Exploring Unknown Image Generation for Zero-Shot Learning via Diffusion Models

Lei Xiang[1,a)] Yasutomo Kawanishi[2,1,b)] Takahiro Komamizu[1,c)] Ichiro Ide[1,d)]

Abstract

Generation-based Zero-Shot Learning (ZSL) methods typically utilize generative models to synthesize features for unseen classes based on semantic attribute descriptions, bridging the gap between seen and unseen data. However, most existing generative approaches focus solely on feature synthesis, making the results difficult to interpret or visualize. To overcome this limitation, we propose a novel method that integrates a diffusion model into the ZSL pipeline to directly generate images of unseen classes. To further improve the generated images' visual quality and semantic alignment, we introduce a semantic neighbor-guided conditioning strategy that enriches unseen class attributes with guidance from the most similar seen class, leveraging the model's ability to generate high-quality seen images to ensure overall superior results. Extensive experiments on three standard ZSL benchmarks demonstrate the effectiveness of our method in both recognition performance and image quality.

1. Introduction

Zero-Shot Learning (ZSL) aims to identify novel classes that are not included in the training data by leveraging shared information, such as attributes [10] and word vectors [16]. Due to the absence of samples of unseen classes, a natural solution is to generate data to enrich the training set. Generative ZSL methods [12], [13], [18], [22] address this challenge by synthesizing visual features for unseen classes conditioned on semantic attributes, enabling conventional classifiers to generalize beyond seen categories.

These approaches have achieved promising results and have received increasing attention in recent years. However, existing methods mainly focus on feature-level synthesis rather than image-level generation, which limits interpretability and practical applicability. To overcome this limitation, we explore the generation of unseen images for ZSL, leveraging the capabilities of powerful Diffusion Models (DMs). Compared with conventional generative models, DMs are capable of generating high-quality and semantically faithful images, making them well-suited for interpretable ZSL. Furthermore, to enhance the alignment of DMs with ZSL objectives, especially when training with only seen data, we introduce a semantic neighbor-guided conditioning strategy. This strategy explicitly controls both primary (target) and secondary (guide) conditions, ensuring the quality of the generated unseen images guided by the most semantically similar seen class. This dual-conditioning mechanism enhances the quality and discriminability of the generated images for unseen categories, promoting more effective and interpretable ZSL.

[1] Graduate School of Informatics, Nagoya University
[2] Guardian Robot Project, RIKEN
[a)] xiangl@cs.is.i.nagoya-u.ac.jp
[b)] yasutomo.kawanishi@riken.jp
[c)] taka-coma@acm.org
[d)] ide@i.nagoya-u.ac.jp

2. Related Work—Diffusion Models

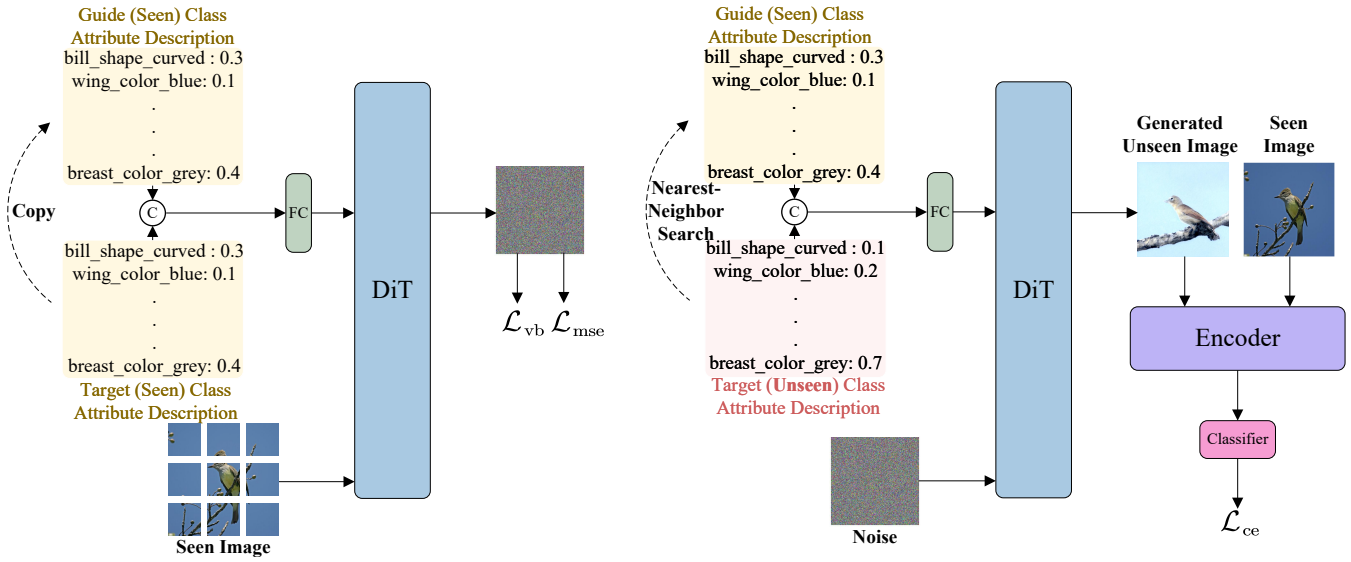
DMs have been extensively studied in the generative modeling community. Some recent works [19], [20] leverage them to generate multiple images from the same prompt for contrastive learning, treating them as positive pairs. Other approaches, such as the Diffusion Classifier [11], apply them directly for zero-shot classification, achieving competitive results across various benchmarks. However, these methods rely on textual prompts as conditioning inputs, which are not typically available in the ZSL setting. Instead, ZSL relies on class-wise continuous attribute descriptions, where each value represents the likelihood of an attribute occurring in images belonging to a specific class.

3. Proposed Method

3.1 Overview

To address this challenge and improve the quality of generated images, we introduce a semantic neighbor-guided conditioning strategy that integrates both primary (target class) and secondary (guide class that are most similar with the target class) attribute descriptions, leveraging the model's ability to generate high-fidelity seen class images, thereby enhancing the relevance and detail of the generated unseen images.

In ZSL, the dataset is divided into two disjoint subsets: a seen dataset and an unseen dataset. The seen dataset is denoted as $D^s = \{(x_i^s, y_i^s)\}_{i=1}^{N^s}$, where x_i^s and $y_i^s \in C^s$ represent the i-th image and its corresponding label from the set C^s of seen classes. In the same way, the unseen dataset is denoted as $D^u = \{(x_j^u, y_j^u)\}_{j=1}^{N^u}$, where x_j^u and $y_j^u \in C^u$ represent the j-th image and its label from the set C^u of unseen classes. The seen and unseen classes are disjoint,



(a) Stage 1: Generative model training (b) Stage 2: Zero-shot classification

Fig. 1: Architecture of the proposed method

i.e., $C^s \cap C^u = \emptyset$. Notably, the attribute descriptions across seen and unseen classes share a common semantic space, *i.e.*, $A = A^s \cup A^u$, where $A^s = \{\mathbf{a}_i^s\}_{i=1}^{|C^s|}$, $A^u = \{\mathbf{a}_j^u\}_{j=1}^{|C^u|}$ and $\mathbf{a}_i^s, \mathbf{a}_j^u \in \mathbb{R}^M$. Each element of matrix $A \in \mathbb{R}^{(|C^s|+|C^u|)\times M}$ is represented by an M-dimensional vector, corresponding to the same set of attributes. The difference between classes lies in the attribute occurrence probabilities.

The overall framework is illustrated in Figure **??**, which consists of two stages: **Stage 1** trains a generative model on the seen dataset using the augmented attribute description. Once trained, the generative model is used in **Stage 2** to generate unlimited images for the unseen classes by providing their corresponding attribute descriptions. This process transforms the zero-shot recognition problem into a conventional supervised image classification task. Then, an image encoder and a classifier in **Stage 2** are trained by a combination of real images from seen classes and generated images of unseen classes. After training, the model can predict the class label for any input image, whether it belongs to a seen or unseen class.

3.2 Stage 1: Generative Model Training

To generate high-quality images for ZSL, we adopt the Diffusion Transformer (DiT) [15], which avoids training on unseen class images, thus aligning with the ZSL setting. Following the design of DiT, given an image x^s from the seen dataset D^s, the input to DiT is a spatial representation $\mathbf{z}_0^s$ through Transformer blocks with conditional modules (AdaLN-Zero [17]). The diffusion forward process is then applied to generate infinite variations of the input by progressively adding noise across timesteps, following the forward diffusion chain as:

$$q(\mathbf{z}_{1:T}^s \mid \mathbf{z}_0^s) = \prod_{t\geq 1} q(\mathbf{z}_t^s \mid \mathbf{z}_{t-1}^s), \tag{1}$$

$$q(\mathbf{z}_t^s \mid \mathbf{z}_{t-1}^s) = \mathcal{N}\left(\mathbf{z}_t^s; \sqrt{1-\beta_t}\mathbf{z}_{t-1}^s, \beta_t \boldsymbol{I}\right), \tag{2}$$

where β_t is a pre-defined variance schedule, and $\mathcal{N}(\cdot)$ denotes a Gaussian distribution with mean $\sqrt{1-\beta_t}\mathbf{z}_{t-1}^s$ and covariance $\beta_t \boldsymbol{I}$. At the final timestep $t = T$, the input becomes nearly pure Gaussian noise. The maximum diffusion step is set to T, and a timestep $t \sim [1, T]$ is randomly sampled during training. The noised inputs at timesteps $t-1$ and t are denoted as $\mathbf{z}_{t-1}^s$ and $\mathbf{z}_t^s$, respectively.

Since only the seen dataset is available during training, the model tends to exhibit a strong bias toward seen classes. Meanwhile, we observe that DiT is capable of generating high-quality images for these seen classes, indicating that the model can effectively capture and reconstruct their primary visual content. To mitigate this bias and improve generalization for unseen classes, we introduce a semantic neighbor-guided conditioning strategy, which enhances the unseen class condition by incorporating guidance from the most semantically similar seen class. Since only seen images are available during training and we employ a concatenation operation to enhance the conditioning input, we construct a combined condition vector $\mathbf{c}_i^s$ by concatenating the attribute descriptions of the same seen classes to train the generative model. Specifically, two differently weighted versions of the attribute description $\mathbf{a}_i^s$ are concatenated and passed through a fully connected layer $\phi(\cdot)$ to produce the augmented condition as:

$$\mathbf{c}_i^s = \phi(\text{concat}(\gamma_1 \mathbf{a}_i^s, \gamma_2 \mathbf{a}_i^s)), \tag{3}$$

where γ_1 and γ_2 are scalar weights that control the contributions of the primary and secondary components of the attribute vector, respectively. During sampling, we further control the relative weights of seen and unseen classes, al-

lowing us to balance between maintaining the quality of generated images and improving diversity across classes. Please see Section 3.3 for details.

The denoising process is then formulated as:

$$p_\theta(\mathbf{z}_t^s|\mathbf{z}_{t-1}^s,\mathbf{c}_i^s)=\mathcal{N}(\mathbf{z}_{t-1}^s;\mu_\theta(\mathbf{z}_t^s,t,\mathbf{c}_i^s),\Sigma_\theta(\mathbf{z}_t^s,t)), \quad (4)$$

$$\mu_\theta(\mathbf{z}_t^s,t,\mathbf{c}_i^s)=\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t^s-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{z}_t^s,t,\mathbf{c}_i^s)\right), \quad (5)$$

where $\alpha_t=1-\beta_t, \bar{\alpha}_t=\sum_{i=1}^{t}\alpha_i$ and θ is the parameters of DiT.

To enhance conditional generation, Classifier-Free Guidance (CFG) [8] is employed. During training, both conditional and unconditional models are jointly optimized using a shared score estimator ϵ_θ. At the inference time, the guidance-modified score is computed as:

$$\hat{\epsilon}_\theta(\mathbf{z}_t^s,t,\mathbf{c}_i^s)=(1+w)\epsilon_\theta(\mathbf{z}_t^s,t,\mathbf{c}_i^s)-w\epsilon_\theta(\mathbf{z}_t^s,t,\emptyset), \quad (6)$$

where w is the guidance scale to adjust the conditioning, and $\emptyset$ denotes a null (unconditional) input. Finally, the generative model is optimized with the following objectives:

$$\mathcal{L}_{\text{vb}}=\sum_t \mathcal{D}_{\text{KL}}\left(q(\mathbf{z}_t^s|\mathbf{z}_{t-1}^s)||p_\theta(\mathbf{z}_{t-1}^s|\mathbf{z}_t^s,\mathbf{c}_i^s)\right), \quad (7)$$

$$\mathcal{L}_{\text{mse}}=\|\epsilon_\theta(\mathbf{z}_t^s)-\epsilon_t\|, \quad (8)$$

where $\mathcal{L}_{\text{vb}}$ is the variational bound loss, $\mathcal{D}_{\text{KL}}$ represents the Kullback–Leibler divergence, and $\mathcal{L}_{\text{mse}}$ minimizes the prediction error of the noise estimator ϵ_θ.

3.3 Stage 2: Zero-Shot Classification

After training, the DiT model generates an unseen image x^u conditioned on attribute descriptions. For brevity, we denote the sampling process of DiT as $G(\cdot)$. The generated unseen image $\tilde{x}^u$ is then obtained as:

$$\tilde{x}^u=G(w\hat{\epsilon}_\theta,\mathbf{c}_j^u), \quad \mathbf{c}_j^u=\phi(\text{concat}(\gamma_1\mathbf{a}_j^u,\gamma_2\mathbf{a}_i^*)), \quad (9)$$

where w is the weight of the condition, $\mathbf{a}_j^u$ and $\mathbf{a}_i^*$ denote the attribute descriptions of the target unseen class and its most similar seen class, respectively. γ_1 and γ_2 control the contribution of the primary (target) and secondary (guide) conditions, respectively. To obtain the reference seen class $\mathbf{a}_i^*$, we perform a nearest-neighbor search to find the class most similar to the target unseen class in the shared semantic space. Since the trained DiT model generates high-quality images of seen classes, adjusting the attribute description weights allows us to preserve the primary characteristics of the generated images.

Once the generated unseen image $\tilde{x}^u$ is obtained, we adopt a standard generative-based ZSL pipeline. To effectively extract image features, we employ ViT-base [6] as the Encoder $E(\cdot)$. Accordingly, the visual features of seen and generated unseen images are extracted as:

$$\mathbf{v}^s=E(x^s), \quad \tilde{\mathbf{v}}^u=E(\tilde{x}^u), \quad (10)$$

where $\mathbf{v}^s$ is the visual feature of the seen image x^s and $\tilde{\mathbf{v}}^u$ is that of the generated unseen image $\tilde{x}^u$. At last, to achieve ZSL prediction, we need to train a Classifier $f(\cdot)$, which predicts attribute probability. Then, $E(\cdot)$ and $f(\cdot)$ are updated using the cross-entropy loss $\mathcal{L}_{\text{ce}}$ as:

$$\begin{aligned}\mathcal{L}_{\text{ce}}=&-\log\frac{\exp(f(\mathbf{v}^s)^\mathsf{T}\cdot\mathbf{a}_y^s)}{\sum_{\mathbf{a}_i^s\in A^s}\exp(f(\mathbf{v}^s)^\mathsf{T}\cdot\mathbf{a}_i^s)}\\&-\log\frac{\exp(f(\tilde{\mathbf{v}}^u)^\mathsf{T}\cdot\mathbf{a}_y^u)}{\sum_{\mathbf{a}_j^u\in A^u}\exp(f(\tilde{\mathbf{v}}^u)^\mathsf{T}\cdot\mathbf{a}_j^u)},\end{aligned} \quad (11)$$

where $f(\mathbf{v}^s)$ and $f(\tilde{\mathbf{v}}^u)$ are predicted attribute probabilities of seen and generated unseen images, respectively, and $\mathbf{a}_y^s$ and $\mathbf{a}_y^u$ are the ground-truth attribute descriptions of seen and generated unseen images, respectively.

After training the model, given an image x, we first extract its visual feature $E(x)$. The prediction is then made by matching the visual feature with the attribute prototypes in the semantic space. The predicted label is given by:

$$\hat{y}=\arg\max_{\mathbf{a}\in\tilde{A}} f(E(x))^T\cdot\mathbf{a}. \quad (12)$$

The definition of $\tilde{A}$ depends on the evaluation setting: In Conventional Zero-Shot Learning (CZSL), where test instances come exclusively from unseen classes, $\tilde{A}=A^u$, and in Generalized Zero-Shot Learning (GZSL), which involves both seen and unseen test classes, $\tilde{A}=A^s\cup A^u$.

3.4 Training Strategy

Training a diffusion model requires substantial computational resources, and even fine-tuning a pre-trained diffusion model can be time-intensive. To overcome this limitation, we follow the approach proposed by Xie et al. [24], which fine-tunes only the bias terms of the model to significantly reduce training time while maintaining performance.

4. Experiments

We conducted extensive experiments on three prominent ZSL benchmark datasets: Animals with Attributes 2 (AWA2) [23], SUN Attribute (SUN) [14], and Caltech-UCSD Birds-200-2011 (CUB) [21]. We followed the Proposed Split (PS) setting [23] to split each dataset into seen and unseen classes.

During inference, *i.e.*, performing CZSL and GZSL classification, we followed the evaluation protocols in [23]. In the CZSL setting, we calculated the average per-class Top-1 accuracy of unseen classes, denoted as *Acc*. For the GZSL scenario, we measured the Top-1 accuracy of the seen and unseen classes, defined as S and U, respectively. We also computed their harmonic mean, defined as $H=2SU/(S+U)$.

To evaluate the effectiveness of the proposed method, we compared it with several state-of-the-art methods, as shown in Table 1, with a particular focus on generative methods. On the CUB dataset, our method achieved a CZSL accuracy of 77.4 and a GZSL harmonic mean of 68.0, indicating effective generalization on fine-grained bird categories. On the SUN dataset, our method achieved 74.7 CZSL accuracy and

Table 1: Performance comparison with state-of-the-art methods on the CUB [21], SUN [14], and AWA2 [23] benchmarks.

Methods	CUB [21]				SUN [14]				AWA2 [23]			
	CZSL	GZSL			CZSL	GZSL			CZSL	GZSL		
	Acc	*U*	*S*	*H*	*Acc*	*U*	*S*	*H*	*Acc*	*U*	*S*	*H*
Embedding-based Methods												
DUET [5]	69.9	63.7	84.7	72.7	64.4	45.7	45.8	45.8	72.3	62.9	72.8	67.5
ZSLViT [3]	78.9	69.4	78.2	73.6	68.3	45.9	48.4	47.3	70.7	66.1	**84.6**	74.2
Feature Generation-based Methods												
FREE [4]	—	60.4	75.4	67.1	—	47.4	37.2	41.7	—	55.7	59.9	57.7
CE-GZSL [7]	70.4	63.1	78.6	70.0	63.3	48.8	38.6	43.1	77.5	63.9	66.8	65.3
FREE+ESZSL [1]	—	51.3	78.0	61.8	—	48.2	36.5	41.5	—	51.6	60.4	55.7
CLSWGAN+DSP [2]	—	60.0	**86.0**	70.7	—	48.3	43.0	45.5	—	51.4	63.8	56.9
VADS [9]	**82.5**	**75.4**	83.6	**79.3**	**76.3**	**64.6**	**49.0**	**55.7**	**86.8**	**74.1**	74.6	**74.3**
Image Generation-based Method												
Ours	77.4	65.8	70.4	68.0	74.7	61.0	36.3	45.5	76.0	66.8	83.7	**74.3**

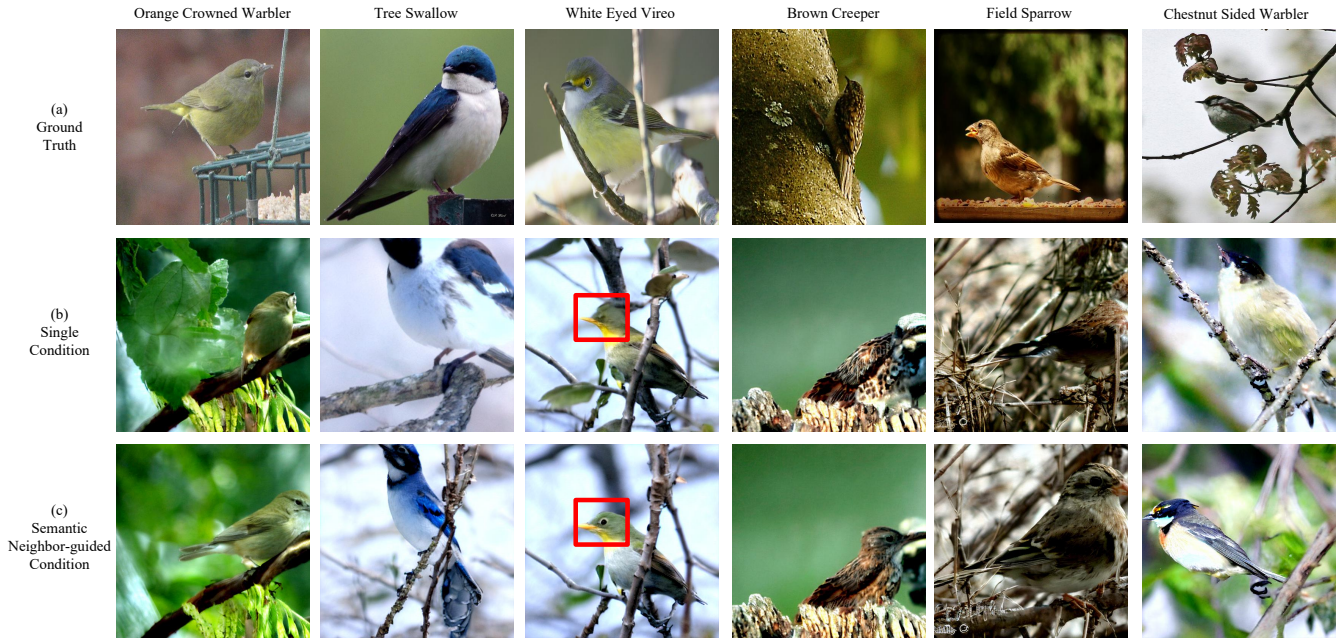


Fig. 2: Comparative results of generated images in different condition settings on CUB [21].

a harmonic mean of 45.5, ranking among the top methods in the classification of unseen classes, although with slightly lower seen-class accuracy. On the AWA2 dataset, our method achieved 76.0 in CZSL and a harmonic mean of 74.3 in GZSL, matching the best-performing method (VADS [9]) and outperforming other generative and embedding-based methods. These results demonstrate that our method effectively enhanced cross-domain alignment while maintaining strong recognition capability for both seen and unseen classes.

To better illustrate the effectiveness of the enhanced condition, we compared images generated with single and augmented conditions as shown in Fig. 2. Semantic neighbor-guided condition inputs produced clearer and more complete images, while single-condition results often lacked structure. Moreover, the augmented condition enabled the generation of more detailed and semantically accurate images, for example, finer visual characteristics were better captured for the "White-Eyed Vireo" class.

5. Conclusion

We proposed a novel generative-based ZSL method that leveraged DiT [15] as the generation model. Unlike existing generative-based ZSL methods that generate features, our method directly generates unseen images, making the results more interpretable and visually observable. To further improve the quality of the generated images, we introduced an augmented conditioning mechanism that enhances unseen image generation.

For future work, we plan to optimize the conditioning mechanism in the style of image editing, allowing more fine-grained and controllable image generation.

Acknowledgements

This work was partly supported by JST BOOST, Japan Grant Number JPMJBS2422, and JSPS Kakenhi JP24H00733. The authors would like to take this opportunity to thank the "THERS Nagoya University, Program for Fostering Next Generation AI Professionals."

References

[1] Cetin, S.: Closed-form sample probing for training generative models in zero-shot learning, Master's thesis, Middle East Technical University, Ankara, Turkiye (2022).

[2] Chen, S., Hou, W., Hong, Z., Ding, X., Song, Y., You, X., Liu, T. and Zhang, K.: Evolving semantic prototype improves generative zero-shot learning, *Computing Research Repository arXiv Preprint,* arXiv:2306.06931 (2023).

[3] Chen, S., Hou, W., Khan, S. and Khan, F. S.: Progressive semantic-guided vision transformer for zero-shot learning, *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23964–23974 (2024).

[4] Chen, S., Wang, W., Xia, B., Peng, Q., You, X., Zheng, F. and Shao, L.: FREE: Feature refinement for generalized zero-shot learning, *Proceedings of the 18th IEEE/CVF International Conference on Computer Vision*, pp. 122–131 (2021).

[5] Chen, Z., Huang, Y., Chen, J., Geng, Y., Zhang, W., Fang, Y., Pan, J. Z. and Chen, H.: Duet: cross-modal semantic grounding for contrastive zero-shot learning, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 1, pp. 405–413 (2023).

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, Sylvain, U. J. and Neil, H.: An image is worth 16x16 words: Transformers for image recognition at scale, *Computing Research Repository arXiv Preprint,* arXiv:2010.11929 (2020).

[7] Han, Z., Fu, Z., Chen, S. and Yang, J.: Contrastive embedding for generalized zero-shot learning, *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2371–2381 (2021).

[8] Ho, J. and Salimans, T.: Classifier-free diffusion guidance, *Computing Research Repository arXiv Preprint,* arXiv:2207.12598 (2022).

[9] Hou, W., Chen, S., Chen, S., Hong, Z., Wang, Y., Feng, X., Khan, S., Khan, F. S. and You, X.: Visual-augmented dynamic semantic prototype for generative zero-shot learning, *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23627–23637 (2024).

[10] Lampert, C. H., Nickisch, H. and Harmeling, S.: Attribute-based classification for zero-shot visual object categorization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 36, No. 3, pp. 453–465 (2013).

[11] Li, A. C., Prabhudesai, M., Duggal, S., Brown, E. and Pathak, D.: Your diffusion model is secretly a zero-shot classifier, *Proceedings of the 19th IEEE/CVF International Conference on Computer Vision*, pp. 2206–2217 (2023).

[12] Mishra, A., Krishna Reddy, S., Mittal, A. and Murthy, H. A.: A generative model for zero shot learning using conditional variational autoencoders, *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2188–2196 (2018).

[13] Narayan, S., Gupta, A., Khan, F. S., Snoek, C. G. and Shao, L.: Latent embedding feedback and discriminative features for zero-shot classification, *Proceedings of the 16th European Conference on Computer Vision, vol.22*, pp. 479–495 (2020).

[14] Patterson, G. and Hays, J.: SUN attribute database: Discovering, annotating, and recognizing scene attributes, *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2751–2758 (2012).

[15] Peebles, W. and Xie, S.: Scalable diffusion models with Transformers, *Proceedings of the 19th IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205 (2023).

[16] Pennington, J., Socher, R. and Manning, C. D.: GloVe: Global Vectors for word representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543 (2014).

[17] Perez, E., Strub, F., De Vries, H., Dumoulin, V. and Courville, A.: FiLM: Visual reasoning with a general conditioning layer, *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, No. 1, pp. 3942–3951 (2018).

[18] Schonfeld, E., Ebrahimi, S., Sinha, S., Darrell, T. and Akata, Z.: Generalized zero- and few-shot learning via aligned variational autoencoders, *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8247–8255 (2019).

[19] Tian, Y., Fan, L., Chen, K., Katabi, D., Krishnan, D. and Isola, P.: Learning vision from models rivals learning vision from data, *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15887–15898 (2024).

[20] Tian, Y., Fan, L., Isola, P., Chang, H. and Krishnan, D.: StableRep: Synthetic images from text-to-image models make strong visual representation learners, *Advances in Neural Information Processing Systems*, Vol. 36, pp. 48382–48402 (2023).

[21] Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S. J. and Perona, P.: Caltech-UCSD Birds 200, *Technical Report CNS-TR-2010-001, California Institute of Technology, Pasadena, CA, USA* (2010).

[22] Xian, Y., Lorenz, T., Schiele, B. and Akata, Z.: Feature generating networks for zero-shot learning, *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5542–5551 (2018).

[23] Xian, Y., Sharma, S., Schiele, B. and Akata, Z.: f-VAEGAN-D2: A feature generating framework for any-shot learning, *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10275–10284 (2019).

[24] Xie, E., Yao, L., Shi, H., Liu, Z., Zhou, D., Liu, Z., Li, J. and Li, Z.: DiffFit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning, *Proceedings of the 19th IEEE/CVF International Conference on Computer Vision*, pp. 4230–4239 (2023).