ChimeraLoRA: Multi-Head LoRA-Guided Synthetic Datasets

Hoyoung Kim[1] Minwoo Jang[1] Jabin Koo[2] Sangdoo Yun[3] Jungseul Ok[1,2]

Graduate School of AI, POSTECH[1], Dept. of CSE, POSTECH[2], NAVER AI Lab[3]

https://cskhy16.github.io/chimeralora

Abstract

Beyond general recognition tasks, specialized domains and fine-grained settings often encounter data scarcity, especially for tail classes. To obtain less biased and more reliable models under such scarcity, practitioners leverage diffusion models to supplement underrepresented regions of real data. Specifically, recent studies fine-tune pretrained diffusion models with LoRA on few-shot real sets to synthesize additional images. While an image-wise LoRA trained on a single image captures fine-grained details yet offers limited diversity, a class-wise LoRA trained over all shots produces diverse images as it encodes class priors yet tends to overlook fine details. To combine both benefits, we separate the adapter into a class-shared LoRA A for class priors and per-image LoRAs $\mathcal{B}$ for image-specific characteristics. To expose coherent class semantics in the shared LoRA A, we propose a semantic boosting by preserving class bounding boxes during training. For generation, we compose A with a mixture of $\mathcal{B}$ using coefficients drawn from a Dirichlet distribution. Across diverse datasets, our synthesized images are both diverse and detail-rich while closely aligning with the few-shot real distribution, yielding robust gains in downstream classification accuracy.

1. Introduction

While general recognition tasks enjoy abundant and class-balanced data, specialized domains often face data scarcity [3, 25, 46] and long-tailed class distributions [5, 27]. For example, class rarity in fine-grained tasks can limit data collection, leaving only a few labeled images per class [12, 35]. Training under such data scarcity often causes models to overfit and learn decision boundaries biased toward majority classes, degrading generalization performance [2]. To supplement limited data, recent work leverages generative priors in pretrained text-to-image diffusion models [38] to synthesize additional training images by conditioning class names with text prompts [15]. However, without guidance from real images, synthetic data easily drifts from the target distribution and lower downstream accuracy [15, 44].

To narrow the real-to-synthetic gap, recent work exploits few-shot real images [6, 15, 20, 21]. Specifically, a training-free image-wise baseline initializes the diffusion process from features of a single reference image to synthesize samples close to that reference [15]. Beyond this, an image-wise variant embeds the reference into a diffusion model by fine-tuning lightweight low-rank adapters (LoRAs) [17], enabling the model to capture fine-grained details [21]. However, these image-wise approaches make it difficult to generate diverse images, as they rely on a single image. To leverage the remaining images of the same class, a class-wise LoRA is fine-tuned on all shots to encode class-level priors and promote diversity, yet the resulting samples often overlook instance-specific details [20]. This trade-off stems from adapting the diffusion model with LoRA at a single granularity, either an image or a class, motivating a unified image- and class-level adaptation.

To generate synthetic images that are both diverse and fine-grained, we adopt a multi-head LoRA architecture [11, 45]. Specifically, LoRA [17] approximates updates to a large weight matrix as the product of two low-rank matrices, LoRA A and LoRA B. We give these two LoRAs distinct roles by sharing a single LoRA A across all few-shot images and assigning each image its own LoRA B. In this way, the shared LoRA A captures class-level priors that drive diverse generation, while the per-image LoRA heads $\mathcal{B}$ encode instance-specific details. For training, we freeze the base diffusion model and jointly fine-tune the LoRA A and B. To promote coherent class semantics in the shared A, we propose a semantic boosting technique that utilizes bounding boxes localized by Grounded Segment Anything Model [22, 26].

At generation time, we fix A and mix multiple heads from $\mathcal{B}$ with nonnegative coefficients sampled from a Dirichlet distribution. Thanks to the shared A and a different Dirichlet-weighted head for each generated image, the synthesized images exhibit both fine-grained details and diversity. In addition, our semantic boosting drives cover-

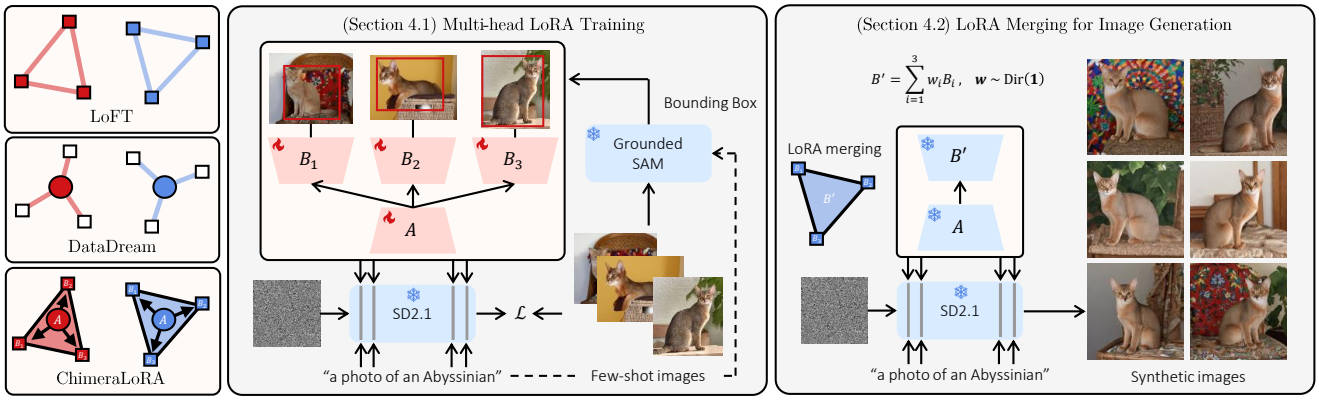


Figure 1. *An overview of the proposed method.* (left) We synthesize images with a multi-head LoRA that integrates the strengths of image-wise LoRA (LoFT [21]) and class-wise LoRA (DataDream [20]). The blue and red regions indicate where LoRA is applied during generation. (center) Given few-shot images, we fine-tune the multi-head LoRA while preserving bounding boxes obtained from Grounded-SAM [22, 26]. (right) We merge LoRA heads using weights sampled from a Dirichlet distribution to obtain diverse synthetic images.

age of the full visible extent of a target class, rather than a partial visibility. Across diverse classification tasks, including realistic medical domains and long-tailed scenarios, our synthetic images are not only distribution-aligned to the few-shot references, yielding robust gains over baselines, but also qualitatively diverse and detailed. Figure 1 compares our method with prior work and outlines the overall pipeline.

Our main contributions are summarized as follows:

- We present a multi-head LoRA framework in which LoRA A encodes class-level priors and LoRA heads $\mathcal{B}$ capture instance-specific details, producing diverse and fine-grained synthetic images (Section 4.1).
- Our synthetic datasets generally improve downstream accuracy across various benchmarks, including specialized domains and long-tailed settings (Section 5.2).
- We generate synthetic images aligned with the real few-shot distribution and analyze the synthetic-to-real gap both qualitatively and quantitatively (Section 5.3).

2. Related Work

Few-shot guided synthetic datasets. Recent methods for synthetic dataset generation leverage real images as guidance rather than relying solely on text prompts. For example, IsSynth [14] conditions on the latent space of real images to improve performance, and DISEF [6] perturbs real-image latents for generation. Recently, LoFT [21] trains image-wise LoRA adapters and mixes their contributions at sampling. While these single-image approaches capture fine-grained details, they overlook the broader class distribution. To improve class-level coverage, DataDream [20] trains class-wise LoRA adapters on all shots of a class, emphasizing generality over image-level fidelity. To combine both strengths, we introduce a shared LoRA A for class-level priors and LoRA heads $\mathcal{B}$ for instance-specific details, unifying class-level generality with image-level fidelity.

Multi-head LoRA architectures. Even in a single-head LoRA [17], recent analysis identifies an asymmetric role: LoRA A acts as a simple projection agnostic to the input distribution, whereas LoRA B tends to capture the input data distribution [49]. Building on this observation, multi-head designs explicitly allocate task or instance specific capacity to multiple LoRA B heads while sharing LoRA A. Specifically, HydraLoRA [45] employs expert routing over LoRA B heads, FedSA-LoRA [11] keeps local LoRA B per client to address heterogeneity in federated learning, and AsymLoRA [47] applies this paradigm to multimodal instruction tuning. Following these developments, we adopt an asymmetric multi-head LoRA architecture to fine tune diffusion models on few shot images.

Semantic preservation in data augmentation. To address the challenge of semantic preservation during data augmentation, several methods propose alternatives beyond standard transformations. For instance, KeepAugment [10] utilizes saliency maps to preserve informative regions, while ObjectCrop [31] and ContrastiveCrop [34] aim to obtain reliable positive samples in contrastive learning by leveraging object proposals and semantic-aware localization, respectively. With the emergence of Segment Anything Model (SAM) [22], SAMAug [48] simply combines raw images with SAM-generated masks for medical image segmentation. However, these approaches do not explicitly ensure object integrity under cropping. In contrast, we focus on generating images with complete objects for synthetic datasets by leveraging bounding boxes from Grounded-SAM [22, 26].

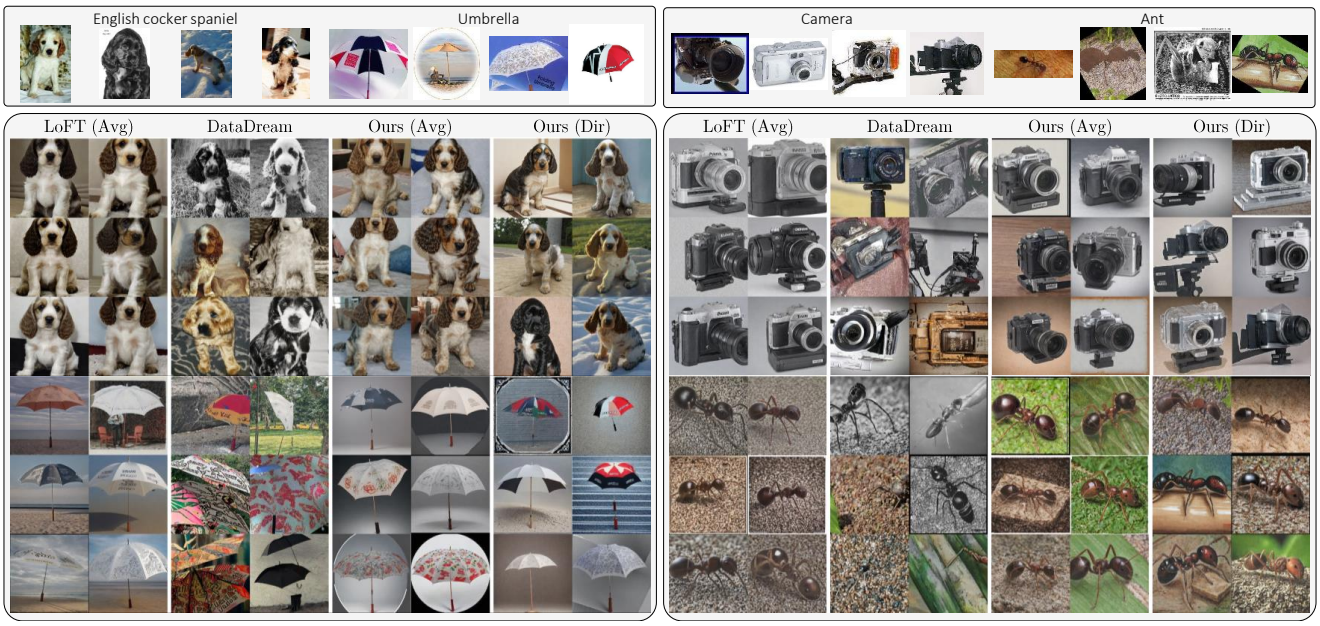


Figure 2. *Qualitative results of synthetic images.* (top) Four real images per class. (bottom) Synthetic images generated with LoRA based methods. For the camera class, LoFT (image-wise LoRA) shows low diversity with near duplicate single viewpoint shots, while DataDream (class-wise LoRA) increases diversity but lowers fidelity, often failing to render a camera. Our multi-head LoRA produces accurate cameras across varied viewpoints. Here, Avg merges heads with uniform weights and Dir uses Dirichlet sampled weights.

3. Preliminaries

To supplement few shot examples, we leverage the generative power of large-scale diffusion models combined with parameter-efficient LoRA adaptation. Therefore, as preliminary background, we first introduce the fundamentals of latent diffusion models in Section 3.1, and then review previous approaches that employ LoRA for synthetic image generation in Section 3.2.

3.1. Latent Diffusion Models

Latent diffusion models (LDMs), such as Stable Diffusion [38], are probabilistic generative models designed to generate high-resolution images conditioned on text prompts y. Let $\mathcal{D}$ be a dataset of image-text pairs. For $(x, y) \in \mathcal{D}$, let an encoder $\mathcal{F}$ map the image x into a latent representation $z = \mathcal{F}(x)$. The forward diffusion process progressively adds Gaussian noise ϵ sampled from a standard normal distribution $\mathcal{N}(0, 1)$ to the latent z over t steps, resulting in increasingly noisy latent variables z_t. The reverse process learns to iteratively remove this noise, conditioned on the text prompt y. Specifically, an intermediate representation $\tau(y)$ obtained from a pretrained text encoder τ is provided to the cross-attention layers of the UNet [39] to guide the denoising process. To achieve this, the conditional LDM parameterized by θ is trained with the objective as follows:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau(y))\|_2^2\right], \quad (1)$$

where t is uniformly sampled from $\{1, \ldots, T\}$ with T the number of diffusion timesteps. To generate a synthetic image, an initial latent noise z_T is iteratively denoised conditioned on the text prompt y, and the resulting latent z_0 is decoded by a decoder $\mathcal{G}$ to produce the final image $x' = \mathcal{G}(z_0)$. While previous works [37, 41] primarily focus on the visual quality of individual generated images, we investigate whether the collection of synthetic images produced by LDMs can serve as effective training datasets for downstream tasks.

3.2. Single-head LoRA-Guided Synthetic Datasets

Generating synthetic datasets solely from a pretrained LDM conditioned on text class prompts $c \in \mathcal{C}$, where $\mathcal{C}$ denotes the set of target classes, often results in significant distribution shifts relative to the downstream target task [8, 15]. Recent work mitigates this issue by few-shot guidance, assuming access to a small labeled dataset $\mathcal{D}_{\text{fs}} = \{(x_i, y_i)\}_{i=1}^{K|\mathcal{C}|}$, containing K examples per class [6]. In this setting, LoRA [17] has been employed to efficiently adapt a pretrained LDM θ to the few-shot dataset $\mathcal{D}_{\text{fs}}$. Specifically, given a pretrained weight matrix $W_0 \in \mathbb{R}^{d_1 \times d_2}$, LoRA introduces two trainable low-rank matrices $B \in \mathbb{R}^{d_1 \times r}$ and $A \in \mathbb{R}^{r \times d_2}$, with rank $r \ll \min(d_1, d_2)$. Keeping the LDM parameters θ fixed, A and B are jointly optimized on a subset $\mathcal{D}' \subseteq \mathcal{D}_{\text{fs}}$ as follows:

$$\min_{A,B} \mathbb{E}_{(x,y)\sim\mathcal{D}',\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_{\theta,A,B}(z_t, t, \tau(y))\|_2^2\right]. \quad (2)$$

Figure 3. *Robust generation using semantic boosting (SB).* Without SB, a LoRA trained on a one-shot image often fails to render a car even when prompted with "a photo of a car". With SB, repeated exposure to the car region during training robustly generates complete cars.

Depending on the choice of the subset $\mathcal{D}'$, the resulting LoRA adapters capture visual variability at different granularities. An image-wise LoRA is trained on a single-image subset $\mathcal{D}' = \{(x, y)\}$, capturing instance-specific features and often yielding high fidelity, yet offering limited coverage of the class distribution. In contrast, a class-wise LoRA is optimized per class c using $\mathcal{D}' = \{(x, y) \in \mathcal{D}_{\text{fs}} \mid y = c\}$, encoding class priors and promoting broader diversity, but it overlooks instance-level details. Figure 2 shows that single granularity methods degrade quality: LoFT [21] (image-wise) yields low diversity, whereas DataDream [20] (class-wise) shows low fidelity.

4. ChimeraLoRA

To generate diverse and fine-grained synthetic images, we propose ChimeraLoRA. We first describe a multi-head LoRA design trained with semantic boosting (Section 4.1). We then introduce a merging strategy that composes multiple heads for image generation (Section 4.2).

4.1. Multi-head LoRA Training

Inspired by asymmetric LoRA architectures such as HydraLoRA [45], we separate the roles of LoRA into two parts: (i) a shared LoRA A that aggregates class-level knowledge and (ii) a set of image-wise LoRA heads $\mathcal{B} = \{B_i\}_{i=1}^{K}$ that capture per-image details, as illustrated in Figure 1. For simplicity, consider K images $\{x_1, ..., x_K\}$ of a single label y. We define a per-image reconstruction loss on image x_i as follows:

$$\mathcal{L}(A, B_i) := \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_{\theta, A, B_i}(z_{i,t}, t, \tau(y))\|_2^2\right], \quad (3)$$

where $z_{i,t}$ is the noisy latent from an augmented view $f_{\text{aug}}(x_i)$ at time step t and $\tau(y)$ denotes the class embedding from the pretrained text encoder τ. To capture the class prior, the shared LoRA A is optimized over total K images by aggregating per-image objectives as follows:

$$\mathcal{L}(A, \mathcal{B}) := \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}(A, B_i) \; . \quad (4)$$

Following previous work [17], we initialize the LoRA A with random Gaussian weights and set each image-wise adapter B_i to zero. We then jointly optimize A and all $\{B_i\}_{i=1}^{K}$ by minimizing (4). For stable training of the shared LoRA A, we use distinct learning rates, setting LoRA A's rate lower than the LoRA B's ones [13].

Additionally, we focus on the noisy latent $z_{i,t}$ in (3). For robust training, practitioners typically employ data augmentation, and $z_{i,t}$ is therefore obtained from an augmented view $f_{\text{aug}}(x_i)$. However, common augmentations may not fully preserve the target class, which can be misaligned with the text prompt and can even hinder generating the target class, as illustrated in Figure 3.

Semantic Boosting with Grounded-SAM. To emphasize class-level semantics shared across few-shot images, we propose a semantic boosting technique based on Grounded-SAM [22, 26]. Specifically, given an image x with label y, we run a text-conditioned object detector using the text prompt for y to produce candidate boxes and define $b^\star$ as the minimal enclosing box of the retained high-confidence targets. We then sample a crop region $\mathcal{R} \subset \mathbb{R}^2$ on x with mild scaling and translation jitter, while enforcing $b^\star \subseteq \mathcal{R}$. To prevent the condition from being violated, we apply zero-padding to the original image, so that $b^\star$ remains fully visible and the crop meets the target size. This semantic cropping enables robust generation of the target class, as shown in Figure 3. Furthermore, since our semantic cropping repeatedly exposes the target class region during training, the model better preserves the target's aspect ratio and fine-grained details under the same training setup, as demonstrated in Figure 4.

4.2. LoRA Merging for Image Generation

While each image-wise LoRA B_i captures instance-specific details, generating synthetic images with a single adapter often fails to cover the full within-class distribution. Instead, we synthesize each image by combining the K image-wise adapters with nonnegative weights sampled from a Dirichlet distribution as follows:

$$B' = \sum_{i=1}^{K} w_i B_i, \quad (w_1, \ldots, w_K) \sim \text{Dirichlet}(\boldsymbol{\alpha}) \; , \quad (5)$$

where $\boldsymbol{\alpha} = \alpha \, \mathbf{1}_K = (\alpha, \ldots, \alpha)$. After forming B', we generate synthetic images by applying the class-adapter LoRA A together with B' to the base diffusion model.

Figure 4. *Effect of semantic boosting.* (a) The input image is used to train a LoRA under varying cropping methods. (b) Without cropping, the generated images exhibit a distorted aspect ratio of the primary object. (c, d) Conventional random and center cropping methods result in outputs where the object is consistently truncated. (e) In contrast, our semantic boosting preserves the object's structural integrity and details, leading to a robust generation.

In the symmetric Dirichlet case with $\boldsymbol{\alpha} \in \mathbb{R}^K$, the expectation and variance are computed as:

$$\mathbb{E}[w_i] = \frac{1}{K}, \quad \mathrm{Var}[w_i] = \frac{K-1}{K^2\,(K\alpha+1)}\,. \tag{6}$$

Here, the concentration α controls how the mixture spreads over the simplex Δ^{K-1}. When $\alpha = 1$, $\mathbf{w}$ is uniformly distributed over the simplex. For $\alpha < 1$, the distribution becomes sparse and typically concentrates most of its mass on a single B_i, which behaves like an image-wise regime. When $\alpha > 1$, the weights cluster near the uniform vector and approximate a class-wise regime. We refer to this framework as ChimeraLoRA, which establishes a class-level backbone from the few-shot references and attaches instance-specific details to produce coherent semantics, yet diverse images.[1]

Remark. When $\alpha = 1$, $\mathbf{w}$ is uniform on the simplex Δ^{K-1}, and we observe that this setting typically yields decent downstream performance. However, sampling a fresh $\mathbf{w}$ for every image can hinder batch-wise generation, as B' must be rebuilt per sample. We consider two practical variants to mitigate this overhead: (i) set $w_i = 1/K$ in (5), which still maintains high fidelity with reasonable coverage, as shown in Figure 2; and (ii) reuse a single $\mathbf{w} \sim \mathrm{Dirichlet}(\mathbf{1})$ to synthesize multiple images, with diffusion stochasticity providing additional variation. However, we note that per-image mixtures with $\mathrm{Dirichlet}(\mathbf{1})$ still provide the broadest coverage. In Appendix B.2, we analyze the trade-off between wall-time and accuracy across the three methods.

5. Experiment

5.1. Experimental Setup

Datasets. We evaluate on 11 publicly available image classification datasets: FGVCAircraft (AIR) [30], Caltech101 (CAL) [9], StanfordCars (CAR) [23], DTD [4], EuroSAT (EUR) [16], Flowers102 (FLO) [32], Food101 (FOD) [1], OxfordPets (PET) [33], Skin Lesions (ISIC) [19], CIFAR-10 [24], and ImageNet100 [40]. Our benchmarks cover diverse fine-grained tasks including cars and pets, and also specialized domains such as satellite imagery, textures, and medical dermatology, reflecting practical few-shot constraints in real applications.

Implementation Details. We adopt CLIP ViT-B/16 [7, 36] as the downstream encoder. During fine-tuning, we attach rank-16 LoRA adapters to both the image and text encoders and train the model on synthetic training datasets derived from the given 4-shot references. Unless otherwise noted, all methods are trained for 60 epochs with AdamW [28] at a learning rate of 1×10^{-4} using a cosine annealing scheduler. All experiments are run with three random seeds per setting, and we report the mean and variance.

Baselines. We compare our method against three methods: IsSynth [14], LoFT [21], and DataDream [20]. IsSynth is train-free and synthesizes data using features extracted from the given 4-shot references. LoFT and DataDream fine-tune diffusion models with LoRA in image-wise and class-wise configurations, respectively. To match trainable-parameter budgets in the 4-shot setting, DataDream uses LoRA rank 16, since it trains a single class-wise LoRA over all four images, whereas LoFT and our method use rank 4, as the adaption is split across the four images. We note that thanks to the shared adapter A, our approach uses 37.5% fewer trainable parameters than both baselines. We employ a multi-head LoRA with distinct learning rates: 1×10^{-4} for LoRA A and 1×10^{-3} for LoRA B. All LoRA-based methods use Stable Diffusion 2.1 [38] as the base diffusion model. For generation, the trained adapters are attached to Stable Diffusion with guidance scale 2, and when composing the per-image adapters we draw mixture coefficients from $\mathrm{Dirichlet}(\mathbf{1})$. Unless otherwise noted, we reproduce the results for all baselines.

[1] In the Appendix A, we expand our discussion to the setting with two concentration parameters, α and β, rather than a single α.

Table 1. *Downstream performance with synthetic datasets under 4-shot scenarios.* Starting with 4-shot labels, we generate 500 additional images per class to train on 504 per class in total, improving accuracy by 2.1 percentage points on average over state-of-the-art baselines across nine datasets. We mark the best in bold and the second best with underlines.

Methods	AIR	CAL	CAR	DTD	EUR	FLO	FOD	PET	ISIC	AVG
CLIP (0-shot)	24.7	93.0	65.2	44.4	47.6	71.4	86.1	89.2	21.1	60.3
CLIP (4-shots)	$41.3_{\pm 0.3}$	$95.5_{\pm 0.1}$	$74.3_{\pm 0.2}$	$62.0_{\pm 0.7}$	$83.5_{\pm 0.6}$	$89.9_{\pm 0.4}$	$86.5_{\pm 0.1}$	$93.3_{\pm 0.4}$	$19.6_{\pm 1.5}$	$71.8_{\pm 0.4}$
IsSynth [14]	$39.9_{\pm 0.1}$	$95.5_{\pm 0.1}$	$71.5_{\pm 0.6}$	$\underline{60.1}_{\pm 0.3}$	$73.4_{\pm 0.6}$	$89.0_{\pm 1.0}$	$85.6_{\pm 0.1}$	$91.6_{\pm 0.2}$	$23.8_{\pm 1.4}$	$70.1_{\pm 0.4}$
DataDream [20]	$\underline{44.3}_{\pm 0.4}$	$\mathbf{96.1}_{\pm 0.1}$	$\mathbf{81.7}_{\pm 0.3}$	$56.0_{\pm 0.6}$	$72.2_{\pm 0.7}$	$\underline{92.9}_{\pm 0.5}$	$\mathbf{86.0}_{\pm 0.1}$	$92.2_{\pm 0.1}$	$20.7_{\pm 1.1}$	$71.3_{\pm 0.3}$
LoFT [21]	$41.7_{\pm 0.6}$	$95.7_{\pm 0.1}$	$78.0_{\pm 0.3}$	$58.0_{\pm 1.5}$	$\underline{85.0}_{\pm 0.7}$	$91.3_{\pm 0.2}$	$85.1_{\pm 0.1}$	$\underline{92.4}_{\pm 0.3}$	$\underline{25.6}_{\pm 0.3}$	$\underline{72.5}_{\pm 0.4}$
ChimeraLoRA	$\mathbf{46.0}_{\pm 0.7}$	$\mathbf{96.1}_{\pm 0.1}$	$\underline{79.6}_{\pm 0.5}$	$\mathbf{61.6}_{\pm 0.5}$	$\mathbf{86.3}_{\pm 0.5}$	$\mathbf{93.4}_{\pm 0.4}$	$\underline{85.7}_{\pm 0.1}$	$\mathbf{93.4}_{\pm 0.1}$	$\mathbf{29.2}_{\pm 0.6}$	$\mathbf{74.6}_{\pm 0.2}$

Table 2. *Downstream performance with synthetic datasets under long-tail scenarios.* When training with only 4 samples per tail class, the classifier's decision boundary skews toward the long classes, yielding poor tail performance. After adding 500 synthetic images per tail class with ChimeraLoRA, accuracy improves by 7.62 percentage points on average relative to the real-only baseline, with a 14.74 percentage-point gain on the tail classes specifically.

Methods	CIFAR10			ImageNet100			DTD			EuroSAT			Flowers102		
	Head	Tail	Avg.	Head	Tail	Avg.	Head	Tail	Avg.	Head	Tail	Avg.	Head	Tail	Avg.
Real	**98.8**	70.1	84.5	78.8	79.1	79.0	**86.5**	46.7	66.6	**99.2**	13.6	56.4	92.8	94.2	93.5
DataDream [20]	98.1	76.3	87.2	88.7	<u>91.3</u>	<u>90.0</u>	78.8	<u>55.5</u>	67.2	<u>98.8</u>	<u>48.7</u>	<u>73.9</u>	<u>93.3</u>	95.5	<u>94.4</u>
LoFT [21]	<u>98.4</u>	<u>76.5</u>	<u>87.4</u>	**88.9**	91.2	<u>90.0</u>	<u>84.3</u>	51.6	<u>67.9</u>	98.7	47.2	73.0	91.5	<u>96.0</u>	93.7
ChimeraLoRA	98.3	**81.0**	**89.6**	<u>88.8</u>	**91.6**	**90.2**	80.2	**56.6**	**68.4**	98.0	**51.3**	**74.5**	**93.9**	**96.9**	**95.4**

5.2. Synthetic Datasets for Downstream Tasks

Few-shot scenarios. We investigate whether synthetic datasets can surpass 4-shot real datasets. Table 1 shows that fine-tuning CLIP with 4-shots per class attains an average accuracy of 71.8% across nine datasets. We then generate 500 synthetic images per class and fine-tune CLIP on 504 images per class. In this setting, our ChimeraLoRA surpasses prior state-of-the-art methods. Notably, many baselines remain below the 4-shot real model even after adding synthetic data, underscoring a synthetic-to-real gap that limits its practical utility. Although our method also fails to surpass the 4-shot real model on DTD and FOD, its drop is smaller than competing approaches, with DataDream decreasing by 11.3 percentage points (pp) on EUR. Overall, thanks to our synthetic images, ChimeraLoRA can build synthetic datasets that outperform real 4-shot datasets.

Long-tail scenarios. In practice, class frequencies are long-tailed rather than uniform with 4-shots per class. We therefore study augmenting only the tail classes with synthetic images. To simulate an extreme long-tailed regime, we split each dataset so that half of the classes are head classes with up to 500 real images and the other half are tail classes with 4-shots each. In this regime, training only on real images produces a model biased toward the long classes, as shown for EuroSAT in Table 2 where the accuracy gap between long and tail classes is 85.6 pp. Table 2 shows that adding synthetic images to tail classes leads to average accuracy gains, and our ChimeraLoRA outperforms baselines across five datasets. Especially, we observe that on ImageNet-100, adding synthetic images to the tail classes not only improves tail accuracy but also increases the accuracy of the long classes.

5.3. Synthetic-to-Real Gap Analyses

Within-class visualization. In Section 4.2, we merge multiple LoRA heads with weights sampled from Dirichlet($\mathbf{1}$) and use the merged adapter for image synthesis. Figure 6 shows that such Dirichlet-weighted mixtures, which lie inside the probability simplex, transfer to actual generations. Specifically, we visualize the *banded* class from DTD by taking four real images and, for each method, 500 synthetic images, for a total of 1,504 images. Then, we compute CLIP image embeddings, reduce dimensionality with PCA [42], and apply t-SNE [29]. Here, red stars mark the real anchors. As intended, ChimeraLoRA spreads its samples roughly uniformly within the real region, whereas the baselines place many samples outside it. In addition, LoFT collapses into a few tight clusters, producing near-duplicates, while DataDream drifts farther from the anchors and often yields lower-fidelity samples with wavy banding and inconsistent colors. For clarity, we recommend zooming in on Figure 6.

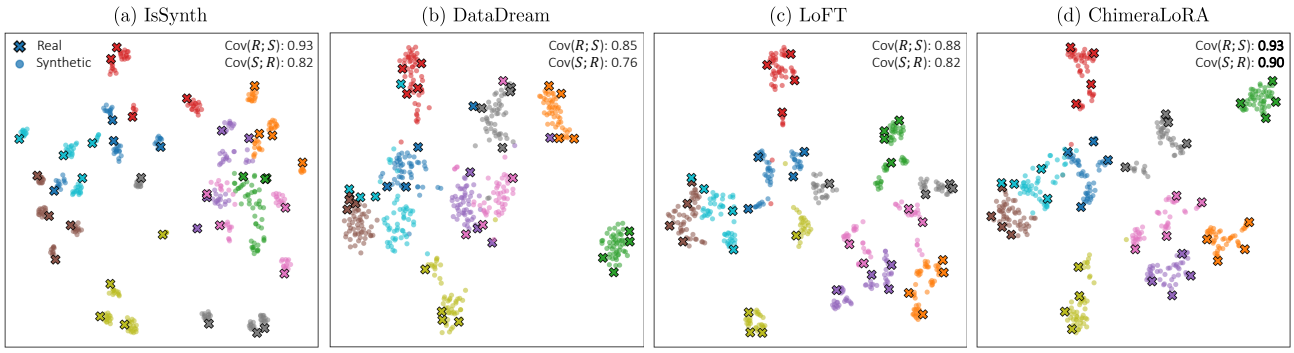


Figure 5. *t-SNE for real and synthetic images.* ChimeraLoRA generates mainly inside the region spanned by the real anchors marked with crosses and attains the highest coverage across methods, with $\mathrm{Cov}(\mathcal{R};\mathcal{S}) = 0.93$ and $\mathrm{Cov}(\mathcal{S};\mathcal{R}) = 0.90$.

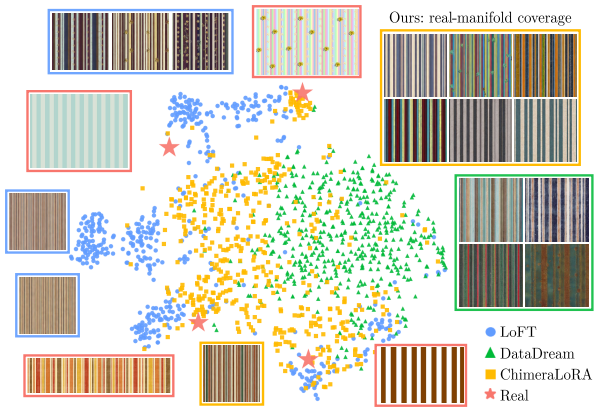


Figure 6. *Real-manifold coverage.* Our ChimeraLoRA samples (yellow rectangles) fall within the region spanned by the four real anchors (red stars), indicating coverage of the real manifold, whereas the baselines drift outside.

Cross-class visualization. Beyond the single-class analysis, Figure 5 presents a cross-class visualization of ten DTD classes. We use 4 real images and 50 synthetic images per class, for 540 images in total. Similarly to the single-class case, we compute CLIP image embeddings and visualize them with t-SNE. To evaluate how well the real and synthetic sets cover one another, we report two directional coverages. Let the real set be $\mathcal{R}$ and the synthetic set be $\mathcal{S}$. Let $\phi(\cdot)$ be the L2-normalized CLIP image embedding and define the cosine distance $\delta(\mathbf{u},\mathbf{v}) = 1 - \langle \mathbf{u},\mathbf{v}\rangle$. Define a class radius ρ from the median real-real nearest-neighbor distance in CLIP space. With $\mathcal{R}$ as the anchor, the coverage of $\mathcal{R}$ by $\mathcal{S}$ is defined as follows:

$$\mathrm{Cov}(\mathcal{R};\mathcal{S}) = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}} \mathbf{1}\left[\exists s \in \mathcal{S} \text{ s.t. } \delta\big(\phi(r),\phi(s)\big) \le \rho\right]. \quad (7)$$

The symmetric measure $\mathrm{Cov}(\mathcal{S};\mathcal{R})$ is defined by swapping $\mathcal{R}$ and $\mathcal{S}$. Figure 5 demonstrates that ChimeraLoRA attains higher scores on both measures than other methods, indicating that real and synthetic images intermix more naturally.

Table 3. *Synthetic-to-real gap analyses.* Across nine datasets, ChimeraLoRA produces synthetic images that most closely match the 4-shot real reference on average, with the lowest FID@4 and the highest CLIP score and centroid similarity.

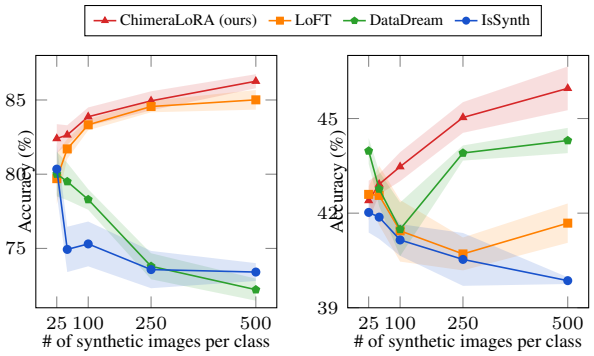
Methods	FID@4 ↓	CLIP score ↑	Centroid Sim. ↑
Real (4-shots)	0.00	29.48	100.0
DataDream [20]	0.23	29.67	87.8
LoFT [21]	0.22	30.04	90.1
ChimeraLoRA	**0.20**	**30.31**	**90.5**

Quantifying the synthetic-to-real gap. We evaluate the synthetic-to-real gap relative to the 4-shots real reference using three metrics computed per class and averaged across classes. First, Fréchet Inception Distance (FID) is computed in CLIP image-embedding space rather than Inception [43], following recent work [18]. As a closer distributional match yields a smaller FID, the Real 4-shot row has FID = 0 in Table 3. Second, CLIP score is the cosine similarity between each image's CLIP embedding and the class text embedding, reported after multiplying by 100. Third, centroid similarity is the cosine similarity in CLIP space between the centroid of 500 synthetic images and the centroid of the 4 real images for each class, normalized so that the Real 4-shots row equals 100.0. Table 3 shows that ChimeraLoRA attains lower FID and higher CLIP score and centroid similarity than the baselines, indicating that our method exhibits the smallest synthetic-to-real gap.

5.4. Further Ablation Studies

We conduct a range of ablation studies on our method, and additional ablation results are provided in the Appendix C.

Ablation on the number of synthetic images. As generating 500 synthetic images per class is costly, we conduct experiments on smaller number of synthetic images. Figure 7 shows that ChimeraLoRA maintains robust performance, and its accuracy increases with the number of



(a) EuroSAT (b) FGVCAircraft

Figure 7. *Robustness under scaling the synthetic budget.* Accuracy rises with more synthetic images per class for ChimeraLoRA. Shaded regions indicate variability across seeds.

Figure 8. *Synthetic images with sharing LoRA B.* For each class, the left three images are from our ChimeraLoRA with a shared LoRA A, and the right three are from a variant with a shared LoRA B. The right images look more diverse but often miss the target object or fine details such as a motorcycle's wheel.

synthetic images as the added synthetic images align with the real distribution induced by the few-shot references. However, on the EUR and AIR datasets, DataDream and LoFT exhibit declining accuracy as more synthetic images are added, revealing a persistent synthetic-to-real gap that worsens with scale.

Ablation on shared LoRA parts. In Section 4.1 we introduce a shared LoRA A to encode class priors instead of sharing LoRA B. Figure 8 compares the results of shared LoRA A on the left and shared LoRA B on the right for each class. With shared B, images often look more diverse yet miss the target object or fine details, such as truncated headphone ends, awkward staplers, and motorbikes lacking inner wheel detail. Conceptually, LoRA A plays an encoder-like role that projects features into a shared rank-r subspace, while LoRA B acts as a decoder that lifts this representation back to the full model space. When the few shot references share the same semantics, sharing the encoder A promotes a class consistent encoding, and instance specific decoders B can then reconstruct high frequency structure. This division yields better object integrity and sharper details than sharing B. Our observation aligns with asymmetric roles reported in previous work [11, 49]: LoRA B is more tightly coupled to the input data distribution, whereas LoRA A aggregates shared knowledge.

Table 4. *Component ablation.* Each proposed components contribute to performance gains.

Multi-Head LoRA	Semantic Boosting	AIR	FLO
✗	✗	41.7	91.3
✓	✗	43.9	93.1
✗	✓	44.4	92.2
✓	✓	**46.0**	**93.4**

Ablation on proposed components. Table 4 isolates the effects of multi-head LoRA and semantic boosting. Removing both components reduces our method to LoFT [21], which uses only per-image LoRAs without a class-shared adapter or box guidance and yields the lowest accuracy. Introducing multi-head LoRA brings a class-shared adapter that coordinates the per-image adapters and delivers clear gains. Incorporating semantic boosting preserves class bounding boxes during fine-tuning and further improves accuracy. Neither component is uniformly superior across datasets. Combined in ChimeraLoRA, their benefits compound and the model attains the best overall results.

6. Conclusion

In this work, we propose a multi-head LoRA guided method for synthetic dataset generation, called ChimeraLoRA. By separating the roles of two low-rank adapters, we use a class-shared adapter to encode class priors and per-image adapters to model instance-level details. To help the shared adapter capture class semantics, we introduce semantic boosting that leverages class bounding boxes during adapter fine-tuning. For image synthesis, we fix the class-shared adapter and merge the per-image adapters to generate images with high diversity and fidelity, which in turn improve downstream task performance. Extensive experiments across diverse classification tasks and practical domains, including medical applications and long-tailed scenarios, show that our method outperforms baselines.

Limitations and future work. We use Grounded-SAM as a general solution across domains for Semantic Boosting, but in medical settings, domain-specific tools such as MedSAM may be more appropriate, and our current medical domain validation remains limited. Future work will strengthen domain-specific evidence by evaluating on additional medical datasets, conducting robustness analyses under clinically relevant perturbations, and expanding the discussion of related work on generative augmentation for long-tailed medical image classification rather than assuming that simple tool substitution is sufficient.

Acknowledgements. This work was partly supported by the IITP grants and the NRF grants funded by Ministry of Science and ICT, Korea (No.RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH); No.RS-2024-00457882, AI Research Hub Project; IITP-2026-RS-2024-00437866; RS-2024-00509258, Global AI Frontier Lab).

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 5

[2] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 1

[3] Arkabandhu Chowdhury, Mingchao Jiang, Swarat Chaudhuri, and Chris Jermaine. Few-shot image classification: Just use a library of pre-trained feature extractors and a simple classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9445–9454, 2021. 1

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5

[5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 1

[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. *arXiv preprint arXiv:2312.03046*, 2023. 1, 2, 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 5

[8] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7382–7392, 2024. 3

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 5

[10] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. Keepaugment: A simple information-preserving data augmentation approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1055–1064, 2021. 2

[11] Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 8

[12] Zhengrui Guo, Conghao Xiong, Jiabo Ma, Qichen Sun, Lishuang Feng, Jinzhuo Wang, and Hao Chen. Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15590–15600, 2025. 1

[13] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models. In *International Conference on Machine Learning*, pages 17783–17806. PMLR, 2024. 4

[14] 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 *The Eleventh International Conference on Learning Representations*, 2023. 2, 5, 6

[15] 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 *The Eleventh International Conference on Learning Representations*, 2023. 1, 3

[16] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 5

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1, 2, 3, 4

[18] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024. 7

[19] Mohamed A Kassem, Khalid M Hosny, and Mohamed M Fouad. Skin lesions classification into eight classes for isic 2019 using deep convolutional neural network and transfer learning. *IEEE access*, 8:114822–114832, 2020. 5

[20] Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. Datadream: Few-shot guided dataset generation. In *European Conference on Computer Vision*, pages 252–268. Springer, 2024. 1, 2, 4, 5, 6, 7

[21] Jae Myung Kim, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. Loft: Lora-fused training dataset generation with few-shot guidance. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA, 2025. 1, 2, 4, 5, 6, 7, 8

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer White-

head, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1, 2, 4

[23] 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. 5

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[25] SuBeen Lee, WonJun Moon, and Jae-Pil Heo. Task discrepancy maximization for fine-grained few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5331–5340, 2022. 1

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 1, 2, 4

[27] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019. 1

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[29] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 6

[30] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[31] Shlok Kumar Mishra, Anshul Shah, Ankan Bansal, Janit K Anjaria, Abhyuday Narayan Jagannatha, Abhishek Sharma, David Jacobs, and Dilip Krishnan. Object-aware cropping for self-supervised learning. *Transactions on Machine Learning Research*, 2022. 2

[32] 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. 5

[33] 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. 5

[34] Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, and Yang You. Crafting better contrastive views for siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16031–16040, 2022. 2

[35] Linhao Qu, Dingkang Yang, Dan Huang, Qinhao Guo, Rongkui Luo, Shaoting Zhang, and Xiaosong Wang. Pathology-knowledge enhanced multi-instance prompt learning for few-shot whole slide image classification. In *Computer Vision – ECCV 2024*, pages 196–212, Cham, 2025. Springer Nature Switzerland. 1

[36] 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. 5

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 5

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[40] 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*, 115:211–252, 2015. 5

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[42] Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014. 6

[43] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 7

[44] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020. 1

[45] Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Cheng-Zhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *Advances in Neural Information Processing Systems*, 37:9565–9584, 2024. 1, 2, 4

[46] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European conference on computer vision*, pages 266–282. Springer, 2020. 1

[47] Xuyang Wei, Chunlin Tian, and Li Li. AsymloRA: Unlocking the power of multimodal LLMs via asymmetric loRA. In *First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models*, 2025. 2

[48] Yizhe Zhang, Tao Zhou, Shuo Wang, Peixian Liang, Yejia Zhang, and Danny Z Chen. Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 129–139. Springer, 2023. 2

[49] Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez De Ocáriz Borde, Rickard Brüel Gabrielsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62369–62385. PMLR, 2024. 2, 8