Beyond Objects: Contextual Synthetic Data Generation for Fine-Grained Classification

Anonymous CVPR submission

Paper ID

Abstract

Text-to-image (T2I) models are increasingly used for synthetic dataset generation, but generating effective synthetic training data for classification remains challenging. Fine-tuning a T2I model with a few real examples can help improve the quality of synthetic training data; however, it may also cause overfitting and reduce diversity in the generated samples. We propose a fine-tuning strategy BOB *(Beyond OBjects) to mitigate these concerns for fine-grained classification. Given a small set of real examples, we first extract class-agnostic attributes such as scene background and object pose. We then explicitly condition on these attributes during fine-tuning of the T2I model and marginalize them out during generation. This design mitigates overfitting, preserves the T2I model's generative prior, reduces estimation errors, and further minimizes unintended inter-class associations. Extensive experiments across multiple T2I models, backbones, and datasets show that our method achieves state-of-the-art performance in low-shot fine-grained classification when augmented with synthetic data. Concretely,* BOB *outperforms DataDream by 7.4% on the Aircraft dataset (from 50.0% to 57.4% when fine-tuning a CLIP classifier with 5 real images augmented with 100 synthetic images). In three of the four benchmarks, fine-tuning downstream models with 5 real images augmented with* BOB *achieves better performance than fine-tuning with 10 real images. Collectively,* BOB *outperforms prior art in 18 of 24 experimental settings, with over 2% accuracy improvements in 14 of these settings.*

1. Introduction

Powerful text-to-image (T2I) models trained on internet-scale datasets [35, 41] have shown promise in the creation of synthetic data for representation learning [42, 43], 3D synthesis [33], and image editing [3, 16, 31]. However, there remains a considerable performance gap when using synthetic data as training data for downstream tasks such as classification [4, 9, 11], particularly for fine-grained classification where subtle visual differences are critical.

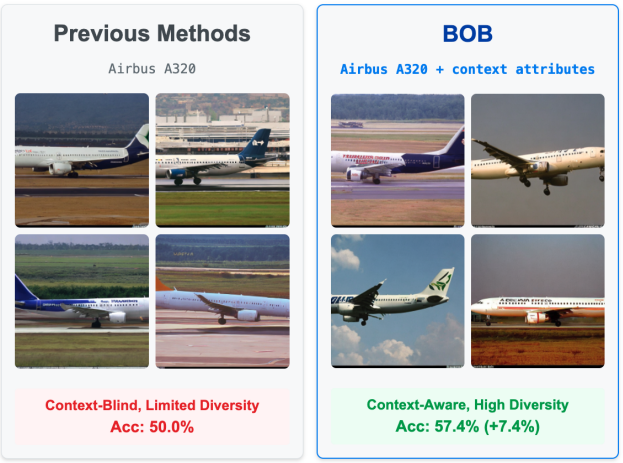

Figure 1. **BOB vs. prior methods.** (*Left*) Prior methods do not include rich contexts during fine-tuning and sampling, leading to reduced diversity in the synthetic dataset and lower classification performance when used for downstream training. (*Right*) BOB mitigates this issue by incorporating rich contexts in both stages.

Ideally, given a target classification task described in language (e.g., "train an aircraft classifier to distinguish a 747-300 from a 747-400"), a T2I model can be directly used to generate training images of the desired classes. A key challenge limiting T2I models from generating informative images is the *model estimation errors* caused by a misalignment between the T2I model's learned distribution and the target task [11]. For example, the visual difference between 747-300 and 747-400 is subtle: the presence of a winglet on a slightly longer wing. In consequence, the presence of low level artifacts and incorrect visual compositions introduces a challenge for the task of fine-grained recognition.

One approach to mitigate model estimation errors for these classes is to provide a few real images to fine-tune the T2I model [21, 47]. The underlying T2I model needs to rely heavily on additional guidance from provided examples to generate not only accurate but also diverse samples

to augment the training of downstream classifiers. However, operating in the few-shot regime requires special considerations [50]. The increased expressivity from fine-tuning can introduce a trade-off where the T2I model starts to overfit to the few examples, losing its strong world prior and hurting the diversity of the synthetic dataset.

In this work, we tackle fine-grained classification with synthetic data generation by introducing BOB (Beyond OBjects). BOB mitigates overfitting during T2I fine-tuning by extracting class-agnostic contextual attributes (background and pose) for each sample via a captioning model and explicitly incorporating them into the conditioning signal. During data generation, background and pose pairs are sampled across the dataset, effectively marginalizing out any unintended inter-class associations. We provide a comprehensive evaluation across three backbones, two T2I models, four datasets, two data scales, and seven existing methods, to demonstrate the effectiveness of BOB.

We observe the most considerable gain on Aircraft, a dataset where T2I models perform poorly and fine-tuning benefits the most. Training the CLIP model [34] using 5 real images augmented with 100 synthetic images per class, BOB achieves 57.4% accuracy compared to 50.0% with DataDream [21], a 7.4% improvement. Importantly, across three of the four datasets (Aircraft, Cars, and CUB), using 5 real images augmented with BOB generated images results in *better* classification performance than using 10 real images: e.g., CLIP fine-tuned on CUB achieves accuracy of 75.8% with 5 real images augmented with BOB generated images and only 74.6% with 10 real images without augmentation. Overall, BOB outperforms existing state-of-the-art methods by at least 2% on 18 of the 24 experimental settings (backbone, dataset source, and dataset size). On the six remaining settings (on the Pets [32] dataset), BOB offers comparable performance with state-of-the-art. To summarize, we make the following contributions:

1. We introduce stronger supervision with more detailed captioning during T2I fine-tuning to mitigate model overfitting and enhance prior preservation (Sec. 3.1).
2. We marginalize out unintended inter-class associations by randomly sampling class-agnostic features (background, pose) across the whole dataset (Sec. 3.2).
3. We provide a comprehensive evaluation (Sec. 4) across seven previous methods, two T2I models, and 24 different experimental settings and demonstrate that our methods outperforms previous methods in 18 of the 24 settings with competitive performance in the rest.

2. Related Work

Personalization. Many personalization methods serve as inspiration for approaches aimed at synthetic data generation for classification. Personalization methods seek to guide and control T2I models beyond text descriptions, typically using a few image exemplars. These approaches have proven effective at tailoring T2I models to reproduce highly specific visual concepts [10, 24, 27, 37, 48, 53, 54, 56]. However, while they enable strong concept-level control, their objectives differ from those required for classification-oriented synthetic data. They emphasize concept fidelity over diversity, often limiting intra-class variation and inter-class separability, which are both essential for robust classifier training. Therefore, such personalization methods are insufficient for addressing the challenges of synthetic data generation for classification.

Synthetic data for classification. The field of synthetic data generation for classification has initiated a significant shift, moving toward leveraging powerful T2I models. This paradigm shift contrasts sharply with traditional data augmentation methods, such as CutMix [51] and Mixup [52]. These techniques interpolate between existing data, which helps smooth the decision function but is limited in sample diversity and fidelity. In contrast, a pre-trained T2I model provides a "world prior" that significantly enhances both sample diversity and fidelity.

Early works demonstrated the utility of T2I models for classification. Real Guidance [14] demonstrated that utilizing these T2I models with simple class descriptions and a few reference images can improve classification performance. Da-fusion [44] incorporate Textual Inversion [10] on the few reference images to generalize to unknown concepts. Subsequent research has largely focused on two parallel areas: improving fine-tuning and enhancing generation. For fine-tuning, methods like Diff-Aug [47], Diff-Gen [47], Diff-Mix [47] and DataDream [21] focused on adapting different components of the T2I model (U-Net and text-encoder) to the target data. On the generation side, previous works primarily focus on better prompt design [40, 49], prompt augmentation with image captions [6, 8], diffusion latents interpolation [46, 55], leveraging vision-language models [30], and hard examples generation [1, 15, 22]. Additionally, Liu et al. [28] demonstrate that unrealistic synthetic attributes often have a negligible impact on fine-tuned classifier performance. Unlike previous work that focuses primarily on fine-tuning or generation, our approach maintains T2I controllability during fine-tuning and shifts the generation objective from simple variety to marginalization to effectively mitigate spurious correlations.

Diffusion classifier. Using diffusion models directly for image classification has shown promising performance [5, 20, 25, 26]. This naturally raises the question of treating T2I models as classifiers which would remove the need to train a separate classifier. However, inference with diffusion classifiers is very computationally expensive, increasing each ImageNet classification from seconds to over 10 minutes [25, 38], motivating research on distilling these capabilities into efficient downstream classifiers.

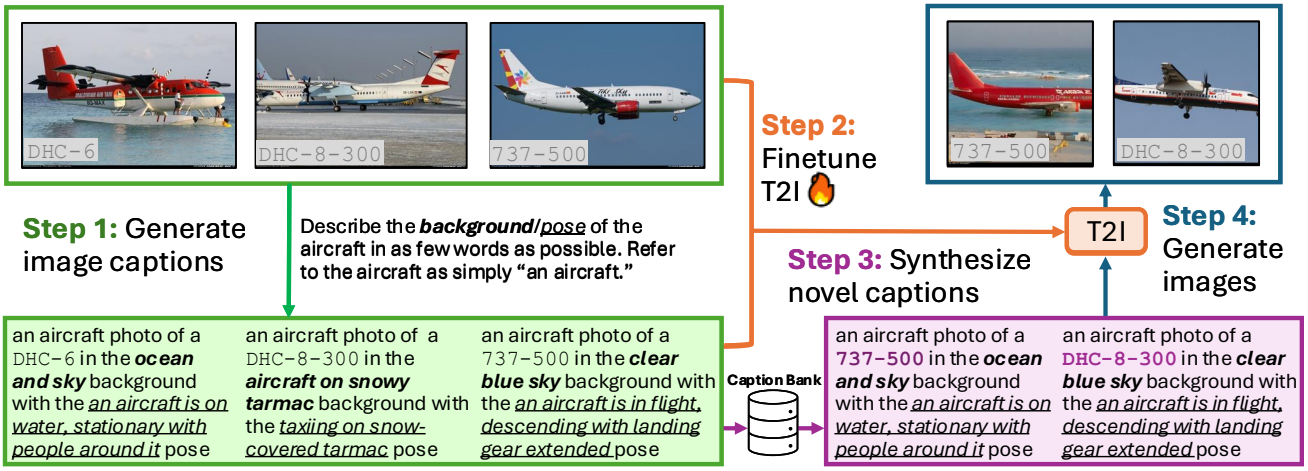

Figure 2. **Overview of BOB.** We extract background and pose attributes from training images using a captioning model (*Step 1*), apply **context preservation** by fine-tuning the T2I model with enriched captions containing class names and context attributes (*Step 2*), and then perform **context marginalization** by generating synthetic data through randomly sampling background-pose pairs across the entire dataset (*Step 3-4*). This preserves class-relevant features while reducing spurious class-context associations.

3. Methodology

We propose an effective approach for fine-grained synthetic data generation that addresses the critical problem of overfitting during T2I model fine-tuning and image generation (Fig. 2). This overfitting exists in both modalities: in the texts, the classification task's ambiguous text-to-image mapping inherently loses T2I controllability by failing to articulate the intra-class visual range; and in the images, the few-shot regime fosters unintended inter-class associations. Our method mitigates both overfitting issues in two stages: *Context Preservation* during fine-tuning and *Context Marginalization* during data generation. We describe both in detail in this section.

3.1. Context Preservation

Our first mitigation strategy *Context Preservation* aims to recover the intra-class visual range lost in the text modality. While T2I models require detailed text-to-image mappings, classification datasets only provide generic labels. Previous approaches use class-specific templates (e.g., "a photo of a `[classname]`") but this simplification reduces diverse visual information to a single description, leading to a loss of T2I controllability. To address this, we propose associating each image with a unique, descriptive text. We achieve this by extracting and explicitly encoding class-agnostic attributes (background and pose) into the text conditioning. This enables the model to learn the specific association between these attributes and the visual context during fine-tuning. Concretely, each image is associated with a unique caption following the template:

a `[descriptor]` photo of a `[classname]` in the `[background]` background with the `[pose]` pose.

The `[descriptor]` is a dataset-level general descriptor such as aircraft or birds. `[classname]` is the name of the class provided by the dataset. The background/pose is extracted by a captioning model for each image. We leverage the Qwen 2.5VL-7B, a state-of-the-art vision-language model [2] to extract the background with the following prompt: **describe the background of the [descriptor] in as few words as possible. Refer to the [descriptor] as simply 'a [descriptor]'**. Similarly, we use the same prompt with 'background' replaced by 'pose' to extract the pose. We also store the extracted background and pose into a caption bank $\mathcal{B} = \{(b_i, p_i)\}_{i=1}^{N}$, where b_i and p_i represent the background and pose attributes of the i-th training image, respectively (Fig. 2, steps 1-2). This prompting approach serves two purposes: (1) it provides necessary context to guide accurate attribute extraction, and (2) it prevents potential leakage of class-specific information by maintaining generic references to the object category.

Once we established the image-text pairing, we fine-tune the diffusion model using the standard diffusion objective. We follow the standard parameter-efficient fine-tuning procedure by using Low-Rank Adaptation (LoRA) [18] to fine-tune the attention layers of both the U-Net [36] and CLIP text encoder [34]. Consider the following notations: θ as parameters of the attention layers, the image as x, the text as y, the CLIP text encoder as $c(y)$, timestep of diffusion process as t, and the U-Net model as $\epsilon_\theta(x, c_\theta(y), t)$. The parameters θ are updated by minimizing the following objective: $\mathbb{E}_{(x,y)\sim D,\, \epsilon\sim\mathcal{N},\, t\sim\mathcal{U}} \|\epsilon - \epsilon_\theta(x, c_\theta(y), t)\|_2^2$.

3.2. Context Marginalization

Our second stage, Context Marginalization, reduces unintended associations between class labels and visual con-

text which are common in data-scarce settings. When only a small number of images are available per class, certain backgrounds or poses may appear only with specific labels due to limited coverage. The central intuition behind Context Marginalization is that the contextual attributes retained during fine-tuning can be reused during generation to prevent the model from reinforcing the dataset idiosyncrasies.

We can formalize this problem more rigorously from a causal perspective, consider the generation process of our training data that our T2I model emulates shown in Fig. 3: image X is generated given the class-relevant attributes Y and the class-agnostic attributes Z. To directly model the relationship between X and Y, we sample from the intervention distribution $P(X|do(Y))$ by invoking the backdoor adjustment: $P(X|do(Y)) = \sum_Z P(X|Y,Z)P(Z)$.

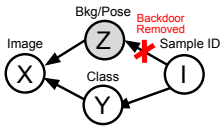

Figure 3. Causal graph

To sample from $P(Z)$, we generate synthetic images using the same prompt template but randomly sample background–pose pairs from our caption bank, independent of the class label (steps 3-4 in Fig. 2). This sampling strategy exposes each class to the variety of backgrounds and poses present in the dataset, helping the model focus on class-defining visual features rather than incidental context.

4. Experiments

In this section, we will first go over the experimental setup for our comprehensive evaluation. Next, we will perform detailed quantitative analysis on the experimental results to demonstrate advantages of our proposed approach with respect to previous methods. Lastly, we present additional analysis showing that our method produces synthetic data that better aligns with the target data distrbution and perform ablations showing the necessity of both the context preservation and content marginalization for generating informative data for the downstream classification task.

4.1. Experimental Setup

Datasets. We follow the standard settings for data-scarce augmentation where we only have 5 or 10 real images per class. We follow the previous evaluation setting [46] where we use Aircraft [29], CUB [45], Car [23], and Pets [32] datasets for evaluation. Additionally, we also use the CUB-LT [39] and Flower-LT [47] dataset to extend BOB to a different data-scarce setting: long-tail classification.

Backbones. We use three different backbones with different degrees of language supervision during pre-training for downstream task fine-tuning. For a backbone with dense language supervision, we use the CLIP VIT-B/16 model [7, 34]. For a backbone with weak language supervision, we use the ImageNet classification-trained ResNet-50 model [12]. Finally, for a backbone with no language supervision, we use the masked auto-encoder (MAE) VIT-B/16 model pre-trained on ImageNet [13]. *Unlike previous works which focus on one of these types, we decided to have all three language supervision settings to provide a wider perspective towards the behavior and usefulness of different methods.*

Baseline methods. We use seven popular existing data generation or augmentation methods for comparisons: RealGuidance [14], Da-fusion [44], Diff-Aug [47], Diff-Gen [47], Diff-Mix [47], DataDream [21], and Diff-II [46]. In the synthetic data generation stage, we use the hyperparameters and procedure provided in the original paper. We utilize the fine-tuned T2I weights provided by the Diff-II paper for synthetic data generation on Stable Diffusion v1.5 [35]. We reproduce Diff-Aug, Diff-Gen, and Diff-Mix using the same T2I model from the paper: Stable Diffusion v1.5. Similar for Datadream, we reproduce their results using the same T2I model: Stable Diffusion v2.1-base. Since RealGuidance and Da-fusion are relatively older methods on older T2I models, we reproduce their results with the more recent T2I model of Stable Diffusion v2.1-base. In summary, when possible, we utilized the relevant T2I model for each prior method to compare with directly.

Implementation details. For fair comparison with existing methods, we fine-tune our method on both Stable Diffusion v1.5 and Stable Diffusion v2.1-base. We utilize the same hyperparameters as DataDream with the exception of longer fine-tuning: 400 epochs instead of 200 epochs. For fair comparisons, we also extend the DataDream method to 400 epochs. We show in the Appendix that both methods improve with the increased training.

In the downstream classification fine-tuning process, we replicate the few-shot examples such that there is close to a 50/50 split of real and synthetic images. The classification objective function is a weighted average on the cross-entropy loss between real (r) and synthetic data (s): $L = \lambda \cdot CE(f_\theta(x_r, y_r) + (1-\lambda) \cdot CE(f_\theta(x_s), y_s)$, where λ is a hyperparameter and $f_\theta(x)$ is the classifier with parameters θ. Mixup (except for the Diff-Mix setting) and Cutmix augmentation is applied separately between the real and synthetic data. For fair comparison, we perform hyperparameter tuning by training only for 10 epochs and evaluating a separate validation set before finally training the downstream classifier with the best hyperparameters for 100 epochs. Details are provided in the Appendix.

4.2. Few-shot Classification

In the few-shot classification setting we use 5 or 10 real images per class which we use to fine-tune the pre-trained T2I model before generating 100 synthetic images per class. Additional experiments using different numbers of real images and synthetic images are shown in the Appendix. In training the downstream classifier, the real images are repli-

Table 1. **Few-shot classification accuracy.** The best performing method is in **bold** and the second best is underlined. Across three different backbones used as downstream classifier, our method outperforms existing methods by a considerable margin on Aircraft (AirC), Car, and CUB. On the Pets dataset, our method obtains similar performance of previous methods.

	Method	SD Ver.	5-shot					10-shot				
			AirC	Car	CUB	Pets	Avg	AirC	Car	CUB	Pets	Avg
CLIP	Real Only		44.37	79.01	67.72	92.76	70.97	55.73	84.87	74.59	93.65	77.21
	Diff-Aug	v1.5	44.67	80.93	68.05	92.27	71.48(+0.51)	57.19	86.07	77.29	93.44	78.50(+1.29)
	Diff-Gen	v1.5	47.54	81.60	69.21	91.69	72.51(+1.54)	58.60	88.43	76.45	93.57	79.26(+2.05)
	Diff-Mix	v1.5	42.09	80.19	67.45	92.78	70.63(-0.34)	52.73	87.31	73.60	93.34	76.75(-0.47)
	Diff-II	v1.5	49.02	82.16	70.41	92.75	73.59(+2.62)	60.25	89.02	77.05	93.02	79.84(+2.63)
	BOB (ours)	v1.5	55.85	88.10	**75.84**	92.24	78.01(+7.04)	**68.88**	**92.42**	**81.26**	93.31	**83.97(+6.76)**
	RealGuidance	v2.1	43.12	80.23	69.93	**92.78**	71.52(+0.55)	52.96	85.36	76.45	92.79	76.89(-0.32)
	Da-fusion	v2.1	42.39	79.83	69.33	92.59	71.035(+0.07)	55.27	79.83	76.02	**94.04**	76.29(-0.92)
	DataDream	v2.1	50.04	84.58	70.74	92.67	74.51(+3.54)	63.89	90.26	78.90	93.90	81.74(+4.53)
	BOB (ours)	v2.1	**57.37**	**88.41**	75.43	92.73	**78.49(+7.52)**	67.61	92.00	80.95	93.77	83.58(+6.37)
ImageNet	Real Only		39.62	56.16	48.22	83.17	56.79	55.48	78.50	68.05	86.75	72.20
	Diff-Aug	v1.5	43.27	70.95	57.24	85.09	64.14(+7.35)	57.91	85.34	72.74	87.40	75.85(+3.65)
	Diff-Gen	v1.5	48.42	80.73	60.91	86.95	69.25(+12.46)	60.32	88.85	72.40	89.93	77.88(+5.68)
	Diff-Mix	v1.5	38.27	76.58	53.28	85.36	63.37(+6.58)	52.21	86.41	68.16	88.63	73.85(+1.65)
	Diff-II	v1.5	52.28	82.95	63.60	**87.63**	71.62(+14.83)	62.81	88.53	73.60	**89.95**	78.72(+6.52)
	BOB (ours)	v1.5	60.02	**88.80**	68.78	86.38	76.00(+19.21)	70.79	92.60	78.62	89.04	82.76(+10.56)
	RealGuidance	v2.1	35.53	68.76	57.34	87.25	62.22(+5.43)	49.23	83.13	70.43	87.23	72.51(+0.31)
	Da-fusion	v2.1	42.60	73.99	59.03	86.17	65.45(+8.66)	56.69	85.80	71.38	88.98	75.71(+3.51)
	DataDream	v2.1	54.58	86.15	67.40	84.85	73.25(+16.46)	67.99	91.29	77.48	88.38	81.29(+9.09)
	BOB (ours)	v2.1	**60.31**	88.64	**71.38**	87.00	**76.83(+20.04)**	**73.78**	**92.52**	**79.52**	89.40	**83.81(+11.61)**
MAE	Real Only		41.13	53.94	39.63	76.81	52.88	57.61	79.12	62.50	82.97	70.55
	Diff-Aug	v1.5	44.28	72.22	55.35	74.76	61.65(+8.77)	60.64	86.79	75.72	84.41	76.89(+6.34)
	Diff-Gen	v1.5	51.79	82.66	62.79	77.32	68.64(+15.76)	63.85	90.92	77.10	85.15	79.26(+8.71)
	Diff-Mix	v1.5	41.46	78.16	52.50	81.28	63.35(+10.47)	60.31	88.29	69.09	86.40	76.02(+5.47)
	Diff-II	v1.5	54.90	82.09	66.53	**88.33**	72.69(+20.08)	65.20	90.39	77.05	89.21	80.46(+9.91)
	BOB (ours)	v1.5	**62.32**	87.73	69.23	87.46	76.69(+23.81)	75.70	**93.16**	80.17	**89.56**	84.65(+14.10)
	RealGuidance	v2.1	38.70	68.78	52.47	80.62	60.14(+7.26)	57.13	84.58	73.33	86.94	75.50(+4.95)
	Da-fusion	v2.1	46.98	73.39	51.90	75.52	61.95(+9.07)	58.57	87.61	73.33	83.21	75.68(+5.13)
	DataDream	v2.1	58.54	85.81	69.07	80.38	73.45(+20.57)	71.20	92.12	79.15	86.35	82.21(+11.66)
	BOB (ours)	v2.1	61.21	**88.48**	**73.21**	86.72	**77.41(+24.53)**	**75.85**	92.96	**81.29**	88.99	**84.77(+14.22)**

cated such that there is a 50/50 split between synthetic and real data. We present the performance of our method compared with seven existing baselines in Tab. 1: Diff-Aug, Diff-Gen, Diff-Mix, Diff-II using Stable Diffusion v1.5 and RealGuidance, Da-fusion and DataDream using Stable Diffusion v2.1. Downstream tasks include Aircraft, a task with lowest maximum starting baseline performance of 44.37%, moderate maximum baseline performance tasks of Car and CUB (79.01% and 67.72%), to Pets, with a relatively high maximum baseline performance of 92.76% in the 5-shot setting. Our method improves performance over all the baseline and the best performing existing method, in all tasks with the exception of Pets. For Aircraft, Car and CUB downstream tasks, BOB improves performance by at least 6.36% and up to 34.54% over the baseline of training with only the real data, and at least 2.77% and up to 10.25% over the best performing existing method. Detailed analysis focusing on specific aspects of these experiments follow.

Aircraft classification task. The pre-trained Stable Diffusion model has the least amount of knowledge about the Aircraft dataset, as indicated by the very poor performance of RealGuidance which is a personalization and fine-tuning free method. Focusing on the 5-shot setting for the FGVC-Aircraft classification task, using the ImageNet trained ResNet-50, augmenting real images with RealGuidance generated images results in a degradation in performance of the ImageNet pretrained model by 4.09% and 6.25% in 5- and 10-shot settings. Improvements by other previous methods range in 3.65-14.96% with DataDream performing the best, while our method, BOB leads to a 20.69% improvement raising the accuracy from 39.62% to 60.31%, 5.73% higher improvement than DataDream. Including the CLIP and MAE backbones for downstream tasks, BOB provides 3.78-7.33% improvement in the 5-shot and 4.65-5.79% in the 10-shot settings over the best performing previous method for this downstream task.

Table 2. **Long-tail classification accuracy.** The best performing method is in **bold** and the second best is underlined. The expected accuracy across all the classes is reported. *Many* reports classes with over 20 (30) examples for CUB-LT (Flower-LT). *Medium* reports classes with between 5-20 (10-30) examples for CUB-LT (Flower-LT). *Few* reports classes with under 5 (10) examples for CUB-LT (Flower-LT). Results from fine-tuning an ImageNet pre-trained ResNet-50 indicates that BOB outperforms existing methods.

Method	SD Ver.	CUB-LT						Flower-LT					
		IF=100				50	10	IF=100				50	10
		Many	Med	Few	All			Many	Med	Few	All		
Real Only		86.00	65.22	17.84	37.73	49.32	60.09	99.45	97.70	60.74	72.08	87.41	93.70
Diff-Gen	v1.5	87.22	68.69	26.06	43.95	59.47	67.78	99.79	96.71	71.25	79.17	92.93	95.12
Diff-Mix	v1.5	87.70	73.12	32.76	49.46	60.61	67.06	99.61	98.47	73.17	80.93	91.99	94.77
Diff-II	v1.5	87.54	72.16	44.05	56.10	64.52	70.28	99.82	98.45	79.51	85.35	95.20	97.62
BOB (ours)	v1.5	**88.48**	75.37	52.24	62.19	70.57	74.54	**100.0**	98.56	**84.13**	**88.60**	95.68	96.13
DataDream	v2.1	87.25	71.23	39.72	53.42	66.05	72.32	100.0	**98.67**	79.96	85.73	94.13	96.08
BOB (ours)	v2.1	88.43	**75.56**	**53.47**	**63.06**	**73.00**	**76.28**	99.45	98.41	83.48	88.07	**96.85**	**97.80**

Pets classification task. Pets classification task has the highest baseline performance of 76.81%, 83.17% and 92.76% with 5 real images and 82.97%, 86.75% and 93.65% with 10 real images for the MAE, ImageNet and CLIP backbones indicating that this downstream dataset distribution is represented much better compared to other datasets. We note two interesting observations. First, it appears that the Stable Diffusion model have additional knowledge of this dataset since RealGuidance can improve performance by more than 4%. Second, there is low variability in performance across all the methods in CLIP and ImageNet backbones. This is most likely due to significant overlap from the pre-training data. We comment on this in more detail in the Appendix. Overall, our method BOB improves performance on par with existing methods in this task reaching a performance of 87.46%, 87% and 92.73% in the 5-shot and 89.56%, 89.40% and 93.77% in the 10-shot setting, within 1% of performance of best existing method.

Comparison of 5-shot and 10-shot performance. We compare the performance of training with 5 real images augmented with synthetic data generated from those 5 images against training with 10 real images alone (without synthetic augmentation). This comparison is particularly revealing as it demonstrates whether our synthetic data generation method can effectively substitute for collecting additional real-world training examples, which is often expensive and time-consuming in practice. As shown in Tab. 1, we observe a consistent and striking pattern across all three backbones (CLIP, ImageNet, and MAE): the addition of synthetic data generated by fine-tuning with only 5 real images using BOB outperforms using 10 real images alone in all datasets with the exception of Pets, where the performance is similar. For the Aircraft dataset, the gains are particularly substantial. With the ImageNet backbone, 5 real images plus BOB-generated synthetic data (v2.1) achieves 60.31% accuracy, substantially outperforming the 55.48% achieved with 10 real images alone—a gain of 4.83%. With the MAE backbone, the improvement is even more pronounced: 61.21% (5-shot + BOB) versus 57.61% (10-shot real only), representing a 3.60% gain. Similarly, for the CUB dataset with the CLIP backbone, our 5-shot augmented approach reaches 75.43-75.84% accuracy compared to 74.59% with 10 real images, and for the Car dataset with the ImageNet backbone, we achieve 88.64% with 5-shot + BOB versus 78.50% with 10-shot real only, which is an impressive 10.14% improvement. The only exception is the Pets dataset, where the 10-shot real-only baseline already achieves very high performance, leaving limited room for improvement and suggesting that the pre-trained models already have substantial knowledge of this domain. These results indicate that BOB allows for highly efficient sampling of training data for fine-tuning the target downstream task, effectively doubling the value of each real training example.

4.3. Long-tail Classification

To demonstrate that our method extends beyond few-shot classification, we perform experiments in a long-tail classification setting using the CUB-LT dataset [39] and Flower-LT [47] using a ImageNet pre-trained ResNet-50 backbone. In the long-tail classification settings, the number of images per class used for fine-tuning is artifically skewed to follow an exponential distribution specified in Samuel et al. [39] for CUB-LT and Wang et al. [47] for Flower-LT. For synthetic data generation, we set a budget of 200 total images per class. Similar to the few-shot examples, we duplicate the real examples by a constant factor c such that the number of images in the head (classes with abundant real images) are close to 200 images, arriving at the number of images defined by the following equation: $200 -$ number of real images $\times\, c$ ($c = 6$ for CUB-LT and $c = 5$ for Flower-LT). Tab. 2 shows BOB outperforms existing methods by a considerable margin on long-tail classes.

Performance gains on CUB-LT. CUB-LT is a relatively challenging datasets for long-tail classification with relatively lower accuracies when training with only real data. In this challenging setting, we observed across every imbal-

Figure 4. **Visualizations.** *left.* 737-400 images from real data and synthetic data generated by Diff-II, DataDream, and BOB (ours). Diff-II generates images with aircrafts with high contrast in simple backgrounds. DataDream generates more realistic aircrafts that are only on the ground. Our method BOB generate realistic aircrafts in very diverse settings such as taking off, flying, or on the ground with mountainous background, resulting in images that are visually similar to real images.

anced factor on CUB-LT datasets, our method BOB outperforms existing ones by a margin of at least 4%. We observe the greatest improvement in performance when there is a large class imbalance. For the setting with the largest imbalance (IF=100), our method improves from previous method by at least 6%: 56.10% with Diff-II to 62.19% with ours and 53.42% with DataDream to 63.06% with ours. Although classes with many examples tend to have similar accuracies (87.54% with Diff-II vs. 88.48% with ours), there is a large increase in performance in the classes with few examples: 44.05% with Diff-II to 52.24% with ours. These findings provide strong evidence that for challenging long-tail tasks such as CUB-LT, BOB generates more informative data than previous methods.

Competitive performance on Flower-LT. In contrast to the dataset CUB-LT, Flower-LT is a relatively easier task with fewer classes and higher accuracies. For the most difficult part of this benchmark with IF=100, BOB achieves a 2%-3% gain. For IF=50 and IF=10, BOB performs competitively against previous methods with accuracies within 1% range. These results show that even for easier long-tail tasks, BOB generates the best synthetic datasets.

4.4. Analysis

Having demonstrated that our proposed method BOB outperforms existing methods for few-shot classification and long-tail classification, we next analyze *why* it works better. We perform analysis using the 10-shot Aircraft data setting with ImageNet-1K pre-trained ResNet-50 backbone to reveal that: (1) our method produces synthetic data that more closely resembles the real data. (2) the performance gains are not due to distilling class knowledge from the captioning model, (3) diversity in the synthetic data is crucial but does not fully explain the performance gains, and (4) both context preservation and marginalization are important for creating high performing data.

Qualitative analysis. Visualization of images generated by our method shown in Fig. 4 produces a sharp contrast compared to existing methods. We observe in Fig. 4 that the previous methods lack realism or diversity. With Diff-II, the aircrafts have high contrast with background that is typically monotonous. DataDream generates realistic looking images with many complexities in the background but the aircrafts are all on the ground. In contrast, our method BOB, produces images that are both realistic and diverse; including images of aircraft on the ground with a mountain in the background, an aircraft taking off, etc. It is clear that the synthetic images generated from BOB resemble the real images closest: suggesting that the source of performance gain is from better alignment with the real data distribution.

Real vs. synthetic distribution. We provide an analysis on the alignment of the synthetic dataset distributions with the target dataset by computing the per-class Frechet Inception Distance (FID) [17] between the whole training dataset and the synthetic datasets. A lower value indicates that the generated dataset is closer to the real data distribution. We reveal that the synthetic dataset produced by BOB perform are better aligned to the training data distribution compared to DataDream and Diff-II. In Fig. 5 *left*, we observe that, on average, the per-class FID is lower for our method BOB with the mode at around 26 compared to 31 with DataDream and 37 with Diff-II. 91 out of 100 classes exhibited a decrease in FID using our method BOB compared to DataDream (see the Appendix for more details). The lower FID further reinforces our qualitative analysis that the generated data is closer to the real data distribution.

Is it distillation? Since the captioning models themselves have some fine-grained classification capabilities, the natural question is whether we are inadvertly distilling these capabilities down to the T2I model, and subsequently, to the downstream classification model. To test this hypothesis, we use two additional captioning model: Qwen VL2.5-3B [2] and GPT-4o [19]. If the source of the performance gains is distillation, then the classifcation capabilities of these captioning model should be a strong indicator of downstream classification performance. However, Fig. 5 *right*

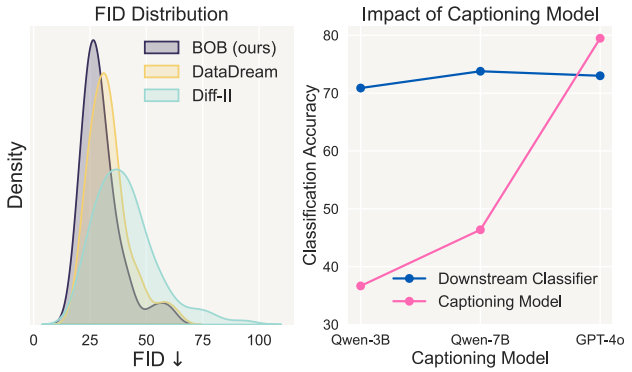

Figure 5. **Where do the performance gains come from?** *(Left)* Per-class FID density plot shows BOB generates data considerably closer to the training distribution than previous methods, explaining the performance gains. *(Right)* Caption model classification accuracy vs. downstream classifier performance shows no correlation, ruling out distillation as the source of gains.

demonstrates that there is no such association at any discernable level. We observe that GPT-4o has considerably higher fine-grained classification capability with accuracy close to 80%. However, when used as a captioning model in our method, the downstream classification performance does not improve. Similarly for the weaker captioning model Qwen-3B, we do not exhibit a considerable decrease with accuracy 70.88%, which is still over 3% higher than the DataDream baseline. These findings suggest that the performance gains we observed are not due to distilling classification capabilities from the captioning model.

Is it prompt diversity? Another question is whether the performance gains from context marginalization arise from mitigating unintended inter-class associations or simply from increasing prompt diversity. To disentangle these factors, we introduce three settings. First, we mimic context marginalization but sample backgrounds and poses only from the same class, yielding low diversity while preserving spurious correlations; this leads to very poor performance (64.38%) in Tab. 3. Next, we use the Diff-II [46] suffices, which has 10 dataset-level caption summaries, offering low diversity but remove spurious correlations; performance improves from 64.38% to 66.53%, indicating the benefit of breaking these correlations, though the accuracy remains low. Finally, we prompt GPT to generate 100 possible aircraft backgrounds and poses, providing high diversity; this achieves strong performance (72.10%). While this highlights the importance of diversity, it still underperforms our proposed method, which uses only the contexts present in the small dataset. Overall, prompt diversity is crucial but cannot fully explain the gains of BOB.

Ablation studies. Finally, ablate on context preservation and marginalization by directly including and excluding them from the pipeline. Results from Tab. 4 reveals both preservation and marginalization is necessary for creating high-performing datasets. Without preservation and marginalization, the algorithm is identical to DataDream baseline, which achieves only 68%. Adding marginalization without preservation result in a 3% improvement in accuracy to 70.13 %. Including both marginalization and preservation results in the best accuracy at 73.78%. Interestingly, preservation without marginalization leads to worse performance. We observe that the performance decreased from 68% to 65.90% when using preservation without marginalization. These ablations reveal the necessity of both context preservation and dataset-wide marginalization for generating highly informative images. Additional ablations on the attributes used is in the Appendix.

Table 3. **Analysis on prompt diversity.** Results show that prompt diversity is very important as GPT generated background and pose performs very well. However, our method BOB is still the best, indicating that it isn't all about prompt diversity.

Captions Source	Accuracy
Background/pose of the same class	64.38
Suffices provided from Diff-II [46]	66.53
100 background/pose extracted from GPT	72.10
All the extracted background/pose (BOB)	**73.78**

Table 4. **Ablations on preservation/marginalization.** Results show both are needed for creating high performant synthetic data.

Preservation	Marginalization	Accuracy
✗	✗	68.00
✓	✗	65.90
✗	✓	70.13
✓	✓	**73.78**

5. Conclusion

We introduce BOB as a fine-tuning strategy for text-to-image (T2I) models that mitigates overfitting and addressing the unique challenges of fine-grained classification. By leveraging more detailed captioning to extract class-agnostic background and pose information, conditioning on these features during fine-tuning, and marginalizing them out during data generation, our approach reduces unintended class associations and narrows the distribution gap between synthetic and real data. Extensive experiments across multiple backbones, datasets, and scales demonstrate consistent and significant performance gains, including over 7% improvement on the Aircraft dataset and best performance in nearly all settings. This work highlights the potential of caption-guided fine-tuning to improve synthetic data for downstream classification tasks and opens avenues for further research on scaling it to broader domains.

Limitation. BOB is designed for fine-grained recognition where class-agnostic attributes can be reliably inferred. How to infer the appropriate invariances for tasks with context-dependent classes remains an open question.

References

[1] Reyhane Askari-Hemmat, Mohammad Pezeshki, Elvis Dohmatob, Florian Bordes, Pietro Astolfi, Melissa Hall, Jakob Verbeek, Michal Drozdzal, and Adriana Romero-Soriano. Improving the scaling laws of synthetic data with deliberate practice. In *Forty-second International Conference on Machine Learning*, 2025. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 7

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402. IEEE, 2023. 1

[4] Max F Burg, Florian Wenzel, Dominik Zietlow, Max Horn, Osama Makansi, Francesco Locatello, and Chris Russell. Image retrieval outperforms diffusion models on data augmentation. *arXiv preprint arXiv:2304.10253*, 2023. 1

[5] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 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. *arXiv preprint arXiv:2312.03046*, 2023. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[8] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *Advances in neural information processing systems*, 36:79024–79034, 2023. 2

[9] 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. 1

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023. 2

[11] Scott Geng, Cheng-Yu Hsieh, Vivek Ramanujan, Matthew Wallingford, Chun-Liang Li, Pang Wei W Koh, and Ranjay Krishna. The unmet promise of synthetic training images: Using retrieved real images performs better. *Advances in Neural Information Processing Systems*, 37:7902–7929, 2024. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 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 *International Conference on Learning Representations*, 2023. 2, 4

[15] Reyhane Askari Hemmat, Mohammad Pezeshki, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Feedback-guided data synthesis for imbalanced classification. *Transactions on Machine Learning Research*, 2024. 2

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[17] 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. 7

[18] 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. 2022. 3

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7

[20] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[21] 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

[22] Soroush Abbasi Koohpayegani, Anuj Singh, KL Navaneet, Hadi Jamali-Rad, and Hamed Pirsiavash. Genie: Generative hard negative images through diffusion. *arXiv preprint arXiv:2312.02548*, 2023. 2

[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. 4

[24] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 2

[25] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023. 2

[26] Alexander Cong Li, Ananya Kumar, and Deepak Pathak. Generative classifiers avoid shortcut solutions. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[27] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023. 2

[28] Yiwen Liu, Jessica Bader, and Jae Myung Kim. Does feasibility matter? understanding the impact of feasibility on synthetic training data. *arXiv preprint arXiv:2505.10551*, 2025. 2

[29] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 4

[30] Eyal Michaeli and Ohad Fried. Advancing fine-grained classification by structure and subject preserving augmentation. *Advances in Neural Information Processing Systems*, 37: 22316–22349, 2024. 2

[31] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 1

[32] 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, 4

[33] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[34] 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. 2, 3, 4

[35] 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, 4

[36] 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

[37] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[38] 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

[39] Dvir Samuel, Yuval Atzmon, and Gal Chechik. From generalized zero-shot learning to long-tail with class descriptors. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 286–295, 2021. 4, 6

[40] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8011–8021, 2023. 2

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 1

[42] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36: 48382–48402, 2023. 1

[43] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15887–15898, 2024. 1

[44] Brandon Trabucco, Kyle Doherty, Max A Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *International Conference on Learning Representations*, 2023. 2, 4

[45] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 4

[46] Yanghao Wang and Long Chen. Inversion circle interpolation: Diffusion-based image augmentation for data-scarce classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25560–25569, 2025. 2, 4, 8

[47] Zhicai Wang, Longhui Wei, Tan Wang, Heyu Chen, Yanbin Hao, Xiang Wang, Xiangnan He, and Qi Tian. Enhance image classification via inter-class image mixup with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17223–17233, 2024. 1, 2, 4, 6

[48] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2

[49] Zhuoran Yu, Chenchen Zhu, Sean Culatana, Raghuraman Krishnamoorthi, Fanyi Xiao, and Yong Jae Lee. Diversify, don't fine-tune: Scaling up visual recognition training with synthetic images. *Transactions on Machine Learning Research*, 2025. Featured Certification. 2

[50] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *Advances in neural information processing systems*, 33:2734–2746, 2020. 2

[51] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2

[52] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 2

[53] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2

[54] Brian Nlong Zhao, Yuhang Xiao, Jiashu Xu, Xinyang Jiang, Yifan Yang, Dongsheng Li, Laurent Itti, Vibhav Vineet, and Yunhao Ge. Dreamdistribution: Learning prompt distribution for diverse in-distribution generation. In *International Conference on Learning Representations*, 2025. 2

[55] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *arXiv preprint arXiv:2305.15316*, 2023. 2

[56] Zhuofan Zong, Dongzhi Jiang, Bingqi Ma, Guanglu Song, Hao Shao, Dazhong Shen, Yu Liu, and Hongsheng Li. Easyref: Omni-generalized group image reference for diffusion models via multimodal llm. In *Forty-second International Conference on Machine Learning*, 2024. 2