Zero-Shot Learning by Harnessing Adversarial Samples

Zhi Chen[1], Pengfei Zhang[1], Jingjing Li[2], Sen Wang[1], Zi Huang[1]
[1]School of Electrical Engineering and Computer Science, The University of Queensland, Australia
[2]School of Computer Science and Engineering, University of Electronic Science and Technology of China
{zhi.chen,pengfei.zhang,sen.wang}@uq.edu.au,lijin117@yeah.net,huang@itee.uq.edu.au

ABSTRACT

Zero-Shot Learning (ZSL) aims to recognize unseen classes by generalizing the knowledge, *i.e.,* visual and semantic relationships, obtained from seen classes, where image augmentation techniques are commonly applied to improve the generalization ability of a model. However, this approach can also cause adverse effects on ZSL since the conventional augmentation techniques that solely depend on single-label supervision is not able to maintain semantic information and result in the *semantic distortion* issue consequently. In other words, image argumentation may falsify the semantic (e.g., attribute) information of an image. To take the advantage of image augmentations while mitigating the semantic distortion issue, we propose a novel ZSL approach by Harnessing Adversarial Samples (HAS). HAS advances ZSL through adversarial training which takes into account three crucial aspects: **(1) robust generation** by enforcing augmentations to be similar to negative classes, while maintaining correct labels, **(2) reliable generation** by introducing a latent space constraint to avert significant deviations from the original data manifold, and **(3) diverse generation** by incorporating attribute-based perturbation by adjusting images according to each semantic attribute's localization. Through comprehensive experiments on three prominent zero-shot benchmark datasets, we demonstrate the effectiveness of our adversarial samples approach in both ZSL and Generalized Zero-Shot Learning (GZSL) scenarios. Our source code is available at https://github.com/uqzhichen/HASZSL.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

zero-shot learning, adversarial training

ACM Reference Format:
Zhi Chen[1], Pengfei Zhang[1], Jingjing Li[2], Sen Wang[1], Zi Huang[1]. 2018. Zero-Shot Learning by Harnessing Adversarial Samples. In *Proceedings of Ottawa '23: ACM MULTIMEDIA (MM '23).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

MM '23, October 29– November 2, 2023, Ottawa, Ontario, Canada

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

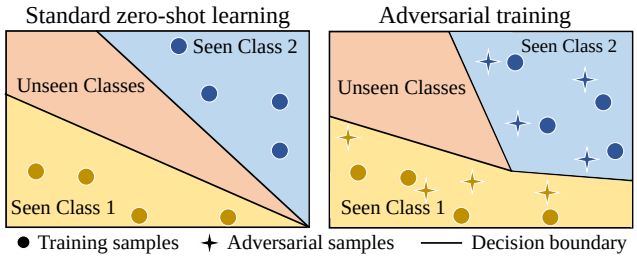


Figure 1: Our method employs adversarial training to direct the ZSL model towards generating augmentations that are conditioned on attributes to diversify the training images.

1 INTRODUCTION

Traditional image recognition systems operate within a predefined label space, predicting categories based on a predetermined set of labels. To expand beyond this limited label space and accommodate emerging labels, one option is to laboriously retrain the recognition model using data associated with the new labels. Zero-Shot Learning (ZSL) [19, 41] offers an alternative solution, presenting a flexible and scalable approach to incorporate new classes into the model without requiring retraining. This adaptability allows ZSL to efficiently accommodate novel categories, making it a more effective solution for expanding the capabilities of image recognition systems.

ZSL entails associating visual patterns with semantic vectors, which are typically annotated attributes. These attributes consist of multiple dimensions, each representing a semantic feature of the objects being analyzed. For example, one dimension could represent the "breast color" of a bird, with the value being "red". Many research efforts [4, 21, 44] have enabled ZSL methods to exhibit a strong ability to localize, effectively allowing them to attend to the regions of the images that are most relevant to the semantic meanings in the attributes. This capability of ZSL methods to localize and attend to the relevant regions of the images is a key factor in their improved accuracy and effectiveness.

ZSL is fundamentally an image recognition task, where image augmentation techniques are commonly utilized to enhance the generalization ability. However, we observe that image augmentations can cause detrimental effects on ZSL. As illustrated in Figure 2, we deduce two attributes (e.g., Blue Breast and Red Breast) for three bird categories using the ZSL model APN [44]. The model accurately predicts the breast color of the three bird categories without augmentation. Nevertheless, when image augmentations are applied to raw images, the model produces incorrect values for the two attributes, causing the categories to become indistinguishable. We have also conducted a series of experiments, involving various types of image augmentation techniques in training a ZSL model, quantitatively observing performance drop. We term the problem caused

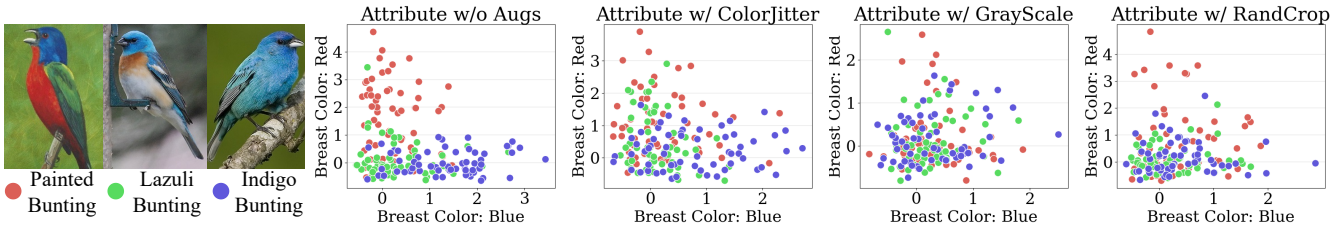


Figure 2: A close look at the semantic distortion problem. We observe that traditional image augmentation techniques lead to confusion in the ZSL model, causing it to generate incorrect attributes. This, in turn, results in indistinguishable attribute space.

by direct pixel manipulation on images as ***semantic distortion***. The causes of this problem can be attributed to two main factors: (1) Conventional image augmentation applies a set of *label-preserving* transformations, which, despite improving generalization ability, can introduce semantic distortion issues due to their lack of *semantic preservation*. This is because these data augmentation strategies [32, 35, 48] apply direct pixel manipulation on images, which can modify the semantic essence of the data, conflicting with ZSL objectives. Unlike supervised learning, which relies on a single label for guidance, ZSL demands dense semantics as strong supervision, making aggressive image augmentation susceptible to creating semantic ambiguity and compromising the semantic coherence prioritized in ZSL. (2) An intriguing property of neural networks is that the entire space of activations, rather than individual units, contains the bulk of semantic information [34], making the entire space less sensitive to pixel manipulation. In ZSL, however, one can inspect the final individual units (i.e., attribute prediction), which usually attend to regions containing a specific semantic meaning [21, 44]. Consequently, subtle pixel changes can lead to incorrect predictions in individual units. Based on these limitations, exploring alternative methods for enhancing ZSL performance without compromising semantic coherence is essential.

In light of the controllable image manipulation achieved by adversarial training [15, 16, 43], we resort to learning adversarial samples as augmentations to diversify the visual space without causing semantic distortion problem, as shown in Figure 1. However, we identify three critical challenges of learning adversarial samples for ZSL. First, conventional adversarial training creates adversarial samples by maximizing the classification cross-entropy loss, which is an explicit way of perturbing images to confuse the model to make wrong predictions. However, the goal of generating adversarial samples for zero-shot learning is not to leverage the adversarial samples to confuse the model, but rather to facilitate the generalization ability on diverse images. Second, the learned visual features correspond to the predicted attributes. The latent space requires explicit constraint to prevent significant shifts from the original space while perturbing the original images. Third, ZSL models exhibit strong localization ability, but how to leverage the localization ability when generating adversarial samples remains unclear. To this end, we propose a novel ZSL approach by Harnessing Adversarial Samples (HAS). Specifically, to address the first robustness issue, bearing in mind that we still want to ensure the model recognizes the adversarial samples, but more importantly, perceives the differences, we propose to make the images visually similar to negative classes based on the model's understanding of differences among classes. This is achieved by maximizing the entropy of the classification probability so that the model could assign higher probabilities to other classes. To ensure the model recognizes the adversarial samples, the classification loss still needs to be minimized, so that the model could still perceive the adversarial images as correct ones. Second, to prevent significant shifts from the original space, we use the visual features of the original images to constrain the adversarial learning, so that the learned adversarial samples cannot produce significantly different visual features that cause the semantic distortion problem. Lastly, considering the attention maps on the last layer exhibit the localization probability of different attributes, we propose to explicitly perturb the attention maps by making the model give weak responses to the attribute regions by applying entropy loss and minimizing the total probability magnitude. Through iterative training with the adversarial samples, we can enhance the localization ability of a ZSL model. Through comprehensive experiments on three prominent zero-shot benchmark datasets, we demonstrate the effectiveness of our adversarial samples approach in both ZSL and GZSL contexts. Our results highlight the ability of HAS in overcoming the limitations associated with traditional augmentation techniques within the ZSL domain. Furthermore, the proposed method provides a more coherent approach to image augmentation, which preserves the semantic information of the data while enhancing the model's generalization ability. The contribution of this work can be summarized as follows:

- We introduce a novel approach, Harnessing Adversarial Samples (HAS), to improve the generalization ability of ZSL models. HAS generates adversarial samples through controllable image perturbation, addressing the semantic distortion issue found in traditional image augmentations.
- We consider three aspects in learning adversarial samples for ZSL. 1. Our method facilitates **robust generation** by maximizing the entropy of classification probabilities while minimizing cross-entropy loss, making images visually similar to negative classes without compromising model recognition. 2. We achieve **reliable generation** by utilizing the visual features of original images to stabilize adversarial learning, preventing significant shifts from the original visual space that could cause semantic distortion. 3. **Diverse generation** is promoted by leveraging the ZSL model's localization ability to perturb attention maps, enhancing the model's ability to focus on attribute regions.
- Through comprehensive experimental study, we demonstrate that adversarial samples with proper learning objectives can effectively improve the generalization ability of ZSL models.

2 RELATED WORK

2.1 Zero-Shot Learning

The ZSL task [7, 19, 28, 40, 41] consists in generalizing the knowledge learned from seen classes to unseen classes. With this goal, embedding-based methods learn a visual-to-semantic regression function by transforming the visual features to semantic information, *e.g.,* attributes[14], documents[13], w2v embeddings[24]. Zhu *et al.* [50] proposed a Semantic-Guided Multi-Attention localization model (SGMA) that learns to discover the local features of the discriminative patches, which demonstrates strong localization performance with human annotations. The multi-attention loss encourages compact and diverse attention distribution by applying geometric constraints over attention maps. Xu *et al.* [44] proposed an Attribute Prototype Network (APN) that learns prototypes for each of the attributes. The learnable attribute prototypes are used as the 1×1 filter to convolve the feature maps with the feature maps from ResNet101 in order to attend to the region of interest. As an extension to APN [44], to learn local features, Liu *et al.* [21] proposed using the GloVe [26] model to learn local features and to extract semantic vectors from the attribute names (*e.g.,* "plain head"), which further represent the attribute prototypes. Chen *et al.* [4] proposed a mutually semantic distillation network to progressively distill the intrinsic semantic representations between and visual and attribute features. The other mainstream ZSL direction leverages generative models to hallucinate unseen visual features conditioned on the semantic information [6, 7, 9, 10, 40, 45]. There are various generative models explored in this area, including GANs [33, 40, 49], VAEs [29, 38], Flows [8, 31]. Despite the success of image augmentation in supervised learning, its effectiveness in ZSL is still an open research problem. ZSL requires explicit supervision, *i.e.,* attributes, which are very sensitive to subtle visual changes. In this context, we propose a novel approach that leverages adversarial samples as augmented images to improve ZSL performance.

2.2 Adversarial Samples

Adversarial samples, formed by adding imperceptible perturbations to images, can lead models to make wrong predictions. Attacks by adversarial samples [12, 34] pose a security concern in the field of recognition tasks. For generating adversarial examples in deep learning models, Fast Gradient Sign Method (FGSM) [15] is a simple and effective method. The paper provides a detailed explanation of FGSM and its applications, and it is considered a foundational work in the field of adversarial machine learning. The Worrisome is the phenomenon of adversarial examples [1], imperceptibly perturbed natural inputs that induce erroneous predictions in state-of-the-art classifiers. While early arts [27, 36] suggest that there is a tradeoff between the robustness and the accuracy in recognition systems, there are works [16, 43] have shown that adversarial samples can also bring performance improvement. Our method further validates that adversarial features are also beneficial for ZSL recognition models, which agree with the conclusions drawn from these aforementioned studies. Shafiee *et al.* [30] proposed Attribute-based Universal Perturbation Generator to generate adversarial samples to attack zero-shot learning models. While our work focuses on leveraging adversarial samples to improve the zero-shot performance.

3 METHODOLOGY

3.1 Preliminaries

Assume a dataset $\mathcal{D}$ could be divided into two sets of classes $\mathcal{D}^s$ and $\mathcal{D}^u$, representing the seen classes and the unseen classes. The seen classes are used for training the ZSL model, i.e., $\mathcal{D}^s = \{I, y, \phi(y) | I \in \mathcal{I}, y \in \mathcal{Y}^s\}$ from seen classes $\mathcal{Y}^s$, where I is an image in the RGB image space $\mathcal{I}$, y is the associated class label and $\phi(y) \in \mathbb{R}^K$ is the class-level semantic embedding annotated with K different semantic attributes. For the unseen classes $\mathcal{D}^u$, the class samples are unknown during training. The training procedure consists of two parts, i.e., standard ZSL training and adversarial training. In standard training, we update the weights of the model to learn to predict the attributes of a given image and further infer the corresponding class, either from seen or unseen classes. In the adversarial training, we fix the model weights and update the training images via gradient ascent so that the adversarial samples can be yielded. The adversarial samples are then fed into the model for training with legitimate samples.

3.2 Learning ZSL Model

We follow the baseline method Attribute Prototype Network (APN) [44] to train a standard ZSL model. Given an image I, the ResNet backbone produces the feature maps $x = f(I) \in \mathbb{R}^{C\times H\times W}$, where C, H and W represent the channel, height, and width respectively. The average pooling is performed on the feature maps to generate dense features $g(x) \in \mathbb{R}^C$. Then, a linear transformation $V \in \mathbb{R}^{C\times K}$ will be applied to predict the attribute vector $\hat{a} = g(x)^T V$. The classification loss function is then formulated as:

$$\mathcal{L}_{CLS} = -\log \frac{\exp(g(x)^T V \phi(y))}{\sum_{\hat{y}\in \mathcal{Y}^s} \exp(g(x)^T V \phi(\hat{y}))}. \tag{1}$$

To further improve the localization ability of the ZSL model, a convolutional layer $CV \in \mathbb{R}^{C\times K}$ is performed on the feature maps $\boldsymbol{x}$. The filter size in the kernel is set to 1×1. This operation results in the attribute attention map $h(x) \in \mathbb{R}^{K\times H\times W}$. This operation localizes the semantic information of each attribute on each attention map. A max pooling operation is further applied on the attribute attention map to generate the attribute vector $\hat{a} \in \mathbb{R}^K$. Lastly, the attention localization loss can be formulated as:

$$\mathcal{L}_{LOC} = \|\hat{a} - \phi(y)\|_2^2, \tag{2}$$

where the mean squared error is calculated between the predicted attribute vectors and the ground truth attributes.

3.3 Generating Adversarial Augmentations

As discussed in the introduction section, the *semantic distortion* problem poses a challenge in diversifying the visual space with traditional image augmentation techniques. Thus, we aim to devise a *semantic-preserving* augmentation strategy, which diversifies the visual space while avoiding causing the semantic distortion problem.

Inspired by Fast Gradient Sign Method (FGSM) [15], learning perturbation on images creates adversarial samples. We develop a semantic-preserving algorithm that performs controllable image augmentations by perturbing images with gradient descent. However, adversarial samples created by FGSM aim to confuse the model to make wrong predictions. In contrast, in our method, the goal is to allow the model to conditionally diversify the images while the

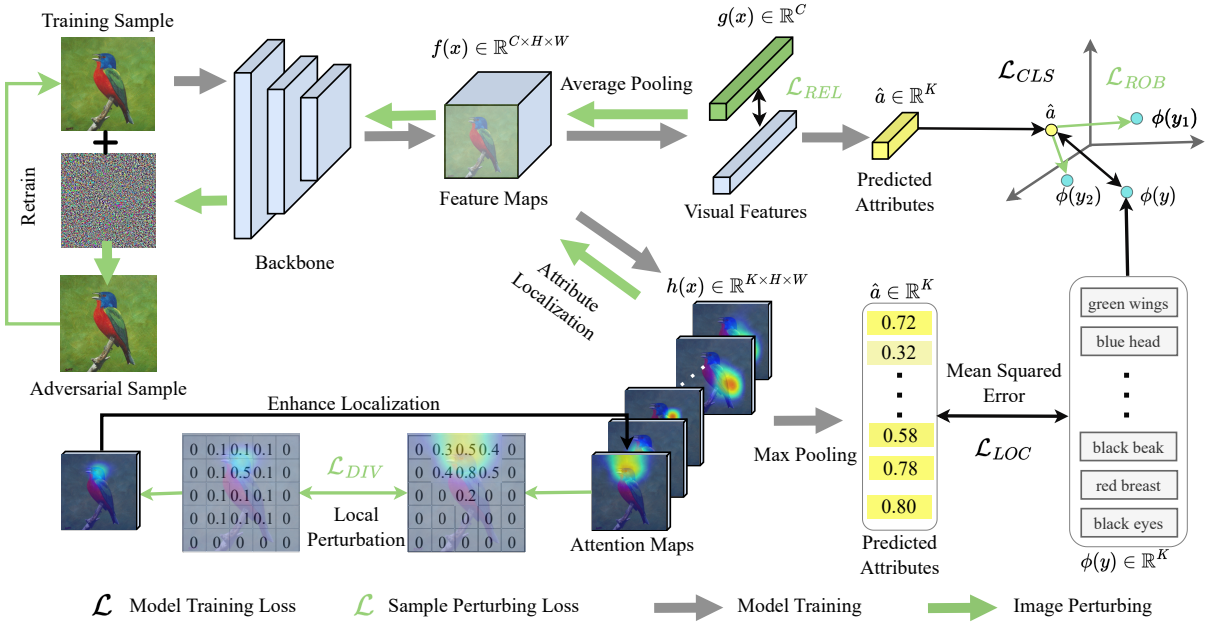


Figure 3: An illustration of our proposed HAS, which leverages adversarial training to guide the ZSL model in creating attribute-conditioned augmentations to diversify the visual space. Using the APN baseline method [44] for standard ZSL training, global visual features are projected into the attribute space for classification, and attribute attention maps enhance localization ability. Combining adversarial training with the baseline method, we generate attribute-conditioned image augmentations, resulting in a robust ZSL model that generalizes better to unseen classes while maintaining performance on known classes.

image should maintain the attribute prediction accuracy. Our method does not involve additional parameters in the model for generating adversarial samples. Instead, the model introduced in Section 3.2 is used to provide guidance with different objective functions to update the input images. The overall loss function is formulated as follows:

$$\mathcal{L}_{HAS} = \underbrace{\mathcal{L}_{CLS} - \lambda_1 \mathcal{L}_{ROB}}_{Robustness} + \underbrace{\lambda_2 \mathcal{L}_{REL}}_{Reliability} - \underbrace{\lambda_3 \mathcal{L}_{DIV}}_{Diversity}, \tag{3}$$

where $\mathcal{L}_{CLS}$ and $\mathcal{L}_{ROB}$ ensure robust generation by maintaining correct label information, $\mathcal{L}_{REL}$ facilitates reliable generation by preventing the images from semantic distortion, and $\mathcal{L}_{DIV}$ promotes diverse generation by perturbing the local attention maps.

Given the objective function $\mathcal{L}_{HAS}$, we iteratively generate the adversarial samples I^{ad} with FGSM:

$$I_0^{ad} = I,\ I_{t+1}^{ad} = I_t^{ad} + \epsilon \operatorname{sign} \nabla_{I_t^{ad}} \mathcal{L}_{HAS}(\boldsymbol{W}, I_t^{ad}, \phi(y)), \tag{4}$$

where ϵ is the perturbation strength, $\boldsymbol{W}$ is the overall model parameters depicted in Section 3.2, sign function applies a max-norm constraint on the gradients.

3.4 Robust Generation

To generate diverse adversarial samples without causing the model to misclassify, we propose to perturb the images to appear similar to negative classes. To do so, we resort to entropy maximization on the classification probabilities. The model's output probabilities, obtained after the Softmax layer, represent the confidence level for the predicted classes. By maintaining the highest confidence for the correct class while maximizing probability entropy across all classes, we enforce the model to update the image until it visually resembles other classes. The loss function that maintains robustness is formulated as follows:

$$\mathcal{L}_{ROB} = -\Sigma_{y\in \mathcal{Y}^s} p_y \log(p_y),\ p_y = \frac{\exp(g(x)^T V \phi(y))}{\sum_{\hat{y}\in \mathcal{Y}^s} \exp(g(x)^T V \phi(\hat{y}))} \tag{5}$$

where p_y represents the probability of assigning the sample to class y. When generating adversarial samples for attacking a model [15], it is common to apply gradient ascent on the images with classification loss, *e.g.,* cross-entropy loss. This learning objective optimizes the images to mislead the model into making incorrect predictions. However, in our work, we aim to gently adjust the decision boundary in response to the visual perturbation, rather than provoke erroneous predictions. To achieve this, we still need to minimize the classification loss $\mathcal{L}_{CLS}$, which helps avoid drastic visual changes that might confuse the model and lead to improper boundary adjustments. By focusing on this objective, our approach promotes the generation of diverse adversarial samples that appear similar to negative classes while preserving the model's ability to classify correctly. This enables us to explore the model's sensitivity to visual perturbations and refine its decision boundaries without compromising its overall performance.

3.5 Reliable Generation

To further facilitate reliable perturbations applied to the images, we address the drift problem occurring in the latent space. The drift problem refers to a situation where an image produces the expected class probabilities according to the supervision provided, but its learned representations in the latent space undergo significant changes. Such

Algorithm 1 Harnessing Adversarial Samples for ZSL

Input: seen dataset $\mathcal{D}^s$, training epoch E, batch size N^b, learning rate η
Initialize: Model weights $\boldsymbol{W}$

for epoch $e = 0, 1, \ldots, E-1$ **do**
  for batch $i = 0, 1, \ldots, B-1$ **do**
    Randomly select a batch $\{I, y, \phi(y)\}^{N^b}$
    Compute $\mathcal{L}_{CLS} + \mathcal{L}_{LOC}$ # Standard ZSL training
    Update model $\boldsymbol{W}_i \leftarrow \boldsymbol{W}_{i-1} - \eta \bigtriangledown (\mathcal{L}_{CLS} + \mathcal{L}_{LOC})$
    for $t = 0, 1, \ldots, T-1$ **do** # Perturb samples for T times
      Compute adversarial loss using $\mathcal{L}_{HAS}$ in Eq. 3
      Update adversarial samples I^{ad}_{t+1} using Eq. 4
    end for
    Compute $\mathcal{L}_{CLS} + \mathcal{L}_{LOC}$ with I^{ad}_{T} # Adversarial training
    Update model $\boldsymbol{W}_i \leftarrow \boldsymbol{W}_{i-1} - \eta \bigtriangledown (\mathcal{L}_{CLS} + \mathcal{L}_{LOC})$
  end for
end for
Return model weights $\boldsymbol{W}$

changes affect the model's ability to generalize effectively and lead to instability in the learning process. To mitigate this issue, we introduce a constraint that aims to prevent the learned representations from deviating too much from the original ones. By doing so, we ensure that the latent features remain reliable and consistent. This constraint helps the model maintain a more robust and meaningful representation of the data in the latent space, which can enhance its overall performance and ability to generalize to unseen classes. The loss function that facilitates reliability can be formulated as:

$$\mathcal{L}_{REL} = \|g(f(I)) - g(f(I^{ad}_t))\|^2_2 \tag{6}$$

where $g(f(I))$ represents the dense visual features generated by the model, and $g(f(I^{ad}_t))$ represents that of the perturbed images.

3.6 Diverse Generation

To further diversify the adversarial samples, we propose perturbing the images based on localized attributes. In attention maps, $h(x) \in \mathbb{R}^{K\times H\times W}$, each map $h(x)_k \in \mathbb{R}^{H\times W}$ corresponds to the presence of a specific attribute. By slightly shifting the presence of attributes within the images, we can achieve perturbation with respect to each individual attribute.

We can interpret the attention map as a probability distribution over regions, where the highest probability represents the attribute value. In other words, the region with the highest probability is the area to which the model exhibits the greatest sensitivity. Keeping the model fixed, an image can be altered locally as local responses change. Similar to $\mathcal{L}_{ROB}$, we employ entropy maximization to equalize the response values of different regions. Moreover, to make the sample more challenging to learn, we propose suppressing the overall attention weights by adding a regularization term to the attention maps. The diversity loss can be formulated as follows:

$$\mathcal{L}_{DIV} = \Sigma_{k\in K}\|h(f(I^{ad}_t))_k\|^2_2 - h(f(I^{ad}_t))_k \log h(f(I^{ad}t))_k, \tag{7}$$

where $h(f(I^{ad}_t))_k$ is the k-th attention map for k-th attribute. By incorporating localized attribute perturbation and entropy maximization, the proposed method enhances the ZSL model's ability to adapt to localized changes in the input images while maintaining robustness to adversarial perturbations.

3.7 Model Training and Zero-Shot Prediction

To facilitate a clear understanding of the overall model training and zero-shot prediction process, we present the training process in Algorithm 1. For each batch of the training data, we initally perform standard training on the model using the original data, adhering to the procedures described in Section 3.2. Subsequently, we fix the model and compute gradients on the images for T steps, generating the adversarial samples. These adversarial samples, containing attribute-conditional visual variance, are then fed into the model for ZSL training. Upon completion of the model training, the model gains ZSL capability.

For zero-shot prediction, given a test image I^u from unseen classes, we can predict the corresponding attributes and compute the compatibility score with the candidate attribute vectors. The label exhibiting the highest compatibility becomes the predicted class label:

$$\hat{y^u} = \underset{y^u\in\mathcal{Y}^u}{\arg\max}[g(f(I))^T \boldsymbol{V}\phi(y^u)] \tag{8}$$

For generalized zero-shot learning (GZSL), test images may originate not only from unseen classes but also from seen classes. As the model is trained exclusively on seen classes only, the prediction will inevitably exhibit bias towards seen classes. Following existing methods[2, 21, 44], calibrated stacking is employed to reduce the probability score on seen classes by a calibration factor μ. Consequently, the GZSL prediction is formulated as:

$$\hat{y} = \underset{y\in\mathcal{Y}}{\arg\max}[(g(f(I))^T \boldsymbol{V}\phi(y)) - \mu\mathbb{1}(y \in \mathcal{Y}^s)], \tag{9}$$

where $\mathbb{1}(\cdot)$ is an indicator function that determines whether a label originates from seen or unseen classes.

4 EXPERIMENTS

4.1 Datasets and Implementation Details

We conduct experiments on three widely used benchmark ZSL datasets of image classification. They are **CUB** [37], **AwA2** [20] and **SUN** [25]. CUB consists of 11,788 images from 200 bird species, of which 150 are selected as seen classes and 50 as unseen classes. Each class is annotated with 312 attributes. AWA2 is a considerably larger dataset with 30,475 images from 50 classes, and they are annotated with 85 attributes. SUN dataset has 14,340 images of 717 scene classes, of which 645 classes and the rest of 72 classes are chosen as seen and unseen classes, each class has 101 attributes.

Following compared methods [4, 21, 44], we adopt ResNet101 pretrained on ImageNet1k as our CNN backbone to extract feature maps. We use Adam optimizer with beta1 set to 0.5 and beta2 set to 0.999. The learning rate is initialized as 10^4 and decreased every ten epochs by a factor of 0.8. The batch size is set to 64. The loss weights λ_1, λ_2, λ_3 and the perturbation strength ϵ is set to between {0.01, 1.0}, {0.1, 10.0}, {1e-5, 1e-3} and {1,8}, respectively.

To avoid the failure of classification accuracy for imbalanced class distributions, we adopt average per-class Top-1 accuracy as the fair evaluation criteria for conventional ZSL and the seen and

Table 1: Performance comparison in accuracy (%) of the state-of-the-art ZSL and GZSL on three datasets. For ZSL, performance results are reported with the average top-1 classification accuracy (T1). For GZSL, results are reported in terms of top-1 accuracy of unseen (U) and seen (S) classes, together with their harmonic mean (H). The best and second-best results are marked in Red and Blue. APN* represents training with an image size of 448x448.

Methods	CUB				AwA2				SUN			
	T1	*U*	*S*	*H*	*T1*	*U*	*S*	*H*	*T1*	*U*	*S*	*H*
Generative Methods												
f-CLSWGAN(CVPR'18) [40]	57.3	43.7	57.7	49.7	65.3	56.1	65.5	60.4	60.8	42.6	36.6	39.4
CADA-VAE(CVPR'19) [29]	60.4	51.6	53.5	52.4	64.0	55.8	75.0	63.9	61.7	47.2	35.7	40.6
f-VAEGAN-D2(CVPR'19) [42]	61.0	48.4	60.1	53.6	71.1	57.6	70.6	63.5	64.7	45.1	38.0	41.3
TF-VAEGAN(ECCV'20) [23]	64.9	52.8	64.7	58.1	72.2	59.8	75.1	66.6	66.0	45.6	40.7	43.0
E-PGN(CVPR'20) [45]	72.4	52.0	61.1	56.2	73.4	52.6	83.5	64.6	-	-	-	-
SDGZSL(ICCV'21) [7]	75.5	59.9	66.4	63.0	72.1	64.6	73.6	68.8	62.4	48.2	36.1	41.3
HSVA(NeurIPS'21) [5]	62.8	52.7	58.3	55.3	-	59.3	76.6	66.8	63.8	48.6	39.0	43.3
Embedding-based Methods												
SP-AEN(CVPR'18) [3]	55.4	34.7	70.6	46.6	58.5	23.3	90.9	37.1	59.2	24.9	38.6	30.3
TCN(ICCV'19) [18]	59.5	52.6	52.0	52.3	71.2	61.2	65.8	63.4	61.5	31.2	37.3	34.0
DVBE (CVPR'20) [22]	-	53.2	60.2	56.5	-	63.6	70.8	67.0	-	45.0	37.2	40.7
APN(NeurIPS'20) [44]	72.0	65.3	69.3	67.2	68.4	56.5	78.0	65.5	61.6	41.9	34.0	37.6
APN*(NeurIPS'20) [44]	75.6	68.6	71.6	70.1	69.8	60.1	86.5	71.0	62.6	42.8	37.7	40.1
GEM(CVPR'21) [21]	77.8	64.8	77.1	70.4	67.3	64.8	77.5	70.6	62.8	38.1	35.7	36.9
MSDN (CVPR'22) [4]	76.1	68.7	67.5	68.1	70.1	62.0	74.5	67.7	65.8	52.2	34.2	41.3
HAS (ours)	76.5	69.6	74.1	71.8	71.4	63.1	87.3	73.3	63.2	42.8	38.9	40.8

Table 2: Ablation study for different components of HAS.

Methods	CUB		AwA2		SUN	
	T1	*H*	*T1*	*H*	*T1*	*H*
Baseline(APN*)	75.6	70.1	69.8	71.0	62.6	40.1
$+\mathcal{L}_{CLS}$	74.3	68.9	68.6	70.2	59.2	37.8
$+\mathcal{L}_{CLS} + \mathcal{L}_{ROB}$	74.7	68.5	68.5	69.3	60.1	37.9
$+\mathcal{L}_{CLS} + \mathcal{L}_{DIV}$	69.2	68.2	67.6	69.9	60.8	38.4
$+\mathcal{L}_{CLS} + \mathcal{L}_{ROB} + \mathcal{L}_{REL}$	76.0	71.5	71.2	72.6	62.9	40.3
$+\mathcal{L}_{CLS} + \mathcal{L}_{DIV} + \mathcal{L}_{REL}$	76.1	70.9	70.7	72.8	61.8	40.7
HAS (ours)	**76.5**	**71.8**	**71.4**	**73.3**	**63.2**	**40.8**

unseen set performance in GZSL:

$$Acc_{\mathcal{Y}} = \frac{1}{|\mathcal{Y}|} \sum_{y=1}^{|\mathcal{Y}|} \frac{\#\ of\ correct\ predictions\ in\ y}{\#\ of\ samples\ in\ y}, \tag{10}$$

where $|\mathcal{Y}|$ is the number of testing classes. A correct prediction is defined as the highest probability of all candidate classes. Following [39], the harmonic mean of the average per-class Top-1 accuracies on seen Acc_S and unseen Acc_U classes are used to evaluate the performance of generalized zero-shot learning. It is computed by:

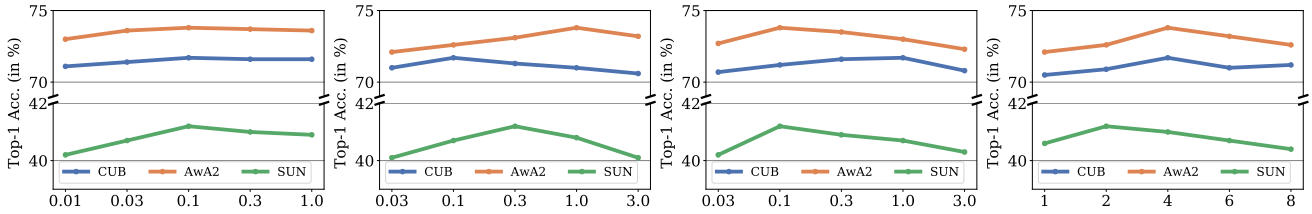
$$H = \frac{2 * Acc_S * Acc_U}{Acc_S + Acc_U}. \tag{11}$$

4.2 Comparison with the State-of-the-Art

We selected recent state-of-the-art ZSL methods for comparison, which include generative methods f-CLSWGAN [40], CADA-VAE [29], f-VAEGAN-D2 [42], TF-VAEGAN [23], E-PGN [45], SDGZSL [7], HSVA [5], and embedding-based methods SP-AEN [3], TCN [18], DVBE [22], APN [44], GEM [21] and MSDN [4]. Table 1 presents the performance comparison with these state-of-the-art ZSL methods on three benchmark datasets. Notably, in the original APN implementation, the image size used is 224x224. To facilitate a fair comparison with GEM and MSDN, which utilize an image size of 448x448, we have reproduced APN with the larger image size and report the performance for this modified version as APN*. For GZSL, our methods HAS achieves the best results of 71.8% and 73.3% on CUB and AwA2 datasets. Compared with our baseline method APN, we achieve consistent performance improvement on all three datasets and both conventional and generalized settings, which demonstrates the effectiveness of adversarial samples. As the SUN dataset involves a substantial number of classes, *i.e.,* 717, generative methods can better generalize to unseen classes than embedding-based methods. However, we still achieve comparable performance with other embedding-based methods.

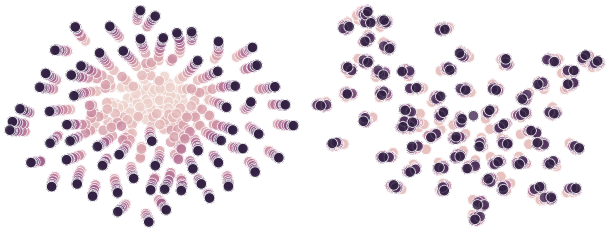
4.3 Ablation Study

To analyze the contribution of each perturbation component, we conduct an ablation study on the complete adversarial training strategy. As shown in Table 2, we decompose our complete training strategy into seven different combinations. Baseline(APN*) is the standard ZSL model APN trained with the image size of 448×448 without adversarial training. In $+\mathcal{L}_{CLS}$, we incorporated adversarial training and generate adversarial samples with the classification loss only. This variant slightly worsens the ZSL performance on the three datasets, which confirms the conclusion drawn from [27, 36] the robustness and accuracy tradeoff in the vanilla adversarial training. In both $+\mathcal{L}_{CLS}+\mathcal{L}_{ROB}$ and $+\mathcal{L}_{CLS}+\mathcal{L}_{DIV}$, the robustness loss cannot improve the performance. This confirms the utility of the reliability loss $\mathcal{L}_{MEA}$, which is necessary to prevent semantic distortion when



(a) Reliability weight λ_1 **(b) Robustness weight λ_2** **(c) Diversity weight λ_3** **(d) Perturbation strength ϵ**

Figure 4: Hyper-parameter sensitivity. The horizontal axis indicates the varying hyper-parameters for (a) Reliability weight λ_1, (b) Robustness weight λ_2, (c) Diversity weight λ_3 and (d) Perturbation strength ϵ.



(a) w/o $\mathcal{L}_{REL}$ **(b) w/ $\mathcal{L}_{REL}$**

Figure 5: Visualization of the visual features. Darker colors represent visual features in the later perturbation stages. (a) The drastic changes without constraint to visual features lead to semantic distortion. (b) $\mathcal{L}_{REL}$ avoids the visual features shifting from the original positions.

perturbing the images. We also visualize the difference of visual features with and without the reliability loss $\mathcal{L}_{REL}$ as shown in Figure 5. We perturb 64 random images for 10 times, darker colors represent later perturbation stages. The visual features significantly shift from the original space without the constraint loss, while this issue is effectively mitigated when applying $\mathcal{L}_{REL}$. We also report the performance of the two variants without $\mathcal{L}_{DIV}$ or $\mathcal{L}_{ROB}$, both of which can achieve performance increase on the baseline method. When combining all the components, we achieve the best performance.

4.4 Hyper-parameter Sensitivity

In a series of experiments, we investigate the influence of the four hyper-parameters on the performance of our proposed method. Figure 4 (a) shows the effect of varying the constraint weight λ_1 from 0.01 to 1.0. As the reliability weight increases, the overall performance also improves. However, the performance saturation occurs at around 0.1, and larger reliability weights tend to decrease performance due to over-controlled perturbation. Figure 4 (b) demonstrates that the robustness weight is a considerably sensitive hyper-parameter. We observe that coarse-grained datasets tend to perform better when the robustness weight is high. In Figure 4 (c), we vary the diversity weight λ_3 and found that AwA2 and SUN both achieve the best performance at 0.1, while CUB peaks at 1.0. The perturbation strength can be considered the learning rate of generating adversarial samples. In Figure 4 (d), CUB and AwA2 achieve the best performance at 4, while SUN peaks at 2.

Table 3: Effects of different traditional augmentation techniques on the CUB dataset.

	CUB		AwA2		SUN	
	CZSL	GZSL	CZSL	GZSL	CZSL	GZSL
Baseline	72.0	67.2	68.4	67.4	61.6	37.6
ColorJitter0.2	66.8	61.1	66.5	65.1	59.9	37.4
ColorJitter0.4	67.3	60.9	66.6	64.9	59.7	37.0
GrayScale0.2	69.6	65.1	68.0	66.8	60.1	37.1
GrayScale0.4	68.4	63.3	66.6	66.0	59.9	37.0
GaussianBlur(L)	71.1	65.9	69.1	65.7	60.8	37.0
GaussianBlur(H)	68.7	62.1	62.5	56.2	58.5	34.3
RandomRotate	65.8	59.5	56.6	54.8	46.7	21.7
RandomCrop	67.0	60.8	60.8	58.8	36.1	14.9
CutOut	70.5	65.5	62.5	60.2	60.1	35.2
MixUp	66.4	59.5	38.9	39.1	51.8	28.2
CutMix	68.5	62.4	53.6	42.0	57.2	29.8
SnapMix	69.0	61.1	51.3	45.4	56.5	30.1

4.5 Traditional Image Augmentation Results

We explore a variety of representative image augmentation techniques in the training of the ZSL model APN, including Color Jitter, Grayscale, Gaussian Blurring, random rotation and random crop. For each augmentation, we apply both mild and strong augmentation to see their effects. For random rotation, the rotation degree ranges from 0 and 360 degrees. For random crop, we crop half the size of the original images. As reported in Table 3, there is a consistent performance drop across all datasets and in both conventional zero-shot learning and generalized zero-shot learning settings. Notably, we take one step further by exploring state-of-the-art data mixing augmentation strategies, going beyond traditional techniques. We examine methods such as CutOut [11], MixUp [47], CutMix [46], and SnapMix [17]. These approaches have demonstrated their effectiveness in enhancing the performance of deep neural networks. In contrast to traditional image augmentation techniques that solely modify an image's appearance, these methods alter the underlying pixel values of the image. CutOut augments data by removing a rectangular region from an image. MixUp combines images and blends labels using linear combinations. CutMix employs a cut-and-paste operation to mix images, and it combines labels based on the area ratio. SnapMix estimates the semantic structure of a synthetic image by leveraging class activation maps. However, unfortunately, semantic space is usually not on images also cause inconsistency between images and attributes, *i.e.,* causing semantic distortion.

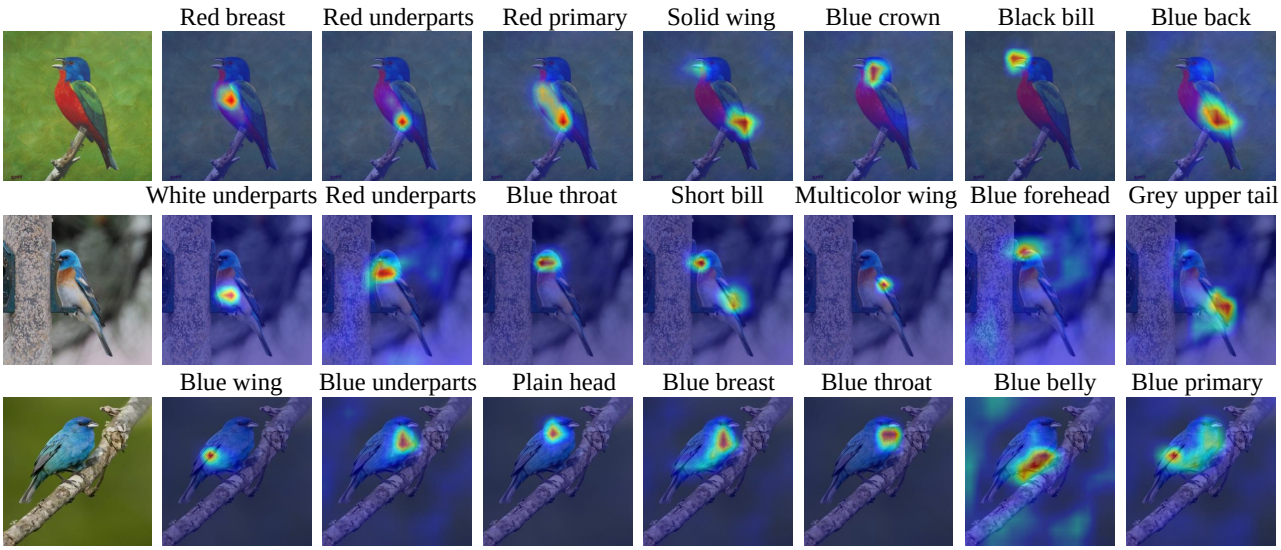
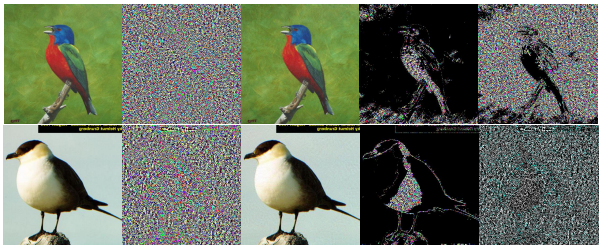


Figure 6: Visualization of the predicted attribute attention maps on CUB dataset.



(a) **(b)** **(c)** **(d)** **(e)**

Figure 7: Visualization of the perturbation components. (a) the original images; (b) the learned overall perturbation; (c) the adversarial samples after perturbation; (d) the foreground perturbation after striping the background perturbation; (e) the perturbation on the background.

4.6 Qualitative Results

4.6.1 Perturbation Visualization. Perturbations on images are often imperceptibly small, making it challenging to distinguish augmentations from the original images. Nonetheless, it is important to note that even such seemingly negligible perturbations can significantly influence the model's behavior [27, 36]. This counterintuitive phenomenon can be attributed to the high-dimensional space of the data and the intricacies of non-linear computations performed by deep learning models. To gain a deeper understanding of adversarial training behavior, we normalize the perturbations applied to the images, i.e., the differences between the augmentations and the original images. Figure 7 illustrates this: (a) displays the original two images, (c) shows the adversarial samples, and (b) presents the normalized differences. By zooming in, subtle differences become more visible. We empirically discover that the perturbation background is dominated by a few similar values, as demonstrated in (e). By removing these background values, we can isolate the foreground perturbations, as shown in (d). This observation confirms that our attribute-conditional adversarial training can effectively perceive the differences between the background and the foreground objects.

4.6.2 Attribute Attention Visualization. We present attention maps for randomly selected attributes on three example images from the CUB dataset. The attention maps are generated by applying max pooling operations to the model output, highlighting the regions where the model is focusing on each attribute. As our local perturbation component specifically focuses on changing the attended areas that ZSL model paying most attention to, which prevents overfitting to the easily attended areas. This visualization in Figure 6 confirms that our method localization ability for each attribute, demonstrating its capacity to recognize and differentiate the specific features within the images. This visualization helps to better understand the model performance and interpretability in the context of zero-shot learning. Furthermore, the attention maps reveal that the attended areas are relatively small due to the local perturbation constraint, which fosters more precise localization. This property contributes to the model overall effectiveness in handling zero-shot learning tasks.

5 CONCLUSION

In this paper, we propose a novel approach to harness the adversarial samples for enhancing the generalization ability of zero-shot learning models. By addressing the challenges of generating adversarial samples for ZSL, we have developed a method that incorporates three key learning properties, enabling effective utilization of adversarial samples to boost ZSL performance. The proposed approach demonstrates the potential of adversarial samples in improving the generalization capabilities of ZSL models, while mitigating the semantic distortion issues inherent in traditional image augmentation techniques. We hope our experimental study will help understand the difference in model behavior between single-label supervision and semantic attributes supervision, and pave the way for developing more robust semantic-condition visual augmentations.

ACKNOWLEDGMENTS

This work was partially supported by Australian Research Council CE200100025, DE200101610.

REFERENCES

[1] B. Battista, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. 2013. Evasion attacks against machine learning at test time. In *ECML PKDD*. 387–402.

[2] W. Chao, S. Changpinyo, B. Gong, and F. Sha. 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*. Springer, 52–68.

[3] L. Chen, H. Zhang, J. Xiao, W. Liu, and S. Chang. 2018. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *IEEE CVPR*. 1043–1052.

[4] S. Chen, Z. Hong, G. Xie, W. Yang, Q. Peng, K. Wang, J. Zhao, and X. You. 2022. Msdn: Mutually semantic distillation network for zero-shot learning. In *IEEE/CVF CVPR*. 7612–7621.

[5] S. Chen, G. Xie, Y. Liu, Q. Peng, B. Sun, H. Li, X. You, and L. Shao. 2021. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *NeurIPS* 34 (2021), 16622–16634.

[6] Z. Chen, J. Li, Y. Luo, Z. Huang, and Y. Yang. 2020. Canzsl: Cycle-Consistent Adversarial Networks for Zero-Shot Learning from Natural Language. In *IEEE/CVF WACV*. 874–883.

[7] Z. Chen, Y. Luo, R. Qiu, S. Wang, Z. Huang, J. Li, and Z. Zhang. 2021. Semantics Disentangling for Generalized Zero-Shot Learning. In *IEEE/CVF ICCV*.

[8] Z. Chen, Y. Luo, S. Wang, J. Li, and Z. Huang. 2022. GSMFlow: Generation Shifts Mitigating Flow for Generalized Zero-Shot Learning. *IEEE Transactions on Multimedia* (2022), 1–12. https://doi.org/10.1109/TMM.2022.3190678

[9] Z. Chen, Y. Luo, S. Wang, R. Qiu, J. Li, and Z. Huang. 2021. Mitigating Generation Shifts for Generalized Zero-Shot Learning. In *Proceedings of the 28th ACM International Conference on Multimedia*.

[10] Z. Chen, S. Wang, J. Li, and Z. Huang. 2020. Rethinking Generative Zero-Shot Learning: An Ensemble Learning Perspective for Recognising Visual Patches. In *ACM International Conference on Multimedia*. 3413–3421.

[11] T. DeVries and G. W Taylor. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. *CoRR* abs/1708.04552 (2017). arXiv:1708.04552 http://arxiv.org/abs/1708.04552

[12] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. 2018. Boosting adversarial attacks with momentum. In *IEEE CVPR*. 9185–9193.

[13] M. Elhoseiny, B. Saleh, and A. Elgammal. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*. 2584–2591.

[14] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. 2009. Describing objects by their attributes. In *CVPR*. 1778–1785.

[15] Ian J. G., Jonathon S., and Christian S. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR*. http://arxiv.org/abs/1412.6572

[16] C. Herrmann, K. Sargent, L. Jiang, R. Zabih, H. Chang, C. Liu, D. Krishnan, and D. Sun. 2022. Pyramid adversarial training improves vit performance. In *IEEE/CVF CVPR*. 13419–13429.

[17] S. Huang, X. Wang, and D. Tao. 2021. Snapmix: Semantically proportional mixing for augmenting fine-grained data. In *AAAI*, Vol. 35. 1628–1636.

[18] H. Jiang, R. Wang, S. Shan, and X. Chen. 2019. Transferable contrastive network for generalized zero-shot learning. In *ICCV*. 9765–9774.

[19] E. Kodirov, T. Xiang, and S. Gong. 2017. Semantic autoencoder for zero-shot learning. In *CVPR*. 3174–3183.

[20] C. H. Lampert, H. Nickisch, and S. Harmeling. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI* 36, 3 (2013), 453–465.

[21] Y. Liu, L. Zhou, X. Bai, Y. Huang, L. Gu, J. Zhou, and T. Harada. 2021. Goal-oriented gaze estimation for zero-shot learning. In *IEEE/CVF CVPR*. 3794–3803.

[22] S. Min, H. Yao, H. Xie, C. Wang, Z. J. Zha, and Y. Zhang. 2020. Domain-aware Visual Bias Eliminating for Generalized Zero-Shot Learning. In *CVPR*. 12664–12673.

[23] S. Narayan, A. Gupta, Fahad S. Khan, C. Snoek, and L. Shao. 2020. Latent embedding feedback and discriminative features for zero-shot classification. In *ECCV*. Springer, 479–495.

[24] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. 2013. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650* (2013).

[25] G. Patterson, C. Xu, H. Su, and J. Hays. 2014. The sun attribute database: Beyond categories for deeper scene understanding. *IJCV* 108 (2014), 59–81.

[26] J. Pennington, R. Socher, and C. D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[27] A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang. 2020. Understanding and Mitigating the Tradeoff between Robustness and Accuracy. In *ICML*. PMLR, 7909–7919.

[28] B. Romera-Paredes and P. Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *ICML*. PMLR, 2152–2161.

[29] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In *IEEE/CVF CVPR*. 8247–8255.

[30] N. Shafiee and E. Elhamifar. 2022. Zero-Shot Attribute Attacks on Fine-Grained Recognition Models. In *ECCV*. Springer, 262–282.

[31] Y. Shen, J. Qin, L. Huang, L. Liu, F. Zhu, and L. Shao. 2020. Invertible zero-shot recognition flows. In *ECCV*. Springer, 614–631.

[32] C. Shorten and T. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data* 6, 1 (2019), 1–48.

[33] H. Su, J. Li, Z. Chen, L. Zhu, and K. Lu. 2022. Distinguishing unseen from seen for generalized zero-shot learning. In *IEEE/CVF CVPR*. 7885–7894.

[34] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. 2014. Intriguing properties of neural networks. In *ICLR*. http://arxiv.org/abs/1312.6199

[35] L. Taylor and G. Nitschke. 2018. Improving deep learning with generic data augmentation. In *2018 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 1542–1547.

[36] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness May Be at Odds with Accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=SyxAb30cY7

[37] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

[38] W. Wang, Y. Pu, V. Verma, K. Fan, Y. Zhang, C. Chen, P. Rai, and L. Carin. 2018. Zero-shot learning via class-conditioned deep generative models. In *AAAI*, Vol. 32.

[39] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE TPAMI* 41, 9 (2018), 2251–2265.

[40] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. 2018. Feature generating networks for zero-shot learning. In *CVPR*. 5542–5551.

[41] Y. Xian, B. Schiele, and Z. Akata. 2017. Zero-shot learning-the good, the bad and the ugly. In *CVPR*. 4582–4591.

[42] Y. Xian, S. Sharma, B. Schiele, and Z. Akata. 2019. f-VAEGAN-D2: A feature generating framework for any-shot learning. In *CVPR*. 10275–10284.

[43] C. Xie, M. Tan, B. Gong, J. Wang, A. L Yuille, and Q. V Le. 2020. Adversarial examples improve image recognition. In *IEEE/CVF CVPR*. 819–828.

[44] W. Xu, Y. Xian, J. Wang, B. Schiele, and Z. Akata. 2020. Attribute Prototype Network for Zero-Shot Learning. In *NeurIPS*.

[45] Y. Yu, Z. Ji, J. Han, and Z. Zhang. 2020. Episode-Based Prototype Generating Network for Zero-Shot Learning. In *CVPR*. 14035–14044.

[46] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF ICCV*. 6023–6032.

[47] H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.

[48] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. 2020. Random erasing data augmentation. In *AAAI*, Vol. 34. 13001–13008.

[49] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal. 2018. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*. 1004–1013.

[50] Y. Zhu, J. Xie, Z. Tang, X. Peng, and A. Elgammal. 2019. Semantic-Guided Multi-Attention Localization for Zero-Shot Learning. In *NeurIPS*. 14917–14927.