Enhance Image Classification via Inter-Class Image Mixup with Diffusion Model

Zhicai Wang[1]*, Longhui Wei[2]†, Tan Wang[3], Heyu Chen[1], Yanbin Hao[1], Xiang Wang[1]†,
Xiangnan He[1], Qi Tian[2]
[1]University of Science and Technology of China, [2] Huawei Inc.,
[3] Nanyang Technological University

Abstract

Text-to-image (T2I) generative models have recently emerged as a powerful tool, enabling the creation of photo-realistic images and giving rise to a multitude of applications. However, the effective integration of T2I models into fundamental image classification tasks remains an open question. A prevalent strategy to bolster image classification performance is through augmenting the training set with synthetic images generated by T2I models. In this study, we scrutinize the shortcomings of both current generative and conventional data augmentation techniques. Our analysis reveals that these methods struggle to produce images that are both faithful (in terms of foreground objects) and diverse (in terms of background contexts) for domain-specific concepts. To tackle this challenge, we introduce an innovative inter-class data augmentation method known as Diff-Mix [1], which enriches the dataset by performing image translations between classes. Our empirical results demonstrate that Diff-Mix achieves a better balance between faithfulness and diversity, leading to a marked improvement in performance across diverse image classification scenarios, including few-shot, conventional, and long-tail classifications for domain-specific datasets.

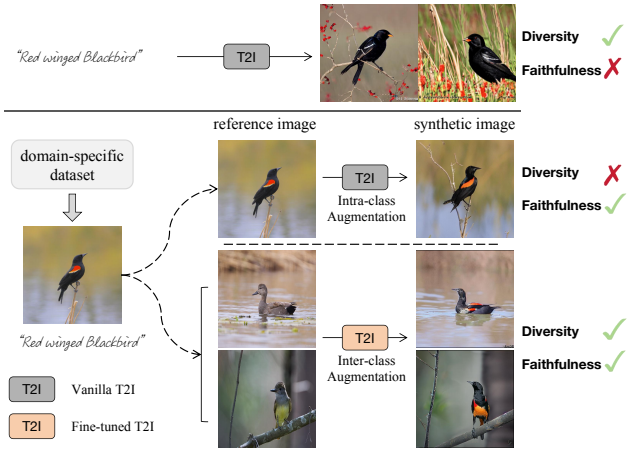


Figure 1. Strategies to expand domain-specific datasets for improved classification are varied. Row 1 illustrates vanilla distillation from a pretrained text-to-image (T2I) model, which carries the risk of generating outputs with reduced faithfulness. Intra-class augmentation, depicted in Row 2, tends to yield samples with limited diversity to maintain high fidelity to the original class. Our proposed method, showcased in Rows 3 and 4, adopts an inter-class augmentation strategy. This involves introducing edits to a reference image using guidance of another class, which significantly enriches the dataset with a greater diversity of samples.

1. Introduction

In comparison to GAN-based models [7, 17, 25], contemporary state-of-the-art text-to-image (T2I) diffusion models exhibit enhanced capabilities in producing high-fidelity images [12, 37, 44, 49]. With the remarkable cross-modality alignment capabilities of T2I models, there is significant potential for generative techniques to enhance image classification [2, 4]. For instance, a straightforward approach entails augmenting the existing training dataset with synthetic images generated by feeding categorical textual prompts to a T2I diffusion model. However, upon reviewing prior approaches employing T2I diffusion models for image classification, it becomes evident that the challenge in generative data augmentation for domain-specific datasets is producing samples with both a **faithful foreground** and a **diverse background**. Depending on whether a reference image is used in the generative process, we divide these methods into two groups:

- Text-guided knowledge distillation [52, 57] involves generating new images from scratch using category-related prompts to expand the dataset. For the off-the-shelf T2I models, such vanilla distillation presume these models have comprehensive knowledge of target domain, which can be problematic for domain-specific datasets. Insufficient domain knowledge easily makes the distillation process less effective. For example, vanilla T2I models

*This work was done during the internship in Huawei Inc..
†Xiang Wang and Longhui Wei are both the corresponding authors.
[1]https://github.com/Zhicaiwww/Diff-Mix

struggle to generate images that accurately represent specific bird species based solely on their names (see Row 1 of Fig. 1).
- Generative data augmentation [1, 69] employs generative models to enhance existing images. Da-fusion [58], for instance, translates the source image into multiple edited versions within the **same class**. This strategy, termed **intra-class** augmentation, primarily introduces intra-class variations. While intra-class augmentation retains much of the original image's layout and visual details, it results in limited background diversity (see Row 2 of Fig. 1). However, synthetic images with constrained diversity may not sufficiently enhance the model's ability to discern foreground concepts.

Based on these observations, a fundamental question emerges: '*Is it feasible to develop a method that optimizes both the diversity and faithfulness of synthesized data simultaneously*?'

In this work, we introduce **Diff-Mix**, a simple yet effective data augmentation method that harnesses diffusion models to perform **inter-class** image interpolation, tailored for enhancing domain-specific datasets. The method encompasses two pivotal operations: personalized fine-tuning and inter-class image translation. Personalized fine-tuning [15, 46] is originally designed for customizing T2I models and enabling them to generate user-specific contents or styles. In our case, we implement the technique to tailor the model, enabling it to generate images with faithful foreground concepts. Inter-class image translation in Diff-Mix entails transforming a reference image into an edited version that incorporates prompts from different classes. This translation strategy is designed to retain the original background context while editing the foreground to align with the target concept. For instance, as depicted in the bottom rows of Fig. 1, Diff-Mix can generate images of land birds in diverse settings, such as maritime environments, enriching the dataset with a variety of counterfactual samples.

Unlike previous non-generative augmentation methods, such as Mixup [68] and CutMix [66], Diff-Mix works in a foreground-perceivable inter-class interpolation manner and shares a different mechanism with the non-generative approaches. Our experiment under the conventional classification setting indicates that incorporating both CutMix and Diff-Mix could further enhance performance. Additionally, when compared with other generative approaches, we conduct experiments under few-shot and long-tail scenarios and observe consistent performance improvements. Our contributions can be summarized as follows:

- We pinpoint the critical factors that affect the efficacy of generative data augmentation in domain-specific image classification: namely, faithfulness and diversity.
- We introduce Diff-Mix, a simple yet effective generative data augmentation strategy that leverages fine-tuned diffusion models for inter-class image interpolation.
- We conduct a comparative analysis of Diff-Mix with other distillation-based and intra-class augmentation methods, as well as non-generative approaches, highlighting its unique features and benefits.

2. Related Works

Text-to-image diffusion models. Following pretraining on web-scale data, the T2I diffusion model has demonstrated robust capabilities in generating text-controlled images [37, 49, 53, 56]. Its versatility has led to diverse applications, including novel view synthesis [6, 63], concept learning [28, 46], and text-to-video generation [22, 55], among others. Recent advancements [29] also highlight the cross-modality features of such generative models, showcasing their ability to serve as zero-shot classifiers.

Synthetic data for image classification. There are two perspectives on the utilization of synthetic data for image classification: knowledge distillation [2] and data augmentation [5, 51, 54, 62]. From the knowledge distillation perspective, SyntheticData [19] reports significant performance gains in both zero-shot and few-shot settings by leveraging off-the-shelf T2I models to obtain synthetic data. The work of [2] has indicated that fine-tuning the T2I model on ImageNet [47] yields improved classification accuracy by narrowing the domain gap. Some works also find that learning from the synthetic data presents strong transferability [19, 57] and robustness [4, 30, 67]. From the data augmentation standpoint, Da-fusion [58] achieves stable performance improvements on few-shot datasets by augmenting from reference images. In a related study [3], the use of StyleGAN [25] for generating interpolated images between two different domains has been shown to enhance classifier robustness for out-of-distribution data. Our work shares similarities with AMR [5], which generates realistic novel examples by interpolating between two images using GAN [16]. The distinction lies in our discussion of interpolation using the T2I diffusion model, where its noise-adding and denoising characteristics enable a smoother implementation of interpolation.

Non-generative data augmentation. Mixup [68] and CutMix [66] stand out as two prominent non-generative data augmentation methods, serving as effective regularization techniques during training. While Mixup achieves augmented samples through a convex combination of two images, CutMix achieves augmentation by cutting and pasting parts of images. However, both methods are constrained in their ability to produce realistic images. In addressing this limitation, the utilization of generative models emerges as a potential solution to alleviate this issue.

3. Method

3.1. Preliminary

Text-to-Image diffusion model. Diffusion models generate images by gradually removing noise from a Gaussian noise source [21]. In a diffusion process with a total of T steps, its forward process, which gradually adds noise, is represented as a Markov chain with a Gaussian transition kernel, where $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I}\right)$, where x_t represents the noisy image at step t. The training objective at step t is to predict the noise to reconstruct $\boldsymbol{x}_{t-1}$. When training a text-conditioned diffusion model, the simplified training objective can be summarized as follows:

$$\mathbb{E}_{\epsilon,\mathbf{x},\mathbf{c},t}\left[\|\epsilon - \epsilon_\theta\left(\mathbf{x}_t, \mathbf{c}, t\right)\|_2^2\right], \tag{1}$$

where ϵ_θ represents the predicted noise, and $\boldsymbol{c}$ is the encoded text caption associated with the image $\boldsymbol{x}$.

T2I personalization. T2I personalization aims to personalize a diffusion model for generating specific concepts using a limited number of concept-oriented images [28, 40, 42]. These concepts are typically represented using identifiers (*e.g.*, "`[V]`"). As a result, we formalize the constructed image-caption set as $\boldsymbol{x}$, "`photo of a [V]`". Various personalization methods differ in their fine-tuning strategies. For instance, Textual Inversion (TI) [15] makes the identifier learnable, but other modules are not fine-tuned, potentially sacrificing some faithfulness in image generation. On the other hand, Dreambooth (DB) [46] fine-tunes the U-Net [45] for more refined personalized generation but faces the challenge of increased computational cost.

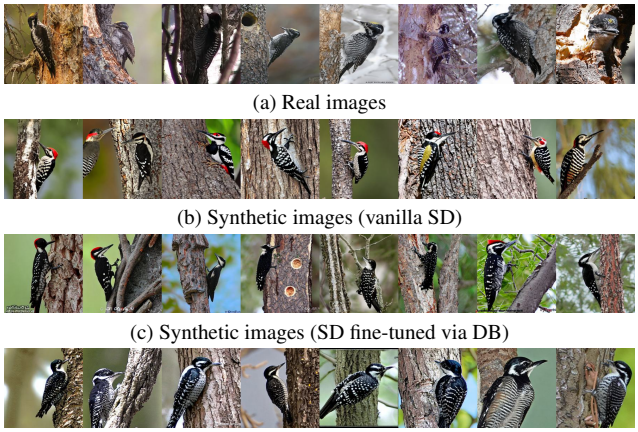
Image-to-image translation. Image translation enables image synthesis and editing using a reference image as guidance [24, 64, 71]. Diffusion-based image translation methods can generate fine edits, which refer to subtle modifications, with varying degrees of shift relative to the reference image [8, 35, 59]. Here, we draw inspiration from SDEdit [35] to perform edits on the reference image, where the target image $\boldsymbol{x}^{\text{tar}}$ is translated from a reference image $\boldsymbol{x}^{\text{ref}}$. During translation, the reverse process does not traverse the full process but starts from a certain step $\lfloor sT \rfloor$, where $s \in [0, 1]$ controls the insertion position of the reference image with noise, as follows,

$$\boldsymbol{x}_{\lfloor sT \rfloor} = \sqrt{\tilde{\alpha}_{\lfloor sT \rfloor}}\boldsymbol{x}_0^{\text{ref}} + \sqrt{1-\tilde{\alpha}_{\lfloor sT \rfloor}}\epsilon. \tag{2}$$

By adjusting the strength parameter s, one can strike a balance between the diversity of the generated images and their faithfulness to the reference image.

3.2. General Framework

The Diff-Mix pipeline consists of two key steps. Firstly, to produce more faithful images for domain-specific datasets, we propose treating it as a T2I personalization problem and fine-tuning the Stable Diffusion (SD). Subsequently, to enhance the diversity of synthetic data beyond the well-fitted training distribution, we employ inter-class image translation. This process produces interpolated images with increased background diversity for each class.



(a) Real images

(b) Synthetic images (vanilla SD)

(c) Synthetic images (SD fine-tuned via DB)

(d) Synthetic images (SD fine-tuned via TI+DB)

Figure 2. Examples of "`American Three toed Woodpecker`". (a) Real images from the training set. (b-d) synthetic images generated using different fine-tuned models with the same number of fine-tuning steps. TI+DB indicates both text embedding and U-Net are fine-tuned. TI+DB achieves a more faithful output compared to DB alone (check the head and wing patterns of the birds).

3.3. Fine-tune Diffusion Model

Vanilla distillation tends to be less effective, especially as the number of training shots increases (refer to Sec. 4.1). In order to mitigate the distribution gap, we propose fine-tuning Stable Diffusion in conjunction with current widely-used T2I personalization strategies.

Dreambooth meets Textual Inversion. Many fine-grained datasets provide terminological names for their categories, like "`American Three toed Woodpecker`" and "`Pileated Woodpecker`". We could construct image-text pairs using category-related prompts and fine-tune the denoising network of SD using Eq. 1, which is analogous to Dreambooth. However, we observe that directly incorporating these specialized terms into the text during fine-tuning can impede convergence and hinder the generation of faithful images. We attribute this challenge to the semantic proximity of terminology within a fine-grained domain, where fine-tuning the vision module alone tends to be less effective at distinguishing two similar classes within the same family, like "`Woodpecker`". Inspired by Textual Inversion [15], we opt to replace the terminological name in the dataset with "`[V`^{i}`] [metaclass]`" where "`[V`^{i}`]`" is a learnable identifier, and i varies from 1 to N, represent-

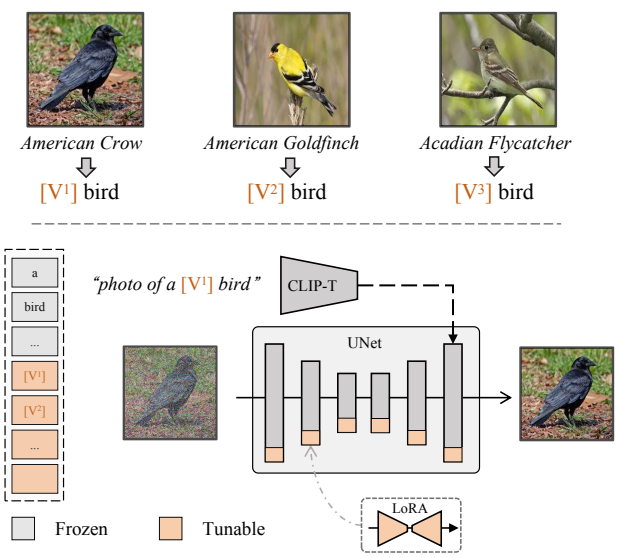


Figure 3. Fine-tuning framework of Diff-Mix operates as follows: Initially, we replace the class name with a structured identifier formatted as "`[V`^{i}`] [metaclass]`", thereby sidestepping the need for specific terminological expressions. Next, we engage in joint fine-tuning of these identifiers and the low-rank residues (LoRA) of U-Net to capture the domain-specific distribution.

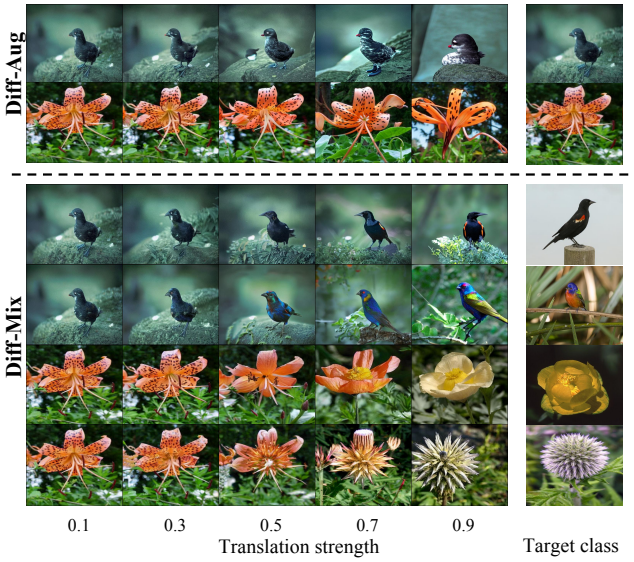


Figure 4. Examples of images translated using Diff-Mix and Diff-Aug across various strengths. Diff-Aug employs the same target and reference image classes, typically resulting in subtle modifications. Diff-Mix progressively adjusts the foreground to align with the target class as the translation strength increases, while preserving the background layout from the reference image.

ing the category index. The illustration of our fine-tuning strategy is presented in Fig. 3. The term "`[metaclass]`" is determined by the theme of the current dataset, such as "`bird`" for a fine-grained bird dataset. By concurrently fine-tuning the identifier and the U-Net, we empower the model to quickly adapting to the fine-grained domain, allowing it to generate faithful images using the identifier (see comparison between Row 3 and Row 4 in Fig. 2).

Parameter efficient fine-tuning. In this context, we embrace the parameter-efficient fine-tuning strategy known as LoRA [23]. LoRA distinguishes itself by fine-tuning the residue of low-rank matrices instead of directly fine-tuning the pre-trained weight matrices. To elaborate, consider a weight matrix $\boldsymbol{W} \in \mathbb{R}^{m\times n}$. The tunable residual matrix $\Delta\boldsymbol{W}$ comprises two low-rank matrices: $\boldsymbol{A} \in \mathbb{R}^{m\times d}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times d}$, defined as $\Delta\boldsymbol{W} = \boldsymbol{A}\boldsymbol{B}^\top$. As a default configuration, we set the rank d to 10.

3.4. Data Synthesis Using Diffusion Model

In generating pseudo data, three strategies can be used with our fine-tuned diffusion model [2]: (1) distillation-based method **Diff-Gen**, (2) intra-class augmentation **Diff-Aug**, and (3) inter-class augmentation **Diff-Mix**.

Diff-Gen and Diff-Aug. For a target class y^i and its textual condition, "`photo of a [V`^{i}`] [metaclass]`", both methods generate synthetic samples annotated with class i. Specifically, Diff-Gen generates samples from scratch by initializing with random Gaussian noise and proceeding through the full reverse diffusion process with T steps. Diff-Gen can produce images aligned with its fine-tuned distribution. In contrast, Diff-Aug sacrifices a portion of diversity and generate images by editing on a reference image. Specifically, it randomly sample a image from the intra-class training set and enhances the image through image translation using Eq. 2. The term "intra-class" means that the conditioning prompts are constructed based on the ground truth categories of images, and such a denoising process tends to introduce less variation, particularly for the foreground concepts (see top rows of Fig. 1).

Diff-Mix. Diff-Mix employs the same translation process as Diff-Aug, but the reference image is sampled from the full training set rather than intra-class set to enable inter-class interpolation. The key difference is that Diff-Mix can generate numerous counterfactual examples, such as a blackbird in the sea (see fourth row of Fig. 1). This necessitates that downstream models make a more refined differentiation of category attributes, thereby reducing the impact of spurious correlations introduced by variations in the background. Denoting the label of the reference image as y^j, by inserting the reference image into the reverse process with the prompt "`photo of a [V`^{i}`] [metaclass]`", we can obtain interpolated images between the i_{th} and j_{th} categories. By controlling the intensity s, we can precisely

[2]We use the prefix "Diff-" denotes the T2I model is fine-tuned and "Real-" denotes the vanilla T2I model.

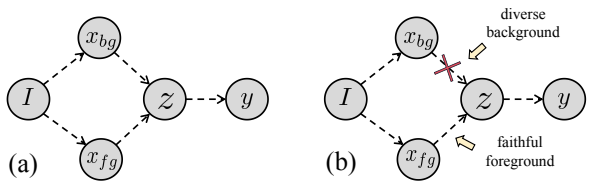


Figure 5. A schematic explanation of Diff-Mix's effectiveness using structural casual model [41]. $\boldsymbol{x}_{fg}$ is the foreground that determines the real class label, $\boldsymbol{x}_{bg}$ denotes the background. $\boldsymbol{x}_{fg} \rightarrow \boldsymbol{z} \rightarrow y$ is the causal path that we are focusing and $\boldsymbol{x}_{fg} \leftarrow I \rightarrow \boldsymbol{x}_{bg} \rightarrow \boldsymbol{z} \rightarrow y$ is the backdoor path that introduces spurious relations between $\boldsymbol{x}_{fg}$ and y.

manage the interpolation process. When annotating the synthetic image, unlike Mixup and Cutmix, we take into account the non-linear nature of diffusion translation. Thus the annotation function is given by

$$\tilde{y} = (1 - s^{\gamma})y^{i} + s^{\gamma}y^{j}, \tag{3}$$

where γ is a hyperparameter introducing non-linearity. Our empirical findings indicate that a γ smaller than 1 is favored. Additionally, in low-shot cases, the samples with higher confidence in the target class are preferred (see details in Sec. 4.5).

Construct synthetic dataset. To construct the synthetic dataset using Diff-Mix, similar to Da-fusion [58], we adopt a randomized sampling strategy ($s \in \{0.5, 0.7, 0.9\}$) for the selection of translation strength. While applying the inter-class editing, we observe that Diff-Mix tends to produce more undesirable samples compared to Diff-Aug. These undesirable samples have incomplete foreground such as fragmented bird bodies. This is caused by the intrinsic shape and pose differences among classes. To mitigate this, we introduce a simple data-cleaning approach to reduce the proportion of such problematic images. We utilize the large vision language model CLIP [43] to assess the confidence in the content, serving as the filtering criterion. Further details can be found in the supplementary materials (SMs).

Analysis. We depict the core insight of Diff-Mix in Fig. 5. To eliminate the spurious correlation introduced by $\boldsymbol{x}_{bg}$, learning on the synthetic set with randomized $\boldsymbol{x}_{bg}$ (background) can cut off spurious correlation, forcing the classification model to infer only from the foreground. The study in Fig. 7 (b) shows that the more diverse the background (larger the referable class number), the better the performance on the CUB test set.

4. Experiments

In this section, we investigate the effectiveness of Diff-Mix in domain-specific datasets. The key questions we aim to address are as follows:

Q1: Can generative inter-class augmentation lead to more significant performance gains in downstream tasks compared to those intra-class augmentation methods and distillation-based methods?
Q2: Is improved background diversity the secret weapon of Diff-Mix for enhancing the performance?
Q3: How to choose the fine-tuning strategy and annotation strategy to boost the performance gains for the inter-class augmentation?

To address Q1, we separately discuss these questions in few-shot settings in Sec. 4.1, conventional classification in Sec. 4.2, as well as long-tail classification in Sec. 4.3. Additionally, to answer Q2, we conduct a test for background robustness in Sec. 4.4 and perform an ablation study focusing on the size and diversity of synthetic data in Sec. 4.5. For Q3, we conduct an ablation study to empirically discover effective strategies for deployment in Sec. 4.5.

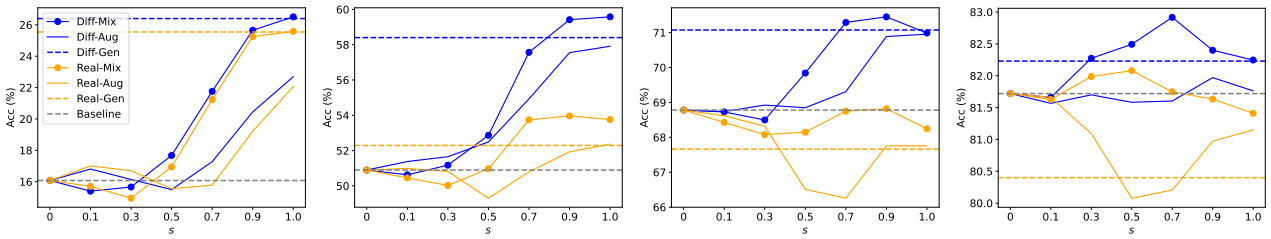
4.1. Few-shot Classification

Experimental Setting. To investigate the impact of different data expansion methods, we conduct few-shot experiments on a domain-specific dataset Caltech-UCSD Birds (CUB) [61], with shot numbers of 1, 5, 10, and all. For augmentation-based methods, the synthetic dataset is constructed using various translation strengths (s), specifically, $s \in \{0.1, 0.3, ..., 1.0\}$. We expand the original training set with a multiplier of 5 and cache the synthesized dataset locally for joint training. Real data are replaced by synthetic data proportionally during training, and the replacement probability p is set as 0.1, 0.2, 0.3, and 0.5 for all-shot, 10-shot, 5-shot, and 1-shot classification, respectively. All experiments use ResNet50 with an input resolution of 224×224. Additional details can be found in the SMs.

Comparison methods. To unveil the trade-off between faithfulness and diversity resulting from different expansion strategies, we compared **Diff-Mix** with **Diff-Gen** and **Diff-Aug**. Furthermore, we conduct experiments on expansion strategies using vanilla SD: **Real-Mix**, **Real-Gen**, and **Real-Aug**, where 'Real' signifies that SD is not fine-tuned.

Main results. To answer Q1 under the few-shot classification setting, we augment CUB using X-Mix, X-Aug, and X-Gen, where 'X' denotes 'Diff/Real' for simplicity. The results are shown in Fig. 6, and we can observe that:

1. Diff-Mix generally outperforms the intra-class competitor X-Aug and distillation competitor X-Gen in various few-shot scenarios. It tends to achieve higher gains when the strength s is relatively large, *i.e.*, $\{0.5, 0.7, 0.9\}$, where the foreground has been edited to match the target class and the background retains similarities to the reference image.
2. Among the Real-X methods, distillation tends to be more effective than the augmentation method when the shot



(a) 1-shot (b) 5-shot (c) 10-shot (d) All-shot

Figure 6. Few-shot classification results on CUB.

number is low, but the trend reverses as the shot number increases (compare Real-Gen with Real-Aug). Real-Gen's samples even show less effective than Real-Aug ($s = 1.0$) under the all-shot case [3]. This indicates that the importance of faithfulness in the trade-off between faithfulness and diversity increases with the shot number. Additionally, Real-Mix exhibits consistent and stable improvement over the other two methods.

3. Diff-Gen consistently outperforms Real-Gen under four scenarios. Notably, Real-Gen's performance declines below that of the baseline as the shot numbers reach 10, showcasing the importance of the fine-tuning process which increases the faithfulness of synthetic samples.

4.2. Conventional Classification

Experimental setting. To test whether Diff-Mix can further boost performance in a more challenging setting, *i.e.*, under the all-shot scenario with high input resolution, we conduct conventional classification on five domain-specific datasets: CUB [61], Stanford Cars [27], Oxford Flowers [38], Stanford Dogs [26], and FGVC Aircraft [34]. Two backbones are employed: pretrained (ImageNet1K [49]) ResNet50 [18] with input resolution 448×448, and pretrained (ImageNet21K) ViT-B/16 [13] with input resolution 384×384. Label smoothing [36] is applied across all datasets with a confidence level of 0.9. For all expansion strategies, the expansion multiplier is 5 and the replacement probability p is 0.1. Besides, we use a randomized sampling strategy ($s \in \{0.5, 0.7, 0.9\}$) and a fixed γ (0.5, and this is specific to Diff-Mix) for all datasets.

Comparison methods. We compare Diff-Mix with (1) Real-Filtering (RF) [19], a variation of Real-Gen that incorporates clip filtering, (2) Real-Guidance (RG) [19], which augments the dataset using intra-class image translation at low strength ($s = 0.1$), (3) Da-Fusion [58], a method that solely fine-tunes the identifier to personalize each class and employs randomized sampling strategy ($s \in \{0.25, 0.5, 0.75, 1.0\}$), and non-generative augmentation methods (4) CutMix [66] and (5) Mixup [68].

Main results. We show the classification accuracy for different data expansion strategies in Table 1, our observations can be summarized: (1) Diff-Mix consistently demonstrates stable improvements across the majority of settings. Its average performance gain across the five datasets exceeds that of baselines employing intra-class augmentation methods (RG and Da-fusion), distillation method (RF), and non-generative data augmentation techniques (CutMix and Mixup). (2) Real-filtering, analogous to the discussion of Real-Gen, exhibits performance degradation on most datasets due to the distribution gap. (3) The combined use of Diff-Mix and CutMix often yields better performance gains. This is attributed to the distinct enhancement mechanisms of the two methods, *i.e.*, vicinal risk minimization [11, 68] and foreground-background disentanglement. (4) Diff-Mix does not exhibit significant performance improvement in the dog dataset. We attribute this lack of improvement to the complexity of the dog dataset, which often contains multiple subjects in a single image, impeding effective foreground editing (refer to the SMs for visual examples).

4.3. Long-Tail Classification

Experiment setting. Following the settings of previous long-tail dataset constructions [9, 32, 39], we create two domain-specific long-tail datasets, CUB-LT [50] and Flower-LT. The imbalance factor controls the exponential distribution of the imbalanced data, where a larger value indicates a more imbalanced distribution. To leverage generative models for long-tail classification, we adopt the approach of SYNAuG [65], which uniformize the imbalanced real data distribution using synthetic data. Translation strength s (0.7) and γ (0.5) are fixed for both Diff-Mix and Real-Mix.

Comparison methods. We compare Diff-Mix with Real-Mix, Real-Gen, the non-generative CutMix-based oversampling approach CMO [39], and its enhanced variant with two-stage deferred re-weighting [9] (CMO+DRW).

Main results. We present the long-tail classification re-

[3] Real-Aug ($s = 1.0$) remains analogous to the reference image (slightly higher faithfulness compared to Real-Gen) because the discrete forward process cannot approximate the ideal normal distribution within a limited number of steps ($T = 25$).

Backbone	Aug. Method	FT Strategy	Dataset					
			CUB	Aircraft	Flower	Car	Dog	Avg
ResNet50@448	-	-	86.64	89.09	99.27	94.54	87.48	91.40
	Cutmix[66]	-	87.23	89.44	99.25	94.73	87.59	91.65$_{+0.25}$
	Mixup[68]	-	86.68	89.41	99.40	94.49	87.42	91.48$_{+0.08}$
	Real-filtering [19]†	✗	85.60	88.54	99.09	94.59	87.30	91.22$_{-0.18}$
	Real-guidance [19]†	✗	86.71	89.07	99.25	94.55	87.40	91.59$_{+0.19}$
	Da-fusion [58]†	TI	86.30	87.64	99.37	94.69	87.33	91.07$_{-0.58}$
	Diff-Mix	TI + DB	87.16	**90.25**	**99.54**	95.12	87.74	91.96$_{+0.56}$
	Diff-Mix + Cutmix	TI + DB	**87.56**	90.01	99.47	**95.21**	**87.89**	**92.03**$_{+0.63}$
ViT-B/16@384	-	-	89.37	83.50	99.56	94.21	92.06	91.74
	Cutmix[66]	-	**90.52**	83.50	99.64	94.83	**92.13**	92.12$_{+0.38}$
	Mixup[68]	-	90.32	84.31	**99.73**	94.98	92.02	92.27$_{+0.53}$
	Real-filtering [19]†	✗	89.49	83.07	99.36	94.66	91.91	91.69$_{-0.05}$
	Real-guidance [19]†	✗	89.54	83.17	99.59	94.65	92.05	91.80$_{+0.06}$
	Da-fusion [58]†	TI	89.40	81.88	99.61	94.53	92.07	91.50$_{-0.24}$
	Diff-Mix	TI + DB	90.05	84.33	99.64	95.09	91.99	92.22$_{+0.48}$
	Diff-Mix + Cutmix	TI + DB	90.35	**85.12**	99.68	**95.26**	91.89	**92.46**$_{+0.72}$

Table 1. Conventional classification in six fine-grained datasets. '†' indicates our reproduced results using SD.

Method	IF=100				50	10
	Many	Medium	Few	All		
CE	79.11	64.28	13.48	33.65	44.82	58.13
CMO [39]	78.32	58.57	14.78	32.94	44.08	57.62
CMO + DRW [10]	78.97	56.36	14.66	32.57	46.43	59.25
Real-Gen	**84.88**	65.23	30.68	45.86	53.43	61.42
Real-Mix (s=0.7)	84.63	66.34	34.44	47.75	55.67	62.27
Diff-Mix (s=0.7)	84.07	**67.79**	**36.55**	**50.35**	**58.19**	**64.48**

Table 2. Long-tail classification in CUB-LT.

Method	IF=100				50	10
	Many	Medium	Few	All		
CE	99.19	94.95	58.18	80.43	90.87	95.07
CMO [39]	99.25	95.19	67.45	83.95	91.43	95.19
CMO+ DRW [10]	**99.97**	95.06	67.31	84.07	92.06	95.92
Real-Gen	98.64	95.55	66.10	83.56	91.84	95.22
Real-Mix (s=0.7)	99.87	96.26	68.53	85.19	92.96	96.04
Diff-Mix (s=0.7)	99.25	**96.98**	**78.41**	**89.46**	**93.86**	**96.63**

Table 3. Long-tail classification in Flower-LT.

sults for CUB-LT in Table 2 and Flower-LT in 3. The observations are as follows: (1) Generative approaches exhibit superior performance in tackling imbalanced classification issues compared to CutMix-based methods (CMO and CMO+DRW). (2) Real-Mix surpasses Real-Gen in performance across various imbalance factors in two datasets. This indicates that tail classes can benefit from the enhanced diversity by leveraging the visual context of majority classes. (3) Diff-Mix generally achieves the best performance among the compared strategies, especially at the low-shot case, highlighting the importance of fine-tuning.

4.4. Background Robustness

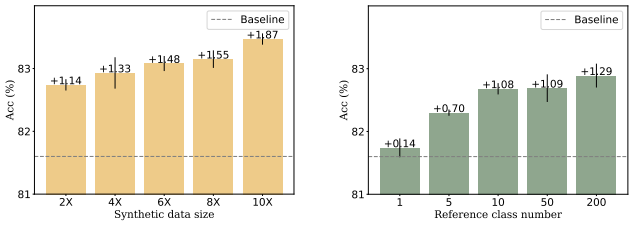
In this section, we aim to evaluate Diff-Mix's robustness to background shifts, specifically, whether synthesizing more diverse samples can improve the classification model's generalizability when the background is altered. To achieve this, we utilize an out-of-distribution test set for CUB, namely Waterbird [48]. We then perform inference on the whole Waterbird set using classifiers that have been trained on either the original CUB dataset or its expanded variations. In Table 4, we present the classification accuracies across the four groups and compare Diff-Mix with other intra-class methods (Da-fusion and Diff-Aug) as well as CutMix. We observe that Diff-Mix generally outperforms its counterparts and achieves a significant performance improvement (+6.5%) in the challenging counterfactual group (waterbirds with land backgrounds). It is important to highlight that the background scenes in the Waterbird dataset are novel to CUB, requiring the classification model to have a stronger perceptual capability for the images' foregrounds.

Group	Base.	CutMix	DA-fusion	Diff-Aug	**Diff-Mix**
(waterbird, water)	59.50	62.46	60.90	61.83	**63.83**
(waterbird, land)	56.70	60.12	58.10	60.12	**63.24**
(landbird, land)	73.48	73.39	72.94	73.04	**75.64**
(landbird, water)	73.97	**74.72**	72.77	73.52	74.36
Avg.	70.19	71.23	69.90	70.28	**72.47**

Table 4. Classification results across four groups in Waterbird [48]. Waterbird is an out-of-distribution dataset for CUB, crafted by segmenting CUB's foregrounds and paste them into the scene images from Places [70]. The constructed dataset can be divided into four groups based on the composition of foregrounds (waterbird and landbird) and backgrounds (water and land) .

4.5. Discussion

In this section, we address Q2 by examining the impact of size and diversity on synthetic data. Furthermore, we perform an ablation study to assess the effects of fine-tuning strategies and training hyperparameters of Diff-Mix, which is aimed at answering Q3. Unless specified otherwise, our discussions are based on experiments conducted using CUB



(a) (b)

Figure 7. Comparison of results across various (a) synthetic data sizes and (b) numbers of referable classes for each target class.

with ResNet50, where inputs are resized to 224×224.

Size and diversity of synthetic Data. The relationship between performance gain and the size of synthetic data is depicted in Fig. 7 (a), where a classification model was trained with synthetic data of varying sizes. We observe a monotonically increasing trend as the multiplier for the synthetic dataset ranges from 2 to 10. Ideally, the combination choices of $(\boldsymbol{x}_i, y_j)$ are in the order of $N|D^{train}|$ ($N = 200$ for CUB). Furthermore, we limit the number of referable classes for each target class, which means the number of referable backgrounds decreases, resulting in a synthetic dataset of relatively lower diversity. The results are shown in Fig. 7 (b), and we observe a consistent improvement in performance as the number of referable classes increases. These results consistently underscore the critical role of background diversity introduced by Diff-Mix.

Impact of fine-tuning strategy. Here we compare three different fine-tuning strategies: TI, DB, and the combined TI+DB. All strategies share the same fine-tuning hyperparameters and training steps (35000). To evaluate the distribution gap, we compute the FID score [20] of synthesized images (Diff-Gen) with the training set. As illustrated in Table 5, we observe that both TI+DB and TI have lower FID scores than DB. This can be attributed to the fact that semantic proximity impedes the convergence process. Additionally, while using TI alone results in a relatively low FID score, the improvements in performance are limited. This limitation stems from TI's inability to accurately reconstruct detailed concept (foreground) information, as it is primarily fine-tuned at the semantic level [28].

Annotation function. In this section, we discuss the impact of the choices of the translation strength s and the non-linear factor γ in Eq. 3. As shown in Table 6, we observe that as the translation strength decreases, the optimal value for γ also decreases, which underscores the non-linearity of Diff-Mix. The comparison between the 5-shot and all-shot settings indicates that the model tends to prefer a more diverse synthetic dataset when the number of training shots is small ($s = 0.9$ for 5-shot, $s = 0.7$ for all-shot). Besides, a larger confidence in the target class is preferred when the shot number is small ($\gamma = 0.1$ for 5-shot, $\gamma = 0.5$ for all-shot). A possible explanation is that the all-shot setting is less tolerant towards unrealistic images, as discussed in Section 6. Empirically, we recommend choosing a higher translation strength ($s \in 0.5, 0.7, 0.9$) and a smaller γ ($\gamma \in 0.1, 0.3, 0.5$) as a conservative option.

		Baseline	TI	DB	TI + DB
5-shot	FID (Diff-Gen)	-	**18.26**	19.55	18.43
	Acc. (Diff-Mix)	50.90	57.64	56.11	**59.41**
All-shot	FID (Diff-Gen)	-	14.13	14.64	**13.99**
	Acc. (Diff-Mix)	81.60	81.86	81.99	**82.85**

Table 5. Comparison of distribution gap and classification accuracy across three fine-tuning strategies. TI solely fine-tunes the identifier, and DB solely fine-tunes the U-Net, and TI+DB.

γ	5-shot			All-shot		
	$s = 0.5$	$s = 0.7$	$s = 0.9$	$s = 0.5$	$s = 0.7$	$s = 0.9$
1.5	-4.50	-0.31	+10.30	-1.08	+0.92	**+0.90**
1.0	-2.31	+2.99	+10.79	+0.25	+1.14	**+0.90**
0.5	+2.35	+8.44	+11.01	+0.92	**+1.30**	+0.86
0.3	+3.94	+9.41	**+11.15**	**+0.97**	+1.24	+0.69
0.1	**+6.18**	**+9.86**	+10.84	+0.50	+0.88	+0.84
0.0	+5.25	+9.41	+11.06	+0.38	+0.63	+0.54

Table 6. Comparison of performance gain across various γ and translation strength s. Lower γ indicates a higher confidence over target class, *e.g.* ($\gamma = 0.1, s = 0.7$) results in $0.04y_i + 0.96y_j$ and ($\gamma = 0.5, s = 0.7$) results in $0.16y_i + 0.84y_j$.

5. Conclusion

In this work, we investigate two pivotal aspects, faithfulness and diversity, that are critical for the current state-of-the-art text-to-image generative models to enhance image classification tasks. To achieve a more effective balance between these two aspects, we propose an inter-class augmentation strategy that leverages Stable Diffusion. This method enables generative models to produce a greater diversity of samples by editing images from other classes and shows consistent performance improvement across various classification tasks.

6. Acknowledgement

This research is mainly supported by the National Natural Science Foundation of China (92270114). This work is also partially supported by the National Key R&D Program of China under Grant 2021ZD0112801.

References

[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017. 2

[2] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 1, 2

[3] Haoyue Bai, Ceyuan Yang, Yinghao Xu, S-H Gary Chan, and Bolei Zhou. Improving out-of-distribution robustness of classifiers via generative interpolation. *arXiv preprint arXiv:2307.12219*, 2023. 2

[4] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *arXiv preprint arXiv:2302.02503*, 2023. 1, 2

[5] Christopher Beckham, Sina Honari, Vikas Verma, Alex M Lamb, Farnoosh Ghadiri, R Devon Hjelm, Yoshua Bengio, and Chris Pal. On adversarial mixup resynthesis. *Advances in neural information processing systems*, 32, 2019. 2

[6] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5968–5976, 2023. 2

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

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

[9] 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. 6

[10] 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. 7

[11] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000. 6

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[13] 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. 6

[14] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 1, 2, 5

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[18] 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. 6

[19] 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? *arXiv preprint arXiv:2210.07574*, 2022. 2, 6, 7, 3

[20] 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. 8

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[22] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 2

[26] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011. 6, 2

[27] 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. 6, 2

[28] 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, 3, 8

[29] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*, 2023. 2

[30] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat seng Chua. Invariant grounding for video question answering. In *CVPR*, 2022. 2

[31] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. *arXiv preprint arXiv:2305.08891*, 2023. 1

[32] 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. 6

[33] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 3

[34] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6, 2

[35] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[36] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019. 6

[37] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2

[38] 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. 6, 2

[39] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6887–6896, 2022. 6, 7, 3

[40] Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Conceptbed: Evaluating concept learning abilities of text-to-image diffusion models. *arXiv preprint arXiv:2306.04695*, 2023. 3

[41] Judea Pearl. Causal inference in statistics: An overview. 2009. 5

[42] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[43] 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, 1

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[46] 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, 3

[47] 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. 2

[48] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 7

[49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 1, 2, 6

[50] 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. 6

[51] Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific reports*, 9(1):16884, 2019. 2

[52] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR 2023–IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[53] 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. 2

[54] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Boosting zero-shot classification with synthetic data diversity via stable diffusion. *arXiv preprint arXiv:2302.03298*, 2023. 2

[55] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[56] Shikun Sun, Longhui Wei, Zhicai Wang, Zixuan Wang, Junliang Xing, Jia Jia, and Qi Tian. Inner classifier-free guidance and its taylor expansion for diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[57] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *arXiv preprint arXiv:2306.00984*, 2023. 1, 2

[58] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 2, 5, 6, 7, 3

[59] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 3

[60] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 3

[61] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5, 6, 2

[62] Zhicai Wang, Yanbin Hao, Tingting Mu, Ouxiang Li, Shuo Wang, and Xiangnan He. Bi-directional distribution alignment for transductive zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19893–19902, 2023. 2

[63] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2

[64] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 3

[65] Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, Nayeong Kim, Suha Kwak, and Tae-Hyun Oh. Exploiting synthetic data for data imbalance problems: Baselines from a data perspective. *arXiv preprint arXiv:2308.00994*, 2023. 6

[66] 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, 6, 7

[67] An Zhang, Wenchang Ma, Xiang Wang, and Tat seng Chua. Incorporating bias-aware margins into contrastive loss for collaborative filtering. In *NeurIPS*, 2022. 2

[68] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2, 6, 7

[69] Chenyu Zheng, Guoqiang Wu, and Chongxuan Li. Toward understanding generative data augmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[70] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 7

[71] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3

Enhance Image Classification via Inter-Class Image Mixup with Diffusion Model

Supplementary Material

(a) Filtered samples

(b) Preserved samples

Figure 8. Examples of (a) filtered samples with the lowest 10% confidence scores and (b) preserved samples in the CUB dataset.

7. Appendix

This appendix is organized as follows:

- In Sec. 7.1, we elaborate on the details of data cleaning design for Diff-Mix.
- In Sec. 7.2, additional visualizations are presented, including the visualization of attention maps and failure examples of complex datasets.
- In Sec. 7.3, few-shot classification results on a general dataset Pascal is provided [14].
- In Sec. 7.4 and Sec. 7.5, implementation details and latency considerations are presented, respectively.

7.1. Data-Cleaning Strategy

Due to the inherent differences in contour and size between the two classes, there is a higher risk of producing less realistic images during inter-class editing. We employ a simple data cleaning strategy that utilizes CLIP [43][4] as the filtering criterion. Specifically, we construct a positive caption, "a photo with a [metaclass] on it", and a negative caption, "a photo without a [metaclass] on it", and evaluate the synthetic data's confidence score towards the positive caption. We filter out the 10% of samples with the lowest confidence scores (we do not synthesize an additional 10% samples after data cleaning), and a subset of the filtered samples is displayed in Fig. 8. The preserved samples constitute the synthetic dataset that participates in the training process of downstream classification tasks.

[4] https://huggingface.co/openai/clip-vit-base-patch32

7.2. Visualizations

Visualizations of attention maps. In Section 3.4, we have shown that Diff-Mix can perform foreground editing while preserving most of the layout of the reference image. To support the claim, we provide evidence that SD can offer weak segmentation through textual conditions and achieve realistic foreground editing. We present visualizations of attention maps in Figure 9 for different datasets. The identifier, class descriptor (e.g., "`bird`", "`car`") and the "`<eot>`" token, which contains the global semantic information, tend to attend to the foreground part in the reference image. For example, the mentioned tokens primarily emphasize the bird rather than the tree branches (refer to Row 1 in the figure). This suggests that textual guidance, can offer a robust foreground prior, aiding in effective foreground editing at intermediate strengths. We posit that this characteristic ensures the generation of challenging samples where the foreground is replaced by the target concept when employing Diff-Mix for inter-class editing.

Visualizations on more datasets. In Fig. 12, we illustrate the editing process of Diff-Mix with varying strength $s \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$ across five datasets. It is worth noting that for strength $s = 1$, the translated images still exhibit a certain degree of similarity to the reference images. This phenomenon may be attributed to the last time-step not guaranteeing a zero signal-to-noise ratio, preserving the style and layout of the reference images [31]. Particularly, when the foreground is distinct against a simple background, Diff-Mix tends to generate high-quality interpolated images. We also observe that for more complex datasets, such as Stanford Dogs, where the foreground is less clear, and there are multiple concepts in a single image, unrealistic images tend to be generated, as seen in Fig. 13 (a) and (b). For general dataset Pascal [14], the dramatic differences in contour and size between two distinct classes lead to the generation of more unrealistic images (*e.g.*, "`bus`" $\longrightarrow$ "`cat`"), especially at intermediate strengths (*e.g.*, 0.7), as seen in Fig 13 (c) and (d).

Real-Gen versus Diff-Gen. To illustrate the distribution gap between domain-specific datasets and the pre-trained T2I model, as well as to demonstrate how fine-tuning can significantly mitigate this gap, we present a comparison in Fig. 14. It is worth noting that Real-Gen sometimes fails to generate correct concepts based on the terminology name of the target class (see "`photo of a chuck`

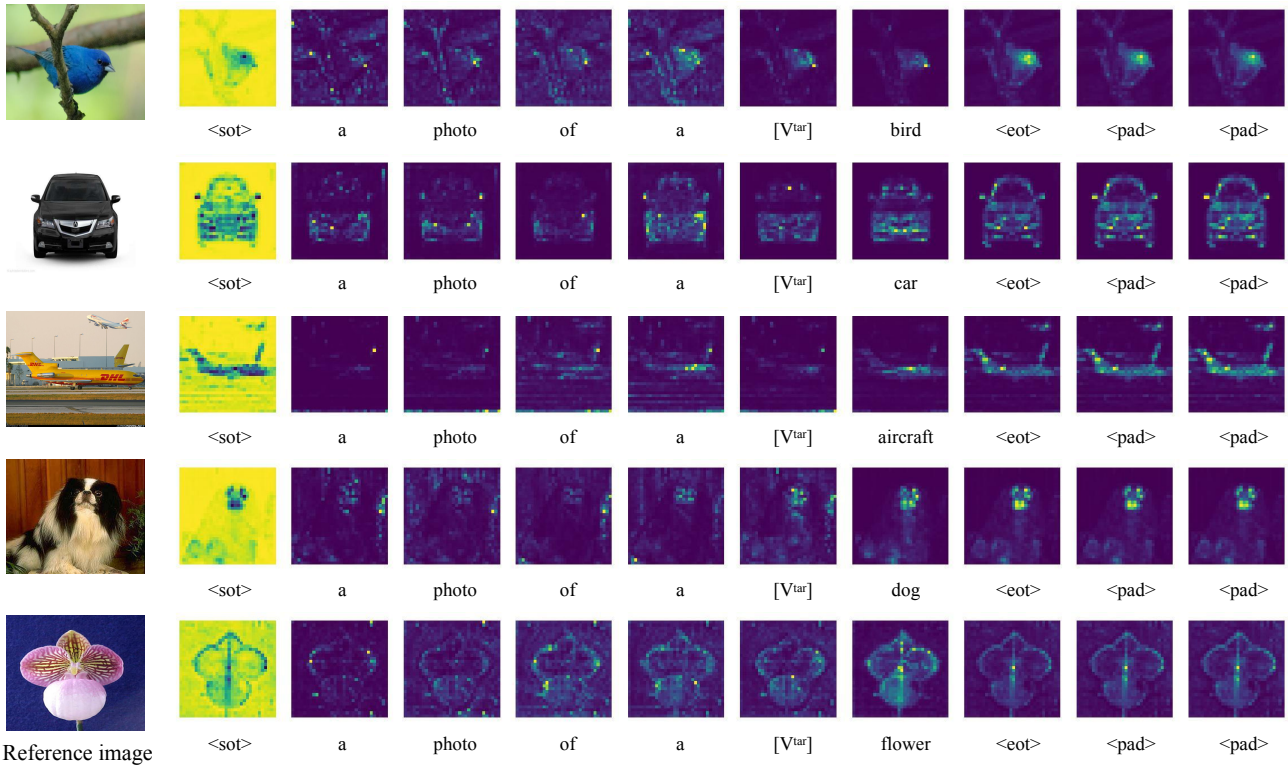


Figure 9. Visualizations of attention maps are shown in different rows for various datasets: CUB (Row 1) [61], Stanford Cars (Row 2) [27], FGVC Aircraft (Row 3) [34], Stanford Dogs (Row 4) [26], and Oxford Flowers (Row 5) [38]. These attention maps were generated during inter-class editing using Diff-Mix.

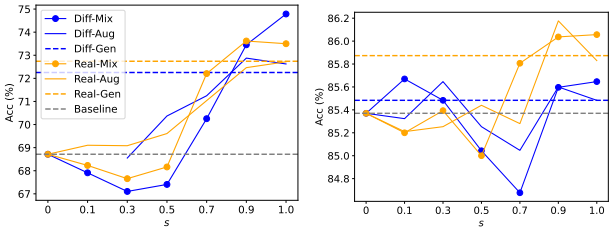
`will widow`" in panel (a)). Diff-Gen tends to generate more faithful outputs, it is noted that the majority of the generated images exhibit a similar layout and closely resemble the training samples. This resemblance is especially pronounced for those training samples characterized by a prominent foreground and a simple background.

Real-Mix versus Diff-Mix. In Fig. 15, we compare the generated samples between Real-Mix and Diff-Mix. We observe that, by conditioning on the reference image, Real-Mix accurately captures the semantic meaning of the terminology name (see "`chuck will widow`" in panel (a)). This feature of Real-Mix is consistent with its superior performance in few-shot classification, as depicted in Fig. 6. Additionally, we observe that Diff-Mix achieves more precise foreground editing (refer to Fig. 15 (c) and (d)). This enhanced accuracy can be attributed to the class descriptor maintaining its focus on the semantic content without being diverted to other extraneous information.

7.3. Experiments

Few-shot classification in Pascal.

In Sec. 7.2, We have demonstrated that inter-class editing for general datasets tends to produce unrealistic images due to the visual gaps between two classes. Here, we present 5-shot and all-shot classification results in Figure 10 for different expansion strategies on the general dataset Pascal [14]. Originally, Pascal is an object class recognition dataset containing 11,530 images and 6,929 object segmentation masks. We construct it into a classification dataset, following the setting of Da-fusion [58], resulting in a training split of 1,464 and a validation split of 1,449 for 20 general classes (*e.g.*, cat and boat). The main observation is that inter-class augmentation tends to be less effective for this general dataset, especially as the shot number increases (compare X-Aug and X-Mix in the figure). The effective-



(a) 5-shot (b) all-shot

Figure 10. 5-shot and all-shot classification results in Pascal.

hyperparameter	DB	TI	TI+DB
Base Model	Stable Diffusion-v1.5	Stable Diffusion-v1.5	Stable Diffusion-v1.5
Optimized	U-Net(LORA)	$[V^i]$	U-Net(LORA) + $[V^i]$
Optimization Steps	35000	35000	35000
Batchsize	8	8	8
Input Resolution	512× 512	512× 512	512× 512
Learning Rate	5e-5	5e-5	5e-5
Placeholder Token	-	$[V^i]$	$[V^i]$
LORA Rank	10	-	10
# if inference steps (T)	25	25	25
Guidance Scale	7.5	7.5	7.5
Noise Scheduler	DPMsolver++[33]	DPMsolver++[33]	DPMsolver++[33]

Table 7. Hyperparameters. This tables summarizes the hyperparameter settings of different fine-tuning strategies.

Dataset	# of classes	# of training	# of val.	Source
CUB	200	5994	5794	Huggingface.co
FGVC Aircraft	100	3334	3333	Huggingface.co
Oxford Flowers	102	4070	4119	Huggingface.co
Stanford Dogs	120	12000	8580	vision.stanford.edu
Stanford Cars	196	8144	8041	Huggingface.co

Table 8. Statistics of datasets.

ness of fine-tuning also decreases, with Real-Gen consistently outperforming Diff-Gen. This suggests that the pre-trained SD is capable of generating sufficiently diverse and faithful samples for these coarse concepts. Diff-Mix excels in handling domain-specific scenarios, where smaller differences in contour and layout between two classes are presented.

7.4. Implementation Details.

Diff-Mix. Diff-Mix comprises two stages: the fine-tuning stage and the sampling stage. The implementation details of Diff-Mix for three different fine-tuning strategies are depicted in Table 7. Note that our fine-tuning strategy heavily relies on the diffuser [60] repository. For DB and TI+DB, we only fine-tune the residual LORA matrices in attention modules in the U-Net. Please note that in the original Dreambooth [46] paper, an unlearnable identifier was introduced to represent user-specific concepts in concept learning. However, in our implementation, we have opted not to use the identifier and have implemented it as a straightforward fine-tuning of text-to-image models. All fine-tuning and sampling processes are conduct on 4 RTX3090 GPUs.

Datasets. We list the statistics of the datasets involved in our experiments in Table 8.

Conventional classification. The hyperparameter settings for the CUB dataset are presented in Table 9. To reproduce Real-filtering (RF) [19], a subset is derived from Real-Gen through data cleaning, as detailed in Section 7.1. Real-guidance (RG) [19] augments the dataset with low-strength intra-class editing, akin to Real-Aug with a strength parameter $s = 0.1$. For the replication of Da-fusion [58], we fine-tune the synthetic data (SD) using our TI strategy over 35,000 steps, with translation strengths randomly selected from the set $0.25, 0.5, 0.75, 1.0$. For CutMix and Mixup, the weight decay is 1×10^{-5}, and the mixup ratios are set to 0.1 and 0.3, respectively.

hyperparameters	ResNet50	ViT-B/16
Source	torchvision	torchvision
# of parameters	25.5M	86.6M
Pre-trained	ImageNet1K	ImageNet21K
Fine-tuned	-	ImageNet1K
Input Resolution	448×448	384×384
Batchsize	64	32
Epochs	128	100
Optimizer	SGD	SGD
Learning Rate	0.02	0.001
Momentum	0.9	0.9
Weight Decay	5e-5	5e-5
Label Smoothing	0.9	0.9

Table 9. hyperparameters. This table summarizes the hyperparameter settings for CUB using two visual backbones in our conventional classification task.

Few-shot classification. The few-shot classification is conducted on CUB with varying shot numbers: 1, 5, 10, and all. The comparison methods encompass: (1) inter-class augmentation strategies, namely Diff-Mix and Real-Mix, (2) intra-class augmentation strategies, namely Diff-Aug and Real-Aug, and (3) distillation-based methods, Diff-Gen and Real-Gen. The backbone model used is ResNet50 with an input resolution of 224^2. We employ the same hyperparameters as in the conventional setting, as detailed in Table 9, albeit with a larger batch size (256) and a higher learning rate (0.05). All experiments are conducted with three trials, and the average results are reported.

Long-tail classification. Thanks to the authors of CMO [39], the reproduced long-tail results are built upon its open-

Dataset	# of classes	# of training	Imbalance Factor (IF)
CUB-LT	200	{1242, 1798, 2238}	{100,50,10}
Flower-LT	102	{847, 1238, 1532}	{100,50,10}

Table 10. Statistics of long-tail datasets CUB-LT and Flower-LT.

source git repository [5]. To construct the imbalanced dataset, an imbalance factor is introduced to control the imbalance level. The imbalance factor ρ is defined as $\rho = \frac{\max_k\{n_k\}}{\min_k\{n_k\}}$, where n_k is the number of samples in the k-th class. Specifically, given a normal dataset, we first sort the classes based on the number of images within classes in descending order, and use k' to denote the sorted class index. A subset of images is randomly sampled from each class to achieve the desired imbalance, ensuring that the number of images for each class corresponds to the calculated target. The number of sampled images is determined by,

$$n_{k'} = \max\left(\bar{n} \times \left(\frac{1}{\rho}\right)^{\frac{k}{N-1}}, 1\right) \tag{4}$$

where $\bar{n}$ is the averaged number of images for each class, and N is the total number of classes. The statistics of constructed CUB-LT and Flower-LT datasets can be found in Table 10. To uniformize the distribution of imbalanced real data, we first fix the number of iterative samples within each epoch and replace real samples with synthetic data with a 50% probability. Note that the synthetic data conforms to the distribution specified by Eq. 4 with reversed class indices, thereby generating more synthetic images for tail classes. We maintain a constant number of training epochs and learning rate for both synthesis-free and synthesis-based approaches to ensure a fair comparison. We further present accuracy results for three distinct subsets when IF is 100: Many-shot classes (classes with over 20/30 training samples for CUB-LT/Flower-LT), medium-shot classes (classes with 5-20/10-30 samples for CUB-LT and Flower-LT), and few-shot classes (classes with fewer than 5/10 samples for CUB-LT/Flower-LT).

7.5. Latency

Compared to non-generative augmentation methods, Diff-Mix's implementation incurs additional computational overhead during fine-tuning and data sampling. For instance, when working with the CUB dataset, which contains approximately 6,000 training samples, the fine-tuning process is completed in about 3 hours. This duration is achieved using an input resolution of 512×512 on 4 NVIDIA RTX 3090 GPUs with a total batch size of 8. During sampling, synthetic samples are generated at the same resolution with a total of $T = 25$ reverse steps. The throughput across various translation strengths is evaluated

[5] https://github.com/naver-ai/cmo

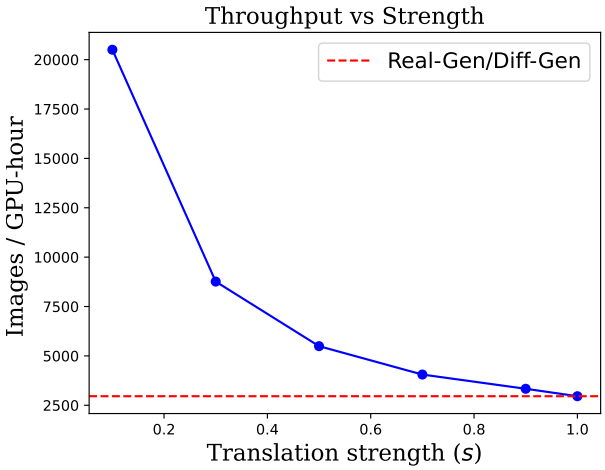


Figure 11. Sampling throughput across various translation strengths in a single RTX 3090 GPU.

Method	Translation strength s	Images per GPU-hour
Real-filtering	$\in\{1.0\}$	2,957
Real-guidance	$\in\{0.1\}$	20,502
Da-fusion	$\in\{0.25,0.5,0.75,1.0\}$	4,952
Diff-Mix	$\in\{0.5,0.7,0.9\}$	4,179

Table 11. Comparison of sampling throughput of different expansion strategies.

in Fig. 11, and a throughput comparison with other synthesis strategies is provided in Table 11. While Diff-Mix is more efficient than generating data from scratch, it is less so than low-strength editing (*e.g.*, Real-guidance). For generating synthetic data for the CUB dataset with a 5x multiplier (resulting in approximately 30,000 images), the process requires roughly 2.5 hours using 4 NVIDIA RTX 3090 GPUs.

8. Limitations

Our inter-class augmentation method shows less effective when applied to general datasets that encompass a broad spectrum of concepts. We are optimistic, however, that integrating an image inpainting strategy or confining Diff-Mix to operate among adjacent classes could address this limitation. Moreover, the current annotation strategy is determined empirically and lacks a robust theoretical foundation, which may limit the generalizability of the strategy.

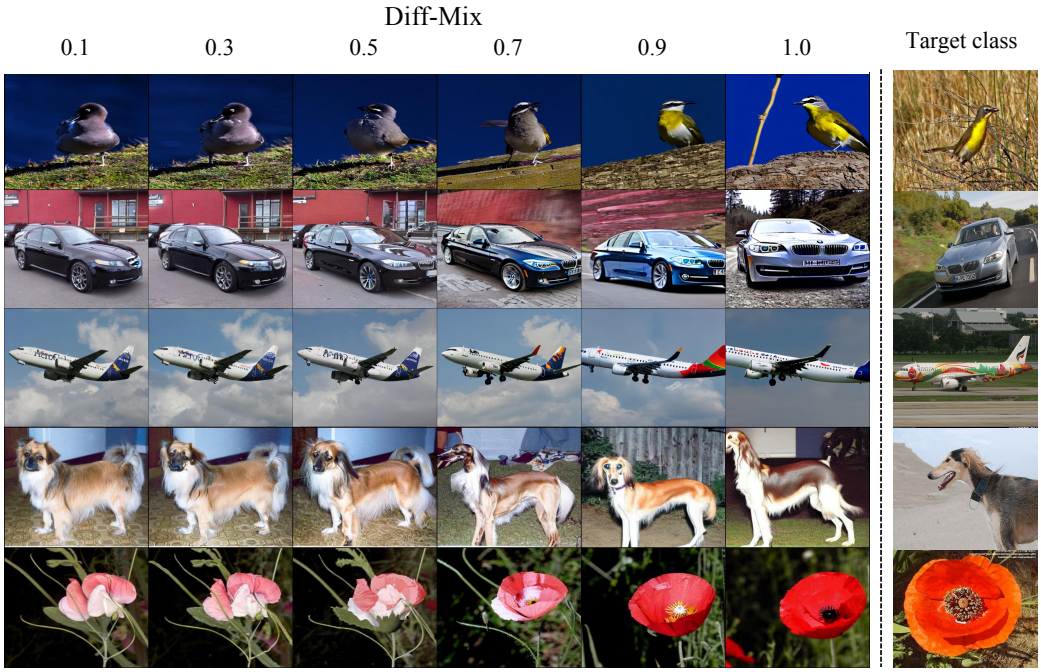


Figure 12. Examples of image generated using Diff-Mix with varying translation strengths in CUB (Row 1), Stanford Cars (Row 2), FGVC Aircraft (Row 3), Stanford Dogs (Row 4) , Oxford Flowers (Row 5).

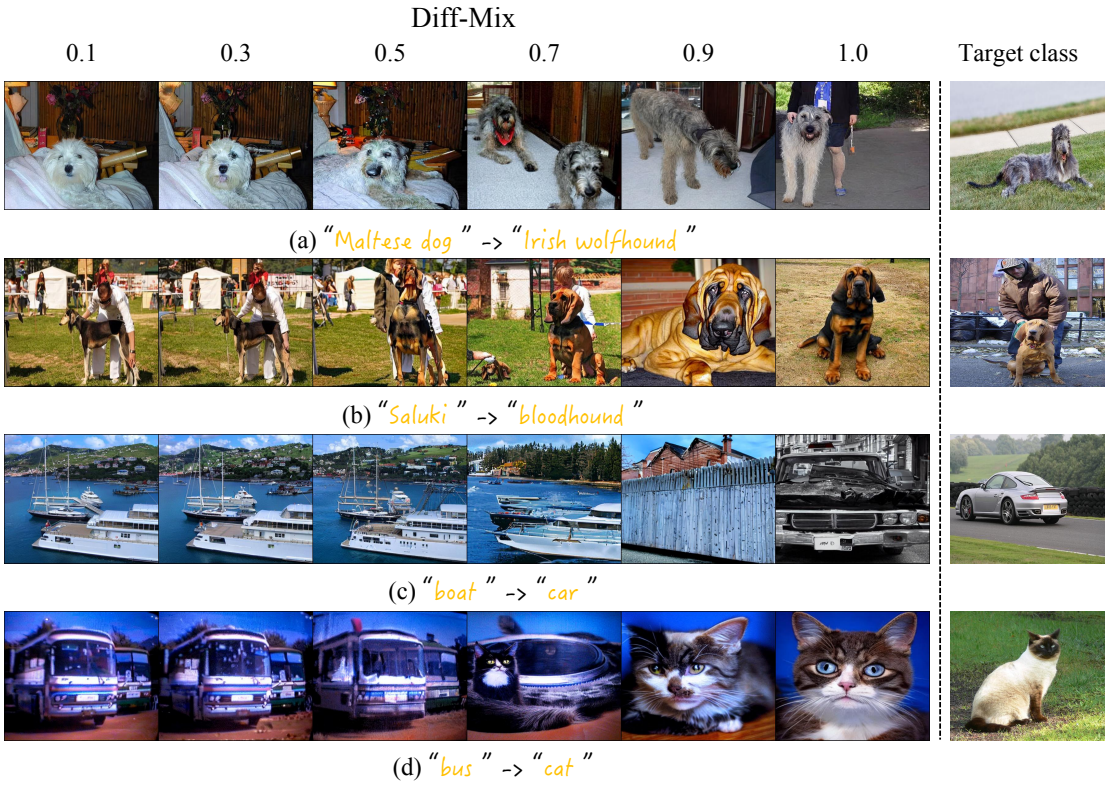


Figure 13. Failure examples generated using Diff-Mix with varying translation strengths are shown in panels (a) and (b) for the complex dataset Stanford Dogs, and in panels (c) and (d) for the general dataset Pascal [14].

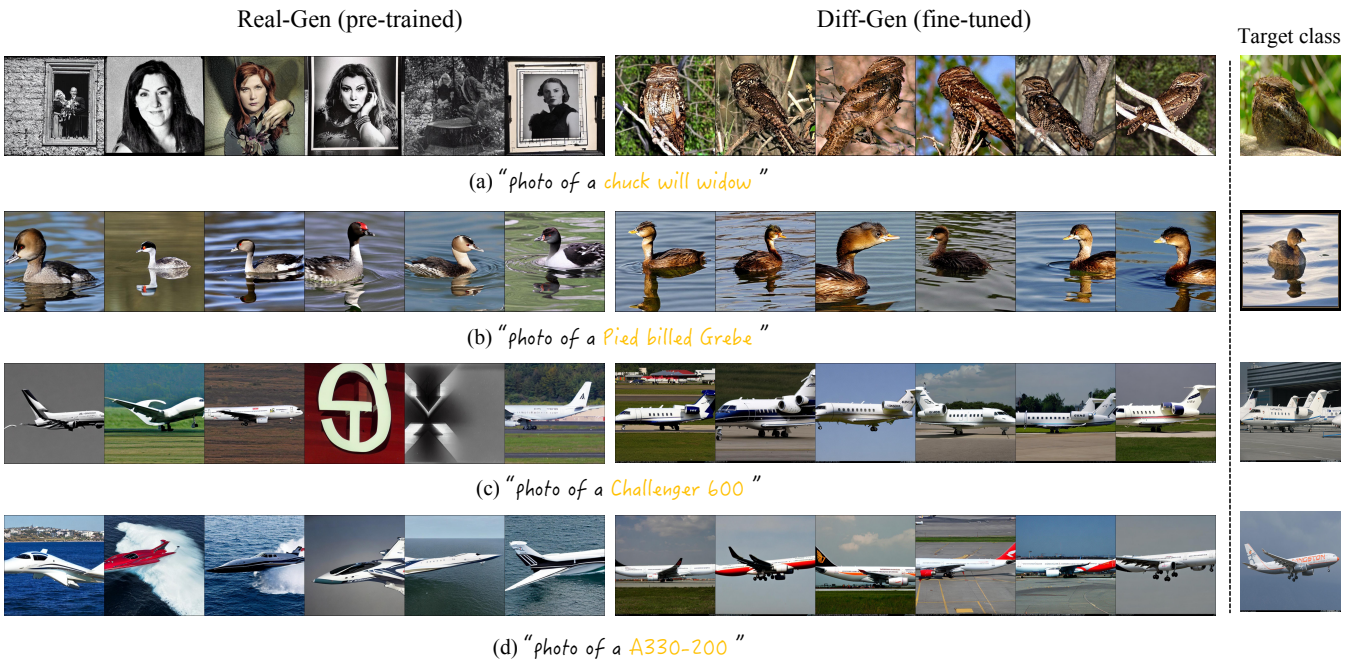


Figure 14. Examples of image generated using Real-Gen and Diff-Gen. The prompts are formatted as "`photo of a [terminology name]`" for Real-Gen and "`photo of a [V^i] [metaclass]`" for Diff-Gen. Panels (a) and (b) depict the samples of CUB dataset, while panels (c) and (d) depict the samples of FGVC Aircraft dataset.

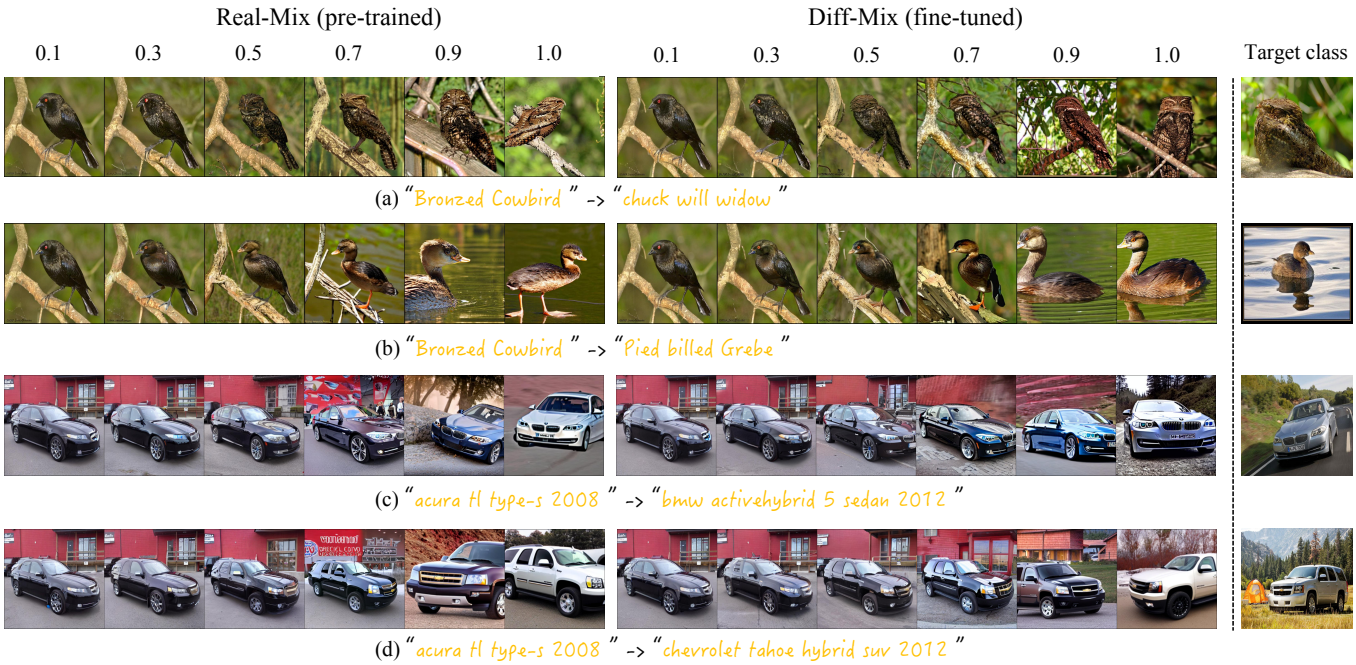


Figure 15. Examples of image generated using Real-Mix and Diff-Mix with varying translation strengths. The prompts are formatted as "`photo of a [terminology name]`" for Real-Mix and "`photo of a [V^i] [metaclass]`" for Diff-Mix. Panels (a) and (b) depict the samples of CUB dataset, while panels (c) and (d) depict the samples of Stanford Car dataset.