DIFFUSEMIX: Label-Preserving Data Augmentation with Diffusion Models

Supplementary Material

Overview

This supplementary document contains additional results and discussions. Summarily, Section 1 provides analysis on varying values of λ which defines the ratio of blending between a fractal image and a hybrid image in DIFFUSEMIX. Section 2 provides comparisons of augmentation overhead between DIFFUSEMIXand existing image augmentation strategies with respect to their generalization performances. Section 3 discusses examples of poorly constructed prompts and their effects on image generation. Section 4 provides experimental results of using different masking strategies of the state-of-the-art methods with DIFFUSEMIX. Section 5 provides the convergence analysis of DIFFUSEMIX. Section 6 provides more visualizations of the augmented training images obtained using DIFFUSEMIX. Section 7 provides a complete list of general and fined-grained results. Section 8 provides some visual examples of the collected fractal image dataset.

1. Fractal Blending Ratio

We experiment and observe the effect of varying fractal blending ratio λ in DIFFUSEMIX and report Top-1 accuracy (%) results on Flower102 dataset in Table 10. The values of λ are varied from 0.1 to 0.5. A higher value of λ indicates a stronger ratio of fractal image blending.

The baseline, ResNet50 without any augmentation, yields a top-1 accuracy of 78.73%. Compared to this, DIFFUSEMIX yields consistent performance gains with all values of λ. The best performance of DIFFUSEMIX is observed at $\lambda = 0.2$, where the top-1 accuracy peaks at 81.30%. However, generally, the performance remains better with a reasonable value of λ. It starts dropping when the value of λ becomes too high. This suggests that higher fractal blending ratios may introduce too much complexity or noise into the original data, which adversely affects the model's performance.

2. Augmentation Overhead

We compare the computational overhead of several existing SOTA image augmentation methods and DIFFUSEMIX with respect to the performance gains. Following Kang *et al.* [21], we define the augmentation overhead $\mathcal{A}_{\mathcal{O}}$ as:

$$\mathcal{A}_{\mathcal{O}} = \frac{\mathcal{T}_{aug} - \mathcal{T}_{van}}{\mathcal{T}_{van}} \times 100(\%),$$

where $\mathcal{T}$ is the total image generation and training time, and $\mathcal{T}_{van}$ is the training time of the baseline network without any augmentation. Although image generation process in DIFFUSEMIX can be expedited by using parallel-processing, We do not utilize it to provide a fair comparison. It can be seen in Figure 6 that DIFFUSEMIX provides a good tradeoff between performance and augmentation overhead by outperforming all existing approaches in terms of accuracy while providing significantly lower augmentation overhead compared to Co-Mixup and SaliencyMix approaches. Moreover, DIFFUSEMIX can also be optimized further by saving the generated images offline once before carrying out any number of subsequent trainings. This may make it significantly faster to perform several experiments on a training model, particularly for optimization and research purposes.

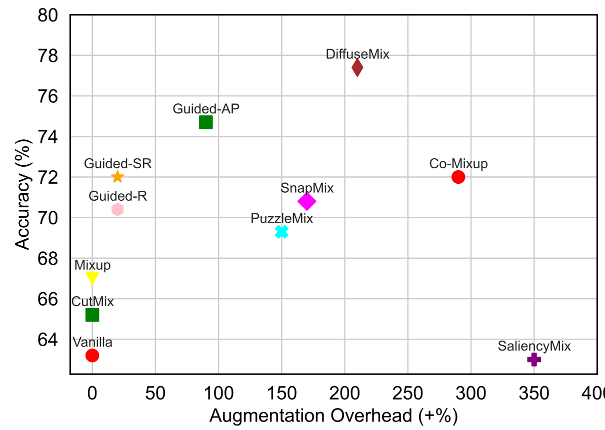


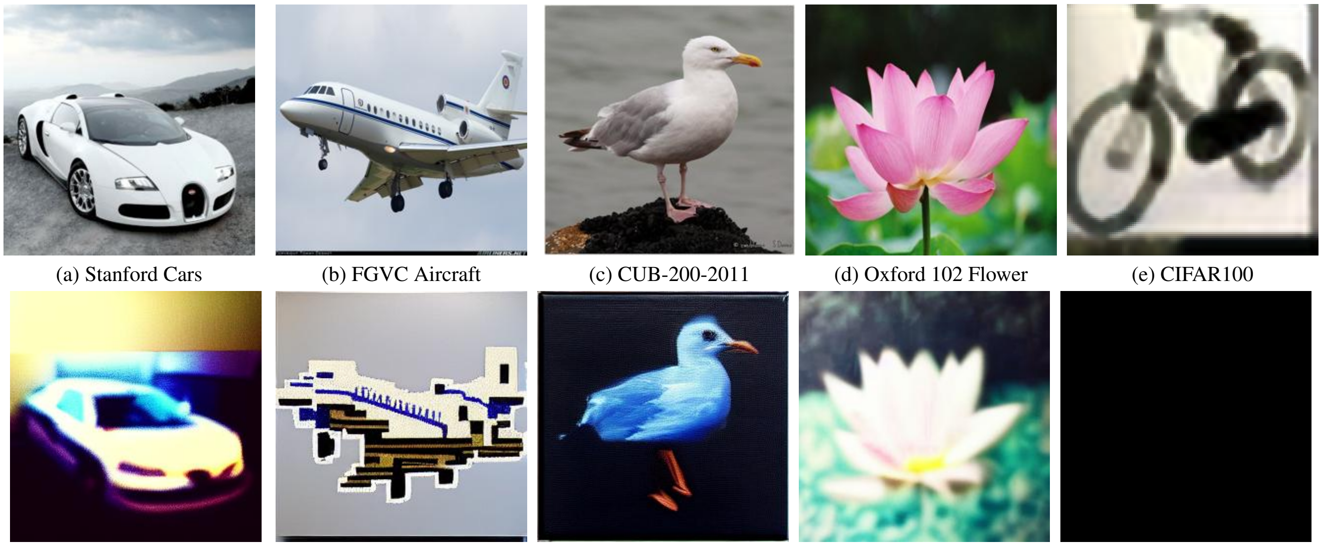
Figure 6. Augmentation overhead (+%) - accuracy (%) plot on CUB-200-2011 dataset with batch size 32.

Table 10. Impact of varying fractal blending ratio in DIFFUSEMIX. Top-1 accuracy is reported using ResNet-50 on Flower102 dataset.

Methods	Top-1 (%)
ResNet50$_{\text{(CVPR'16)}}$ [14]	78.73
+ *DiffuseMix* ($\lambda = 0.1$)	79.81
+ *DiffuseMix* ($\lambda = 0.2$)	**81.30**
+ *DiffuseMix* ($\lambda = 0.3$)	80.97
+ *DiffuseMix* ($\lambda = 0.4$)	79.16
+ *DiffuseMix* ($\lambda = 0.5$)	78.97

3. Prompt Selection

Diffusion Models rely heavily on prompts [12]. Therefore, the intuition behind designing our *bespoke conditional*



(a) Stanford Cars (b) FGVC Aircraft (c) CUB-200-2011 (d) Oxford 102 Flower (e) CIFAR100

Sunset photo that looks like it's taken with 1990s camera | *A painting that is too small for its body* | *A cityscape that is too dark to see anything* | *Photograph that is washed it looks like it is made of paper* | *A distorted, warped painting of a landscape*

Figure 7. **First row:** original training image samples from different datasets *such as* Oxford-102 Flower [36], Stanford Cars [24], and Aircraft [33], CUB-200-2011, and CIFAR100. **Second row:** Corresponding generated images show that the usage of descriptive prompts (blue text) results in poor images not feasible for training. When generating images on the CIFAR100 dataset, several additional challenges may occur due to the small size of the images. For example, the image in the last column taken from CIFAR-100 with its corresponding prompt results in a ***black*** image containing no visible output.

prompts is to introduce the type of prompts that may edit the image in a way that preserves structural information and can easily be applied to a range of diverse datasets. To this end, as described in the manuscript, we propose to use *filter-like* prompts such as snowy, sunset, rainbow, etc. and demonstrate their effectiveness in training robust classifiers.

Conversely, in this section, we discuss *bad* prompts that may not be a good fit for the image generation step of DIFFUSEMIX. Some examples of such prompts are shown in Figure 7. More descriptive and overly complicated prompts generate images that may be too different from the original distribution. The resultant images contain unrealistic foregrounds and backgrounds, rendering these useless for the training of a classifier. This reiterates the importance of our proposed *filter-like bespoke conditional prompts* that do not induce unwanted changes to the training images.

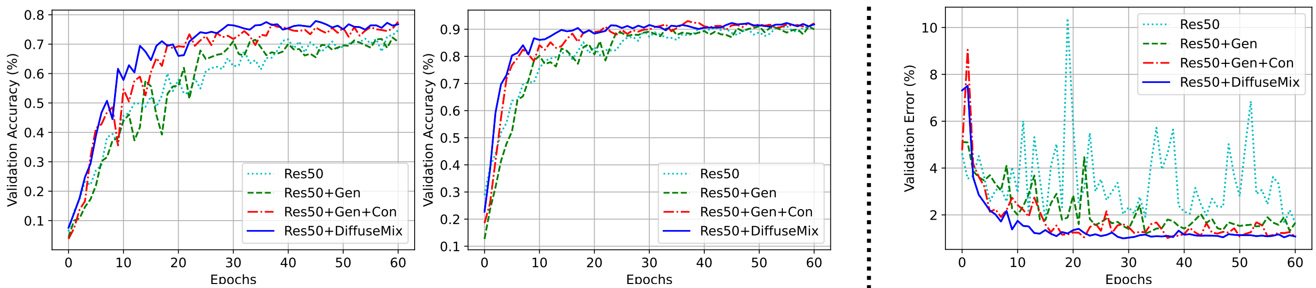
4. DIFFUSEMIX with SOTA Methods

In a series of experiments, we combine DIFFUSEMIX with existing image augmentation approaches [46, 49] to see if any performance gain is observed. Particularly, We replace our masking approach with the masking used in the existing methods while retaining the rest of the pipeline of DIFFUSEMIX same.

For CutMix + DIFFUSEMIX, we replace the concatenation step of DIFFUSEMIX with the random cropping of CutMix. To this end, we randomly crop a patch from the generated image and paste it onto the original image whereas the other stages remain the same. For Mixup + DIFFUSEMIX, we replace concatenation with the pixel blending of original and generated images as proposed in [49] while the rest of the steps remain intact. The results are summarized in Table 11. Using CutMix [46] or Mixup [49] methods yields improvements over baseline ResNet50 training. However, when our proposed approach is added to the existing methods, further performance gains are observed. Top performance is finally observed with our DIFFUSEMIX, which demonstrates the importance of forming hybrid images by concatenating original and generated images.

Table 11. Combining DIFFUSEMIX with SOTA image augmentation methods by replacing the image concatenation technique of DIFFUSEMIX with the masking techniques proposed in [46] & [49]. While DIFFUSEMIX provides consistent gains in these settings, the best performance of 81.30% is achieved when our originally proposed method is used.

Method	Top-1 (%)
ResNet50$_{\text{(CVPR'16)}}$ [14]	78.73
+ CutMix [46]	79.22
+ CutMix [46] + DIFFUSEMIX	79.58
+ Mixup [49]	79.34
+ Mixup [49] + DIFFUSEMIX	80.20
+ DIFFUSEMIX	**81.30**



(a) Top-1 Accuracy (%)　(b) Top-5 Accuracy (%)　(c) Validation Error (%)

Figure 8. ***On left side:*** The curves of top-1 and top-5 accuracy show an increasing trend during initial 60 epochs and remain stable towards the end on Flower102 dataset. This same behavior can also be seen in top-5 accuracy. —our enables smoother training and better convergence while avoiding overfitting. ***On right side:*** Similar to the accuracy plots, using DIFFUSEMIX demonstrates a smoother decrease in validation error compared to ResNet50 or other variants. *Best viewed in color.*

5. DIFFUSEMIX Convergence

Analysis on Top-1 and Top-5 Accuracy: In a series of experiments, we carry out an ablation to observe the top-1 and top-5 accuracies of DIFFUSEMIX and its variants formed by removing the components *(generation, concatenation, and fractal blending)* one by one.

As seen in Figures 8a and 8b, the RES50+DIFFUSEMIX demonstrates generally better performance with convergence at 77.26% accuracy, closely followed by DIFFUSEMIX+GEN+CON at 75.79%, and Res50 at 76.41%. The DIFFUSEMIX+GEN model performs significantly lower yielding 73.96% accuracy. As discussed in the manuscript Section 4, using generated images directly for the training may lead to deteriorated performance, which is re-validated in these experiments. This also shows the importance of each step proposed in DIFFUSEMIX towards robust training more robust classifiers. Overall, similar trends are observed in Top-5 accuracy results (Figure 8b).

Analysis on validation loss: As seen in (Figure 8c), it is clearly noticeable that DIFFUSEMIX helps in model convergence and overall smooth decrease in validation loss during training. Res50 baseline shows a good start with lower initial loss. However, its loss starts fluctuating once the training is continued indicating a potential plateau in learning or its limitation in capturing more complex patterns. Compared to all variants, Res50+DIFFUSEMIX benchmarks better convergence.

6. DIFFUSEMIX Visualizations

In this section, we provide more visual examples of training images obtained using DIFFUSEMIX. As seen in Figure 9, visualizing examples from Flower102 dataset, DIFFUSEMIX enhances the overall variation of the images while retaining the interpretability of each example. For Caltech-UCSD Birds-200-2011 (Figure 10), compared to original images, the augmented images obtained using DIFFUSEMIX exhibit greater clarity and diverse contexts. Similar visual features can be observed in Figure 11 and Figure 12 showcasing examples from Cars and Aircraft datasets, respectively.

Table 12. Top-1 and Top-5 general classification accuracies comparison using PreActResNet-18. Compared numbers are taken either from the original papers or from [21].

Method	Tiny-ImageNet Top-1 Acc (%)	Tiny-ImageNet Top-5 Acc (%)	CIFAR-100 Top-1 Acc (%)	CIFAR-100 Top-5 Acc (%)
Vanilla [14]	57.23	73.65	76.33	91.02
Mixup [49]	56.59	73.02	76.84	92.42
Manifold [42]	58.01	74.12	79.02	93.37
CutMix [46]	56.67	75.52	76.80	91.91
AugMix [15]	55.97	74.68	75.31	91.62
PixMix [17]	-	-	79.70	-
SaliencyMix [41]	56.54	76.14	79.75	94.71
Guided-SR [21]	55.97	74.68	80.60	94.00
PuzzleMix [23]	63.48	75.52	80.38	94.15
Co-Mixup [22]	64.15	-	80.15	-
Guided-AP [21]	64.63	82.49	81.20	94.88
DIFFUSEMIX	**65.77**	**83.66**	**82.50**	**95.41**

7. Performance Evaluation

Extended versions of the performance tables are provided in this section.

7.1. General Classification

In Table 12, Vanilla method serves as our baseline, achieving Top-1 accuracies of 57.23% and 76.33% on Tiny-ImageNet and CIFAR-100 respectively, setting a benchmark for subsequent comparisons. For Mixup we observe a slight decline in performance on Tiny-ImageNet to 56.59% Top-1 accuracy but a marginal improvement on CIFAR-100, reaching 76.84%. Conversely, Manifold Mixup marks

(a) Poinsettia (b) Barbeton Daisy (c) Gazania (d) Dandelion (e) Magnolia

(f) Poinsettia Autumn (g) Barbeton Daisy Snowy (h) Gazania Crayon Sketch (i) Dandelion Crayon Sketch (j) Magnolia Crayon Sketch

Figure 9. Illustration of original training images and DIFFUSEMIX augmented images from the Oxford Flower102 dataset. **First row:** showcases original, unaltered images of various flowers, including *poinsettia, barbeton daisy, gazania, dandelion,* and *Magnolia* classes. **Second row:** illustrates the transformative effects of the DIFFUSEMIX augmentation method. The effects of our custom-tailored prompts-based generation are visible on the generated portion of each image. Overall, DIFFUSEMIX results in a diverse array of images with sufficient structural complexity and diversity to train robust classifiers.

notable performance gains, especially on CIFAR-100 with a Top-1 accuracy of 79.02%. CutMix slightly improves over the baseline on Tiny-ImageNet, whereas AugMix shows a decrement, particularly on CIFAR-100 with a 75.31% Top-1 accuracy.

PixMix introduces variations in the source image instead of mixing two input images. Compared to baseline, PixMix excels on CIFAR-100 with a 79.70% Top-1 accuracy. SaliencyMix, which uses saliency to mix different portions of images, also shows promising results. Particularly on CIFAR-100, it achieves a Top-1 accuracy of 79.75%. The Guided-SR method performs slightly lower compared to AugMix on Tiny-ImageNet but stands out on CIFAR-100 with 80.60% Top-1 accuracy, indicating its effectiveness. PuzzleMix and Co-Mixup introduce more complex ways to augment data, with PuzzleMix reaching a notable 63.48% Top-1 accuracy on Tiny-ImageNet. Co-Mixup tops these methods on Tiny-ImageNet with 64.15% Top-1 accuracy but does not maintain this lead on CIFAR-100. Guided-AP pushes the performance boundaries further by achieving superior accuracies among its predecessors, e.g., 81.20% Top-1 accuracy on CIFAR-100.

DIFFUSEMIX, our proposed method, which surpasses all prior techniques by securing the highest accuracies: 65.77% Top-1 on Tiny-ImageNet and 82.50% Top-1 on CIFAR-100. Our approach not only surpasses the conventional mixup strategies but also sets a new standard in enhancing the generalization of deep learning models. The performance of DIFFUSEMIX stays consistent across the compared datasets, underlining its superior capability and efficiency.

Table 13. Top-1 and Top-5 accuracies comparison on ImageNet using ResNet-50. Compared numbers are taken either from the original papers or from [21].

Method	Top-1 Acc.	Top-5 Acc.
Vanilla(CVPR'16) [14]	75.97	92.66
AugMix (ICLR'20) [15]	76.75	93.30
Manifold(ICML'19) [42]	76.85	93.50
Mixup(ICLR'18) [49]	77.03	93.52
CutMix(ICCV'21) [46]	77.08	93.45
Guided-SR(AAAI'23) [23]	77.20	93.66
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	-
Azizi et al.(arXiV'23) [1]	78.17	-
DIFFUSEMIX	**78.64**	**95.32**

In Table 13, we provide a comparison of various methods in terms of Top-1 and Top-5 accuracies on ImageNet, specifically when training ResNet-50 as per the training configuration of in Kang and Kim [21]. It starts with the baseline Vanilla ResNet model, showing accuracies of 75.97% for Top-1 and 92.66% for top-5. Various techniques, including Azizi et al., AugMix, Manifold, Mixup, CutMix, Guided-SR, PixMix, PuzzleMix, GuidedMixup,

Table 14. Top-1 accuracy comparison on fine-grained visual classification task while training from scratch on ResNet-50.

	Methods	Top-1 Accuracy (%)		
		CUB	Aircraft	Cars
	Vanilla [14]	65.50	80.29	85.52
automated	Auto Aug [8]	-	82.28	88.04
	Fast AA [31]	-	82.20	87.19
	DADA [28]	-	81.16	87.14
	RA [9]	-	82.30	87.79
	AdaAug [5]	-	82.50	88.49
mixup family	Mixup [49]	71.33	82.38	88.14
	CutMix [46]	72.58	82.45	89.22
	SaliencyMix [41]	66.66	83.14	89.04
	Guided-SR [21]	74.08	83.51	89.23
	SnapMix [19]	75.53	82.96	90.10
	PuzzleMix [23]	74.85	82.66	89.68
	Co-Mixup [22]	72.83	83.57	89.53
	GuidedMixup [21]	77.08	84.32	90.27
	DIFFUSEMIX	**79.37**	**85.76**	**91.26**

Co-Mixup, YOCO, and another entry from Azizi et al., display a range of improvements, with Top-1 accuracies spanning from 69.24% to 78.17% and top-5 accuracies (when provided) ranging up to 93.86%. The most notable performance is observed in the DIFFUSEMIX method, which outperforms the others by achieving the highest accuracies at 78.64% for top-1 and 95.32% for top-5.

7.2. Fine-Grained Visual Classification

Table 14 presents a comparison of Top-1 accuracy of various methods on a fine-grained visual classification task, using ResNet-50. The methods are categorized into two main groups: automated methods and the mixup family, and are evaluated across three datasets: CUB, Aircraft, and Cars.

In the automated data augmentation, the Vanilla method achieves 65.50%, 80.29%, and 85.52% accuracy on CUB, Aircraft, and Cars respectively. Other automated methods like Auto Aug, Fast AA, DADA, RA, and AdaAug show varied performance, with AdaAug topping this category with accuracies of 82.50% for Aircraft and 88.49% for Cars. The mixup family methods show a notable performance improvement, particularly GuidedMixup demonstrating the accuracies of 77.08%, 84.32%, and 90.27% on the three datasets respectively. Nevertheless, our DIFFUSEMIX stands out by outperforming all compared methods significantly, achieving the highest accuracies of 79.37% for CUB, 85.76% for Aircraft, and 91.26% for Cars.

This indicates that while both categories of methods enhance performances, mixup family methods demonstrate superior capability in handling fine-grained visual classification tasks. DIFFUSEMIX, in particular, showcases exceptional improvements, suggesting its effectiveness in extracting nuanced features from the images.

Table 15. Top-1 accuracy on data scarcity experiment using Flower102 dataset where only 10 random images per class are used. Experiments are performed with ResNet-18 network.

Methods	Valid (%)	Test (%)
Vanilla(CVPR'16) [14]	64.48	59.14
Mixup(ICLR'18) [49]	70.55	66.81
CutMix(ICCV'19) [46]	62.68	58.51
SaliencyMix(ICLR'21) [41]	63.23	57.45
Guided-SR(AAAI'21) [21]	72.84	69.31
SnapMix(AAAI'21) [19]	65.71	59.79
PuzzleMix(ICML'20) [23]	71.56	66.71
Co-Mixup(ICLR'21) [22]	68.17	63.20
GuidedMixup(AAAI'23) [21]	74.74	70.44
DIFFUSEMIX	**77.14**	**74.12**

7.3. Data Scarcity

Table 15 presents the Vanilla method as a baseline with 64.48% accuracy on the validation set and 59.14% on the test set. SOTA techniques like Mixup and PuzzleMix show improved accuracies, with Mixup achieving 70.55% on validation and 66.81% on the test set, and PuzzleMix reaching 71.56% and 66.71%, respectively.

Notably, the Guided-SR and GuidedMixup methods significantly outperform other approaches, with GuidedMixup achieving the highest accuracies of 74.74% on validation and 70.44% on the test set. Our DIFFUSEMIX, which surpasses all compared methods, demonstrates remarkable accuracies of 77.14% on the validation and 74.12% on the test set, showcasing its superior ability to generalize well from significantly limited data. This evidence suggests that data augmentation and mixing techniques, especially DIFFUSEMIX, are highly beneficial in enhancing model performance under stringent data constraints.

7.4. Self-Supervised Learning

Table 16 showcases the Top-1 accuracy of self-supervised learning methods, specifically comparing the performance of MoCo v2 and SimSiam.

Initially, MoCo v2 exhibits accuracies of 80.31%, 40.82%, and 51.36% on Flower102, StanfordCars, and Aircraft datasets. After applying DIFFUSEMIX augmentation, it performs better by demonstrating accuracies of

Table 16. Top-1 (%) accuracy of self-supervised learning methods. Adding DIFFUSEMIX yields better performance.

Method	Flower102	Stanford Cars	Aircraft
MoCo v2	80.31	40.82	51.36
+ DIFFUSEMIX	**82.15**	**41.73**	**53.28**
SimSiam	86.93	48.34	40.37
+ DIFFUSEMIX	**89.24**	**49.17**	**42.63**

	Mixup	CutMix	AugMix	Outlier	PixMix	DiffuseMix
Corruptions	48.0	51.5	35.4	51.5	30.5	**28.5**
Consistency	9.5	12.0	6.5	11.3	5.7	**5.1**
Adversaries	97.4	97.0	95.6	97.2	92.9	**90.2**
Calibration	13.0	29.3	18.8	15.2	8.1	**7.7**
Anomaly Det.	71.7	74.4	84.9	**90.3**	89.3	88.3

Table 17. On CIFAR-100, DiffuseMix outperforms SOTA on 4 of the 5 distinct safety metrics. Lower is better except for anomaly detection. (SOTA method results are taken from PixMix [17]).

82.15%, 41.73%, and 53.28%. SimSiam starts with accuracies of 86.93%, 48.34%, and 40.37%. Adding DiffuseMix as an augmentation method improves the performance to 89.24%, 49.17%, and 42.63%. This clearly illustrates that integrating DiffuseMix significantly boosts the performance, demonstrating its effectiveness in enhancing self-supervised learning models. The systematic gains across different datasets and on multiple methods highlight the robustness of our approach and its potential to improve the accuracies of different machine learning models.

7.5. Safety Measures

Table 17 showcases a comparative analysis of several data augmentation methods on the CIFAR-100 dataset, focusing on their performance across five different safety metrics. The methods evaluated include Mixup, CutMix, AugMix, Outlier, PixMix, and DiffuseMix. The results highlight DiffuseMix's superior performance, as it outperforms the state-of-the-art (SOTA) previously established by PixMix in four out of the five categories. DiffuseMix demonstrates better performance in cases of corruptions, consistency, adversaries, and calibration. In the case of Anomaly detection task, our approach demonstrates comparable performance.

8. Fractal Dataset

We collected a dataset of 100 fractal images containing complex patterns and scales. Blending these images to the training images introduces a level of abstraction and complexity not commonly found in regular training images. Some of the example fractal images are provided in Figure 13. As discussed extensively in the manuscript, fractal blending in DiffuseMix helps the network generalize better by adding *contained* noise or perturbations. The ablation studies reported in our manuscript and supplementary suggest that utilizing fractal blending with the generated images helps stabilizing the training and improves the overall convergence.

8.1. Fractal with SOTA Methods

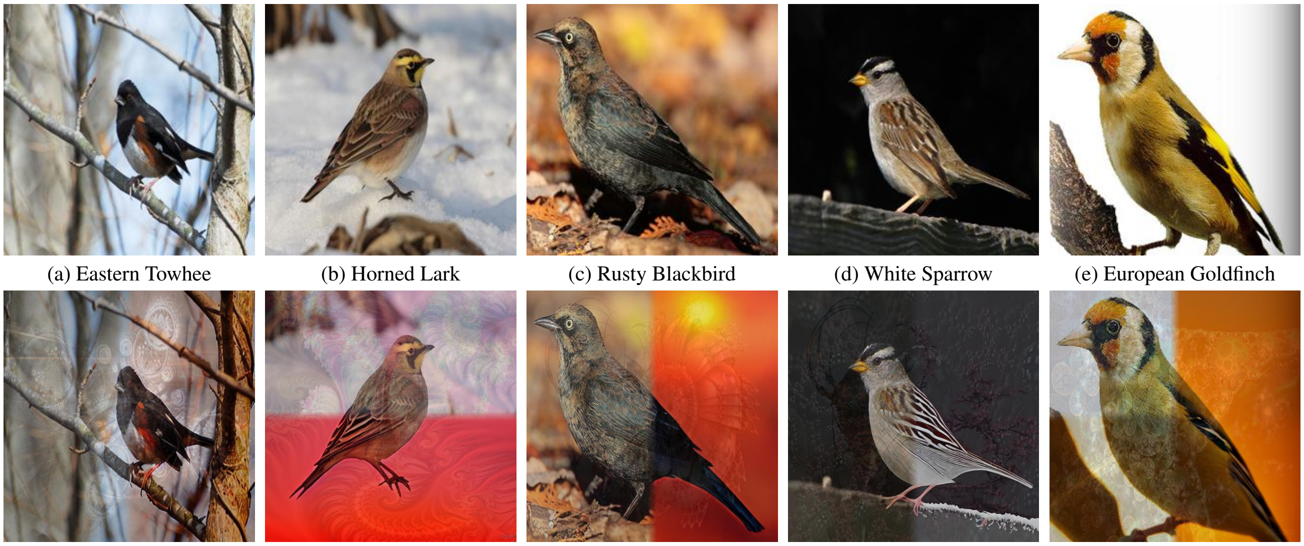
Table 18 presents a performance analysis, particularly focusing on the impact of blending fractals with different augmentation methods using CUB-Birds, Aircraft, Stanford Cars, and Flower102 datasets.

Adding fractal blending to the baseline results in performance improvements on CUB-Birds, Aircraft, and Stanford Cars, but a slight decrease in accuracy on the Flower dataset. The baseline method shows performances of 65.50% on CUB-Birds, 80.29% on Aircraft, 85.52% on Cars and 78.34% on Flower102. The Mixup, CutMix, and PuzzleMix methods, when used without fractal, generally show higher accuracy than the baseline, especially on the Stanford Cars and Aircraft datasets. However, the integration of fractal blending with these methods leads to a significant drop in performance across all datasets, suggesting that fractal blending may not be properly aligned with these particular augmentation techniques.

Table 18. Performance comparison (%) of fractal blending with baseline and other augmentation methods, it is more effective when fractals are blended with our hybrid images H_{iju}.

Method	CUB-Birds	Aircraft	Cars	Flower
Baseline	65.50	80.29	85.52	78.73
+ Fractal	**66.17**	**81.27**	**86.73**	**78.34**
Mixup	71.33	82.38	88.14	78.12
+ Fractal	**43.25**	**44.27**	**54.25**	**57.27**
CutMix	72.58	82.45	89.22	74.36
+ Fractal	**46.74**	**41.47**	**56.37**	**52.28**
PuzzleMix	74.85	82.66	89.68	71.68
+ Fractal	**51.61**	**53.38**	**61.42**	**63.73**
Hybrid (H_{iju})	80.27	85.31	90.59	79.22
+ Fractal	**79.37**	**85.76**	**92.56**	**80.20**

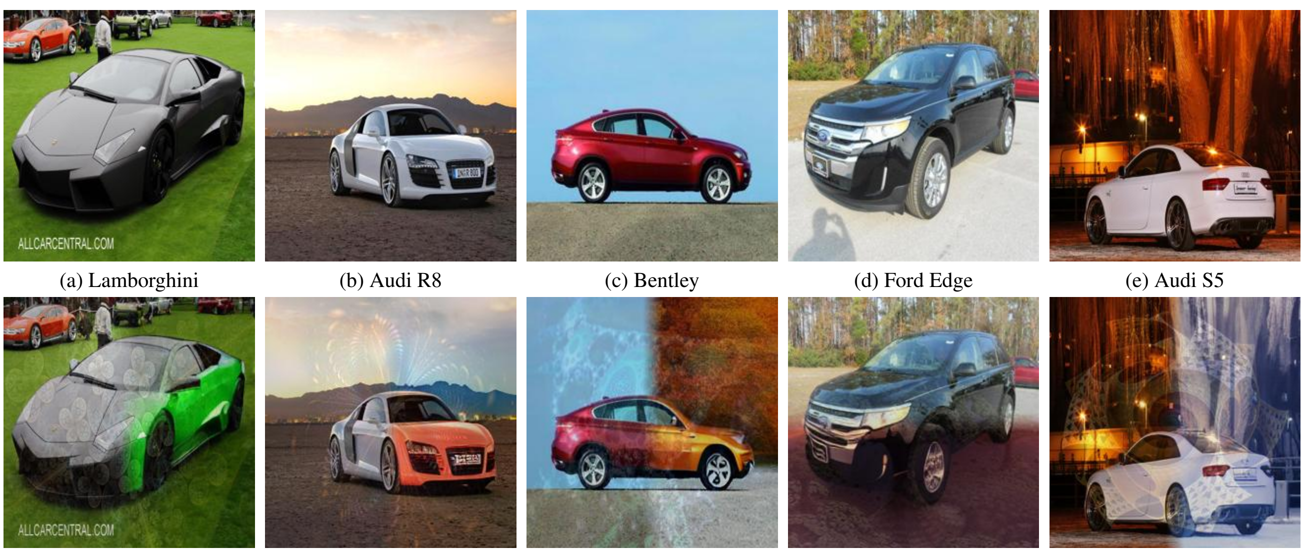
In contrast, when fractals are blended with the hybrid images (Hybrid H_{iju}) in our approach, performance improvements are notably observed in three of the four datasets including Aircraft, Stanford Cars, and Flower102. datasets, this combination leads to improvements in accuracy, indicating a positive synergy between the hybrid images and fractal blending. However, there's a slight decrease in accuracy for the CUB-Birds dataset.



(a) Eastern Towhee (b) Horned Lark (c) Rusty Blackbird (d) White Sparrow (e) European Goldfinch

(f) Eastern Towhee Sunset (g) Horned Lark Sunset (h) Rusty Blackbird Autumn (i) White Sparrow Snowy (j) Europe Goldfinch Snowy

Figure 10. Original and DIFFUSEMIX augmented bird images from the Caltech-UCSD Birds-200-2011 dataset. **Top row:** displays a selection of original, high-resolution bird images, capturing the natural beauty and diversity of species such as the *eastern towhee, horned lark, rusty blackbird, white sparrow,* and *european goldfinch*. **Bottom row:** demonstrates the augmented images obtained using DIFFUSEMIX. The augmented images are visually striking and contextually varied representations of the original subjects.



(a) Lamborghini (b) Audi R8 (c) Bentley (d) Ford Edge (e) Audi S5

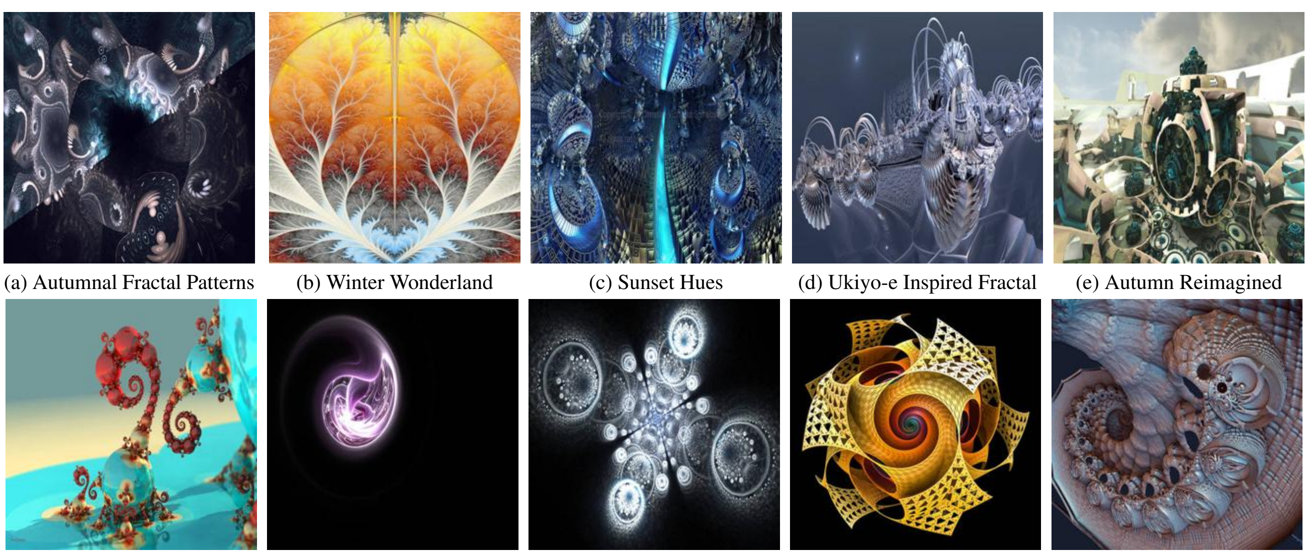
(f) Lamborghini Aurora (g) Audi R8 Rainbow (h) Bentley Autumn (i) Ford Edge Sunset (j) Audi S5 Snowy

Figure 11. **First row:** showcases original images from the Stanford Cars benchmark dataset, featuring unaltered depictions of various car models including a *lamborghini, audi R8, bentley, ford edge* and *audi S5*. **Second row:** presents the images transformed using our DIFFUSEMIX method. The effects of prompts are visible in the generated portions of the images. For example, *lamborghini* is changed to green when *aurora* prompt is applied, creating a vibrant image. The front side of *audi R8* becomes more color-rich when it is generated with *rainbow* prompt. The ambiance (background context) of *bentley* transforms significantly when *autumn* prompt is used. Similar diverse transformations are observed in other examples. These augmented images demonstrate the capability of DIFFUSEMIX in generating visually enriched augmented images for better generalization.

(a) 737-200 (b) 727-200 (c) 737-700 (d) 777-200 (e) A330-300

(f) 737-200 Sunset (g) 727-200 Autumn (h) 737-700 Snowy (i) 777-200 Ukiyo (j) A330-300 Autumn

Figure 12. Illustration of original and DIFFUSEMIX augmented Aircraft images from the FGVC-Aircraft benchmark dataset. **Top row:** presents original aircraft images, each portraying a distinct airplane including the $737-200$, $727-200$, $737-700$, $777-200$, and $A330-300$. These images highlight the design resemblance of various aircraft models, serving as a challenging resource for aircraft fined-grained image classification studies. **Bottom row:** showcases the augmented images obtained using DIFFUSEMIX for each corresponding input image. As seen, DIFFUSEMIX reimagined each aircraft with unique prompts such as *sunset, autumn, snowy* and *ukiyo* resulting in a rich visual appearance with diverse contexts. This also illustrates how image augmentation can be used to simulate different environmental and stylistic scenarios, potentially enhancing the robustness and versatility of the dataset for training robust neural networks.



(a) Autumnal Fractal Patterns (b) Winter Wonderland (c) Sunset Hues (d) Ukiyo-e Inspired Fractal (e) Autumn Reimagined

(f) Snowflake Elegance (g) Dusk's Fractal Canvas (h) East Meets West (i) Seasonal Shifts (j) Frozen Fractal Patterns

Figure 13. Some samples taken from our collected fractal dataset. Each subfigure represents a unique fractal image, demonstrating the diversity and complexity inherently present in fractal geometry.