NoiseCutMix: A Novel Data Augmentation Approach by Mixing Estimated Noise in Diffusion Models

Shumpei Takezaki* Ryoma Bise Shinnosuke Matsuo*
Kyushu University, Fukuoka, Japan
{shumpei.takezaki,shinnosuke.matsuo}@human.ait.kyushu-u.ac.jp

Abstract

In this study, we propose a novel data augmentation method that introduces the concept of CutMix into the generation process of diffusion models, thereby exploiting both the ability of diffusion models to generate natural and high-resolution images and the characteristic of CutMix, which combines features from two classes to create diverse augmented data. Representative data augmentation methods for combining images from multiple classes include CutMix and MixUp. However, techniques like CutMix often result in unnatural boundaries between the two images due to contextual differences. Therefore, in this study, we propose a method, called NoiseCutMix, to achieve natural, high-resolution image generation featuring the fused characteristics of two classes by partially combining the estimated noise corresponding to two different classes in a diffusion model. In the classification experiments, we verified the effectiveness of the proposed method by comparing it with conventional data augmentation techniques that combine multiple classes, random image generation using Stable Diffusion, and combinations of these methods. Our codes are available at: https://github.com/shumpei-takezaki/NoiseCutMix.

1. Introduction

In recent years, data augmentation has been widely employed to improve the performance of deep learning [11, 17, 21]. Commonly used data augmentation techniques include methods that apply perturbations to a single image, such as adding noise, rotating, and scaling.

Moreover, techniques that combine images from two different classes, such as CutMix [23] and MixUp [24], have also been proposed. CutMix augments data by cropping a random rectangle patch from one class image and pasting it onto another class image, yielding the mixed sample as shown in Figure 1(b). The augmented class label is obtained by weighting the two class labels in proportion to the patch area. Mixing images from different classes boosts data diversity and generalization. However, the unnatural boundaries between pasted regions can introduce structural inconsistencies that hinder feature learning.

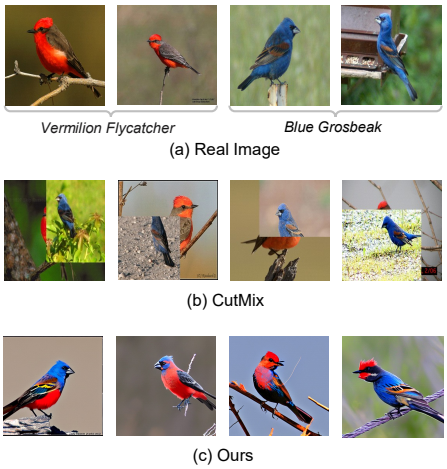

Figure 1. Comparison between images generated by CutMix and our data augmentation method.

A possible way to generate images that naturally merge the features of two source images is to use text-conditioned diffusion models [6] such as Stable Diffusion (SD) [16]. These models can synthesize natural and high-resolution images in response to class labels or text prompts, and they readily produce scenes containing multiple objects. However, when the goal is to blend two classes, prompt engineering alone does not give fine-grained control over the relative contribution of each class, so achieving the precise class ratio required by CutMix remains difficult.

Hence, in this study, we propose NoiseCutMix, which integrates the CutMix idea into the generation process of a diffusion model. During each denoising step, we replace a spatial region of the estimated noise with the noise estimated for another class, using a binary mask whose area ratio directly controls how much each class contributes. This simple mechanism inherits the high-resolution syn-

*Equal contribution

thesis ability of diffusion models while retaining the data-augmentation diversity of CutMix. As illustrated in the left-most image of Figure 1(c), our approach can generate a bird whose red breast resembles that of a "Vermilion Flycatcher" and whose blue head resembles that of a "Blue Grosbeak." NoiseCutMix yields smoother class boundaries than conventional CutMix while allowing precise control over the mixing ratio.

In the classification experiments, we compared the proposed method with conventional data augmentation techniques that combine images from multiple classes, such as CutMix and MixUp, random image generation using Stable Diffusion, and the application of conventional data augmentation to the randomly generated images. We conducted evaluations using three datasets to verify the effectiveness of the proposed method. In particular, we confirmed that the images generated by our method effectively fuse features from different classes while reducing unnaturalness at the boundaries.

2. Related Work

Data augmentation by single-image perturbations, e.g., random rotations, rescaling, color shifts, or added noise, has been studied extensively [11, 17, 21]. These low-cost tricks help combat overfitting, but they change appearance rather than data structure, so the accuracy gains are modest.

Stronger techniques mix two images from different classes [23, 24]. MixUp [24] linearly blends the two images, whereas CutMix [23] pastes a rectangular patch from one onto the other. Although such cross-class mixing improves diversity and generalization, the pasted regions can look unnatural, and the resulting structural mismatch may hurt feature learning.

Recently, several data augmentation methods leveraging diffusion models have been proposed [1, 25]. Fine-tuning-based approaches [2, 18, 22] retrain Stable Diffusion on the data used to train the classifier, then use its generated images as data augmentation. Image-editing-based approaches [8, 18, 20] make direct edits to real images using Stable Diffusion to enhance dataset diversity. Methods that mix real and diffusion-generated images for a single class with fractal blending have been proposed [9, 10].

Our method, NoiseCutMix, enables the generation of natural and diverse images that fuse features from two different classes by leveraging the diverse data augmentation properties of CutMix within the generation process of Stable Diffusion. Furthermore, our approach is effective yet simple, making it easy to plug into existing Stable Diffusion-based data augmentation methods.

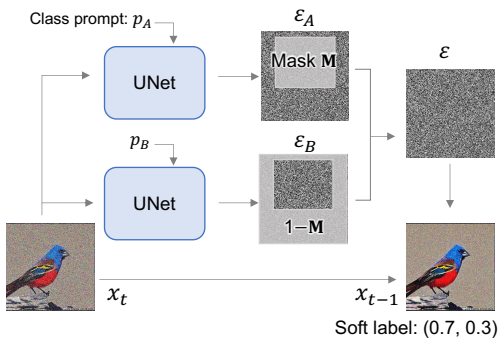

Figure 2. The proposed *NoiseCutmix* mixes estimated noise of diffusion models in the denoising process.

3. NoiseCutMix: Data Augmentation by Mixing Estimated Noise in Diffusion Models

We propose NoiseCutMix, a novel data augmentation method that leverages a pre-trained diffusion model, such as Stable Diffusion, to generate highly natural and diverse samples of a desired class, conditioned on text prompts. As shown in Figure 2, NoiseCutmix blends the estimated noise from two different classes in a CutMix-like manner during the denoising process of a diffusion model. This fusion inherits two complementary strengths: (i) diffusion models generate high-resolution, natural images, while (ii) CutMix synthesizes diverse images by mixing features from different classes. Consequently, NoiseCutMix achieves augmented images that combine realism with variety.

3.1. Mixing Estimated Noise of Diffusion Models

The proposed method mixes the noise estimates of two class prompts during the reverse denoising process of a diffusion model to generate images that blend the visual traits of both classes. In a standard class-conditioned diffusion model, an image of a single target class is obtained by iteratively removing noise from random Gaussian noise while conditioning the UNet on a text prompt p. In contrast, our method takes two class prompts, p_A and p_B corresponding to class A and B, predicts their step-t noises, and mixes them to generate an image that fuses the features of both classes.

Figure 2 shows the overview of the proposed method at an arbitrary reverse step t. Given the current image x_t, the Unet f outputs two noise estimates, $\varepsilon_A = f(x_t, p_A, t)$ and $\varepsilon_B = f(x_t, p_B, t)$. We then combine them with a binary mask $\mathbf{M} \in \{0,1\}^{W \times H}$ sampled uniformly at random over a rectangular region[1]:

$$\varepsilon = \mathbf{M} \odot \varepsilon_A + (\mathbf{1} - \mathbf{M}) \odot \varepsilon_B, \tag{1}$$

where $\odot$ denotes element-wise multiplication, and $\mathbf{1}$ is a binary mask of all ones.

[1] W and H are the width and height of an image (or latent) features in a diffusion model

The mixed noise ε is used to denoise the current image x_t into x_{t-1}, which is fed to the next reverse step. Through this process, the final image inherits characteristics from both classes. Details of how $\mathbf{M}$ is generated are described later.

3.2. Class Labels for Generated Images

Following the CutMix [23], the class label $\tilde{\mathbf{y}}$ (expressed in one-hot format) for the generated image is computed according to the region ratio λ of the mask $\mathbf{M}$:

$$\tilde{\mathbf{y}} = \lambda \mathbf{y}_A + (1-\lambda)\mathbf{y}_B, \tag{2}$$

where λ is sampled from a Beta distribution $\mathrm{Beta}(\alpha, \alpha)$ with a hyperparameter α, as in CutMix. The vectors $\mathbf{y}_A$ and $\mathbf{y}_B$ denote the one-hot vector corresponding to class A and B, respectively.

3.3. Binary Mask for Mixing Estimated Noise

We create the binary mask $\mathbf{M}$ for mixing estimated noise ε_A and ε_B by sampling a rectangular region with an aspect ratio proportional to the image (or latent) resolutions. The box coordinates of the mask are uniformly sampled according to a uniform distribution:

$$r_x \sim \mathrm{Unif}(0, W), \quad r_w = W\sqrt{1-\lambda}, \tag{3}$$

$$r_y \sim \mathrm{Unif}(0, H), \quad r_h = H\sqrt{1-\lambda}. \tag{4}$$

Here, (r_x, r_y, r_w, r_h) denotes the position and size of the rectangle, and the corresponding region is set to 0 in a mask $\mathbf{M}$. This sampling makes the cropped ratio $\frac{r_w r_h}{WH} = 1-\lambda$.

4. Experiments

4.1. Comparison Methods

To evaluate the effectiveness of the proposed method, we compared it with the following five methods.

CutMix [23]: A data augmentation method that replaces part of an image with a region from another image. Specifically, a rectangular region is sampled at random and cut out, then pasted onto the corresponding region of a different image, thus blending two images. The label is computed as $\tilde{\mathbf{y}} = \lambda \mathbf{y}_A + (1-\lambda)\mathbf{y}_B$ based on the proportion of the cut region. We set the Beta distribution parameter α to 1.0 for sampling λ and applied augmentation probability with 0.5.

MixUp [24]: A method that linearly interpolates two images at the pixel level. Specifically, two images are blended in proportion to λ, and the label is also interpolated as $\tilde{\mathbf{y}} = \lambda \mathbf{y}_A + (1-\lambda)\mathbf{y}_B$. We set the Beta distribution parameter α to 0.2 for sampling λ and also applied augmentation probability with 0.5.

SD-random: Augment the dataset by randomly generating images with Stable Diffusion [16]. Specifically, we used the text prompt "a photo of a (*class name*)" for class condition.

SD-random + CutMix/MixUp: We took the images generated by SD-random above and applied CutMix or MixUp to them. In other words, this is a simple combination of Stable Diffusion augmentation with the conventional data augmentation methods.

4.2. Datasets

We used the following three datasets to evaluate our proposed method. During training, 20% of the training images were randomly selected as validation data.

CUB [19]: Caltech-UCSD Birds (CUB) is a dataset for detailed bird classification, consisting of 11,788 images across 200 bird species. The split is 5,994 images for training and 5,794 for evaluation.

Flower [15]: Oxford Flowers is a dataset for detailed flower classification, including 8,189 images of 102 flower types. The split is 6,149 images for training and 2,040 for evaluation.

Aircraft [14]: FGVC-Aircraft is a dataset for detailed airplane classification, comprising 10,000 images of 102 airplane classes. The split is 6,667 images for training and 3,333 for evaluation.

4.3. Implementation Details

We used a ResNet50 [4] pretrained on ImageNet as the classifier. The batch size was set to 64, the learning rate to 0.001, and we used Adam [12] as the optimizer. The number of epochs was set to 100, and we adopted the parameters from the epoch with the highest accuracy on the validation set for evaluation.

Stable Diffusion v1.5 (SD) was employed as the diffusion model, and we used the text prompt "a photo of (*class name*)" for a class condition. We also used Classifier-free guidance [5] with a guidance scale of 7.5. Additionally, we employed DPMsolver++ [13] as the denoising sampler and set the inference steps to 25.

We performed fine-tuning for SD on each target dataset, following prior work [20], which references LoRA [7] and Textual Inversion [3].[2] In our experiment, we generated the same number of augmented images as the original dataset (i.e., 100% of the dataset size).

5. Experimental Results

5.1. Quantitative Evaluation

Table 1 shows the classification results on the three datasets. "Original" refers to the result without any advanced data augmentation. We observe that our proposed data augmentation method outperforms conventional methods that combine multiple images (CutMix, MixUp) on CUB and Flower. In particular, we see improvements of 2.27% on

[2] In practice, we used publicly available fine-tuned weights: https://github.com/Zhicaiwww/Diff-Mix

Table 1. Quantitative Evaluation: Accuracy [%] on three datasets. We show the mean and standard deviation over five trials.

Method	CUB [19]	Flower [15]	Aircraft [14]
Original	67.78 (± 1.39)	91.71 (± 1.09)	81.69 (± 0.90)
CutMix [23]	66.51 (± 1.20)	89.93 (± 0.75)	**83.28 (± 0.32)**
MixUp [24]	67.57 (± 0.88)	90.64 (± 0.52)	82.71 (± 0.84)
SD-random [16]	67.60 (± 0.87)	92.52 (± 1.26)	78.27 (± 0.33)
SD-random + CutMix	62.90 (± 0.70)	87.38 (± 1.01)	77.78 (± 0.83)
SD-random + MixUp	64.57 (± 0.63)	89.15 (± 0.72)	78.36 (± 0.52)
Ours	**68.78 (± 0.33)**	**92.91 (± 0.32)**	79.69 (± 0.62)

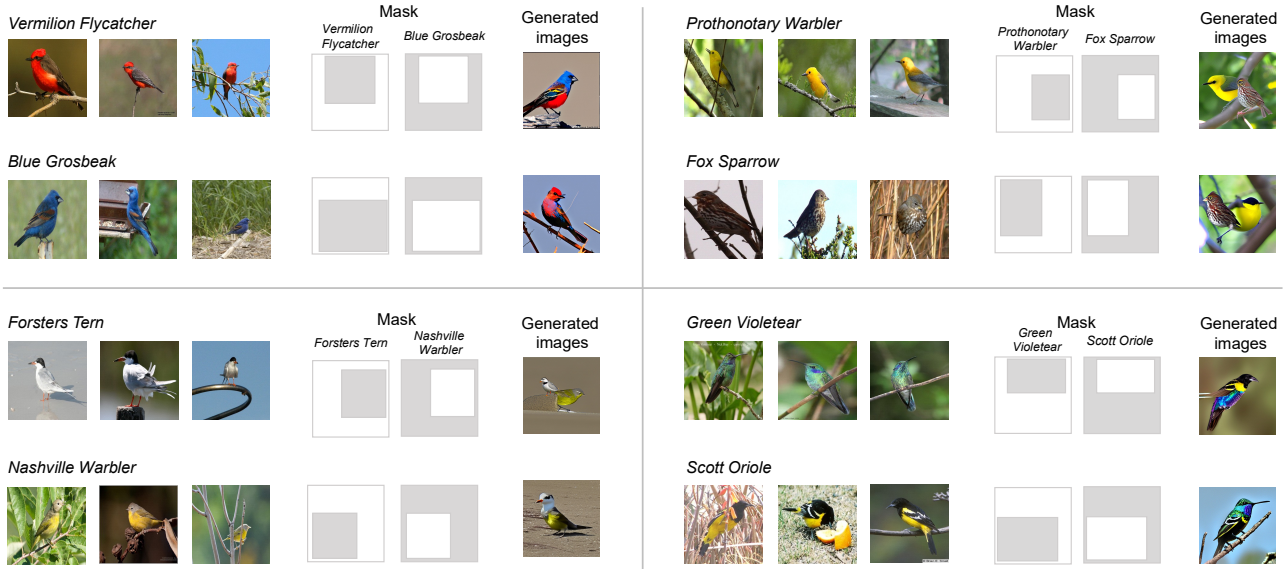

Figure 3. Examples from the CUB dataset: real images, images generated by our data augmentation method, and the masks used.

CUB and 2.98% on Flower over CutMix. This is because our method mixes the estimated noise of two classes at each denoising step of Stable Diffusion, resulting in more natural boundaries between the combined images compared to CutMix. For Aircraft, neither NoiseCutMix nor SD-random improved accuracy because Stable Diffusion failed to match the dataset distribution, already degrading the original baseline. As NoiseCutMix depends on SD, no improvements were observed.

We also confirmed that our proposed method outperforms pure random image generation using Stable Diffusion (SD-random) on all datasets. This implies that simply generating random images with SD is less effective than merging multiple classes to increase data diversity. Moreover, our method also surpasses SD-random+CutMix, indicating that simply combining SD-generated images with conventional CutMix is insufficient; partial merging of noise estimates is crucial.

5.2. Qualitative Evaluation

Figure 3 shows real images, images generated by our proposed method, and the masks used, for four different class label pairs from the CUB dataset. For instance, the top-left example applies our data augmentation to classes "Vermilion Flycatcher" and "Blue Grosbeak." These observations indicate that our method can generate images naturally blending features from two classes.

6. Conclusion, Limitation, and Future Work

NoiseCutMix blends estimated class-conditioned noise within diffusion models to synthesize images that coherently fuse two classes while preserving realism. Experiments show consistent gains over standard augmentation. A key limitation is dependence on Stable Diffusion (SD) aligning with the target data distribution, which can be mitigated by restricting training to in-distribution samples. As future work, our approach, being both effective and simple, could be integrated into existing Stable Diffusion-based data augmentation pipelines.

Acknowledgements: This work was supported by JSPS KAKENHI Grant Number JP23KJ1723, JP24KJ1805, and JP25K22846, and JST ACT-X Grant Number JPMJAX23CR, and ASPIRE Grant Number JPMJAP2403.

References

[1] Panagiotis Alimisis, Ioannis Mademlis, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, and Georgios Th Papadopoulos. Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions. *Artificial Intelligence Review*, pages 1–55, 2025. 2

[2] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *Transactions on Machine Learning Research*, 2023. 2

[3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023. 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *Advances in Neural Information Processing Systems Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 1

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[8] Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *Computer Vision and Pattern Recognition*, pages 27621–27630, 2024. 2

[9] Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *Computer Vision and Pattern Recognition*, pages 27621–27630, 2024. 2

[10] 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*, 2024. 2

[11] Nour Eldeen Khalifa, Mohamed Loey, and Seyedali Mirjalili. A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artificial Intelligence Review*, pages 2351–2377, 2022. 1, 2

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. 3

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

[14] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 3, 4

[15] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 3, 4

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 3, 4

[17] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 1, 2

[18] Brandon Trabucco, Kyle Doherty, Max A Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *International Conference on Learning Representations*, 2024. 2

[19] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200 dataset. Technical report, California Institute of Technology, 2011. 3, 4

[20] 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 *Computer Vision and Pattern Recognition*, pages 17223–17233, 2024. 2, 3

[21] Suorong Yang, Weikang Xiao, Mengchen Zhang, Suhan Guo, Jian Zhao, and Furao Shen. Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*, 2022. 1, 2

[22] Jianhao Yuan, Jie Zhang, Shuyang Sun, Philip Torr, and Bo Zhao. Real-fake: Effective training data synthesis through distribution matching. In *International Conference on Learning Representations*, 2024. 2

[23] 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 *International Conference on Computer Vision*, pages 6023–6032, 2019. 1, 2, 3, 4

[24] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 1, 2, 3, 4

[25] Yue Zhou, Chenlu Guo, Xu Wang, Yi Chang, and Yuan Wu. A survey on data augmentation in large model era. *arXiv preprint arXiv:2401.15422*, 2024. 2