Why Training with Synthetic Data Fails for OOD: Distribution Gap Amplifies Noise Misalignment in Diffusion Models

Ying Hua[1], Jessica Bader[1,2,3], Jae Myung Kim[2,3,4], Zeynep Akata[1,2,3]
[1]Technical University of Munich [2]Munich Center for Machine Learning
[3]Helmholtz Munich [4]University of Tübingen

Abstract

Generative models can produce high-fidelity and diverse images, motivating growing interest in using synthetic data to augment training sets for downstream computer vision tasks such as image classification. However, the benefits of using synthetic data diminish when the target data is out-of-distribution (OOD) relative to the generative model's pretraining data, such as satellite imagery. We analyze this limitation and find that significant distribution gaps amplify the effect of signal leakage (containing channel mean information) in noisy latent during diffusion training. Consequently, it amplifies the mismatch between noise distribution during training and inference and degrades synthetic quality. To address this issue, we propose two strategies: ***Mean-Perturbed Noise*** *and* ***Mean-Aligned Noise*** *to align the noise distribution. Experiments demonstrate that our methods correct low-frequency characteristics (*e.g. *color) of synthetic satellite images and improve the performance of classifiers trained on synthetic data in OOD scenarios.*

1. Introduction

Deep vision models and tasks rely on large-scale training datasets to achieve strong performance. Many recent models require millions or even billions of training images. However, collecting and annotating such large-scale training datasets poses significant time and labor challenges. Given the remarkable performance of text-to-image models in synthesizing photorealistic images, several works have explored augmenting real training data with synthetic data to address data scarcity in downstream model training [6, 11, 14]. Specifically, they use few-shot real data as guidance to improve the image generation with diffusion models and find that synthetic data generated in this way is beneficial for classifier training. However, the classification performance gain achieved with synthetic data varies across different benchmarks. For instance, while DataDream [14] achieves classification performance comparable to training on real few-shot data when using synthetic data alone across the majority of classification benchmarks, it shows a larger performance gap on EuroSAT [12] and DTD [5]. Intuitively, this inconsistency may arise from the distribution gap between the target data and the training data of the generative model. A large data distribution gap makes it difficult for the generative model to learn from limited few-shot target-domain data and capture the key features needed to generate high-quality images.

To investigate whether and how the distribution gap between the generative model's pretraining data and the downstream data correlates with the classification performance gap, we analyze the benchmarks used in DataDream [14]. This reveals a significant correlation: downstream datasets with a larger distribution shift to the pretraining dataset LAION [29] (*e.g.* satellite datasets), exhibit a larger classification performance gap between classifiers trained on solely real few-shot data and synthetic data. This indicates that a large distribution gap hinders the model's ability to adapt to the target domain in few-shot fine-tuning setting and to produce synthetic images effective for classifier training.

To address the impact of distribution gap on generative performance, we focus on noise initialization. Previous work reveals that under common noise schedules, signal leakage containing channel mean can remain in the final noising step, leading to a mismatch between training and inference that degrades image quality [19]. This discrepancy becomes more pronounced when the generative model is applied to out-of-distribution data. To mitigate this, we propose strategies to bridge this gap between training and inference. First, we propose Mean-Perturbed Noise to mitigate the impact of signal leakage in noisy latent during diffusion training. In addition, we propose Mean-Aligned Noise to directly estimate the leaked channel mean information and calibrate the inference noise distribution. Our experiments demonstrate that the proposed methods improve the generation quality in OOD satellite images, and consequently enhance the performance of classifiers trained on them.

Table 1. Distribution gaps between downstream datasets and LAION. Downstream datasets: ImageNet (IN) [27], Caltech 101 (CAL) [17], DTD [5], EuroSAT (EuSAT) [12], FGVC Aircraft (AirC) [20], Oxford Pets (Pets) [24], Stanford Cars (Cars) [16], SUN397(SUN) [33], Food101 (Food) [2], Flowers 102 (FLO) [22], PatternNet (Patt) [34], NWPU-RESISC45 (RES) [4]. LAION includes 411 shards from LAION-400m [29].

metric	IN	CAL	DTD	EuSAT	AirC	Pets	Cars	SUN	Food	FLO	Patt	RES
density	0.38	0.41	0.36	**0.07**	0.25	0.56	0.49	0.48	0.52	0.76	**0.03**	**0.07**
rarity	15.53	13.96	**18.50**	**18.83**	12.73	14.45	11.15	14.38	15.14	15.06	**21.04**	**20.06**

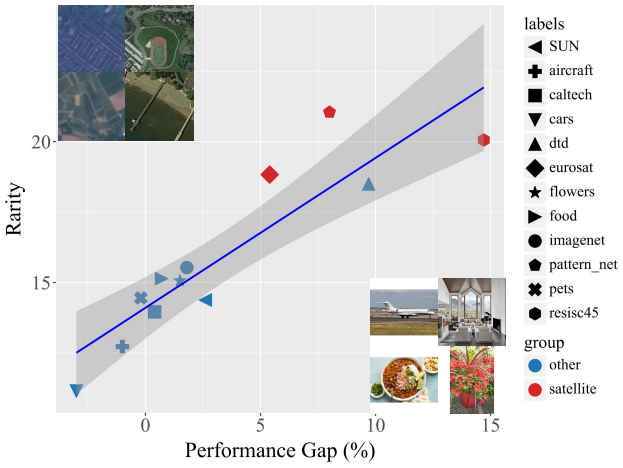

Figure 1. Correlation plot between distribution gap (represented by rarity [10]) and classification performance gap. Pearson coefficient [30] $R = 0.9$, p-value$= 7.14e-5$. Synthetic training images yield smaller performance gains when target data is out-of-distribution (*e.g.* satellite samples in the upper left) for the generative model (trained with LAION-data shown in the lower right).

2. Related Work

Training with Synthetic Data. Synthetic images have gained attention as training data for augmenting real datasets [1, 3, 7, 35] or pre-training models on synthetic data followed by few-shot fine-tuning on real data [9, 32]. Other research has investigated training exclusively on synthetic datasets [9, 28]. Efforts generally center on two key properties: *faithfulness* and *diversity*. For faithfulness, there are techniques like filtering [7, 11, 18] and including class information [28]. Diversity can be enhanced by reducing guidance scale [28], generating varied prompts with language models [9, 11], incorporating domain or background details [7, 28, 31], or using multiple prompt templates [3].

3. Methods

3.1. Distribution-Performance Gap Correlation

To validate our hypothesis that significant distribution gaps between pretraining and downstream data lead to larger classification performance gaps between classifiers trained on synthetic data and those trained on real few-shot data, we empirically analyze by first quantifying the distribution gap and then examining its correlation with the corresponding performance gap. First, we measure the distribution gaps between the pretraining dataset LAION [29] used by Stable Diffusion [26] and classification benchmarks evaluated in DataDream [14], along with two additional satellite datasets: PatternNet [34] and RESISC45 [4]. We extract features with DINOv2 [23] to provide a faithful representation of the distributions. To quantify the distribution gap, we employ *density* [21] and *rarity* [10] for pairwise distribution comparisons between downstream data and LAION.

In Table 1, we observe that all satellite datasets show significantly lower density and higher rarity, indicating a larger distribution shift to LAION compared to other downstream datasets, while DTD only exhibits significantly higher rarity. Given the measured distribution gaps between downstream datasets and LAION, we jointly analyze these results with the classification performance gaps reported in DataDream [14]. As illustrated in Figure 1, we observe a significant linear correlation between the distribution gap, represented by rarity, and the performance gap. A similar correlation is also observed when using density as the metric to compute the distribution gap. This suggests that large distribution shifts from the pretraining data hinder the generative model's ability to adapt to the target domain. As a result, the model struggles to produce high-quality synthetic images, leading to a noticeable performance gap between classifiers trained on synthetic and real data.

3.2. Impact of distribution gap in diffusion process

To understand why the distribution gap affects synthetic data, we examine the role of the diffusion process in this relationship. Diffusion model training consists of the forward process and the reverse process: the former gradually adds Gaussian noise to the input image, while the latter learns to denoise and reconstruct the image. Given a pre-defined noise schedule $\beta_1, ..., \beta_T$, the forward process can be defined as follows with the reparameterization trick [15]:

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0, \quad \boldsymbol{\epsilon}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \qquad (1)$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\boldsymbol{x}_0$ is the latent representation of sample $\boldsymbol{x}$. Ideally, the noise schedule is

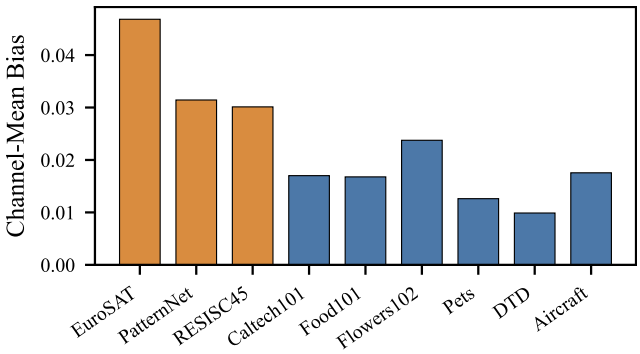

Figure 2. Channel mean bias of $\boldsymbol{x}_T$. As a reference, channel mean bias of pure noise ≈ 0.

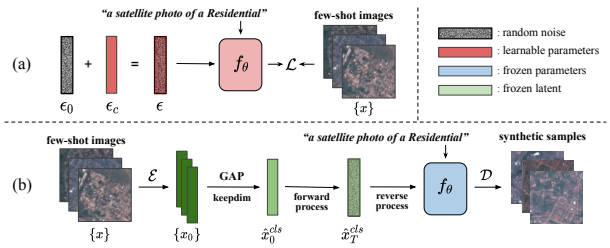

Figure 3. Method overview. (a) Mean-Perturbed Noise. $\boldsymbol{\epsilon}_c$ is channel-wise offset noise providing additional channel mean perturbation. (b) Mean-Aligned Noise. Global Average pooling averages the latent representations of few-shot data over pixels to obtain a channel-wise mean embedding for each few-shot sample and averages across samples to obtain a class-wise representation.

chosen such that $\bar{\alpha}_T = 0$ and $\boldsymbol{x}_T$ at the last diffusion step is pure Gaussian noise. In practice, however, $\bar{\alpha}_T$ is defined as nonzero to avoid dividing by zero, which leaves a signal leakage $\sqrt{\bar{\alpha}_T}\boldsymbol{x}_0$ containing low-frequency information (*e.g.* channel mean) in $\boldsymbol{x}_T$ [19]. As the reverse model is trained on $\boldsymbol{x}_T$, it can rely on the leaked channel mean in the noise for reconstruction instead of learning the low-frequency feature directly. During inference, however, the model is initialized from pure Gaussian noise, presenting a discrepancy between diffusion training and inference. Consequently, the model may fail to accurately represent low-frequency features, reducing the generation quality.

In our scenario, the impact of discrepancy between diffusion training noise and inference noise depends on how well the target distribution aligns with the pretraining distribution. As shown in Figure 2, we observe that satellite datasets with larger distribution shifts from LAION exhibit significantly larger channel mean bias in the noisy latent $\boldsymbol{x}_T$ compared to LAION-like datasets. Consequently, the misalignment between the noise distribution encountered during diffusion training and inference in terms of channel mean becomes more pronounced, making it harder for the model to reconstruct the low-frequency features.

3.3. Mean-Perturbed Noise during training

To address the training-inference misalignment, we first introduce Mean-Perturbed Noise to calibrate the noise distribution during training, based on offset noise [8], a technique originally designed to resolve diffusion model's limitation in generating images with extreme brightness or darkness. We extend it to mitigate signal leakage in diffusion training.

As shown in Figure 3 (a), Mean-Perturbed Noise adds channel-wise Gaussian offset $\boldsymbol{\epsilon}_c$ to standard Gaussian noise $\boldsymbol{\epsilon}_0$ during training. Consequently, noisy image at the last diffusion step T is written as follows:

$$\boldsymbol{x}_T = \sqrt{\bar{\alpha}_T}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_T}(\boldsymbol{\epsilon}_0 + \boldsymbol{\epsilon}_c) \tag{2}$$

where $\boldsymbol{\epsilon}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\epsilon}_c \sim \mathcal{N}(\mathbf{0}, \lambda\boldsymbol{\Sigma}_\mathbf{c})$. Here $\boldsymbol{\Sigma}_\mathbf{c}$ is a block-diagonal covariance matrix with ones within each channel and the offset $\boldsymbol{\epsilon}_c$ is then constant over all pixels within the same channel and varies across different channels. The magnitude of injected offset noise is controlled by hyperparameter λ. As per-channel mean of the standard Gaussian noise $\boldsymbol{\epsilon}_0$ has very small variance after averaging over $H \times W$ (64×64 for input image in resolution 512×512) pixels, making it unable to perturb the signal leakage containing channel-mean information. In contrast, the channel-wise offset explicitly increases the variance of the per-channel mean, injecting a controllable channel-mean shift with the parameter λ. This channel-mean shift varies randomly during training - randomly canceling or amplifying the signal leakage, preventing the model from relying on the channel-mean information in the signal leakage.

3.4. Mean-Aligned Noise during inference

Alternatively, we keep the noise distribution with signal leakage during diffusion training untouched and estimate it to define the inference noise. We use the noise schedule defined in Stable Diffusion [26] to compute $\sqrt{\bar{\alpha}_T} = 0.0682373$ and $\sqrt{1-\bar{\alpha}_T} = 0.997558$ at $T = 1000$. As shown in Figure 3 (b), $\hat{\boldsymbol{x}}_0$ is estimated from the same input few-shot data used for fine-tuning the model. Concretely, we estimate $\hat{\boldsymbol{x}}_0^{\text{cls}}$ per-class by computing the channel-means of the input few-shot data. As a result, instead of pure Gaussian noise, we define the inference noise as follows:

$$\hat{\boldsymbol{x}}_T^{\text{cls}} = \sqrt{\bar{\alpha}_T}\hat{\boldsymbol{x}}_0^{\text{cls}} + \sqrt{1-\bar{\alpha}_T}\boldsymbol{\epsilon}_0, \quad \boldsymbol{\epsilon}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3}$$

This allows the trained model to leverage channel-mean information during inference and better align the low-frequency features of the generated image to the real image.

4. Experiments

4.1. Experimental Setup

We evaluate on three OOD satellite datasets: EuroSAT [12], PatternNet [34] and NWPU-RESISC45 [4]. We finetune

Table 2. Accuracy of CLIP ViT-B/16 fine-tuned on 500 synthetic images per class generated with 16 real examples. Real fine-tuned refers to training the base classifier with 16 real few-shot images. Results are computed on three random seeds.

	EuroSAT	RESISC45	PatternNet
zero-shot	40.26	27.26	34.01
DataDream [14]	$87.22_{\pm 1.52}$	$67.94_{\pm 0.91}$	$86.90_{\pm 1.24}$
Mean-Perturbed Noise	$\mathbf{90.46}_{\pm 1.59}$	$\mathbf{68.94}_{\pm 0.22}$	$\mathbf{87.46}_{\pm 1.68}$
Mean-Aligned Noise	$\mathbf{88.93}_{\pm 1.68}$	$\mathbf{70.00}_{\pm 0.58}$	$\mathbf{87.30}_{\pm 0.91}$
real fine-tuned	$92.60_{\pm 0.41}$	$82.61_{\pm 0.55}$	$94.91_{\pm 0.53}$

Table 3. FID scores between real and synthetic images on satellite datasets. A lower FID indicates that the synthetic data distribution is closer to the real data distribution.

	EuroSAT ↓	RESISC45 ↓	PatternNet ↓
DataDream [14]	62.20	26.18	28.23
Mean-Perturbed Noise	**31.91**	**19.74**	**20.10**
Mean-Aligned Noise	**28.53**	**15.82**	**17.03**

Stable Diffusion 2.1-base [26] with 16-shot samples per class to generate synthetic images. Few-shot finetuning configurations are: 200 training epochs, batch size 8, learning rate $= 1e-4$ for a cosine scheduler. For Mean-Perturbed Noise, we tune the offset scale $\lambda \in [0.01, 0.1]$ and find $\lambda = 0.05, \lambda = 0.07, \lambda = 0.035$ give the best results for EuroSAT, PatternNet and NWPU-RESISC45, respectively. We use 50 inference steps and guidance scale 2.0 to generate 500 images per class. Synthetic images are then used to train CLIP VIT-B/16 [25] classifiers to evaluate their performance in adapting downstream models.

4.2. Qualitative Comparisons

Figure 4 shows qualitative examples of images generated by our methods and DataDream [14], and the few-shot real images used for fine-tuning on one representative class per satellite benchmark. Compared to DataDream, Mean-Perturbed Noise and Mean-Aligned Noise produce synthetic satellite images with more accurate color tones, as shown in "SeaLake". Beyond color fidelity, our methods also improve foreground-background separation in the generated images: "River" and "Bridge" are more distinct than in DataDream. The improved visual fidelity achieved by our methods is further supported by the FID [13] shown in Table 3. Both our methods achieve significantly lower FID than DataDream across all datasets, indicating improved alignment between the synthetic and real distributions.

4.3. Quantitative Comparisons

Table 2 compares classifiers trained on synthetic data and those trained on real few-shot data. We demonstrate that synthetic data generated by Mean-Perturbed Noise and Mean-Aligned Noise significantly outperform SOTA method DataDream across all three datasets. On EuroSAT, Mean-Perturbed Noise achieves 90.46% top-1 accuracy, improving over DataDream by 3.24%. On NWPU-RESISC45, Mean-Aligned Noise performs best with 70% top-1 accuracy. On PatternNet, Mean-Perturbed Noise and Mean-Aligned Noise perform comparably, with 0.56 and 0.4 percentage point improvement over DataDream.

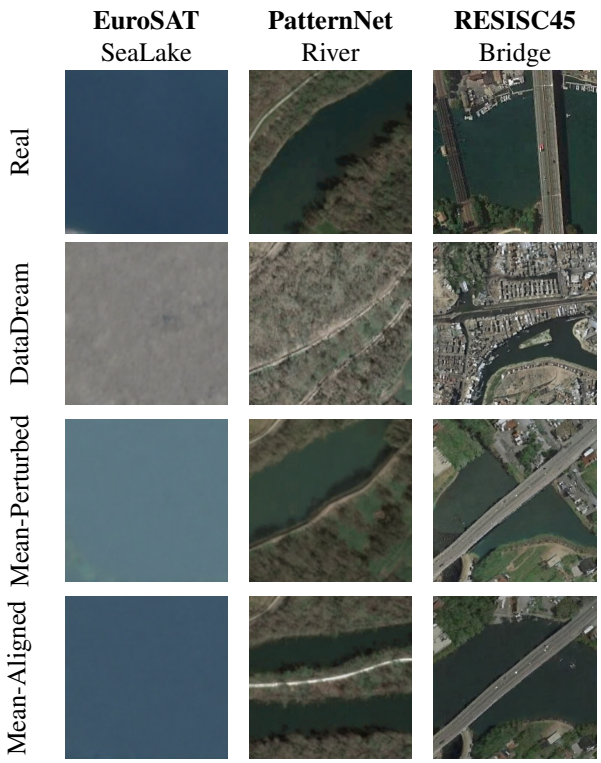

Figure 4. Synthetic image examples from different methods, compared with real samples.

5. Conclusion

We investigate how distribution gaps between generative pretraining data and downstream domains hinder few-shot fine-tuning and synthetic quality. Our analysis reveals a strong correlation between distribution gap and the performance gap between classifiers trained on synthetic and real data, stemming from noise misalignment in diffusion models caused by signal leakage containing channel mean during training. We propose two solutions: Mean-Perturbed Noise, which increases the channel mean variance to mitigate the effect of signal leakage during training, and Mean-Aligned Noise, which calibrates inference noise to match training conditions. Experiments on OOD satellite data show consistent improvements in synthetic quality and downstream performance.

6. Acknowledgments

This work was partially funded by the ERC (853489 - DEXIM) and the Alfried Krupp von Bohlen und Halbach Foundation, which we thank for their generous support. The authors gratefully acknowledge the scientific support and resources of the AI service infrastructure LRZ AI Systems provided by the Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities (BAdW), funded by Bayerisches Staatsministerium fur Wissenschaft und Kunst (StMWK).

References

[1] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *arXiv*, 2023. 2

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 2

[3] Max F Burg, Florian Wenzel, Dominik Zietlow, Max Horn, Osama Makansi, Francesco Locatello, and Chris Russell. Image retrieval outperforms diffusion models on data augmentation. *TMLR*, 2023. 2

[4] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 2, 3

[5] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 1, 2

[6] Victor G. Turrisi da Costa, Nicola Dall'Asen, Yiming Wang, Nicu Sebe, and Elisa Ricci. Diversified in-domain synthesis with efficient fine-tuning for few-shot classification, 2023. 1

[7] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E. Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation, 2023. 2

[8] Nicholas Guttenberg. Diffusion with offset noise, 2023. 3

[9] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv*, 2024. 2

[10] Jiyeon Han, Hwanil Choi, Yunjey Choi, Junho Kim, Jung-Woo Ha, and Jaesik Choi. Rarity score: A new metric to evaluate the uncommonness of synthesized images. *arXiv preprint arXiv:2206.08549*, 2022. 2

[11] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and XIAOJUAN QI. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023. 1, 2

[12] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *AEROS*, 2019. 1, 2, 3

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 4

[14] Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. Datadream: Few-shot guided dataset generation. *ECCV*, 2024. 1, 2, 4

[15] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 2

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 2

[17] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, 2022. 2

[18] Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the power of synthetic data on few-shot object detection. In *CVPR*, 2023. 2

[19] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *WACV*, 2024. 1, 3

[20] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2

[21] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *ICML*, pages 7176–7185. PMLR, 2020. 2

[22] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 2

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[24] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 2

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2

[28] Mert Bulent Sariyildiz, Alahari Karteek, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transfer-

able representations from synthetic imagenet clones. *CVPR*, 2022. 2

[29] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1, 2

[30] Philip Sedgwick. Pearson's correlation coefficient. *Bmj*, 345, 2012. 2

[31] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion, 2023. 2

[32] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *NeurIPS*, 2024. 2

[33] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 2

[34] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145:197–209, 2018. 2, 3

[35] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *arXiv*, 2023. 2