Language-Guided Imaginative Walks: Generative Random Walk Deviation Loss for Unseen Class Recognition using Text Descriptions

Kai Yi, Divyansh Jha, Ivan Skorokhodov, Mohamed Elhoseiny
King Abdullah University of Science and Technology (KAUST), Kingdom of Saudi Arabia
{divyansh.jha, kai.yi, ivan.skorokhodov, mohamed.elhoseiny}@kaust.edu.sa

Abstract

We introduce a language-guided loss for generative models, dubbed as GRaWD (Generative Random Walk Deviation), to improve learning representations of unseen visual classes using purely textual descriptions. Quality learning representation of unseen classes is critical for the better generative understanding of unseen visual classes, i.e., zero-shot learning (ZSL). By generating visual representations of unseen classes from their text descriptions, generative ZSL attempts to differentiate unseen from seen categories. GRaWD loss is defined by constructing a dynamic graph that includes the seen class centers and generated samples in the current minibatch. Our loss initiates a random walk probability from each center through visual generations produced from hallucinated descriptions of unseen classes. As a deviation signal, we encourage the random walk to land after T steps in a representation that is difficult to classify as any seen class. We demonstrate that our loss can inductively improve unseen class representation quality on text-based ZSL benchmarks and achieve state-of-the-art performance on CUB and NABirds datasets.

1. Introduction

Generative models like GANs [15] and VAEs [23] are excellent tools for generating realistic images due to their ability to represent high-dimensional probability distributions. However, they are not explicitly trained to go beyond distribution seen during training. In recent years, generative models have been adopted to go beyond training data distributions and improve unseen class recognition (also known as zero-shot learning) [17, 29, 18, 25, 51, 42]. These methods train a conditional generative model $G(s_k, z)$ [30, 33], where s_k is the semantic description of class k (text descriptions) and z represents within-class variation (e.g., $z \in \mathcal{N}(0, I)$). After training, $G(s_k, z)$ is used to generate imaginary data for unseen classes transforming ZSL into a traditional classification task trained on the generated data. Understanding unseen classes is leveraged by the generative model's improved ability to produce discriminative visual representations using $G(s_u, z)$ from their corresponding text descriptions s_u.

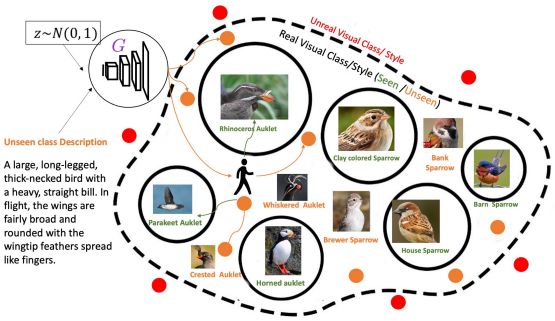

Figure 1: GRaWD loss encourages generatively visiting the orange realistic space, aiming to deviate from the seen classes and avoid the less real red space. Our loss starts from each seen class (in green) and performs a random walk through generated examples of hallucinated unseen classes (in orange) for T steps. We then encourage the landing representation to be distinguishable from seen classes and with this property, the loss helps improve generative ZSL performance.

To generate likable novel visual content, GANs' training has been augmented with a loss that encourages careful deviation from existing style classes [8, 38, 21, 22]. Such models were shown to have some capability to produce unseen aesthetic art [8], fashion [38, 44], and design [32]. In a generalized ZSL context, CIZSL [9] showed an improved performance by modeling a similar deviation to encourage discrimination explicitly between seen and unseen classes. These losses improve unseen representation quality by encouraging the produced visual generations to be distinguishable from seen classes.

Contribution. We propose Generative Random Walk Deviation (*GRaWD*) loss as a language-guided graph-based loss

to improve learning representation of unseen classes; see Fig. 1. Our loss is *parameter-free* and starts from each seen class (in green) and performs a random walk for T steps through examples of hallucinated unseen classes (in orange) that are generated conditioned on text. Then, we encourage the landing representation to be distant and distinguishable from the seen class centers. GRaWD loss is computed over a similarity graph involving seen class centers and generated examples in the current minibatch of hallucinated unseen classes. Thus, GRaWD takes a global view of the data manifold compared to existing deviation losses that are local/per example (e.g., [38, 8, 9]). In contrast to transductive methods (e.g., [42]), our loss is purely inductive; therefore, does not require real descriptions of unseen classes during training. Our work can be connected to recent advances in semi-supervised learning (e.g., [50, 36, 19, 27, 3]) that leverage unlabeled data within the training classes. In these methods, unlabeled data are encouraged to be attracted to existing classes. Our goal is the *opposite*, deviating from seen classes. Also, our loss operates on generated data of hallucinated unseen classes instead of provided unlabeled data. In our experiments, we show that GRaWD loss improves unseen class recognition on challenging ZSL benchmarks.

2. Related Work

Most current ZSL methods can be classified into two branches. One branch casts the task as a visual-semantic embedding problem [14, 40, 28]. akata2015evaluation,akata2016label proposed Attribute Label Embedding(ALE) to model visual-semantic embedding as a bilinear compatibility function between the image space and the attribute space. In [49], deep ZSL methods were presented to model the non-linear mapping between vision and class descriptions. In the context of ZSL from noisy textual descriptions, an early linear approach for Wikipedia-based ZSL was proposed in [10]. Orthogonal to these improvements, generative models like GANs [15] and VAEs [23] have been adopted to formulate multi-modality in zero-shot recognition by synthesizing visual features of unseen classes given its semantic description(text/attributes), e.g., [25, 51, 39, 31, 20, 7]. [51] introduced a GAN model with a classification head with the standard real/fake head to improve text-based ZSL. [39] proposed a cross and distribution aligned VAE to better leverage the seen and unseen relationships. [20] utilized a generative network along with a multi-level supervised contrastive embedding strategy to learn images and semantic relationships. We focus on Wikipedia/Text-based ZSL in this work and our GRaWD loss helps improve the out-of-distribution performance of generative ZSL models.

3. Approach

Let's denote the set of seen and unseen class labels as $\mathcal{S}$ and $\mathcal{U}$, where $\mathcal{S} \cap \mathcal{U} = \emptyset$. We denote the text representations of unseen classes and seen classes as $s_u = \psi(T_u)$ and $s_i = \psi(T_i)$ respectively, $\psi(\cdot)$ is function that extract representation from raw text article describing a class (T_u or T_i). Let's denote the seen data as $D^s = \{(x_i^s, y_i^s, s_i)\}$, where $x_i^s \in \mathcal{X}$ denotes the visual features of the i^{th} image, $y_i^s \in \mathcal{S}$ is the corresponding seen category label. For unseen classes, we are given only their semantic representations, s_u. In Generalized ZSL (GZSL), we aim to predict the label $y \in \mathcal{U} \cup \mathcal{S}$ at test time given x that may belong to seen or unseen classes.

Fig. 2 illustrates the approach overview. We denote the generator as $G(s, z)$ with parameters θ_G, where s is the text description and $z \in \mathbb{R}^Z$ is a random vector sampled from a Gaussian distribution $p_z = \mathcal{N}(0, 1)$. $G(s, z)$ is then used to sample the generated visual feature of a class given its description s_k.We denote the discriminator as D and its parameters as θ_D. The discriminator has two heads. The first head is for binary real/fake classification to predict "real" for images from the training set and "fake" for generated ones. We denote the real/fake probability produced by D for an input image as $D^r(\cdot)$. The second head is a K^s-way classifier over the seen classes. We denote the classification score of a seen class $k \in \mathcal{S}$ given the image as $D^{s,k}(\cdot)$. We denote seen class centers that we aim unseen classes to deviate from as $C = \{\mathbf{c}_1 \cdots \mathbf{c}_{K^s}\}$,

$$\mathbf{c}_i = \phi(G(s_i, z = \mathbf{0})), i \in \{1 \rightarrow K\} \quad (1)$$

where s_i is the text description of seen class i. $X_u = \{x_1^u \cdots x_{N_u}^u\}$ are sampled by $\phi(G(s_u, z))$ where $z \sim p_z = \mathcal{N}(0, I)$, $s_u \sim p_u$ is a text description of a hallucinated unseen class, $\phi(\cdot)$, a feature extraction function that we define as the activations from the last layer of the Discriminator D followed by scaled L2 normalization $L2(\mathbf{v}, \beta) = \beta \frac{\mathbf{v}}{\|\mathbf{v}\|}$. The scaled factor is mainly to amplify the norm of the vectors to avoid the vanishing gradient problem inspired from [4, 48], $\beta = 3$.We explore the unseen/imaginative space of the generator G with a hallucinated semantic representation $s_u \sim p^u$, where p^u is a probability distribution over unseen classes, aimed to be likely hard negatives to seen classes. Similar to [9], we pick two seen semantic descriptions at random $s_a, s_b \in \mathcal{S}$. We then sample $s^u = \alpha s_a + (1 - \alpha) s_b$, where α is uniformly sampled between 0.2 and 0.8 to avoid sampling descriptions close to seen classes ($\alpha \rightarrow \{0, 1\}$).

Let $B \in \mathbb{R}^{N_u \times K^s}$ be the similarity matrix between each of the features of the generations ($x^u \in X_u$) and seen class centers ($c_i \in C$). Similarly, let $A \in \mathbb{R}^{N_u \times N_u}$ compute the similarity matrix between the generated points. In particular, we use the negative Euclidean distances between the embeddings as a similarity measure: $B_{ij} =$

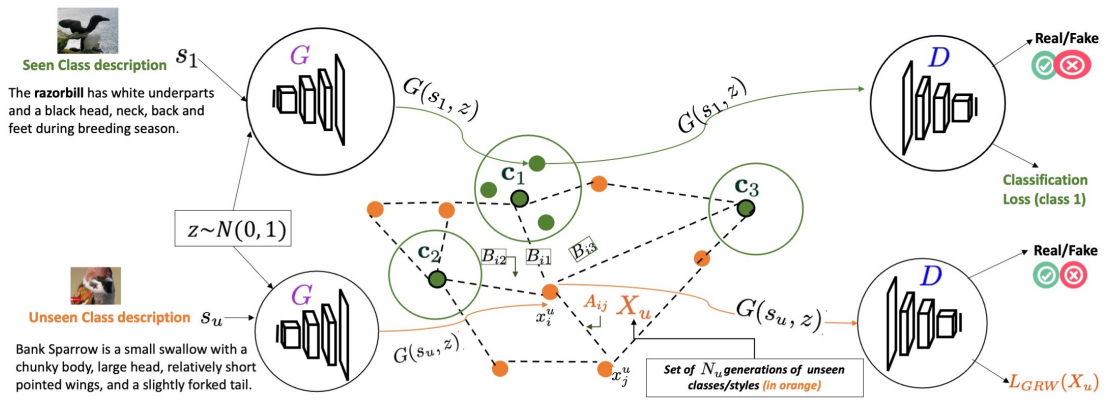

Figure 2: Our loss starts from each seen class center (i.e., $\mathbf{c}_i$), computed from their text descriptions. It then performs a random walk through generated examples of hallucinated unseen classes using $G(s_u, z)$ for T steps. The landing probability distribution of the random walk is encouraged to be uniform over the seen classes. For careful deviation from seen classes, the generations are encouraged to be classified as real by the Discriminator D; see Eq. 3.

$-\|x_i - \mathbf{c}_j\|^2, A_{i,j} = -\|x_i^u - x_j^u\|^2$, where x_i^u and x_j^u are i^{th} and j^{th} features in the set X_u; see Fig. **??**. To avoid self-cycle, The diagonal entries $A_{i,i}$ are set to a small number ϵ. We then define three transition probability matrices: $P^{\mathbf{C}\to X_u} = \sigma(B^{\mathbb{T}}), P^{X_u\to\mathbf{C}} = \sigma(B), P^{X_u\to X_u} = \sigma(A)$, where σ is the softmax operator is applied over each row of the input matrix, $P^{\mathbf{C}\to X_u}$ and $P^{X_u\to\mathbf{C}}$ are the transition probability matrices from each seen class over the N_u generated points and vice-versa respectively. $P^{X_u\to X_u}$ is the transition probability matrix from each generated point over other generated points. We hence define our generative random walker probability matrix as:

$$P^{\mathbf{C}\to\mathbf{C}}(t, X_u) = \sigma(B^{\mathbb{T}}) \cdot (\sigma(A))^t \cdot \sigma(B) \tag{2}$$

where $P_{i,j}^{\mathbf{C}\to\mathbf{C}}(t, X_u)$ denotes the probability of ending a random walk of a length t at a seen class j given that we have started at seen class i; t denotes the number of steps taken between the generated points, before stepping back to land on a seen class.

Loss. To boost the deviation of unseen visual spaces from seen ones, we encourage each row in $P^{\mathbf{C}\to\mathbf{C}}(t)$ to be hard to classify to seen classes:

$$\begin{aligned} L_{GRW}(X_u) = &- \sum_{t=0}^{T} \gamma^t \cdot \sum_{i=1}^{K^s}\sum_{j=1}^{K^s} U_c(j) log(P_{i,j}^{\mathbf{C}\to\mathbf{C}}(t, X_u)) \\ &- \sum_{j=1}^{N_u} U_x(j) log(P_v(j)) \end{aligned} \tag{3}$$

where first term minimizes cross entropy loss between every row in $P^{\mathbf{C}\to\mathbf{C}}(t, X_u) \forall t = 1 \to T$ and uniform distribution over seen classes $U_c(j) = \frac{1}{K^s}, \forall j = 1 \cdots K^s$, where T is a hyperparameter and γ is exponential decay set to 0.7 in our experiments. Note that, if we replaced U_c by an identity matrix to encourage landing to the starting seen class, the loss becomes an attraction signal similar to [19], which defines its conceptual difference to GRaWD. *We call this version GRaWT, T for aTraction*. In second term, we adapt the "visit loss", introduced in [19], to encourage random walker to visit a large set of our generations X_u to extract more learning signals; see the visit loss details in Appendix A. We then integrate $L_{GRW}(X_u)$ into the *Generator G loss* as the first term here

$$\begin{aligned} L_G = &\lambda \mathbb{E}_{X_u\sim\phi(G(s_u,z))), z\sim p_z, s_u\sim p^u}[L_{GRW}(X_u)] \\ &- \mathbb{E}_{z\sim p_z, s_u\sim p^u}[D^r(G(s_u, z))] \\ &- \mathbb{E}_{z\sim p_z, (s_k,y^s)\sim p^s}[D^r(G(s_k, z)) \\ &+ \sum_{k=1}^{K^s} y_k^s log(D^{s,k}(G(s_k, z)))] \end{aligned} \tag{4}$$

The second and the third terms trick the discriminator into classifying the visual generations from both the seen text descriptions s_k and unseen text descriptions s_u, as real. The fourth term encourages the generator to discriminatively generate visual features conditioned on a given seen class description. We then define the *Discriminator D loss* as

$$\begin{aligned} L_D = &\ \mathbb{E}_{z\sim p_z, s_u\sim p^u}[D^r(G(s_u, z))] \\ &+ \mathbb{E}_{z\sim p_z, (s_k,y^s)\sim p^s}[D^r(G(s_k, z))] - \mathbb{E}_{x\sim p_d}[D^r(x)] \\ &+ L_{Lip} - \frac{1}{2}\mathbb{E}_{x,y\sim p_d}[\sum_{k=1}^{K^s} y_k log(D^{s,k}(x))] \\ &- \frac{1}{2}\mathbb{E}_{z\sim p_z, (s_k,y^s)\sim p^s}[\sum_{k=1}^{K^s} y_k^s log(D^{s,k}(G(s_k, z)))] \end{aligned} \tag{5}$$

Here, image x and corresponding class one-hot label y are sampled from the data distribution p_d. s_k and y^s are features of a semantic description and the corresponding one-hot label sampled from seen classes p^s. The first three terms

Table 1: Ablation studies on CUB Dataset (text).

Setting	CUB-Easy		CUB-Hard	
	Top-1 Acc (%)	SU-AUC (%)	Top1-Acc (%)	SU-AUC (%)
Baseline Deviation losses on GAZSL [51]				
+ GRaWT (T=0)	44.0	39.5	13.7	11.8
+ GRaWT (T=3)	43.4	38.8	13.2	11.4
+ Classify $G(s_u, z)$ as class K^{s+1}	43.2	38.3	11.31	9.5
+ CIZSL[9]	44.6	39.2	14.4	11.9
GRaWD Walk length on GAZSL [51]				
+ GRaWD (T=1)	45.41	39.62	13.79	12.58
+ GRaWD (T=3)	45.11	39.25	14.21	13.22
+ GRaWD (T=5)	45.40	40.51	14.00	13.07
+ GRaWD (T=10)	**45.43**	**40.68**	**15.51**	**13.70**

Table 2: Zero-Shot Recognition from textual description on **CUB** and **NAB** datasets (Easy and Hard Splits) showing that adding GRaWD loss can improve the performance. *tr* means the transductive setting.

Metric	Top-1 Accuracy (%)				Seen-Unseen AUC (%)			
Dataset	CUB		NAB		CUB		NAB	
Split-Mode	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
ZSLNS [35]	29.1	7.3	24.5	6.8	14.7	4.4	9.3	2.3
SynC$_{fast}$ [5]	28.0	8.6	18.4	3.8	13.1	4.0	2.7	3.5
ZSLPP [11]	37.2	9.7	30.3	8.1	30.4	6.1	12.6	3.5
FeatGen [47]	43.9	9.8	36.2	8.7	34.1	7.4	21.3	5.6
LsrGAN (*tr*) (Vyas et al. 2020)	45.2	14.2	36.4	9.0	39.5	12.1	23.2	6.4
+GRaWD	**45.6**$^{+0.4}$	15.1$^{+0.9}$	37.8$^{+1.4}$	9.7$^{+0.7}$	39.9$^{+0.4}$	13.3$^{+1.2}$	24.5$^{+1.3}$	6.7$^{+0.3}$
GAZSL [51]	43.7	10.3	35.6	8.6	35.4	8.7	20.4	5.8
+CIZSL [9]	44.6	14.4	36.6	9.3	39.2	11.9	24.5	6.4
+ GRaWD	45.4$^{+1.7}$	**15.5** $^{+5.2}$	**38.4** $^{+2.8}$	**10.1** $^{+1.5}$	**40.7** $^{+5.3}$	**13.7**$^{+5.0}$	**25.8**$^{+5.4}$	**7.4** $^{+1.6}$

approximate Wasserstein distance of the distribution of real features and fake features, and fourth term is the gradient penalty to enforce the Lipschitz constraint; see [16]. The last two terms are the classification losses of the real and generated data to their corresponding classes.

4. Experiments

We performed experiments on existing ZSL benchmarks with text descriptions as semantic class descriptions. Text-based ZSL is more challenging because the descriptions are at the class-level and are extracted from Wikipedia, which is noisier. We found that random walk steps T easy to tune using the validation set. We performed our experiments on Caltech UCSD Birds-2011 (CUB) [43] containing 200 classes with 11, 788 images and North America Birds (NAB) [41] which has 1011 classes with 48, 562 images. We use two metrics widely used in evaluating ZSL recognition performance: standard zero-shot recognition with the Top-1 unseen class accuracy and Seen-Unseen Generalized Zero-shot performance with Area under Seen-Unseen curve [6]. We follow [6, 51, 9] in using the Area Under SUC to evaluate the generalization capability of class-level text zero-shot recognition on four splits (CUB Easy, CUB Hard, NAB Easy, and NAB Hard). The hard splits are constructed such that unseen bird classes from super-categories do not overlap with seen classes. For text representation function $\psi(\cdot)$, we used the TF-IDF[37] representation of the input text followed by an FC noise suppression layer.

Our proposed loss function improves over older methods on all datasets on both Easy and SCE(hard) splits, as shown in Table 2. We show improvements in the range of 0.8-1.8% Top-1 accuracy. We also show improvements in AUC, ranging from 1-1.8%. From Table 2, we show that GAZSL [51]+GRaWD has an average relative Seen-Uneen AUC improvement over GAZSL [51]+CIZSL [9] and GAZSL [51] only of 9.29% and 30.89%. We achieved SOTA results for text datasets. In Table 1, we performed an ablation study where we show that longer random walks performed better hence giving higher accuracies and AUC scores for both easy and hard split for CUB Dataset. With longer walks, the model could have a more holistic view of the generated visual representation in a way that enables better deviation of unseen classes from unseen classes. Therefore, we used T=10 for our experiments.

GRaWD Loss for Transductive ZSL. We also apply our GRaWD loss to transductive ZSL setting where text descriptions of unseen classes are used during training. We choose LsrGAN [42] as the baseline model. Our loss can also improve LsrGAN on text-based datasets on most metrics ranging from 0.3%-3.6%. Despite that our loss does not use unseen class descriptors, it can still improve on average on LsrGAN (transductive) by 2.91% on text-based datasets. However, as we expected, the former improvement in the purely inductive/more realistic setting is more significant. More ablations and experiments can be found in the Appendix.

5. Conclusion

We propose Generative Random Walk Deviation (GRaWD) loss and showed that it improves generative models' capability to better understand unseen classes on text-based zero-shot learning benchmarks. We believe the improvement is due to our learning mechanism's global nature, which operates at the minibatch level producing language guided generations that are message-passing to each other to facilitate better deviation of unseen classes from seen ones.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *PAMI*, 38(7):1425–1438, 2016. 8

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015. 8

[3] Ahmed Ayyad, Yuchen Li, Raden Muaz, Shadi Albarqouni, and Mohamed Elhoseiny. Semi-supervised few-shot learning with prototypical random walks. In *AAAI Workshop*

on Meta-Learning and MetaDL Challenge, pages 45–57. PMLR, 2021. 2

[4] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2874–2883, 2016. 2

[5] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, 2016. 4, 8

[6] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, pages 52–68. Springer, 2016. 4

[7] Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. Free: Feature refinement for generalized zero-shot learning. *arXiv preprint arXiv:2107.13807*, 2021. 2

[8] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 2017. 1, 2

[9] Mohamed Elhoseiny and Mohamed Elfeki. Creativity inspired zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5784–5793, 2019. 1, 2, 4, 7, 8

[10] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, 2013. 2

[11] Mohamed Elhoseiny, Yizhe Zhu, Han Zhang, and Ahmed Elgammal. Link the head to the "beak": Zero shot learning from noisy text description at part precision. In *CVPR*, July 2017. 4

[12] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR 2009.*, pages 1778–1785. IEEE, 2009. 7

[13] Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, pages 21–37, 2018. 8

[14] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013. 2

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 1, 2

[16] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017. 4

[17] Yuchen Guo, Guiguang Ding, Jungong Han, and Yue Gao. Synthesizing samples for zero-shot learning. In *IJCAI*, 2017. 1

[18] Yuchen Guo, Guiguang Ding, Jungong Han, and Yue Gao. Zero-shot learning with transferred samples. *IEEE Transactions on Image Processing*, 2017. 1

[19] P. Haeusser, A. Mordvintsev, and D. Cremers. Learning by association — a versatile semi-supervised training method for neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 626–635, July 2017. 2, 3, 7

[20] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2371–2381, 2021. 2

[21] Aaron Hertzmann. Can computers create art? In *Arts*, volume 7, page 18. Multidisciplinary Digital Publishing Institute, 2018. 1

[22] Aaron Hertzmann. Visual indeterminacy in gan art. *Leonardo*, 53(4):424–428, 2020. 1

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2

[24] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. *arXiv preprint arXiv:1704.08345*, 2017. 8

[25] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, 2018. 1, 2

[26] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009. 7

[27] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. In *Advances in Neural Information Processing Systems*, pages 10276–10286, 2019. 2

[28] Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9273–9281, 2020. 2

[29] Yang Long, Li Liu, Ling Shao, Fumin Shen, Guiguang Ding, and Jungong Han. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. In *CVPR*, 2017. 1

[30] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[31] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 479–495. Springer, 2020. 2

[32] Amin Heyrani Nobari, Muhammad Fathy Rashad, and Faez Ahmed. Creativegan: editing generative adversarial networks for creative design synthesis. *arXiv preprint arXiv:2103.06242*, 2021. 1

[33] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017. 1

[34] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE, 2012. 7

[35] R. Qiao, L. Liu, C. Shen, and A. v. d. Hengel. Less is more: Zero-shot learning from online textual documents with noise suppression. In *CVPR*, June 2016. 4

[36] Mengye Ren, Sachin Ravi, Eleni Triantafillou, Jake Snell, Kevin Swersky, Josh B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations*, 2018. 2

[37] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988. 4

[38] Othman Sbai, Mohamed Elhoseiny, Antoine Bordes, Yann LeCun, and Camille Couprie. Design: Design inspiration from generative networks. In *ECCV workshop*, 2018. 1, 2

[39] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019. 2

[40] Ivan Skorokhodov and Mohamed Elhoseiny. Class normalization for (continual)? generalized zero-shot learning. In *International Conference on Learning Representations*, 2021. 2

[41] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, 2015. 4

[42] Maunil Vyas, Hemanth Venkateswara, and Sethuraman Panchanathan. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *European Conference on Computer Vision*, pages 70–86. Springer, 2020. 1, 2, 4, 8

[43] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 4

[44] Qiang Wu, Baixue Zhu, Binbin Yong, Yongqiang Wei, Xuetao Jiang, Rui Zhou, and Qingguo Zhou. Clothgan: generation of fashionable dunhuang clothes using generative adversarial networks. *Connection Science*, 33(2):341–358, 2021. 1

[45] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016. 8

[46] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *PAMI*, 2018. 7

[47] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 4, 8

[48] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9185–9194, 2019. 2

[49] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2016. 2, 8

[50] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *Advances in Neural Information Processing Systems*, pages 2371–2380, 2018. 2

[51] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, 2018. 1, 2, 4, 7, 8

A. Visit Loss

The second term is called 'visit loss' was proposed in [19] to encourage random walker to visit a large set of unlabeled points. We applied it here to our generated deviation examples X_u. We compute the overall probability that each generated point would be visited by any of the seen class $P_v = \frac{1}{N_u}\sum_{i=0}^{N_c} P_i^{\mathbf{C}\to X_u}$, where $P_i^{\mathbf{C}\to X_u}$ represents the i^{th} row of the $P^{\mathbf{C}\to X_u}$ matrix; see Fig. 2 in the main content. The visit loss is then defined as the cross-entropy between P_v and the uniform distribution $U_x(j) = \frac{1}{N^u}, \forall j = 1 \cdots N^u$. Hence, visit loss encourages to visit as many examples as possible from X_u and hence improves learning representation.

B. Training Algorithm

To train our model, we consider visual-semantic feature pairs, images and text, as a joint observation. Visual features are produced either from real data or synthesized by our generator. We illustrate in Algorithm 1 how G and D are alternatively optimized with an Adam optimizer. The algorithm summarizes the training procedure. In each iteration, the discriminator is optimized for n_d steps (lines $6-11$), and the generator is optimized for 1 step (lines $12-14$).

Algorithm 1: Training procedure of our approach. We use default values of $n_d = 5$, $\alpha = 0.001$, $\beta_1 = 0.5$, $\beta_2 = 0.9$

1: **Input:** the maximal loops N_{step}, the batch size m, the iteration number of discriminator in a loop n_d, the balancing parameter λ_p, Adam hyperparameters α_1, β_1, β_2.
2: **for** iter $= 1, ..., N_{step}$ **do**
3: Sample random text minibatches s_a, s_b, noise z
4: Construct s_u with different α for each row in the minibatch
5: $x_u \leftarrow G(s_u, z)$
6: Compute the similarity matrix B between x_u and the cluster centers $c \in C$. Then compute the generative random walker matrix using Eq. 2.
7: **for** $t = 1, ..., n_d$ **do**
8: Sample a minibatch of images x, matching texts s, random noise z
9: $\tilde{x} \leftarrow G(s, z)$
10: Compute the discriminator loss L_D using Eq. 5
11: $\theta_D \leftarrow \text{Adam}(\bigtriangledown_{\theta_D} L_D, \theta_D, \alpha_1, \beta_1, \beta_2)$
12: **end for**
13: Sample a minibatch of class labels c, matching texts T_c, random noise z
14: Compute the generator loss L_G using Eq. 4
15: $\theta_G \leftarrow \text{Adam}(\bigtriangledown_{\theta_G} L_G, \theta, \alpha_1, \beta_1, \beta_2)$
16: **end for**

C. Test Details

During *Test*, the visual features of unseen classes can be synthesized by the generator conditioned on a given unseen text description s_u, as $x_u = G(s_u, z)$. We generate 60 generated visual features by sampling different z for the same semantic description s_u. We then use the generated features to apply a simple nearest neighbor classifier. We launched every ZSL experiment on a single NVIDIA V100 GPU, and more details can be found in our code attached separately.

D. Ablation Study

In this section, we perform an ablation study to investigate the value of removing the visit loss. We found adding the visit loss is important for stability. Without it, training failed with NaN gradients in 5% of the times. The visit loss adds a bit of stability (previously, we had 0% failures and the performance is better). We show the results in Table 3.

Metric	Top-1 Accuracy (%)				Seen-Unseen AUC (%)			
Dataset	CUB		NAB		CUB		NAB	
Split-Mode	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
GAZSL [51]	43.7	10.3	35.6	8.6	35.4	8.7	20.4	5.8
GAZSL [51]+**GRaWD**	**45.4**	**15.5**	**38.4**	10.1	**40.7**	**13.7**	**25.8**	**7.4**
GAZSL [51]+**GRaWD**- Visit loss	45.3	14.8	38.2	**10.3**	40.1	12.8	25.8	7.4

Table 3: Ablation study using Zero-Shot recognition on **CUB** & **NAB** datasets with two split settings. We experiment with and without the visit loss (second and last row). The first loss is without any deviation loss.

E. Attribute-Based ZSL Results

We performed experiments on the widely used GBU [46] setup, where we use class attributes as semantic descriptors. We performed these experiments on the AwA2 [26], aPY [12], and SUN [34] datasets. In Table 5, we see that GRaWD outperforms all of the existing methods on seen-unseen harmonic mean for AwA2, aPY, and SUN datasets. In the case of the AwA2 dataset, it outperformed the compared method by a significant margin, i.e., 15.1%. It is also competent with existing methods in Top-1 accuracy while improving on AwA2 4.8%. From Table 5, GAZSL [51]+GRaWD has an average relative improvement over GAZSL [51]+CIZSL [9] and GAZSL [51] of 24.92% and 61.35% in harmonic mean.

Tab. 1 in the main content and Tab. 4 show that deviation signal in GRaWD is critical to achieve better performance since the calculated metrics are much better for GRaWD compared to GRaWT for both text-based and attribute based ZSL. The performance can severely degrade without the deviation signal. Tab. 1 in the main content (bottom section) shows that longer walk lengths benefits the training as

model is encouraged to globally explore larger segments of unseen representations' manifold.

Table 4: Attribute based ZSL on AwA2, aPY and SUN. Compared with Haeusser et al. (2017).

	AwA2			aPY			SUN		
	H	S	U	H	S	U	H	S	U
GRaWT (T=0)	32.3	80.5	20.2	23.0	78.9	13.4	26.0	31.6	22.2
GRaWT (T=3)	31.6	80.7	19.7	22.4	75.8	13.1	25.8	31.1	22.1
GRaWD	**39.0**	**88.3**	**25.0**	**27.2**	**83.2**	**16.3**	**27.9**	**37.3**	**22.3**

Table 5: Zero-Shot Recognition on class-level attributes of **AwA2**, **aPY** and **SUN** datasets, showing that GRaWD loss can improve the performance on attribute-based datasets.

	Top-1 Accuracy(%)			Seen-Unseen H		
	AwA2	aPY	SUN	AwA2	aPY	SUN
SJE [2]	61.9	35.2	53.7	14.4	6.9	19.8
LATEM [45]	55.8	35.2	55.3	20.0	0.2	19.5
ALE [1]	62.5	39.7	58.1	23.9	8.7	26.3
SYNC [5]	46.6	23.9	56.3	18.0	13.3	13.4
SAE [24]	54.1	8.3	40.3	2.2	0.9	11.8
DEM [49]	67.1	35.0	61.9	25.1	19.4	25.6
FeatGen [47]	54.3	42.6	60.8	17.6	21.4	24.9
cycle-(U)WGAN [13]	56.2	44.6	60.3	19.2	23.6	24.4
LsrGAN (*tr*) [42]	60.1*	34.6*	62.5	48.7*	31.5*	44.8
+ **GRaWD**	63.7$^{+3.6}$	35.5$^{+0.9}$	**64.2**$^{+1.7}$	**49.2**$^{+0.5}$	**32.7**$^{+1.2}$	**46.1**$^{+1.3}$
GAZSL [51]	58.9	41.1	61.3	15.4	24.0	26.7
+ CIZSL [9]	67.8	42.1	63.7	24.6	25.7	27.8
+ **GRaWD**	**68.4**$^{+9.5}$	**43.3**$^{+2.2}$	62.1$^{+0.8}$	39.0$^{+23.6}$	27.2$^{+3.2}$	27.9$^{+1.2}$