DIFFUSEMIX: Label-Preserving Data Augmentation with Diffusion Models

Khawar Islam[1] Muhammad Zaigham Zaheer[2] Arif Mahmood[3] Karthik Nandakumar[2]
[1]FloppyDisk.AI [2]Mohamed bin Zayed University of Artificial Intelligence [3]Information Technology University, Punjab
[1]khawarr.islam@gmail.com [2]{zaigham.zaheer, karthik.nandakumar}@mbzuai.ac.ae [3]arif.mahmood@itu.edu.pk

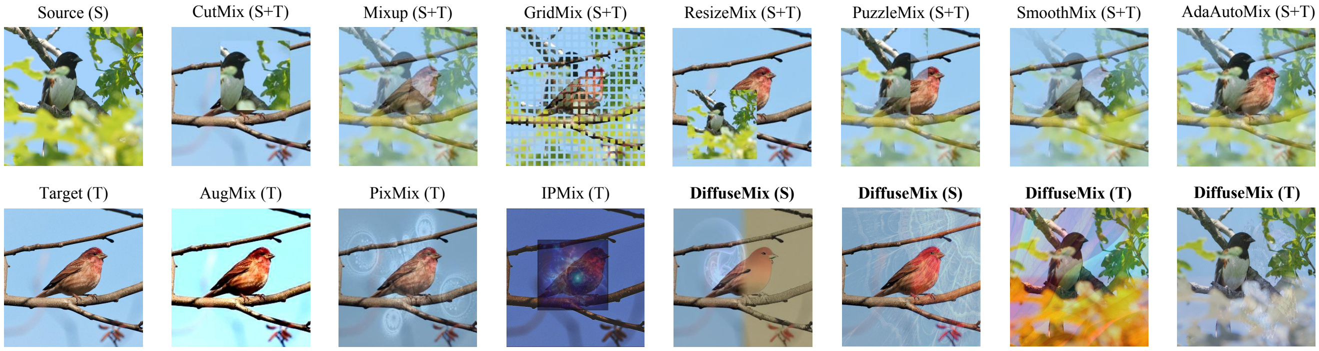


Figure 1. **Top row:** existing mixup methods *interpolate* two different training images [22, 48]. **Bottom row:** label-preserving methods. For each input image, DIFFUSEMIX employs *conditional prompts* to obtain generated images. The input image is then concatenated with a generated image to obtain a hybrid image. Each hybrid image is blended with a random fractal to obtain the final training image.

Abstract

Recently, a number of image-mixing-based augmentation techniques have been introduced to improve the generalization of deep neural networks. In these techniques, two or more randomly selected natural images are mixed together to generate an augmented image. Such methods may not only omit important portions of the input images but also introduce label ambiguities by mixing images across labels resulting in misleading supervisory signals. To address these limitations, we propose DIFFUSEMIX*, a novel data augmentation technique that leverages a diffusion model to reshape training images, supervised by our bespoke conditional prompts. First, concatenation of a partial natural image and its generated counterpart is obtained which helps in avoiding the generation of unrealistic images or label ambiguities. Then, to enhance resilience against adversarial attacks and improves safety measures, a randomly selected structural pattern from a set of fractal images is blended into the concatenated image to form the final augmented image for training. Our empirical results on seven different datasets reveal that* DIFFUSEMIX *achieves superior performance compared to existing state-of-the-art methods on tasks including general classification, fine-grained classification, fine-tuning, data scarcity, and adversarial robustness. Augmented datasets and codes are available here: https://diffusemix.github.io/*

1. Introduction

In the era of deep learning, image-mixing-based data augmentation techniques stand out for their simplicity and effectiveness in addressing the generalization of learning models toward testing scenarios [3, 6, 19, 21–23, 23, 34, 40, 46, 50]. These techniques ingeniously mix randomly selected natural images and their respective labels from the training dataset using a number of *mixing* combinations to synthesize new augmented images and labels. Such a process often implies linear interpolation of data, resulting in the generation of novel training images. These approaches have been proven to be effective in improving the performance of deep models [13, 19, 36, 45, 48],.

However, these techniques may face a number of challenges such as the omission of salient image regions (Figure 1) and label ambiguities due to random placements of images [23]. A few researchers have attempted to alleviate these issues by introducing saliency-based mixup strategies in which important regions of one image are pasted onto the less important portions (mainly context) of another image [21, 23, 40]. These methods not only suffer from the costs but also the shortcomings of saliency detection methods. Moreover, as these methods still rely on mixing of two or more images belonging to different classes, the underlying issue of omitting the important context still persists.

Recently, Diffusion Models [11, 12, 31, 37, 38] have emerged as transformative approaches offering image-to-

Table 1. Comparison of different image mixing techniques: most methods utilize natural images as source and target except [41] using hidden state. DIFFUSEMIX uses a *generated* image produced by a diffusion model leveraging *conditional prompts* and a fractal image for augmentation.

	Input	Mixup [48]	ManifoldMixup [41]	CutMix [45]	SaliencyMix [40]	StyleMix [18]	PuzzleMix [23]	CoMixup [22]	PixMix [17]	GuidMixup [21]	**DIFFUSEMIX**
Components	Source image	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Target image	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗
	Fractal image	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓
	Textual Prompts	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
	Interpolation	✓	✓	✗	✗	✓	✗	✗	✓	✓	✓
	Concatenation	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓
Tasks	Adversarial Robustness	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓
	General Classification	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Fine Grained	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓
	Transfer Learning	✗	✗	✓	✓	✗	✗	✗	✗	✗	✓
	Data Scarcity	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓

image generation and editing processes. Although the idea of using images generated by diffusion models directly as augmented images for training a classifier has been studied by some researchers [1, 39], this way of data augmentation does not result in significant performance gains. In fact, as reported in [1], the model trained using generated images directly as augmentation may even result in lower performance than the baseline trained without any augmentation. The underlying problem can be attributed to the limited control that these diffusion models offer over generated images. Owing to the sensitivity of diffusion models to conditional prompts, generation of desired complex *scenes*, *layouts*, and *shapes* in an image is a cumbersome task [49]. Thus, poorly constructed prompts pose the risk of producing images that may not be suitable for data augmentation as the generated images may deviate drastically from the actual data distribution. Training on such images may result in overfitting of the learning model on wrong data distribution, consequently resulting in performance degradation. Therefore, careful selection of prompts for data augmentation needs further investigation. Moreover, to mitigate the risk of poorly generated images affecting the overall training, a more efficient way of utilizing the generated images is necessary.

To this end, we propose a novel data augmentation method, DIFFUSEMIX, that leverages the capabilities of a Stable Diffusion model to generate diverse samples based on our tailored conditional prompts. In contrast to Trabucco et al. [39], rather than solely relying on Stable Diffusion for augmentation, we propose an effective approach which utilizes both original and generated images to create hybrid images. This way, visual diversity obtained by the diffusion models is *infused* with the original images while retaining the key semantics. In addition, to increase overall structural diversity, we blend self-similarity-fractals with the hybrid images to create the final training images. This blending has previously been found useful for ML safety measures [17, 20] while in our approach, this added diversity helps in avoiding overfitting on the generated contents resulting in performance improvements. Our experimental results show that DIFFUSEMIX benchmarks better generalization as well as increased adversarial robustness compared to the existing state-of-the-art (SOTA) augmentation methods. Moreover, it offers compatibility with a broad spectrum of datasets and can be incorporated into the training of various existing architectures. Some notable aspects of this research work are as follows:

- We introduce a new data augmentation method driven by a diffusion model, which generates diverse images via our bespoke conditional prompts.
- We propose to *concatenate a portion of the natural image with its generative counterpart* to obtain hybrid images. The combination brings richer visual appearances while preserving key semantics.
- We collect a fractal image dataset and blend into the hybrid images. This improves the overall structural complexity of the augmented images and helps to avoid overfitting on generated images, thus resulting in better generalization.
- Extensive experiments on seven datasets for various tasks including *general classification, fine-grained classification, adversarial robustness, transfer learning, and data scarcity* demonstrate the superior performance of our proposed method compared to the existing SOTA image augmentation techniques.

2. Related Work

Data augmentation has become indispensable in enhancing the diversity of training datasets, thereby mitigating the risks of overfitting. Traditional approaches employed strategies—such as horizontal and vertical translations, affine transformations, scaling, and squeezing—when training a model. This not only improves the performance but also improves the generalization of the model on test datasets.

Diffusion Models for Augmentation: Recently, several re-

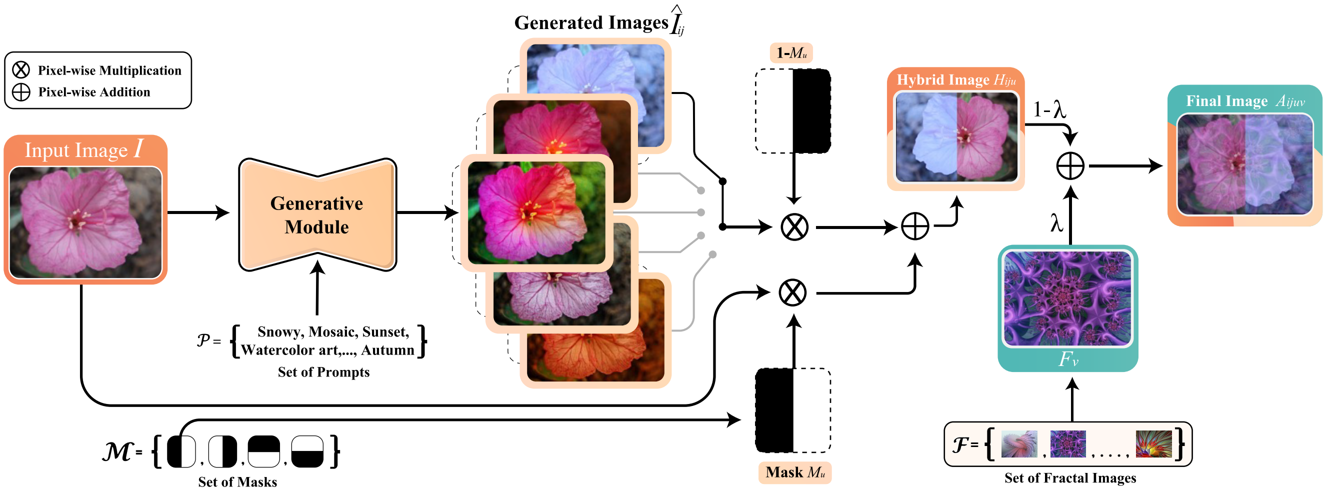


Figure 2. **Architecture of the proposed DiffuseMix approach**. An input image and a randomly selected prompt are input to a diffusion model to obtain a generated image. Input and generated images are concatenated using a binary mask to obtain a hybrid image. A random fractal image is finally blended with this hybrid image to obtain the augmented image.

searchers have explored the possibility of data augmentation with diffusion models. Azizi et al. [1] proposed the utilization of fine-tuned text-to-image diffusion models on ImageNet classification, revealing that augmenting the training set with these synthetic samples may boost classification performance. Similarly, Trabucco et al. [39] investigated diffusion models to create more diverse and semantically varied datasets, aiming to improve outcomes in tasks such as image classification. Li et al. [28] further explored diffusion models-based augmentation for knowledge distillation without real images.

Image Mixing Augmentation: Image mixing is a prominent class of augmentation methods for training robust CNN models [4, 29, 33, 44]. Some of these methods include Mixup, CutMix, and AugMix. Mixup [47] generates synthetic images by linearly interpolating pixel values from two randomly selected images. In contrast, CutMix [46] involves pasting a random patch from one image onto another. AugMix [15] employs a stochastic combination of data augmentation operations on an input image. SaliencyMix [40] utilizes saliency maps to concentrate the augmentation on the image's most vital regions, ensuring overall image integrity. Manifold Mixup [41] enhances representation by interpolating network hidden states during training. This entails blending two hidden states with a random weight to produce an interpolated manifold-based hidden state. PuzzleMix [23], an improvement over the traditional mixup, factors in image saliency, and local statistics during image blending. This method segments an image into patches, allocates weights based on saliency and local statistics, and merges patches from different images in accordance with their weights. PixMix [17] have studied mixing of input images with fractal and feature visualization images to improve ML safety measures. A detailed summary of several image-mixing-based methods along with their components and application tasks is provided in Table 1.

Automated Augmentation: AutoAugment [7], for instance, employed reinforcement learning to pinpoint optimal data augmentation policies, while RandAugment [9] integrates a suite of random data augmentation operations to improve model generalization. AdaAug[5] is proposed to efficiently learn adaptive augmentation policies in a class-dependent and potentially instance-dependent manner.

In contrast to previous methods, our approach emphasizes the concatenation of original and generated images, using a pre-defined library of conditional prompts. The obtained hybrid images are blended with fractal images to further improve the overall performance.

3. DiffuseMix

3.1. Background and Overview

Existing image-mixing-based methods may induce label ambiguity by placing one image on top of the other and consequently overlapping either some portions of the object or its context [21, 45]. In contrast, the core idea of DiffuseMix is to concatenate a portion of the original image with its counterpart generated image in such a way that basic image semantics are preserved while providing diverse object details and contexts for better augmentation.

The proposed method as illustrated in Figure 2 comprises of three pivotal steps: ***generation***, ***concatenation***, and ***fractal blending***. Firstly, conditional prompts are used with a diffusion model to obtain a generative counterpart of the input image. Then, a portion of the original image is concatenated with the rest of the portion taken from the generated image forming a hybrid image. This step is to ensure that the training network always has access to the

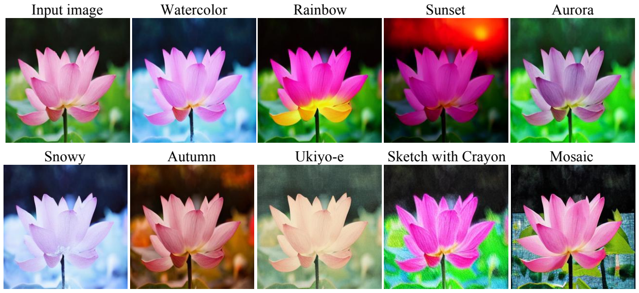


Figure 3. A set of *bespoke conditional prompts* are used to obtain generated images preserving important features and adding rich visual appearance to the input images.

original data along with the generated one. Subsequently, a random fractal image is blended into the hybrid image to obtain the final training image with a diverse structure. Blending fractal images has proven to be effective towards ML safety [16, 17]. In our work, we study the effectiveness of blending fractal images mainly towards improved performance.

3.2. Method

The proposed DIFFUSEMIX is an effective data augmentation technique which can be used to enhance the robustness and generalization of the deep learning models. Formally, $I_i \in \mathbb{R}^{h\times w\times c}$ is an image from the training dataset, $\mathcal{D}_{\text{mix}}(\cdot) : \mathbb{R}^{h\times w\times c} \rightarrow \mathbb{R}^{h\times w\times c}$ denotes our data augmentation method. To obtain the final augmented image A_{ijuv}, input image I_i goes through proposed generation using prompt p_j, concatenation using mask M_u, and blending using fractal image F_v. The overall augmentation process, as also seen in Algorithm 1, can be represented as $A_{ijuv} = \mathcal{D}_{\text{mix}}(I_i, p_j, M_u, F_v, \lambda)$.

Generation: Our generation step $\mathcal{G}(.)$ consists of a *pre-trained* diffusion model that takes a prompt p_j from a predefined set of k prompts, $P = \{p_1, p_2, \dots, p_k\}$ where $j \in [1, k]$ along with the input image I_i and produces an augmented counterpart image $\hat{I}_{ij}$. Image editing process in conventional diffusion models is often open-ended and guided by text prompts to obtain diverse image-to-image or text-to-image translations. In our case, as the goal is to achieve a slightly modified but not too different version of I_i, *filter-like* prompts are curated in $\mathcal{P}$ which do not alter the image drastically. Examples of the prompts used in DIFFUSEMIX are shown in Figure 3. The overall generation step can be represented as: $\hat{I}_{ij} = \mathcal{G}(I_i, p_j)$, where p_j is a randomly selected prompt.

Concatenation: We concatenate a portion of the original input image I_i with its counterpart generated image $\hat{I}_{ij}$ using a randomly selected mask M_u from the set of masks to create a hybrid image H_{iju}:

$$H_{iju} = (\hat{I}_{ij} \odot M_u) + (I_i \odot (\mathbf{1} - M_u)). \tag{1}$$

Algorithm 1 DIFFUSEMIX

Require: $I_i \in \mathcal{D}$ training images dataset, m: number of augmented images, $p_j \in \mathcal{P}$ set of prompts, $M_u \in \mathcal{M}$ set of masks, $F_v \in \mathcal{F}$ library of fractal images, λ: blend ratio

Ensure: $\mathcal{D}'$: m Augmented images

```
$\mathcal{D}' \leftarrow \emptyset$
for each image $I_i$ in $\mathcal{D}$ do
    for $a$ in $\{1 : m\}$ do
        Randomly select prompt $p_j$ from $\mathcal{P}$
        Generate image: $\hat{I}_{ij} \leftarrow \mathcal{G}(I_i, p_j)$
        Randomly select mask $M_u$ from $\mathcal{M}$
        Hybrid image: $H_{iju} \leftarrow M_u \odot I_i + (1 - M_u) \odot \hat{I}_{ij}$
        Randomly select $F_v$ from $\mathcal{F}$
        Blended image: $A_{ijuv} \leftarrow (1 - \lambda)H_{iju} + \lambda F_v$
        Add $A_{ijuv}$ to $\mathcal{D}'$
    end for
end for
return $D'$
```

The mask M_u consists of zeros and ones only and $\odot$ is a pixel-wise multiplication operator. The set of masks contains four kinds of masks including horizontal, vertical and flipped versions. Such masking ensures the availability of the semantics of the input image to the learning network while reaping the benefits of the generated images.

Fractal Blending: A fractal image dataset[*] $\mathcal{F}$ is collected and used for inducing structural variations in the hybrid images. A randomly selected fractal image $F_v \in \mathcal{F}$ is blended to the hybrid image H_{iju} with a blending factor λ as:

$$A_{ijuv} = \lambda F_v + (1 - \lambda) H_{iju}, \tag{2}$$

where λ is the blending factor. This results in the final augmented image A_{ijuv} used to train or fine-tune a deep learning model. The overall augmentation process of DIFFUSEMIX can be represented as:

$$A_{ijuv} = (1 - \lambda)(I_i \odot M_u + \hat{I}_{ij} \odot (\mathbf{1} - M_u)) + \lambda F_v, \tag{3}$$

4. Experiments and Results

In this section, we present the experimental details, datasets used to evaluate our approach, and analyses of the results.

Datasets. To provide comparisons with existing studies on image augmentation [5, 8, 9, 15, 19, 21–23, 30, 40, 41, 45, 48], we evaluate our approach on several *general image classification* and *fine-grained image classification* datasets. In the general image classification category, we employ three datasets including ImageNet [10], CIFAR100 [25] and

[*] Examples of fractal images are provided in Appendix 8.

Tiny-ImageNet-200 [26]. In fine-grained image classification category, we employ four datasets including Oxford-102 Flower [35], Stanford Cars [24], Aircraft [32], and Caltech-UCSD Birds-200-2011 (CUB) [42]. These datasets offer a diverse array of scenarios where images contain a wide range of objects such as plants and animals in various scenes, textures, transportation modes, human actions, satellite imagery, and general objects.

Implementation Details. We utilize InstructPix2Pix [2] diffusion model to generate images with the help of our introduced textual library. For the generation of Mask M in Eq. 1, a template image is divided into two equal parts, either horizontally or vertically. Randomly, one half is turned on and the second half is turned off. In all experiments [†], $\lambda = 0.20$ is used for blending the fractal image in Eq. (2) & (3). Analysis on λ values is provided in Appendix 1.

Textual Prompt Selection. In order to ensure that only appropriate prompts are applied, a bespoke textual library of *filter-like* global visual effects is predefined: *'autumn', 'snowy', 'sunset', 'watercolor art', 'rainbow', 'aurora', 'mosaic', 'ukiyo-e', and 'a sketch with crayon'*. These prompts are selected because of their generic nature and applicability to a wide variety of images. Secondly, these do not alter the image structure significantly while producing a global visual effect in the image. Each prompt in the textual library is appended with a template 'A transformed version of image into *prompt*' to form a particular input to the diffusion model. Examples of images generated through these prompts are shown in Figure 3. More visual examples and discussions on appropriate prompt selection are provided in Appendix 3.

Visualizing Intermediate Steps. Figure 4 depicts images obtained in each step of DiffuseMix. It can be observed that DiffuseMix yields a broader spectrum of augmented images derived from the training set. These images contain full object with no portions omitted and provide suitable variations for training.

5. Performance Evaluation

5.1. General Classification

General Classification (GC) can be a notable way to quantify the effectiveness of an image augmentation method. A higher GC accuracy would mean that an augmentation method successfully provided plausible data variations to improve the learning process. We evaluate our approach for the GC task on three challenging datasets including Tiny-ImageNet-200, CIFAR-100, and ImageNet. Similar to the existing SOTA methods, PreActResNet18 network is employed for the experiments on Tiny-ImageNet-200 and CIFAR-100 datasets [19, 21, 22, 40], while ResNet50 is employed for ImageNet dataset [15, 23, 45].

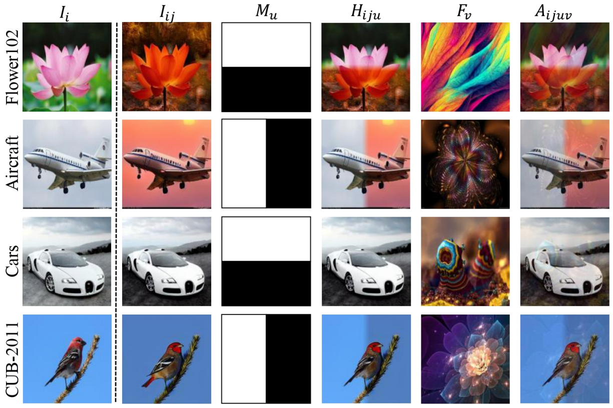


Figure 4. Example images from different stages of DiffuseMix: input image (I_i), generated image ($\hat{I}_{ij}$), mask (M_u), hybrid image (H_{iju}), fractal image (F_v), and final augmented image (A_{ijuv}).

In Table 2 Top-1 and Top-5 accuracy on Tiny-ImageNet and CIFAR-100 datasets is compared with the existing SOTA methods. The proposed DiffuseMix technique demonstrates better performance gains compared to the existing image augmentation approaches. On Tiny-ImageNet dataset, compared to Vanilla model, our DiffuseMix results in notable Top-1 and Top-5 accuracy gains of 8.54% and 10.01%. Moreover, compared to the second best performer, Guided-AP [21], Top-1 and Top-5 accuracy gains of our approach are 1.14% and 1.17%. Similar trends are observed on CIFAR100 dataset, where DiffuseMix demonstrates Top-1 and Top-5 accuracy gains of 6.17% and 4.39% over vanilla and 1.3% and 0.53% over Guided-AP.

We also evaluate DiffuseMix on large-scale ImageNet dataset offering more challenging scenarios where existing SOTA methods [15, 23, 45] only report Top-1 accuracy. Compared to the second best performer PuzzleMix [23], our approach demonstrates a performance gain of 1.13% (Table 3). Compared to Vanilla implementation, our approach demonstrates a performance gain of 2.95%. The GC results on these challenging and diverse benchmark datasets highlight the effectiveness of DiffuseMix in enabling better learning. It also suggests its capability to combat overfitting and achieve better generalization.

5.2. Adversarial Robustness

Following existing SOTA methods [15, 18, 21–23, 40, 41, 45], we evaluate the robustness of our approach against adversarial attacks and input perturbations. In these experiments, fast adversarial training [43] is adapted to create adversarially perturbed input images. The goal of these experiments is to evaluate whether an augmentation approach can demonstrate better resilience against adversarial attacks. FGSM [43] error rates are computed to evaluate the performance against adversarial attacks.

[†] Following AugMix [15], we employ JS-divergence during training.

Table 2. Top-1 and Top-5 accuracy on ***general classification task*** of PreactResNet-18 trained from scratch for 300 epochs following the results of Kang and Kim [21]. Extended table can be seen in Appendix 7 Table 12.

Method	Tiny-ImageNet-200		CIFAR-100	
	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
Vanilla(CVPR'16) [14]	57.23	73.65	76.33	91.02
SaliencyMix(ICLR'21) [40]	56.54	76.14	79.75	94.71
Guided-SR(AAAI'23) [21]	55.97	74.68	80.60	94.00
PuzzleMix(ICML'20) [23]	63.48	75.52	80.38	94.15
Co-Mixup(ICLR'21) [22]	64.15	-	80.15	-
Guided-AP(AAAI'23) [21]	64.63	82.49	81.20	94.88
DIFFUSEMIX	**65.77**	**83.66**	**82.50**	**95.41**

Table 3. Top-1 / Top-5 performance on ImageNet-1K dataset benchmark when trained on ResNet-50 for 100 epochs for ***general classification task***. An extended version of this table is provided in Appendix 7 Table 13.

Method	Top-1 (%)	Top-5 (%)
Vanilla(CVPR'16) [14]	75.97	92.66
PixMix(CVPR'22) [17]	77.40	-
PuzzleMix(ICML'20) [23]	77.51	93.76
GuidedMixup(AAAI'23) [21]	77.53	93.86
Co-Mixup (ICLR'21) [22]	77.63	93.84
YOCO(ICML'22) [13]	77.88	-
DIFFUSEMIX	**78.64**	**95.32**

Table 4. FGSM error rates on CIFAR-100 and Tiny-ImageNet-200 datasets for PreactResNet-18, following [23].

Method	FGSM Error Rates (%)	
	CIFAR-100	Tiny-ImageNet-200
Vanilla(CVPR'16) [14]	23.67	42.77
Mixup (ICLR'18) [48]	23.16	43.41
Manifold (ICML'19) [41]	20.98	41.99
CutMix (ICCV'19) [45]	23.20	43.33
AugMix (ICLR'20) [15]	43.33	-
PuzzleMix(ICML'20) [23]	19.62	36.52
DIFFUSEMIX	**17.38**	**34.53**

As shown in Table 4, DIFFUSEMIX demonstrates an error rate of 17.38% on CIFAR-100, which is lower than all compared methods. PuzzleMix is the second best performer obtaining 19.62% error rate. Similarly on Tiny ImageNet-200, DIFFUSEMIX outperforms SOTA by a notable margin obtaining 34.53% error rate while PuzzleMix remained the second best performer with 36.52% error rate. These results demonstrate that even under adversarial perturbations, DIFFUSEMIX remains resilient surpassing the performance of existing SOTA approaches.

5.3. Fine-Grained Visual Classification

Compared to the general classification, the task of Fine-Grained Visual Classification (FGVC) is notably challenging since it is difficult for a learning model to identify subtle differences between two different objects belonging to the same general class. A robust image augmentation technique should preserve these subtle albeit critical details for the learning model to successfully train on the fine-grained classification task. To evaluate DIFFUSEMIX on this task, we conduct experiments on three datasets including CUB [35], Stanford Cars [24], and Aircraft [32] utilizing ResNet-50 [14] network.

As shown in Table 5, DIFFUSEMIX significantly enhances the generalization capability of ResNet-50, yielding *superior* performances on all benchmark datasets. Specifically, DIFFUSEMIX outperforms widely-acknowledged methods such as Mixup and CutMix, with notable margins. On CUB dataset, DIFFUSEMIX achieved an accuracy of 79.37%, which is a significant leap from the 65.50% Vanilla accuracy. It also outperformed recent methods including Guided-AP 77.08%, and PuzzleMix 72.83%. Similarly, on the Aircraft dataset, DIFFUSEMIX obtained an accuracy of 85.76% surpassing the second best, GuidedMixup 84.32%. The Stanford Cars dataset further validates the remarkable performance of DIFFUSEMIX registering 91.26% accuracy. On the same dataset, other methods such as SnapMix and PuzzleMix achieved 90.10% and 89.68% respectively. The consistent performance gains on various challenging datasets iterate the effectiveness of DIFFUSEMIX in retaining critical salient information necessary to perform fine-grained classification.

Table 5. Top-1 (%) performance comparison on ***fine-grained task*** of ResNet-50. Extended comparisons are provided in Appendix 7 Table 14.

Method	Birds	Aircraft	Cars
Vanilla(CVPR'16) [14]	65.50	80.29	85.52
RA(NIPS'20) [9]	-	82.30	87.79
AdaAug(ICLR'22) [5]	-	82.50	88.49
Mixup(ICLR'18) [48]	71.33	82.38	88.14
CutMix(ICCV'19) [45]	72.58	82.45	89.22
SnapMix(AAAI'21) [19]	75.53	82.96	90.10
PuzzleMix(ICML'20) [23]	74.85	82.66	89.68
Co-Mixup(ICLR'21) [22]	72.83	83.57	89.53
Guided-AP(AAAI'23) [21]	77.08	84.32	90.27
DIFFUSEMIX	**79.37**	**85.76**	**91.26**

5.4. Transfer Learning

Transfer learning or fine-tuning is a widely used way of customizing large architectures with limited computational resources as well as for quick experiments. Most image-mixing-based augmentation methods [15, 15, 19, 21, 21–23, 40, 41, 45, 48] have not reported performance in this important scenario. Nevertheless, we evaluate the perfor-

Table 6. Top-1 (%) accuracy on ***data scarcity*** task of ResNet-18 on Flower102 dataset where only 10 random images per class are used. Extended comparisons are provided in Appendix 7 Table 15.

Method	Valid	Test
Vanilla (CVPR'16) [14]	64.48	59.14
SnapMix (AAAI'21) [19]	65.71	59.79
PuzzleMix (ICML'20) [23]	71.56	66.71
Co-Mixup (ICLR'20) [22]	68.17	63.20
GuidedMixup (AAAI'23) [21]	74.74	70.44
DiffuseMix	**77.14**	**74.12**

mance of our approach on fine-tuning the baseline model using three different datasets including Flower102, Aircraft, and Stanford Cars on ImageNet-pretrained ResNet-50 provided by PyTorch and report results in Table 8.

For the Flower102 dataset, DiffuseMix achieved an accuracy of 98.02%, which is higher than the second best method, AdaAug 97.19%. A similar trend is observable in the Aircraft 85.65% and Cars 93.17% datasets. We also observe that the accuracy obtained by DiffuseMix with fine-tuning (Table 8) is comparable to the performance when training is done from scratch (Table 5). Since fine-tuning consumes significantly less computational resources compared to training from scratch, this experiment elaborates the practical significance of DiffuseMix.

5.5. Data Scarcity

Data scarcity is one of the major issues when training deep neural networks. If the training examples per class are limited, deep networks may not learn meaningful patterns by leading to overfitting and loss of generalization. Augmentation methods are often used to overcome these challenges by generating more data for training. Under this setting, we compare the accuracy of ResNet-18 [14] trained using only 10 images per class of the original Flower102 dataset. As shown in Table 6, our method consistently outperforms other mixup methods in situations where data is limited yielding accuracies of 77.14% Top-1 and 74.12% Top-5. Furthermore, our approach is designed to enhance the diversification of the training dataset. By leveraging our bespoke conditional prompts, DiffuseMix artificially expands and enriches the overall data landscape, enabling a more robust learning of neural networks.

5.6. Analysis and Discussions

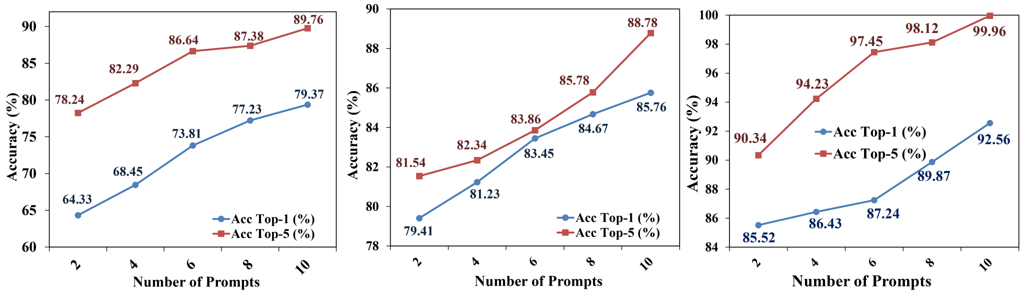
In this section, we provide a detailed ablation study and further analysis of various design choices in DiffuseMix.

Ablation Studies: To evaluate the importance of each component we conduct an ablation study by removing each component and report the observed performance on ResNet-50 using Stanford Cars and Flowers-102 datasets in Table 7.

When all of the components of DiffuseMix are removed, the baseline (vanilla) using only original images I_i obtains Top-1 and Top-5 accuracies of 85.52% & 90.34% on Cars dataset and 78.83% & 94.38% on Flowers dataset. Next, we add fractal blending (F_v) to the input images resulting in slight performance gains on both datasets. Further, we remove both concatenation (H_{iju}) and fractal blending (F_v) by conducting experiments using generated images ($\hat{I}_ij$) directly as augmented images to train the network. This setting brings the method closer to the approaches proposed in [1, 39] which utilize the images generated using diffusion models directly as augmented images. In this experiment, the accuracies obtained on Cars dataset are 87.63% Top-1 and 90.23% Top-5. Similarly, we observe the accuracies of 77.38% and 93.15% on Flowers102 dataset. The results are consistent with the findings in [1] that direct use of generated images may not yield significant performance gains over the vanilla method. Further, we carry out experiments by using generated images with fractal blending to train the model. On the Cars dataset, we observe slight performance gains over the vanilla model with accuracies of 89.42% Top-1 and 91.57% Top-5. However, in Flowers dataset, we observe a slight performance degradation. This demonstrates that to unleash the maximum benefits of fractal blending, it should be accompanied by more diversity in the training dataset. Next, we just remove the fractal blending from DiffuseMixwhile incorporating concatenation between generated ($\hat{I}_ij$) and original (I_i) images to create hybrid images (H_{iju}) used as augmented training examples. This setting increases the accuracies to 90.59% Top-1 & 96.73% Top-5 on Cars dataset and 79.22% Top-1 & 94.38% Top-5 on Flowers dataset. These performance improvements are due to the availability of both generated and original image contents in each augmented image, highlighting the importance of the concatenation step in DiffuseMix. Finally, when all components including fractal blending are used, the best accuracies of 91.26% Top-1 & 99.96% Top-5 on Cars dataset and 80.20% Top-1 & 95.40% Top-5 on Flowers dataset are achieved.

Table 7. Ablation study using Stanford Cars (cars) and Flowers102 (Flow) datasets. Top-1 and Top-5 accuracies are reported with ***different combinations*** of I_i: Input image, $\hat{I}_{ij}$: Generated images using prompts p_j, H_{iju}: Hybrid images using random mask M_u, and F_v: fractal images used to obtain final blended image A_{ijuv}.

	I_i	✓	✓	-	-	-	-
	$\hat{I}_{ij}$	-	-	✓	✓	-	-
	H_{iju}	-	-	-	-	✓	✓
	F_v	-	✓	-	✓	-	✓
Cars	Top-1	85.52	86.73	87.63	89.42	90.59	**91.26**
	Top-5	90.34	92.38	90.23	91.57	96.73	**99.96**
Flow	Top-1	78.73	78.34	77.38	77.81	79.22	**80.20**
	Top-5	94.38	94.91	93.15	93.24	94.38	**95.40**



(a) CUB-200-2011 (b) FGVC Aircraft (c) Stanford Cars

Figure 5. Effect of the number of prompts on overall performance. A detailed ablation study showcases the gains in Top-1 (%) and Top-5 (%) accuracy across CUB Birds-200, Aircraft, and Stanford Cars datasets with an increase in the number of prompts in DIFFUSEMIX.

Table 8. Top-1 (%) accuracy of DIFFUSEMIX on ***fine-tuning*** experiments using ImageNet pretrained ResNet-50.

Method	Flower102	Aircraft	Cars
Vanilla(CVPR'16) [14]	94.98	81.60	88.08
AA(CVPR'19) [8]	93.88	83.39	90.82
RA(NIPS'20) [9]	95.23	82.98	89.28
Fast AA(NIPS'19) [30]	96.08	82.56	89.71
AdaAug(ICLR'22) [5]	97.19	83.97	91.18
DIFFUSEMIX	**98.02**	**85.65**	**93.17**

Overall, consistent gains achieved with each added component signifies the design choices in DIFFUSEMIX towards an effective image augmentation technique.

Number of Prompts Vs. Performance: The variety of augmented images is determined by the number of random prompts which is an important factor in DIFFUSEMIX. A higher number of prompts corresponds to more diverse training samples. Figure 5 shows the effect of increasing the number of prompts on the performance using three different datasets including Birds, Aircraft and Cars. Increasing the number of prompts consistently yields performance gains on all three datasets using both Top-1 and Top-5 accuracy. However, the peak performances are achieved when all ten prompts proposed in our approach are used for training. In almost all datasets, the increasing accuracy trends can be seen even at 10 prompts which shows that the addition of more prompts may further improve the performance. However, it will also incur more computational costs.

Increasing Masks Vs Performance: We conduct experiments using Oxford Flower102 dataset to study the impact of various masks on the overall performance of DIFFUSEMIXand report the results in Table 9. Using even only one kind of mask, vertical in this example, our approach achieves significantly higher accuracies than the vanilla baseline. When both vertical and horizontal masks are used, the accuracies improve further. However, the best accuracies are achieved when both horizontal and vertical masks are used along with random flipping between the positions of input and generated images adding more diversity to the training data. It is also possible to use the masking techniques from previous approaches, such as [27, 45], in DIFFUSMIX. We provide additional analysis on this in Appendix 4.

Table 9. Ablation on the ***effects of masking*** in DIFFUSEMIXon Flower102 dataset. All variants yield notably superior results compared to vanilla on ResNet-50. However, best results are achieved when all four vertical and horizontal masks are used.

Mask	Top-1 (%)	Top-5 (%)
Vanilla(CVPR'16) [14]	89.74	94.38
Ver Mask (◧)	94.02	98.42
Hor + Ver Masks (◧, ⬓)	94.27	99.03
Hor + Ver + Flipping (◧ , ◨, ⬓, ⬒)	95.37	99.39

6. Conclusion

In this paper, we introduced DIFFUSEMIX, a data augmentation technique based on diffusion models to increase diversity in the data while preserving the original semantics of the input image. It involves *generation, concatenation, and fractal blending* steps to create the final augmented image. On multiple tasks such as general classification, fine-grained classification, data scarcity, fine-tuning, and adversarial robustness involving several benchmark datasets including *ImageNet-1k, Tiny-ImageNet-200, CIFAR-100, Oxford Flower102, Caltech Birds, Stanford-Cars, and FGVC Aircraft*, DIFFUSEMIXdemonstrates consistent performance gains and outperforms existing SOTA image augmentation methods.

Limitations: DIFFUSEMIXhas two notable limitations: (1) Image generation relies heavily on text prompts and a wrong textual input may lead to unrealistic results. We address this issue by proposing a set of *filter-like* prompts generally applicable to a wide range of natural images. (2) DIFFUSEMIXrequires additional overheads for generating images (more on this in Appendix 2). This is a small price to pay for a guaranteed better convergence of large-scale classification models and can be mitigated by generating and storing augmented images once.

Acknowledgement: We are thankful to Hamza Saleem for the fruitful insights. Arif Mahmood is funded by the Information Technology University of Punjab, Pakistan.

References

[1] 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.

[2] 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] Jie-Neng Chen, Shuyang Sun, Ju He, Philip HS Torr, Alan Yuille, and Song Bai. Transmix: Attend to mix for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12135–12144, 2022.

[4] Yuan-Chih Chen and Chun-Shien Lu. Rankmix: Data augmentation for weakly supervised learning of classifying whole slide images with diverse sizes and imbalanced categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23936–23945, 2023.

[5] Tsz-Him Cheung and Dit-Yan Yeung. Adaaug: Learning class-and instance-adaptive data augmentation policies. In *International Conference on Learning Representations*, 2021.

[6] Hyeong Kyu Choi, Joonmyung Choi, and Hyunwoo J Kim. Tokenmixup: Efficient attention-guided token-level data augmentation for transformers. *Advances in Neural Information Processing Systems*, 35:14224–14235, 2022.

[7] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[8] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019.

[9] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

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

[12] Chengbin Du, Yanxi Li, Zhongwei Qiu, and Chang Xu. Stable diffusion is untable. *arXiv preprint*, 2023.

[13] Junlin Han, Pengfei Fang, Weihao Li, Jie Hong, Mohammad Ali Armin, , Ian Reid, Lars Petersson, and Hongdong Li. You only cut once: Boosting data augmentation with a single cut. In *International Conference on Machine Learning (ICML)*, 2022.

[14] K. He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[15] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*.

[16] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

[17] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16783–16792, 2022.

[18] Minui Hong, Jinwoo Choi, and Gunhee Kim. Stylemix: Separating content and style for enhanced data augmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14862–14870, 2021.

[19] Shaoli Huang, Xinchao Wang, and Dacheng Tao. Snapmix: Semantically proportional mixing for augmenting fine-grained data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1628–1636, 2021.

[20] Zhenglin Huang, Xiaoan Bao, Na Zhang, Qingqi Zhang, Xiao Tu, Biao Wu, and Xi Yang. Ipmix: Label-preserving data augmentation method for training robust classifiers. *Advances in Neural Information Processing Systems*, 36, 2024.

[21] Minsoo Kang and Suhyun Kim. Guidedmixup: an efficient mixup strategy guided by saliency maps. In *AAAI*, pages 1096–1104, 2023.

[22] JangHyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. In *International Conference on Learning Representations*, 2020.

[23] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*, pages 5275–5285. PMLR, 2020.

[24] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *CVPRw*, pages 554–561, 2013.

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

[26] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[27] Jin-Ha Lee, Muhammad Zaigham Zaheer, Marcella Astrid, and Seung-Ik Lee. Smoothmix: a simple yet effective data augmentation to train robust classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 756–757, 2020.

[28] Zheng Li, Yuxuan Li, Penghai Zhao, Renjie Song, Xiang Li, and Jian Yang. Is synthetic data from diffusion models ready for knowledge distillation? *arXiv preprint arXiv:2305.12954*, 2023.

[29] Wen Liang, Youzhi Liang, and Jianguo Jia. Miamix: Enhancing image classification through a multi-stage augmented mixied sample data augmentation method. *arXiv preprint arXiv:2308.02804*, 2023.

[30] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. *Advances in Neural Information Processing Systems*, 32, 2019.

[31] Xue-Jing Luo, Shuo Wang, Zongwei Wu, Christos Sakaridis, Yun Cheng, Deng-Ping Fan, and Luc Van Gool. Camdiff: Camouflage image augmentation via diffusion model. *arXiv preprint arXiv:2304.05469*, 2023.

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

[33] Thomas Mensink and Pascal Mettes. Infinite class mixup. *arXiv preprint arXiv:2305.10293*, 2023.

[34] Ning Miao, Tom Rainforth, Emile Mathieu, Yann Dubois, Yee Whye Teh, Adam Foster, and Hyunjik Kim. Instance-specific augmentation: Capturing local invariances. 2022.

[35] 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.

[36] Huafeng Qin, Xin Jin, Yun Jiang, Mounim A El-Yacoubi, and Xinbo Gao. Adversarial automixup. *arXiv preprint arXiv:2312.11954*, 2023.

[37] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[38] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

[39] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

[40] AFM Uddin, Mst Monira, Wheemyung Shin, TaeChoong Chung, Sung-Ho Bae, et al. Saliencymix: A saliency guided data augmentation strategy for better regularization. *arXiv preprint arXiv:2006.01791*, 2020.

[41] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR, 2019.

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

[43] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *ICLR*, 2020.

[44] Lingyu Yan, Yu Ye, Chunzhi Wang, and Yun Sun. Locmix: local saliency-based data augmentation for image classification. *Signal, Image and Video Processing*, pages 1–10, 2023.

[45] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *CoRR*, abs/1905.04899, 2019.

[46] 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.

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

[48] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[49] 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.

[50] Qihao Zhao, Yangyu Huang, Wei Hu, Fan Zhang, and Jun Liu. Mixpro: Data augmentation with maskmix and progressive attention labeling for vision transformer. In *The Eleventh International Conference on Learning Representations*, 2022.