\sim\mathcal{D}}[(\mathbf{1}(f \neq h)) - \mathbf{1}(h^* \neq h)]| \\ &\leq \mathbb{E}_{x\sim\mathcal{D}}[\mathbf{1}(h^* \neq f] = \epsilon_{\mathcal{D}}(h^*, f) \end{aligned} \tag{13}$$

We now proceed to prove Theorem.

$$\begin{aligned} \epsilon_u(h, f) =& \epsilon_s(h, f) + (\epsilon_u(h^*, f) - \epsilon_{ug}(h^*, f)) + (\epsilon_s(\psi \circ R^*, f) - \epsilon_u(\psi \circ R^*, f)) \\ & + (\epsilon_u(h, f) - \epsilon_u(h^*, f)) + \epsilon_{ug}(h^*, f) - \epsilon_s(h, f) + \epsilon_u(\psi \circ R^*, f) - \epsilon_s(\psi \circ R^*, f) \\ \leq& \epsilon_s(h, f) + |\epsilon_u(h^*, f) - \epsilon_{ug}(h^*, f)| + |\epsilon_s(\psi \circ R^*, f) - \epsilon_u(\psi \circ R^*, f)| \\ & + |\epsilon_u(h, f) - \epsilon_u(h^*, f)| + \epsilon_{ug}(h^*, f) - \epsilon_s(h, f) + \epsilon_u(\psi \circ R^*, f) - \epsilon_s(\psi \circ R^*, f) \\ \leq& \epsilon_s(h, f) + d_{h^*}(\mathcal{D}^{ug}, \mathcal{D}^u) + d_{R^*}(\mathcal{D}^u, \mathcal{D}^s) + |\epsilon_u(h, h^*) - \epsilon_s(h, h^*)| \\ & + |\epsilon_s(h, h^*) - \epsilon_s(h, f)| + \epsilon_{ug}(h^*, f) + \epsilon_u(\psi \circ R^*, f) - \epsilon_s(\psi \circ R^*, f) \\ \leq& \epsilon_s(h, f) + d_{h^*}(\mathcal{D}^{ug}, \mathcal{D}^u) + d_{R^*}(\mathcal{D}^u, \mathcal{D}^s) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}^s, \mathcal{D}^u) \\ & + [\epsilon_s(h^*, f) + \epsilon_{ug}(h^*, f) + \epsilon_u(h^*, \psi \circ R^*)] + |\epsilon_s(\psi \circ R^*, f)) - \epsilon_u(\psi \circ R^*, h^*))| \\ & - \epsilon_s(\psi \circ R^*, f)) \\ \leq& \epsilon_s(h, f) + d_{h^*}(\mathcal{D}^{ug}, \mathcal{D}^u) + d_{R^*}(\mathcal{D}^u, \mathcal{D}^s) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}^s, \mathcal{D}^u) + \lambda \\ & + |\epsilon_u(\psi \circ R^*, h^*) - \epsilon_u(h^*, \psi \circ R^*)| - \epsilon_s(\psi \circ R^*, f) \\ \leq& \epsilon_s(h, f) + d_{h^*}(\mathcal{D}^{ug}, \mathcal{D}^u) + d_{R^*}(\mathcal{D}^u, \mathcal{D}^s) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}^s, \mathcal{D}^u) + \lambda + \epsilon_u(h^*, f) \\ \leq& \hat{\epsilon}_s(h, f) + d_{h^*}(\mathcal{D}^{ug}, \mathcal{D}^u) + d_{R^*}(\mathcal{D}^u, \mathcal{D}^s) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}^s, \mathcal{D}^u) + \lambda + \epsilon_u(h^*, f) \\ & + \sqrt{\frac{1}{2m} log \frac{1}{2\delta}} \end{aligned} \tag{14}$$

9 Proof of Theorem 3

$$\begin{aligned} \mathcal{R}_B =& \mathbb{E}_{x'\sim\mathcal{B}_\rho, x\sim\mathcal{D}}[\mathbf{1}h(x) \neq h(x'))] \\ =& \mathbb{E}_{x'\sim\mathcal{B}_\rho, x\sim\mathcal{D}}[\mathbf{1}(h(x) \neq h(x'))] - \mathbb{E}_{x'\sim\mathcal{B}_\rho, x\sim\mathcal{D}}[\mathbf{1}(h(x') \neq \psi \circ R(x))] \\ & + \mathbb{E}_{x'\sim\mathcal{B}_\rho, x\sim\mathcal{D}}[\mathbf{1}(h(x') \neq \psi \circ R(x))] - \mathbb{E}_{x\sim\mathcal{D}}[\mathbf{1}(\psi \circ R(x) \neq f(x))] \\ & + \mathbb{E}_{x\sim\mathcal{D}}[\mathbf{1}(\psi \circ R(x) \neq f(x))] \\ \leq& |\mathbb{E}_{x'\sim\mathcal{B}_\rho, x\sim\mathcal{D}}[\mathbf{1}(h(x) \neq h(x'))] - \mathbb{E}_{x'\sim\mathcal{B}_\rho, x\sim\mathcal{D}}[\mathbf{1}(h(x') \neq \psi \circ R(x))]| \\ & + |\mathbb{E}_{x'\sim\mathcal{B}_\rho, x\sim\mathcal{D}}[\mathbf{1}(h(x') \neq \psi \circ R(x))] - \mathbb{E}_{x\sim\mathcal{D}}[\mathbf{1}(\psi \circ R(x) \neq f(x))]| \\ & + \mathbb{E}_{x\sim\mathcal{D}}[\mathbf{1}(\psi \circ R(x) \neq f(x))] \end{aligned} \tag{15}$$

$$
\begin{aligned}
\mathcal{R}_B \leq & \mathbb{E}_{x\sim\mathcal{D}^u}[\mathbf{1}(h(x) \neq \psi \circ R(x))] + \mathbb{E}_{x'\sim\mathcal{B}_\rho(x), x\sim\mathcal{D}^u}[\mathbf{1}(h(x') \neq f(x))] \\
& + \mathbb{E}_{x\sim\mathcal{D}^u}[\mathbf{1}(\psi \circ R(x) \neq f(x))] + \mathbb{E}_{x\sim\mathcal{D}}[\mathbf{1}(\psi \circ R(x) \neq f(x))]
\end{aligned} \tag{16}
$$

10 Dataset Details

Dataset	**#samples**	**Attribute size**	$\#\mathcal{X}^s$	$\#\mathcal{X}^u$
AWA-2	37,322	85	40	10
CUB	11,788	312	150	50
SUN	14,340	102	645	72

Table 4: Details of benchmark datasets

Hyper-parameters	AWA-2	CUB	SUN
Batch size	16	32	16
lr(classifier)	10^{-4}	10^{-3}	10^{-3}
lr (f_a)	10^{-3}	10^{-3}	10^{-3}
lr (Expert-VAEGAN)	10^{-3}	10^{-3}	10^{-3}
λ_1 100	100	100	100
λ_2	1000	1000	1000
λ_3	100	100	100
β	0.5	0.5	0.5

Table 5: Hyper-parameter settings for the datasets. lr stands for learning rate.

We demonstrate the performance of our method on three publicly available benchmark datasets, including AWA-2, SUN and CUB. Animals with Attributes-2 (AWA-2) contains total 37,322 samples belonging to 50 different classes. Each class is represented attribute vector of dimension 85. Out of 50 classes, 40 of them are available as as seen categories and rest of them are test labels. Caltech UCSD Bird (CUB) dataset contains 11,788 samples of 200 different bird species with attribute size of 312. The SUN scene classification consists of 14,340 samples for 717 types of scenes and has attribute size of 102. Details are also provided in table 4.

10.1 Implementation and training details

We perform our experiments in PyTorch 1.13 using Nvidia A100 GPU. For ChebNet, we directly use Pytorch-geometric in the implementation. The hyper-parameters settings for all three datasets are provided in Table 5. We use Adam optimizer to train the models with β_1 = 0.5 and β_2 =0.99.