CONTEXT-GUIDED RESPONSIBLE DATA AUGMENTATION WITH DIFFUSION MODELS

Khawar Islam, Naveed Akhtar
School of Computing and Information Systems, The University of Melbourne
Melbourne, Parkville VIC 3010, Australia
khawar.islam@student.unimelb.edu.au, naveed.akhtar1@unimelb.edu.au

ABSTRACT

Generative diffusion models offer a natural choice for data augmentation when training complex vision models. However, ensuring reliability of their generative content as augmentation samples remains an open challenge. Despite a number of techniques utilizing generative images to strengthen model training, it remains unclear how to utilize the combination of natural and generative images as a rich supervisory signal for effective model induction. In this regard, we propose a text-to-image (T2I) data augmentation method, named DiffCoRe-Mix, that computes a set of generative counterparts for a training sample with an explicitly constrained diffusion model that leverages sample-based context and negative prompting for a reliable augmentation sample generation. To preserve key semantic axes, we also filter out undesired generative samples in our augmentation process. To that end, we propose a hard-cosine filtration in the embedding space of CLIP. Our approach systematically mixes the natural and generative images at pixel and patch levels. We extensively evaluate our technique on *ImageNet-1K*, *Tiny ImageNet-200*, *CIFAR-100*, *Flowers102*, *CUB-Birds*, *Stanford Cars*, and *Caltech* datasets, demonstrating a notable increase in performance across the board, achieving up to $\sim 3\%$ absolute gain for top-1 accuracy over the state-of-the-art methods, while showing comparable computational overhead. Our code is publicly available at **DiffCoRe-Mix**.

1 INTRODUCTION

Mixup data augmentation methods (Kim et al., 2020b; Kang & Kim, 2023) are widely used to augment training data of neural models to achieve better generalization. Approaches under this paradigm devise sophisticated mechanisms to mix different images using apriori or saliency information (Qin et al., 2024; 2023; Han et al., 2022; Chen et al., 2022; Choi et al., 2022). Though effective, these techniques must overcome a critical inherent limitation of the paradigm, which requires deciding on an appropriate supervisory signal for the added augmentation samples (Islam et al., 2024a). Ambiguity in this signal can even lead to reducing model generalization instead of improving it (Azizi et al., 2023).

Recently, Diffusion Models (DMs) have shown remarkable abilities of generating high quality realistic images (Rombach et al., 2022; Liang et al., 2024; Meral et al., 2024). Conditioned on an image or text, DMs can generate multiple new images for a given class by using class-label information in their prompts. Using such images as added training data has emerged as an effective alternate to the conventional data augmentation strategy of using input transformations as the added samples (Islam et al., 2024a; Trabucco et al., 2024; Tian et al., 2024; Fu et al., 2024; Luo et al., 2023). Nevertheless, this alternative comes with its own challenges - the central problem being the inadequate control over the content of the generated images, which can lead to ineffective or even detrimental samples.

Currently, gaining better control over the generative content in DMs is emerging as an active parallel research direction (Mou et al., 2024; Shi et al., 2024; Huang et al., 2021; Xu et al., 2024). However, it is yet to focus on achieving semantic coherence and appropriate alignment with the original data samples for the purpose of data augmentation, which is still widely open. The early DM based augmentation methods (He et al.; Trabucco et al., 2024; Wang et al., 2024) mainly trusted

Table 1: Comparison of representative mixup, generative methods and generative mixup data augmentation methods.

	Mixup Methods		Generative Methods			Generative Mixup Methods	
	CoMixup	Guided-AP	Real-Guid	DA-Fusion	Diff-Mix	DiffuseMix	**DiffCoRe-Mix**
Mixing	Saliency	Saliency	—	—	—	Mask-Wise	P- & P-Wise
Prompt (P)	—	—	Label Description	Derived from Intra-Class	Derived from Inter-Class	Style Prompt	General
Negative P	—	—	—	—	—	—	✓
Contextual P	—	—	—	—	—	—	✓

the impromptu generative outputs for augmentation. Addressing this inadequacy, there are recent attempts to use image editing with DMs for augmentation (Brooks et al., 2023; Tian et al., 2024). However, these techniques largely overlook the advances in the traditional image-mixing paradigm, thereby falling short on fully exploiting them. An exception to that is (Islam et al., 2024a;b), which proposes to leverage traditional image-mixing with an image-to-image (I2I) generative model for data augmentation.

With our text prompts, we ensure an improved control over the outputs by contextualizing the prompt with the original image label. As an additional supervisory signal, we also employ Negative Prompts to restrict the generative output space of the diffusion model. Moreover, we additionally filter out any undesirable generative outputs in a fully automated manner. To that end, we develop a hard cosine filtration mechanism that is deployed in the embedding space of the CLIP-encoder. This filtration affirms appropriateness of the generative image set. We employee patch-level regularization and pixel-level sensitivity based mixing of the original and the generative images to construct the augmented data for improved model performance. Our Context-guided Diffusion based method enables Responsible image-Mixing in that the augmentation samples align well with the original data - hence termed DiffCoRe-Mix. Over the closest technique (Islam et al., 2024a), it provides a strong advantage of avoiding unrealistic or ill-formed augmentation samples - see Fig. 1, which results from foundational technical differences. Over other data augmentation methods, it provides different advantages - see Tab. 1, along with stronger performance. Our main contributions are summarized below.

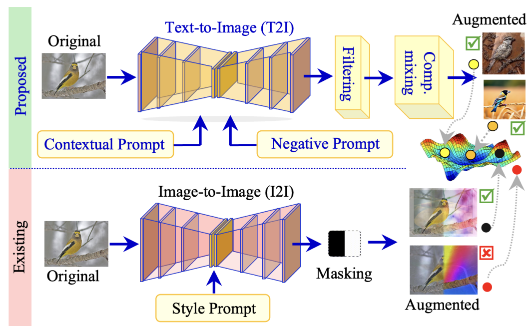

Figure 1: (**Top**) The proposed DiffCoRe-Mix employs a T2I model constrained with contextual and negative prompts. The output of the T2I model is filtered, and image-mixing is employed to introduce better generalization and robustness.(**Bottom**) The closest generative image-mixing method (Islam et al., 2024a) uses an I2I model with style prompt to edit the image by concatenating original and generative image.

- We propose T2I generative data augmentation that ensures semantic alignment of the generative image with the original image while preserving fine-grained details.
- We introduce contextual and negative prompting to ensure domain-specific generative images while restricting undesired samples, and also devise a hard cosine similarity filtration for the CLIP embedding space to further semantically align the generative images to the original samples.
- We incorporate real and generative image into pixel wise approach to reduce the memorization of neural network, and patch-wise to enhance regularization.
- We establish notable efficacy of our approach with extensive experiments on six datasets for the tasks including general classification, fine-grained classification, fine-tuning, and data scarcity; outperforming the state-of-the-art methods across the board by a considerable margin.

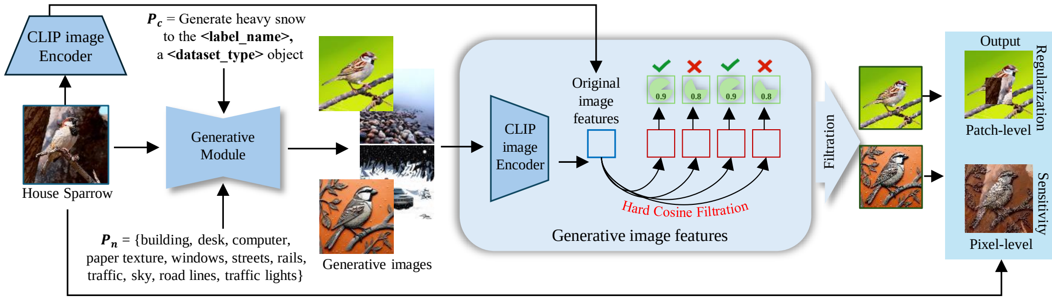

Figure 2: Overview of DiffCoRe-Mix data augmentation method. It takes an input image from dataset to generate a image guided by our contextual and negative prompts. CLIP-based image encoder is utilized to extract features from original and generative image. Then, our hard-cosine filtration approach is used to verify the semantic alignment between the original and generative features. We filter out unaligned images, and mix pixel- and patch-level real and generative images.

2 Related Work

Below, recent advances in data augmentation for vision models are discussed while focusing on the key methods related to our approach.

Mixup Methods: It is widely known that data augmentation helps in model generalization. Deep learning methods commonly apply basic transformations to inputs, e.g., rotation, flipping, to construct augmentation samples. However, more recently, dedicated approaches have emerged for a sophisticated data augmentation (Kang & Kim, 2023; Qin et al., 2024; 2023; Han et al., 2022; Chen et al., 2022; Choi et al., 2022). Among them, image-mixing is the paradigm that mixes a given input sample with other training samples or their sub-parts to create augmentation samples. For instance, Mixup (Zhang et al., 2018) mixes two random images to reduce memorization and increase generalization of classifiers. Similarly, CutMix (Yun et al., 2019) randomly cuts and pastes portions of images to improve model performance on out-of-distribution samples. ResizeMix (Qin et al., 2020) modifies CutMix (Yun et al., 2019) by resizing image sections instead of cutting them, offering a smoother blending. SmoothMix (Lee et al., 2020) makes patch boundaries smoother for better blending to improve mixing. In another line of work, methods like SaliencyMix (Uddin et al., 2020) and Attentive-CutMix (Walawalkar et al., 2020) use saliency extraction to blend the most crucial parts of the images. PuzzleMix (Kim et al., 2020a), GuidedMixup (Kang & Kim, 2023) and Co-Mixup (Kim et al., 2020b) take this a step further by isolating important regions from both source and target images. Co-Mixup (Kim et al., 2020b) introduces more sophistication by mixing three images instead of two. AutoMix (Liu et al., 2022) and SAMix (Li et al., 2021) explore the balance between hand-crafted and saliency-based mixing, breaking the process into sub-tasks. Verma et al. (Verma et al., 2019) extended the Mixup concept to the model hidden layers, mixing feature maps instead of image pixels.

Generative Augmentation Methods: Though effective, the image-mixing paradigm faces an intrinsic limitation of ambiguous supervisory signal for the augmentation samples. Contemporary generative visual models can now generate remarkable high-quality synthetic samples (Hoe et al., 2024; Qi et al., 2024; Mahajan et al., 2024; Miao et al., 2024). Leveraging that, in self-supervised learning, StableRep (Tian et al., 2024) uses generative diffusion models for representation learning and augmentation, focusing on stable representations of real-world objects. Similarly, a popular work in supervised learning, DA-Fusion (Trabucco et al., 2024) directly uses generative and real image instead of parametric transformations to augment training data. Other works (Trabucco et al., 2024; Fu et al., 2024) have also demonstrated excellent potential of diffusion models in various applications generating diverse samples for data augmentation, also considering foreground enhancement and background diversity for domain-specific concepts (Wang et al., 2024).

Mixup with Generative Models: To take advantage of both image-mixing and generative modeling, Islam et al. (Islam et al., 2024a) recently proposed mixing images with their Image-to-Image generative counterparts. However, the lose control over the generative content leads to low-quality augmentation samples in their approach - see Fig. 1. As compare to (Islam et al., 2024a), we propose employing a Text-to-Image generative model where the generative content is explicitly tailored and filtered for semantic alignment with the original data. Moreover, our method also uses more sophisticated mixing mechanisms, enabled by the high-quality generative content of our method.

3 Proposed Method

Overview: Existing use of diffusion models in visual data augmentation relies on image-to-image (I2I) generation (Islam et al., 2024a; Trabucco et al., 2024), which lacks in control over the generative content. The central motivation of our technique is to enable a better control over the generative content to align it with the original data. We achieve this as the first approach that combines text-to-image (T2I) generative modeling with the image-mixing paradigm. Figure 2 provides an overview of the proposed method. Our T2I generative model is constrained with contextual and negative prompts to generate synthetic counterparts of a sample. Encoding the original and synthetic image pairs in CLIP encoder space (Radford et al., 2021), we estimate the semantic alignment of the generated images with the original sample. This is followed by filtering out the unaligned images and augmenting the data with patch- and pixel-level mixing with generated content.

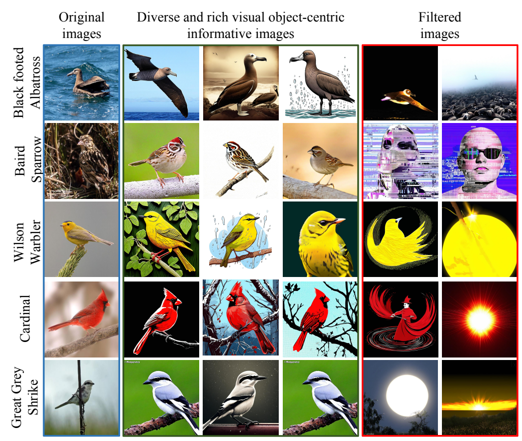

Figure 3: Representative context guided generative images. Despite strong (positive and negative) context guidance, generated images may contain a small fraction ($\sim 10\%$ as confirmed by results in § 6) of samples that semantically do not align well with the original images.

Context Guidance for Generation: In DiffCoRe-Mix, we propose to guide the generative process of a text-to-image (T2I) model with a combination of contextual prompt $\mathcal{P}_c$ and negative prompt $\mathcal{P}_n$. Herein, we term the collective guidance by $\mathcal{P}_c$ and $\mathcal{P}_n$ as contextual guidance. The standard forward diffusion process (Rombach et al., 2022; Dhariwal & Nichol, 2021) adds noise to the input image x_0 in a step-by-step manner. This process is typically modeled as a Markov chain where noise is added at each step following the Gaussian distribution as

$$p(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where x_t is the image at time step t, α_t controls the noise scale, $\mathbf{I}$ is the identity matrix and β_t is the variance of the added noise. Parameterized by a neural model, the reverse diffusion process denoises the sample, progressively generating an image from noise. In DiffCoRe-Mix the reverse process is conditioned on $\mathcal{P}_c$ as well as $\mathcal{P}_n$. We let

$$\begin{aligned} p_\theta(x_{t-1}|x_t, \mathcal{P}_c, \mathcal{P}_n) \propto \mathcal{N}\Big(&x_{t-1}; \mu_\theta(x_t, \mathcal{P}_c, \mathcal{P}_n), \\ &\Sigma_\theta(x_t, \mathcal{P}_c, \mathcal{P}_n)\Big), \end{aligned} \tag{2}$$

where $\mu_\theta(x_t, \mathcal{P}_c, \mathcal{P}_n)$ and $\Sigma_\theta(x_t, \mathcal{P}_c, \mathcal{P}_n)$ are the predicted mean and variance at step t, $\mathcal{P}_c$ guides the generated content toward the intended concept, and $\mathcal{P}_n$ specifies contents that should not be present in the generative output.

In a typical classifier-free T2I model, cross-attention (Chen et al., 2021) is used to provide context information to the reverse diffusion process. In that case, the predicted error of the image is computed as the difference between the conditional and unconditional error with a balancing term γ, i.e.,

$$\tilde{\epsilon}_\theta = (1+\gamma)\epsilon_\theta(x_t, \psi(q), t) - \gamma\, \epsilon_\theta(x_t, \psi(\varnothing), t), \tag{3}$$

where ϵ_θ is the approximation error, $\psi(q)$ computes the conditional signal for the text string q, and $\psi(\varnothing)$ denotes that the signal is computed by passing an empty string to the encoder. In our case, owing to the intended objective of explicitly aligning the generative content with the original data semantics by providing a stronger context, the predicted error is altered to the following

$$\tilde{\epsilon}_\theta = (1+\gamma)\epsilon_\theta(x_t, \psi(\mathcal{P}_c), t) - \gamma\, \epsilon_\theta(x_t, \psi(\mathcal{P}_n), t). \tag{4}$$

Using the cross-attention, our T2I model is able to additionally leverage the negative prompt $\mathcal{P}_c$ along the (positive) prompt $\mathcal{P}_c$ to provide a comprehensive context guidance to the generation process. The intuition behind targeting Eq. (4) in our method comes from the insights of Ban et al. (2024) who affirm that the negative sign of the second term in Eq. (3) encourages removal of the content pertaining to the conditional signal from the generative output.

Hard Cosine Filtration: Despite providing strong contextual guidance to the generative model, we observe semantic misalignment between the original samples and the generated outputs - see Fig. 3. Using misaligned images for data augmentation is detrimental. Hence, we devise a *hard-cosine filtration* to detect and ignore such undesired images. Let $\mathcal{I}, \mathcal{G} \subseteq \mathbb{R}^{h \times w \times c}$ be the sets of corresponding original and generated images, where h, w and c are height, width and channel dimensions. For the filtration, we first compute the semantic similarity $S(.,.)$ between the i^{th} original sample $\mathcal{I}_i$ and its generative counterpart as

$$S(\mathcal{I}_i, \mathcal{G}_i) = \frac{\psi_{\text{CLIP}}(\mathcal{I}_i) \cdot \psi_{\text{CLIP}}(\mathcal{G}_i)}{\|\psi_{\text{CLIP}}(\mathcal{I}_i)\| \|\psi_{\text{CLIP}}(\mathcal{G}_i)\|},$$

where ψ_{CLIP} denotes CLIP encoding of the image. Then, we let

$$R(\mathcal{G}_i) = \begin{cases} 1 & \text{if } S(\mathcal{I}_i, \mathcal{G}_i) > \tau, \\ 0 & \text{otherwise.} \end{cases}$$

We retain $\mathcal{G}_i$ if $R = 1$ and discard it otherwise, subsequently generating another image and consider that as potential $\mathcal{G}_i$. Here, τ is the threshold whose value is computed automatically, as discussed below. Let $\mathcal{I}^k \subset \mathcal{I}$ be a subset of the original images for the k^{th} class, where $|\mathcal{I}^k| = N_k$. We can create $C_2^{N_k}$ pairs $(\mathcal{I}_i^k, \mathcal{I}_j^k)_{i \neq j}$ of these images. Since all these pairs are between real images, we can expect a measure for their semantic similarities $S(\mathcal{I}_i^k, \mathcal{I}_j^k)_{\forall (i,j)}$ to be a reliable handle over the semantics of the class data. Hence, we compute the Expected value of the similarity as

$$\mathbb{E}[S(\mathcal{I}_i^k, \mathcal{I}_j^k)] = \frac{\sum_{i=1, j=1}^{N_k} S(\mathcal{I}_i^k, \mathcal{I}_j^k)_{i \neq j}}{C_2^{N_k}}, \tag{5}$$

and let $\tau \approx \mathbb{E}[S(\mathcal{I}_i^k, \mathcal{I}_j^k)]$. Since $C_2^{N_k}$ becomes sizable even for mildly large class, e.g., 4,950 for $N_k = 100$, we approximate the Expectation values by randomly selecting $z < C_2^{N_k}$ pairs and letting $N_k = z$ in Eq. (5).

Image Mixing: As a result of the above filtration we get a semantically well-aligned generated sample $\mathcal{G}_i$ for an original sample $\mathcal{I}_i$. We mix the two by employing a composite approach that considers both pixel-wise and patch-wise mixing. For the former, we use a mixing ratio variable λ in the range $[0, 1]$, and compute the resulting image as

$$\mathcal{M}_i^{\text{pixel}} = \lambda \mathcal{I}_i + (1 - \lambda) \mathcal{G}_i. \tag{6}$$

For patch-wise mixing, our method can be interpreted as randomly cutting a patch from $\mathcal{I}_i$ and paste it on $\mathcal{G}_i$ or do the vice versa. Concretely, the mixed image is computed as

$$\mathcal{M}_i^{\text{patch}} = \mathbf{M}_p \odot \mathcal{I}_i + (1 - \mathbf{M}_p) \odot \mathcal{G}_i, \tag{7}$$

where $\mathbf{M}_p$ is a binary mask of varying size indicating the patch to be cut from the image, $\odot$ represents element-wise multiplication. Finally, we model the selection of mixing for a given sample as a Bernoulli trial with success probability $\pi = 0.5$ due to two choices. For that, we sample $\eta \sim \text{Bernoulli}(\pi)$ and compute the mixed sample $\mathcal{M}_i$ as

$$\mathcal{M}_i = \begin{cases} \mathcal{M}_i^{\text{pixel}} = \lambda \mathcal{I}_i + (1 - \lambda) \mathcal{G}_i, & \text{if } \eta = 1 \\ \mathcal{M}_i^{\text{patch}} = \mathbf{M}_p \odot \mathcal{I}_i + (1 - \mathbf{M}_p) \odot \mathcal{G}_i, & \text{if } \eta = 0 \end{cases} \tag{8}$$

4 EXPERIMENTS

Implementation Details. Our experiments are conducted using PyTorch on NVIDIA Tesla V100 and RTX 3090Ti GPUs, with training performed in both single-GPU and distributed data-parallel

Method	ImageNet-1K		Tiny ImageNet-200		CIFAR-100	
	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
Mixup (Zhang et al., 2018)	77.03	93.52	56.59	73.02	76.84	92.42
AugMix (Hendrycks et al.)	76.75	93.30	55.97	74.68	75.31	91.62
Manifold Mixup (Verma et al., 2019)	76.85	93.50	58.01	74.12	79.02	93.37
CutMix (Yun et al., 2019)	77.08	93.45	56.67	75.52	76.80	91.91
PixMix (Hendrycks et al., 2022)	77.40	-	-	-	79.70	-
PuzzleMix (Kim et al., 2020a)	77.51	93.76	63.48	75.52	80.38	94.15
GuidedMixup (Kang & Kim, 2023)	77.53	93.86	64.63	82.49	81.20	94.88
Co-Mixup (Kim et al., 2020b)	77.63	93.84	64.15	-	80.15	-
DiffuseMix (Islam et al., 2024a)	78.64	95.32	65.77	83.66	82.50	95.41
DiffCoRe-Mix-50	**79.47**	**96.32**	**65.91**	**84.24**	**82.84**	**96.24**
DiffCoRe-Mix-100	**80.48**	**98.21**	**67.81**	**87.38**	**83.37**	**97.62**

Table 3: Comparison of Top-1 and Top-5 classification performance across three popular datasets ImageNet-1K, Tiny ImageNet-200, and CIFAR-100 using different data augmentation techniques. For ImageNet-1K, ResNet-50 is used as the backbone, while PreActResNet-18 is employed for Tiny ImageNet-200 and CIFAR-100. *-50 and *-100 variants of our method respectively augment 50% and 100% training data.

settings. The initial T2I prompt used to generate a contextual image was `Generate heavy snow to the <lab_name>, a <dataset_type> object`. We augment data based on the number of images present in each class, utilizing Cosine-Continuous Stable Diffusion XL[1] in addition to the original one. The negative prompts varied depending on the dataset type. We set the batch size to 16, and models were trained for 300 epochs using the SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. The initial learning rate was set to 0.01, which decayed by a factor of 0.1 at epochs 150 and 225.

Datasets. The used datasets are grouped into two categories of general and fined-grained classification tasks. For the general classification, we employ three popular datasets including ImageNet-1K (Deng et al., 2009) that contains diverse images, CIFAR100 (Krizhevsky et al., 2009); which is a 32x32 image size dataset and Tiny-ImageNet-200 (Le & Yang, 2015) - a 64x64 image size dataset. For the fine-grained image classification category, we employ Flower-102 (Nilsback & Zisserman, 2008) that contains 10 images per class. We use this dataset for data scarcity. We also used Stanford Cars (Krause et al., 2013), which contains 196 classes of different cars and models having different fined-grained details. We also use Birds-200-2011 (CUB) (Wah et al., 2011), which consists of 200 classes of different bird species. These datasets cover a wide range of image distribution for a comprehensive evaluation.

Baselines. To benchmark our method, we compare with SOTA methods that can be organized into three groups. (a) Image-mixing methods (Kang & Kim, 2023; Kim et al., 2020a; Yun et al., 2019; Zhang et al., 2018; Uddin et al., 2020; Huang et al., 2021; Hendrycks et al.; Verma et al., 2019). These are SOTA methods that mix source and target images in pixel-, patch- or saliency-wise manner. (b) Generative methods (He et al.; Trabucco et al., 2024; Wang et al., 2024). Instead of using basic image transformations to construct augmentation samples, they generate samples for data augmentation using generative models. (c) A method using image-mixing with generative content (Islam et al., 2024a). Conceptually, this method is closest to our approach.

Table 2: Validation and testing set performance of the DiffCoRe-Mix on Flower102 using PreActResNet34 backbone.

Method	Valid Set (%)	Test Set (%)
Mixup (Zhang et al., 2018)	66.18	61.05
CutMix (Yun et al., 2019)	62.45	56.30
SaliencyMix (Uddin et al., 2020)	63.73	58.89
PuzzleMix (Kim et al., 2020a)	66.27	60.74
Co-Mixup (Kim et al., 2020b)	65.10	59.02
Guided-AP (Kang & Kim, 2023)	62.06	55.10
DiffuseMix (Islam et al., 2024a)	67.28	60.82
DiffCoRe-Mix-50	**68.73**	**61.07**
DiffCoRe-Mix-100	**69.84**	**62.58**

[1] https://huggingface.co/stabilityai/cosxl

5 STATE-OF-THE-ART COMPARISON

General Classification: General Classification (GC) serves as a critical benchmarking task to assess the impact of data augmentation techniques on the models (Dosovitskiy et al., 2020; He et al., 2016). For GC, we evaluate DiffCoRe-Mix on ImageNet-1K (Deng et al., 2009), Tiny-ImageNet-200 (Le & Yang, 2015) and CIFAR-100 (Krizhevsky et al., 2009), highlighting the performance gains over the existing methods. Following (Islam et al., 2024a; Kang & Kim, 2023), we trained ResNet-50 on ImageNet-1K with two variants. DiffCoRe-Mix-50 augments 50% training data and DiffCoRe-Mix-100 augments 100% training data. The results are summarized in Table 3.

For Top-1 and Top-5 accuracies, our both variants clearly outperform all previous methods. On ImageNet-1K, DiffCoRe-Mix-100 achieves a very strong performance, outperforming Co-Mixup - a popular saliency-guided image-mixing method - by nearly 3% in Top-1 accuracy and by nearly 5% in Top-5 accuracy. For the DiffCoRe-Mix-50 variant, the gains are relatively low, which is intuitive. However, we still outperform DiffuseMix (Islam et al., 2024a) by an absolute 1% and 0.83% for Top-5 and Top-1 accuracies.

Table 4: Top-1 (%) performance comparison for the fine-grained visual categorization task using DenseNet121.

Method	CUB	Cars	Caltech
Mixup (Zhang et al., 2018)	74.23	89.06	91.47
CutMix (Yun et al., 2019)	74.30	88.84	91.36
SaliencyMix (Uddin et al., 2020)	68.75	88.91	90.78
PuzzleMix (Kim et al., 2020a)	77.27	90.10	91.47
Co-Mixup (Kim et al., 2020b)	77.05	90.23	90.44
Guided-AP (Kang & Kim, 2023)	77.52	90.23	91.84
DiffuseMix (Islam et al., 2024a)	77.82	90.83	92.03
DiffCoRe-Mix-50	**78.12**	**91.17**	**92.23**
DiffCoRe-Mix-100	**79.52**	**92.71**	**93.72**

On Tiny ImageNet-200, as compared to the second best performer (Islam et al., 2024a), DiffCoRe-Mix-100 gains an absolute improvement of 2.04% for Top-1 accuracy and 3.72% for Top-5 accuracy. The trends remain for the CIFAR-100 dataset. As compared to the popular Mixup baseline (Zhang et al., 2018) which also works on image-mixing principles, our notable absolute performance gains for Tiny ImageNet-200 are 11.22% and 14.36%. This highlights the benefit of using T2I models for image mixing.

Fined-Grained Visual Classification: Fine-grained visual categorization (FGVC) tasks, such as distinguishing between fined-grained categories where intra-class is similar, are inherently challenging due to the subtleness of differences between the categories (Wei et al., 2023; Tang et al., 2023). They demand models capable of extracting indistinguishable features, requiring sophisticated data augmentation techniques to enhance generalization performance (Islam et al., 2024a). In Table 4, we summarize the results for FGVC tasks on three datasets. DiffCoRe-Mix leads to across the board improvements over the SOTA methods. On the CUB dataset, known for its demanding task of bird species recognition, DiffCoRe-Mix is able to push the performance to an absolute 1.7%, showcasing strong ability to capture fine-grained details. For the Cars dataset, an absolute gain of 1.88% over SOTA and 3.06% over the popular baseline (Zhang et al., 2018) is visible.Finally, on Caltech, which encompasses a broader range of object categories, our method continues to outperform the existing method in a similar fashion.

Table 5: Top-1 (%) performance of DiffCoRe-Mix on fine-tuning experiments using ImageNet pre-trained Wide ResNet-101.

Method	CUB-Birds	Flowers-102
ResNet-50	78.61	88.26
Mixup (Zhang et al., 2018)	79.37	89.63
CutMix (Yun et al., 2019)	79.42	90.84
SaliencyMix (Uddin et al., 2020)	79.73	91.43
SnapMix (Huang et al., 2021)	79.80	91.64
DiffuseMix (Islam et al., 2024a)	80.23	93.45
DiffCoRe-Mix-50	**81.57**	**94.38**
DiffCoRe-Mix-100	**82.37**	**95.10**

Data Scarcity: Data scarcity is a prevalent challenge for deep learning, especially in the domains pertaining to fine-grained visual classification where data labeling is costly (Nilsback & Zisserman, 2008). GuidedMixup (Kang & Kim, 2023) is among the first augmentation methods to report performance for the data scarcity problem using paring algorithm. Islam et al. (Islam et al., 2024a) followed (Kang & Kim, 2023) in their comprehensive benchmarking. We also compare DiffCoRe-Mix with the SOTA methods, e.g., (Islam et al., 2024a), (Kang & Kim, 2023), (Zhang et al., 2018) under a limited data regime on the Flower102 dataset, using 10 images per class. As summarized in Table 2, DiffCoRe-Mix achieves high performance gains for this problem. As compared to the second best performer DiffuseMix (Islam et al., 2024a), our DiffCoRe-Mix-100 achieves absolute performance gains of $\sim 2.56\%$ and $\sim 1.76\%$ on the validation and test sets respectively. Even our DiffCoRe-Mix-50 variant comprehensively outperforms all existing methods.

Transfer Learning: Transfer learning enables leveraging pre-trained weights to improve model performance on new datasets with relatively smaller training sizes (Islam et al., 2024a; Kang & Kim, 2023). In our transfer learning experiments, we fine-tuned a Wide ResNet-101 model pretrained on ImageNet-1K (Deng et al., 2009) to evaluate the effectiveness of various data augmentation methods. The results are summarized in Table 5.

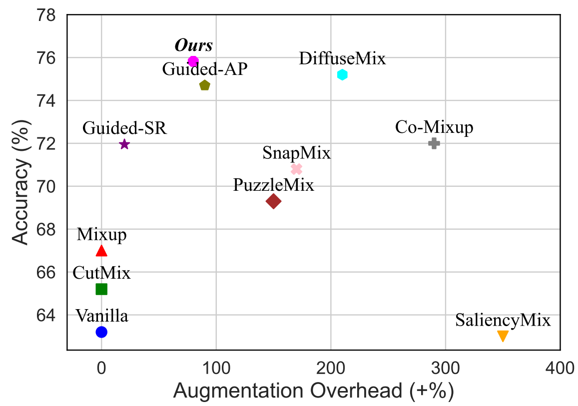

Figure 4: Augmentation overhead (+%) - accuracy (%) plot on CUB dataset with batch size 16. The closer the value to the upper left corner, the better the augmentation strategy.

Comparing DiffCoRe-Mix with the best performing exiting method DiffuseMix (Islam et al., 2024a), we observed notable gains. On the CUB-Birds dataset, DiffCoRe-Mix-50 achieves an accuracy gain of 1.34%. DiffCoRe-Mix-100 further improves the performance to 82.37%, yielding a gain of 2.14% over DiffuseMix. On the Flowers-102 dataset, DiffCoRe-Mix-50 outperforms DiffuseMix by 0.93% (94.38% vs. 93.45%), and DiffCoRe-Mix-100 extends this margin to 1.65%, reaching a Top-1 accuracy of 95.10%. These consistent improvements show that our proposed method significantly enhances the fine-tuning performance compared to the baseline approaches.

Computational Overhead: We analyze the computational overhead against the performance gain of our method and compare it with that of the SOTA methods in Fig. 4. Following (Islam et al., 2024a) and (Kang & Kim, 2023), we define the overhead $\mathcal{A}_{\mathcal{O}}$ as:

$$\mathcal{A}_{\mathcal{O}} = \frac{\mathcal{T}_{aug} - \mathcal{T}_{van}}{\mathcal{T}_{van}} \times 100(\%), \qquad (9)$$

where $\mathcal{T}_{aug}$ denotes the training time after image generation, and $\mathcal{T}_{van}$ is the training time of the baseline model (He et al., 2016) without augmentation. We generate images beforehand and use them for the rest of the training. DiffCoRe-Mix demonstrates a remarkable trade-off between performance and augmentation overhead, outperforming all other methods in accuracy while keeping overhead significantly lower than Co-Mixup (Kim et al., 2020b) and SaliencyMix (Uddin et al., 2020). We also report the training overhead with varying batch sizes in Table 6. As the results show, our method retains one of the lower overheads. This is in addition to the fact that DiffCoRe-Mix achieves considerable performance gains across the board.

Table 6: Augmentation overhead (%) for different methods with varying batch sizes.

Method	Augmentation Overhead (+%) 16	32	64
Mixup	0.9	0.6	0.4
CutMix	1.5	1.0	0.6
SaliencyMix	353.3	701.8	923.3
SnapMix	67.4	64.9	60.2
PuzzleMix	138.5	139.9	134.1
Co-Mixup	292.1	490.2	716.6
Guided-AP (Random)	87.8	81.9	70.1
Guided-AP (Greedy)	89.2	83.0	77.5
DiffCoRe-Mix	**68.2**	**59.3**	**58.9**

Saliency Visualization: To evaluate the impact of augmentation on model's attention in the decision-making process, we compare the saliency maps for DiffCoRe-Mix with the popular data augmentation methods (He et al., 2016; Uddin et al., 2020; Yun et al., 2019; Zhang et al., 2018). As shown in Fig. 5, the typical saliency maps for DiffCoRe-Mix appear more consistently around the regions of foreground object. The outputs show more concentrated activation around salient regions of the birds, especially around their distinguishing features like the head or unique color patterns. This indicates that DiffCoRe-Mix effectively emphasizes on the key features while encouraging the network to learn discriminative regions under diverse augmentations. DiffCoRe-Mix offers a balanced approach, enhancing robustness while preserving key discriminative features, making it suitable for scenarios demanding a balance between precision and generalization.

Batch Processing Time: We also compare DiffCoRe-Mix with other data augmentations methods (Zhang et al., 2018; Yun et al., 2019; Uddin et al., 2020; Kim et al., 2020b; Kang & Kim, 2023; Islam et al., 2024a) in terms of batch processing time. Results are summarized in Fig. 6. For a batch size of 16, Mixup demonstrates a faster training time of 0.39 seconds compared to DiffCoRe-Mix 1.37 seconds, indicating that mixup is more efficient for small batch training.

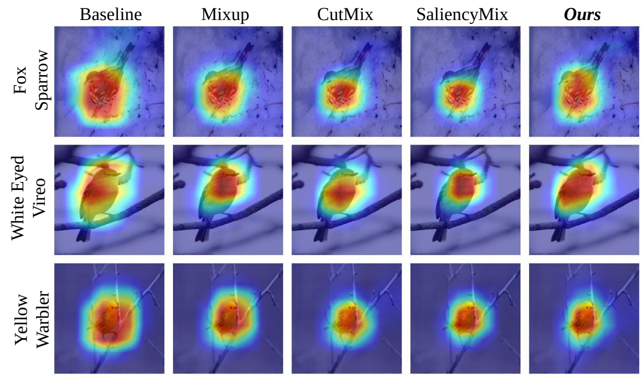

Figure 5: **Representative saliency visualizations** on original data samples. Our method guides the model to more precisely focus on the target object in the image.

However, as the batch size increases to 32, the time difference narrows, with Mixup at 0.52 seconds and DiffCoRe-Mix at 0.54 seconds. This convergence suggests that our method scales more effectively with batch size compared to Mixup. The pattern continues with a batch size of 64, where Mixup records 0.74 seconds, while DiffCoRe-Mix is slightly faster at 0.73 seconds.

6 Further Analysis and Discussions

We provide a detailed ablation study to evaluate important aspects of our proposed method and design choices.

Ablation Studies: We investigate the efficacy of pixel-wise and patch-wise design choices in our DiffCoRe-Mix. To show the effectiveness of our selection, we also report the individual performance of the two variants. All ablation studies are conducted using ResNet-50 (He et al., 2016) backbone. In Table 7, we first use ResNet-50 as the baseline, which is solely trained on traditional augmentation. It achieves a Top-1 accuracy of 85.86% and a Top-5 accuracy of 91.19% on the Stanford Cars dataset. We observe that introducing our Generative Augmentation (GenAug) alone brings 1.82% and 5.60% absolute gains for Top-1 and Top-5 accuracies.

These results are consistent with the findings in (Wang et al., 2024; Trabucco et al., 2024). Further, we individually examine the pixel-wise and patch-wise technique. By adding the pixel-wise mixing, the Top-1 accuracy increases to 88.08% and the Top-5 accuracy to 97.15%, highlighting the effectiveness of pixel-wise augmentation in improving model generalization. Similarly, in our experiment with the patch-wise approach, we again observe a slight performance improvement, where Top-1 accuracy increases to 88.92% and the Top-5 accuracy to 97.73%. Combining GenAug with pixel-wise mixing further improves the performance. This suggests a synergistic effect A similar trend is visible for the patch-wise mixing with GenAug.

Table 7: Performance comparison on Stanford Cars dataset.

Mixing Strategies	Top-1 (%)	Top-5 (%)
ResNet-50 (He et al., 2016)	85.86	91.19
R50 + GenAug	87.68	96.79
R50 + Pixel-wise	88.08	97.15
R50 + Patch-wise	88.92	97.73
R50 + GenAug + Pixel-wise	89.50	97.86
R50 + GenAug + Patch-wise	90.62	98.49
DiffCoRe-Mix-50	**91.85**	**98.83**
DiffCoRe-Mix-100	**92.74**	**99.46**

In our eventual approach, we use two variants of DiffCoRe-Mix-50 (which uses 50% generative augmentation along with original data) and DiffCoRe-Mix-100 (which utilizes 100% generative augmentation with real data). They incorporate both pixel-wise and patch-wise mixing. They achieve

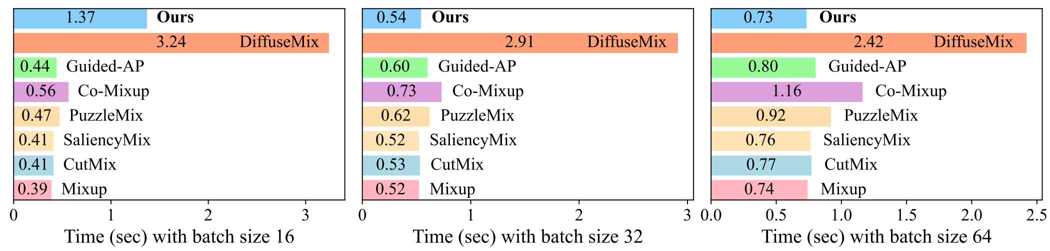

Figure 6: Comparison of batch processing time (sec.) on CUB dataset. Batch sizes of 16, 32, and 64 are used on the same hardware.

impressive performances with a Top-1 accuracy of 91.85% and a Top-5 accuracy of 98.83% for DiffCoRe-Mix-50. Our DiffCoRe-Mix-100 shows even stronger performance, achieving further absolute gains of 0.89% and 0.63% for Top-1 and Top-5 performances.

Generative Inference Computational Cost: Our method allows generating augmentation samples in varying percentages of the original data. Figure 7, summarizes the computational overhead associated with the generative inference for CUB-Birds dataset, where we vary the percentages of 512x512 augmentation samples from 10 to 100.

We also report the percentage of images automatically discarded by our filtration method. It can be observed that this percentage remains consistent in the range [10.1, 11.1]. This ensures a linear generative inference complexity incurred by our technique, which is desirable for scalability. Our total compute timings (hours) show that even 100% augmentation is fully feasible for a small dataset.

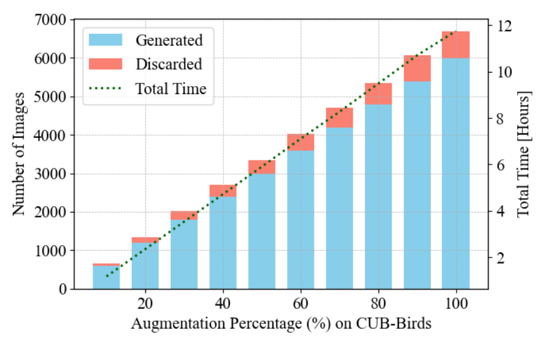

Figure 7: Ablation studies on the percentage of generative augmentation via CosXL.

7 Conclusion

In this work, we proposed DiffCoRe-Mix, a reliable T2I based data augmentation approach that employs diffusion models and various prompts to generate domain-specific class-relevant samples to increase diversity in the training dataset. Our method is intended to provide augmentation samples that are responsibly generated to align with the training data. Our technique mixes real and generative images following a systematic approach that considers both patch and pixel level mixing. On multiple tasks; such as fine-grained classification, general classification, data scarcity, finetuning, and augmentation overhead, we demonstrate notable performance gains on several benchmark datasets including ImageNet-1K, StanfordCars, Tiny-ImageNet-200, CIFAR-100, Flower-102, Caltech Birds. We also demonstrate string computational efficacy for large training batches. Moreover, our result show that the our augmentation samples lead to more precise saliency maps for the induced models.

Acknowledgments

Dr. Naveed Akhtar is a recipient of the Australian Research Council Discovery Early Career Researcher Award (project number DE230101058) funded by the Australian Government. This work is also partially supported by Google Research under Google Research Scholar Program.

REFERENCES

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.

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? arXiv preprint arXiv:2406.02965, 2024.

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, pp. 18392–18402, 2023.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 357–366, 2021.

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, pp. 12135–12144, 2022.

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.

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.

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

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. In International Conference on Learning Representations, 2020.

Yunxiang Fu, Chaoqi Chen, Yu Qiao, and Yizhou Yu. Dreamda: Generative data augmentation with diffusion models. arXiv preprint arXiv:2403.12803, 2024.

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.

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

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and XIAOJUAN QI. Is synthetic data from generative models ready for image recognition? In The Eleventh International Conference on Learning Representations.

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.

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, pp. 16783–16792, 2022.

Jiun Tian Hoe, Xudong Jiang, Chee Seng Chan, Yap-Peng Tan, and Weipeng Hu. Interactdiffusion: Interaction control in text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6180–6189, 2024.

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, volume 35, pp. 1628–1636, 2021.

Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 27621–27630, 2024a.

Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, Karthik Nandakumar, and Naveed Akhtar. Genmix: effective data augmentation with generative diffusion model image editing. arXiv preprint arXiv:2412.02366, 2024b.

Minsoo Kang and Suhyun Kim. Guidedmixup: an efficient mixup strategy guided by saliency maps. In AAAI, volume 37, pp. 1096–1104, 2023.

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, pp. 5275–5285. PMLR, 2020a.

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, 2020b.

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

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

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

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, pp. 756–757, 2020.

Siyuan Li, Zicheng Liu, Zedong Wang, Di Wu, Zihan Liu, and Stan Z Li. Boosting discriminative visual representation learning with scenario-agnostic mixup. arXiv preprint arXiv:2111.15454, 2021.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.

Zicheng Liu, Siyuan Li, Di Wu, Zihan Liu, Zhiyuan Chen, Lirong Wu, and Stan Z Li. Automix: Unveiling the power of mixup for stronger classifiers. In European Conference on Computer Vision, pp. 441–458. Springer, 2022.

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.

Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6808–6817, 2024.

Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9005–9014, 2024.

Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10844–10853, 2024.

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8488–8497, 2024.

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, pp. 722–729. IEEE, 2008.

Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8693–8702, 2024.

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

Huafeng Qin, Xin Jin, Hongyu Zhu, Hongchao Liao, Mounîm A El-Yacoubi, and Xinbo Gao. Sumix: Mixup with semantic and uncertain information. European Conference on Computer Vision (ECCV), 2024.

Jie Qin, Jiemin Fang, Qian Zhang, Wenyu Liu, Xingang Wang, and Xinggang Wang. Resizemix: Mixing data with preserved object information and true labels. arXiv preprint arXiv:2012.11101, 2020.

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, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695, 2022.

Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8839–8849, 2024.

Zhenchao Tang, Hualin Yang, and Calvin Yu-Chian Chen. Weakly supervised posture mining for fine-grained classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23735–23744, 2023.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. Advances in Neural Information Processing Systems, 36, 2024.

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In The Twelfth International Conference on Learning Representations, 2024.

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.

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, pp. 6438–6447. PMLR, 2019.

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

Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3642–3646. IEEE, 2020.

Zhicai Wang, Longhui Wei, Tan Wang, Heyu Chen, Yanbin Hao, Xiang Wang, Xiangnan He, and Qi Tian. Enhance image classification via inter-class image mixup with diffusion model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17223–17233, 2024.

Qi Wei, Lei Feng, Haoliang Sun, Ren Wang, Chenhui Guo, and Yilong Yin. Fine-grained classification with noisy labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11651–11660, 2023.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with language-guided diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9452–9461, 2024.

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.

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