

Exploring the Impact of Dataset Characteristics on CutMix Augmentation Effectiveness

Sung-Hoon Kim

AIFFEL Research 13th

Abstract

Data augmentation is a crucial technique that helps models learn more generalized representations from limited data. To this end, various methods have been proposed, including region-based (also known as patch-based) augmentation strategies. CutMix is widely considered one of the most effective region-based augmentation methods. While CutMix has shown remarkable performance improvements across various benchmarks, it is not universally beneficial. In this study, we investigate the conditions under which CutMix leads to performance gains, as well as scenarios where it may hinder learning. Our findings suggest that the effectiveness of CutMix is highly dependent on the characteristics of dataset.

1. Introduction

1.1. The role of Data Augmentation

Convolutional neural network (CNN) classifiers generally exhibit better performance when trained on large-scale datasets. However, when only limited data is available, models tend to overfit to specific features of the classes in the training set, resulting in poor generalization to unseen data. Data augmentation serves as a key technique to mitigate this overfitting issue and enhance the model's generalization capability. Accordingly, this study aims to investigate whether the augmentation method CutMix can effectively fulfill its intended role of improving generalization across different dataset characteristics.

1.2. CutMix Method

Among various data augmentation techniques, region-based augmentation methods manipulate specific regions of an image—such as by removing, mixing, or replacing them. These methods help reduce the model's reliance on partial features and encourage learning a broader set of discriminative cues, thereby providing a regularization effect. Among such approaches, **CutMix** has been widely recognized for its effectiveness.

CutMix generates augmented samples by cutting a patch from one image and pasting it onto another. The corresponding labels are also mixed in proportion to the area of the patch, encouraging the model to learn from more diverse and informative regions.





	ResNet-50	Mixup	Cutout	CutMix
Image				
Label	Dog 1.0	Dog 0.5 Cat 0.5	Dog 1.0	Dog 0.6 Cat 0.4
ImageNet Cls (%)	76.3 (+0.0)	77.4 (+1.1)	77.1 (+0.8)	78.6 (+2.3)
ImageNet Loc (%)	46.3 (+0.0)	45.8 (-0.5)	46.7 (+0.4)	47.3 (+1.0)
Pascal VOC Det (mAP)	75.6 (+0.0)	73.9 (-1.7)	75.1 (-0.5)	76.7 (+1.1)

Table 1: Overview of the results of Mixup, Cutout, and CutMix on ImageNet classification, ImageNet localization, and Pascal VOC 07 detection (transfer learning with SSD finetuning) tasks. Note that CutMix significantly improves the performance on various tasks.

1.3. Motivation and Research Objectives

While the effectiveness of CutMix is evident—as demonstrated in Table 1—it does not always yield positive results. In fact, there have been multiple cases, particularly observed through the Aiffel research classes, where CutMix hinders training and even degrades model performance.

This study aims to investigate the conditions under which CutMix enhances performance, as well as the conditions that lead to negative outcomes. In particular, we focus on how the characteristics of the dataset influence the effectiveness of CutMix.

By doing so, we hope to prevent counterproductive and inefficient outcomes that may arise from the inappropriate use of CutMix. Furthermore, our findings are expected to contribute to the development of more informed and effective strategies for applying augmentation techniques.

2. Method

2.1. Datasets

In this study, the dataset itself serves as the primary experimental variable. We conducted our experiments on two datasets: Stanford Dogs and Food-101. Datasets such as CIFAR and ImageNet, where the effectiveness of CutMix has already been widely demonstrated, were excluded from this study.

Stanford Dogs requires distinguishing between classes based on fine-grained visual features. Since all classes represent dog breeds, they share many common attributes, and the discriminative cues are often limited to very small regions such as the face or even smaller parts. We aim to investigate whether CutMix, which overlays patches between images, disrupts learning by obscuring these key features, or instead encourages the model to learn from a broader set of features, thereby improving performance.

Food-101, like Stanford Dogs, also has many classes that look similar to each other. However, the features that help distinguish between classes are not limited to small areas like the face or eyes. We chose these two datasets because they are both fine-grained, but they differ in how the important features for classification are spread across the images.

2.2. Architecture

We applied the ResNet-50 architecture for our experiments. Since the performance of CutMix reported in Table 1 was based on ResNet-50, we used the same architecture to maintain consistency. ResNet is widely recognized as a standard benchmark model for image classification, and given that our experiments also focus on classification tasks, ResNet-50 is an appropriate choice to validate the effectiveness of CutMix. By using the same model architecture as the original paper, we ensure that any differences in results are not due to architectural variations.

2.3. Experimental Setup

In this experiment, the patch size used in CutMix was determined according to the following procedure:

$$\begin{aligned} r_w &\sim \mathcal{U}(0, 1), \quad r_h \sim \mathcal{U}(0, 1) \\ W &= \left\lfloor W_{\text{img}} \times \sqrt{1 - r_w} \right\rfloor \\ H &= \left\lfloor H_{\text{img}} \times \sqrt{1 - r_h} \right\rfloor \end{aligned}$$

r is a random real number sampled from a uniform distribution between 0 and 1. W_{img} and H_{img} represent the width and height of the original image, respectively. The width and height of the patch are calculated by multiplying the original dimensions by $\sqrt{1-r}$ and then taking the integer part.

The location of the patch is also randomly determined. A random center point (x, y) is selected within the image dimensions, and the patch is positioned around this point. To ensure the patch does not exceed the image boundaries, its coordinates are clipped as follows:

$$\begin{aligned} x_{\min} &= \max(0, x - \frac{W}{2}) \\ y_{\min} &= \max(0, y - \frac{H}{2}) \\ x_{\max} &= \min(W_{\text{img}}, x + \frac{W}{2}) \\ y_{\max} &= \min(H_{\text{img}}, y + \frac{H}{2}) \end{aligned}$$

This ensures that the entire patch remains within the bounds of the image, regardless of the randomly chosen size and position.

The learning rate was set to 0.01, and the optimizer used was stochastic gradient descent (SGD), following the original CutMix paper. No learning rate decay was applied. All experiments were conducted using TensorFlow version 2.10.1.

2.4. Experiment

We trained and compared models on datasets augmented with four different strategies:

No Augmentation: This serves as the baseline model trained on data without any augmentation.

Basic Augmentation: To provide a reference for augmentation effects, basic data augmentations including random left right flipping and random brightness adjustment were applied.

Mixup: This augmentation technique blends images and labels similarly to CutMix. Previous studies have reported that CutMix overcomes some limitations of Mixup. In this study, Mixup was included to observe under which conditions such similar augmentation methods might negatively impact performance. Basic Augmentation was also applied alongside Mixup.

CutMix: Models were trained on data augmented using the CutMix method described in Section 2.3. Basic Augmentation was applied in conjunction.

2.5. Metric

In classification tasks, common evaluation metrics include Accuracy and F1 Score. The F1 Score is particularly useful to address the issue of skewed class distributions, where Accuracy can be misleading. However, since the datasets used in this study have balanced class distributions, we adopt Accuracy as the sole evaluation metric for its simplicity and interpretability.

3. Result

3.1. Stanford Dogs dataset

In the Stanford Dogs dataset, each model was trained for 50 epochs.



Figure 1: The change in accuracy on both the training and validation datasets during the training process for each augmentation method in Stanford Dogs Dataset.

The No Augmentation model (*No Aug*) and the Basic Augmentation model (*Basic Aug*) produced similar results. Both models reached 100% training accuracy in less than 10 epochs. However, no improvement in validation accuracy was observed after achieving this level of training accuracy. The gap between training and validation accuracy remained substantial, with a difference of approximately 25 percentage points or more that was not narrowed throughout the training process.

Similarly, the Mixup and CutMix models exhibited comparable performance to each other. For both models, validation accuracy plateaued after approximately 20 epochs, with no further improvement observed thereafter. The gap between training and validation accuracy also remained substantial. Notably, the performance of both Mixup and CutMix was significantly lower than that of the *No Aug* and *Basic Aug* models.

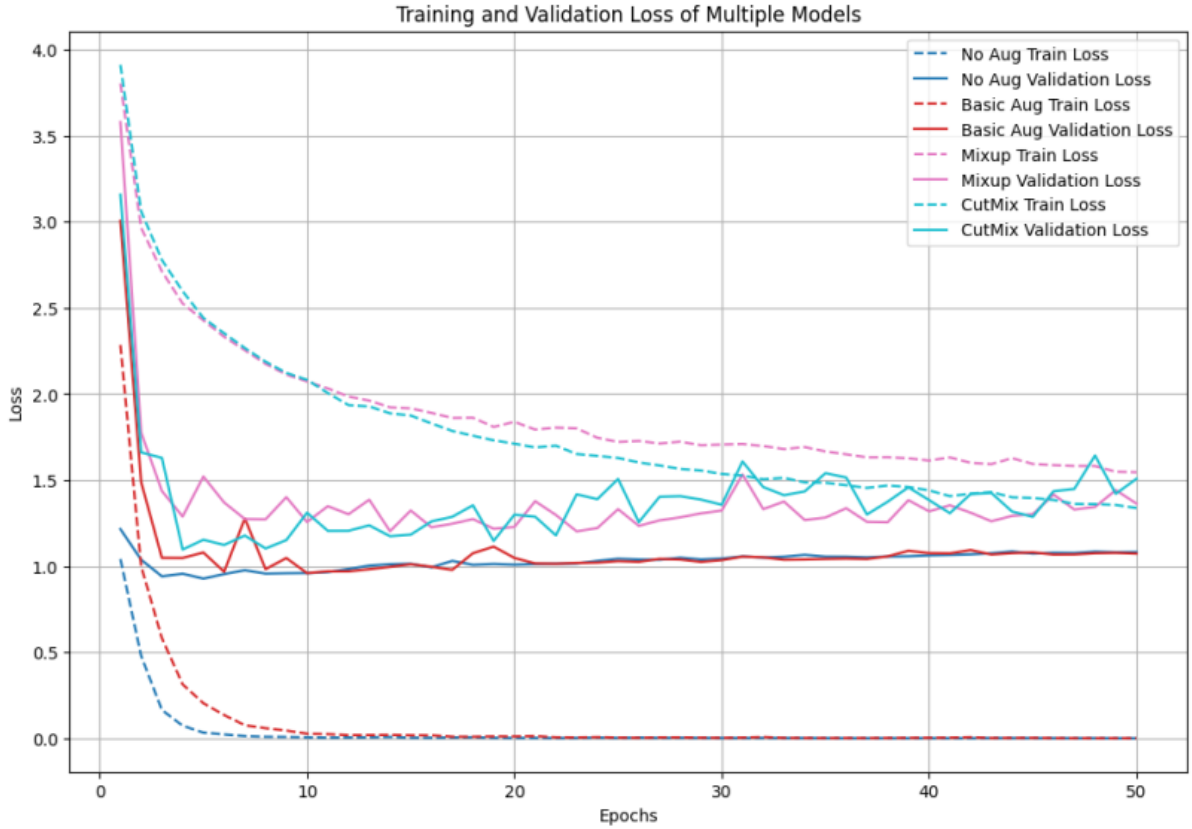


Figure 2: This shows the change in loss on both the training and validation datasets during the training process for each augmentation method in Stanford Dogs Dataset.

For all models, validation loss began to increase before the 10th epoch, indicating early signs of overfitting. In the case of this dataset, the CutMix strategy failed to mitigate overfitting or improve the model's generalization performance. On the contrary, it appears to have negatively impacted the training process.

3.2. Food-101 dataset

For the Food101 dataset, each model was trained for a total of 15 epochs. This decision was based on the observation that, in the Stanford Dogs dataset, performance plateaued after 15 epochs, making further training less meaningful.

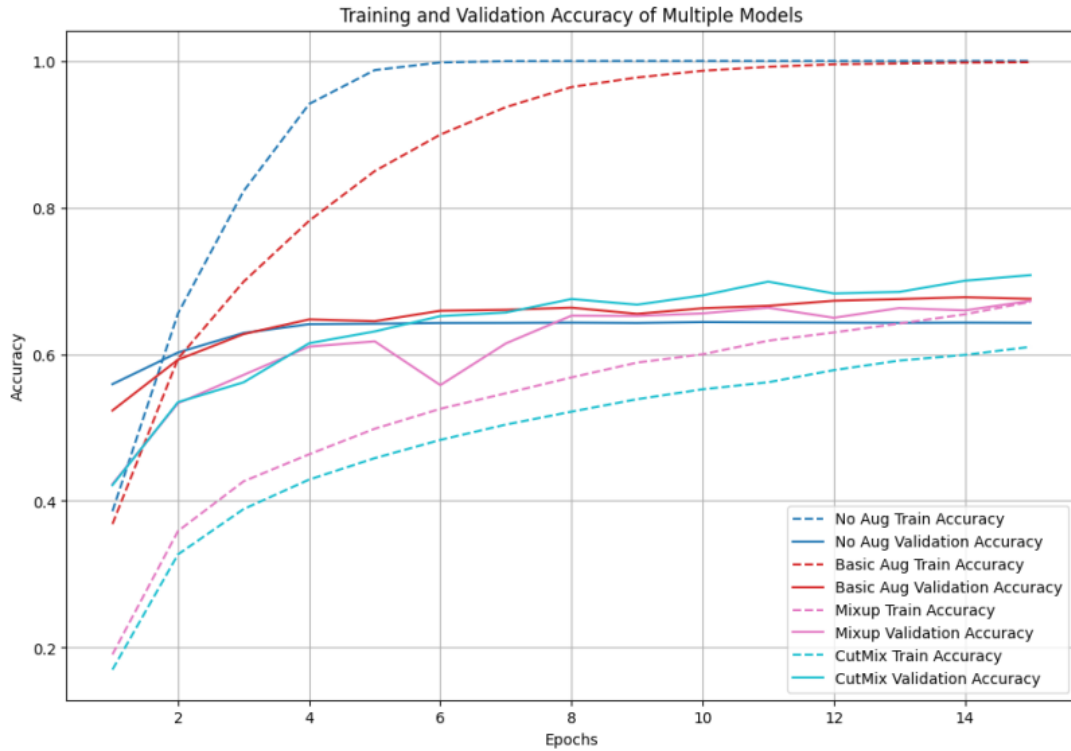


Figure 3: The change in accuracy on both the training and validation datasets during the training process for each augmentation method in Food-101 Dataset.

A noteworthy finding is observed in Figure 3: the models trained with Mixup and CutMix exhibit almost no gap between training and validation accuracy. In fact, the training accuracy remained relatively low until epoch 13. This suggests that these augmentation strategies effectively prevented overfitting and contributed to better generalization performance.

Meanwhile, the *No Aug* and *Basic Aug* models showed similar training patterns in both the Stanford Dogs and Food-101 datasets. In contrast, Mixup and CutMix yielded opposite results between the two datasets, which indicates that their effects are not merely a consequence of dataset characteristics but rather reflect a fundamental influence on the model's learning behavior.

Furthermore, the accuracy of the Mixup and CutMix models was either comparable to or higher than that of the *No Aug* and *Basic Aug* models, providing clear evidence of performance improvement.

In the case of *No Aug* and *Basic Aug*, the gap between training and validation accuracy remained notably large—over 30% point. In contrast, the Mixup and CutMix models showed almost no such gap, further supporting their effectiveness in enhancing model generalization.

As shown in Figure 4, the *No Aug* and *Basic Aug* models begin to exhibit increasing validation loss after approximately the third epoch, similar to what was observed in the Stanford Dogs dataset. This indicates that both models enter an overfitting phase early in the training process, limiting their potential for further performance improvement.

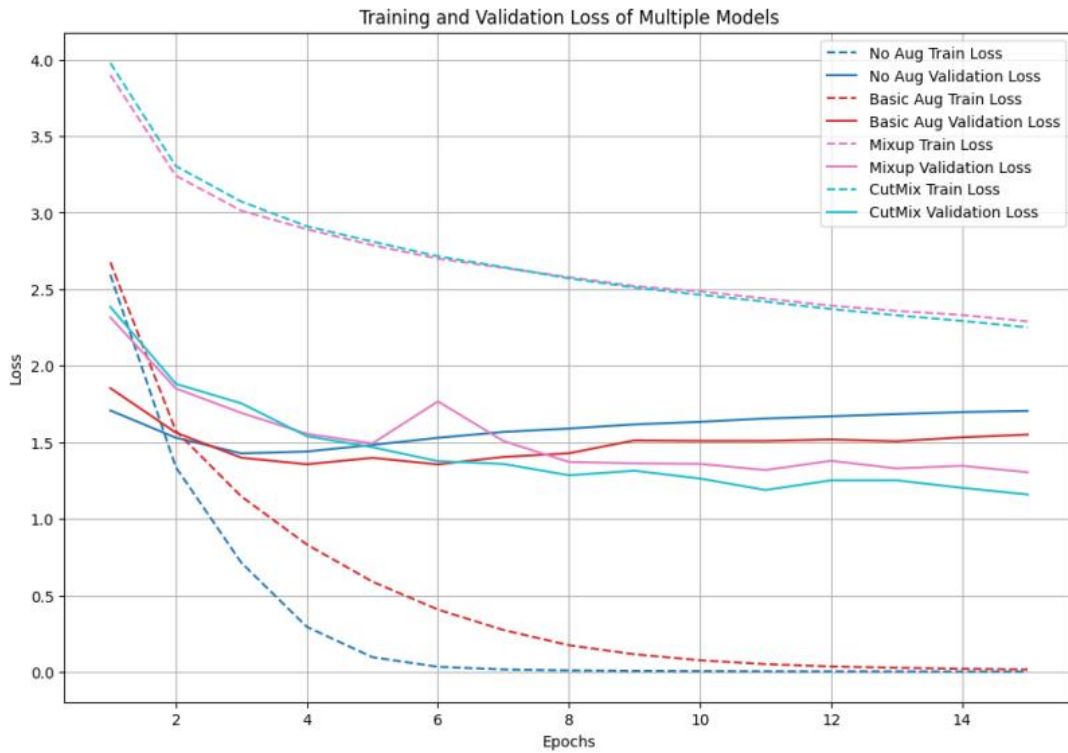


Figure 4: This shows the change in loss on both the training and validation datasets during the training process for each augmentation method Food-1010 Dataset.

In contrast, the Mixup and CutMix models show a consistent decline in validation loss throughout the training period. This suggests that these models are still in an underfitting state and thus retain room for further performance enhancement with additional training.

Unlike in the Stanford Dogs dataset, the CutMix method in the Food-101 dataset effectively mitigated overfitting and contributed to improved generalization performance.

4. Discussion

In order to further explore the causes of the differing results between datasets, we investigate the characteristics of each dataset.

4.1. Limitations of CutMix in the Stanford Dogs Dataset

The Stanford Dogs dataset consists entirely of dog breeds, which inherently share many visual characteristics. As a result, distinguishing between classes often relies on subtle and localized features. In such scenarios, small regions of the image may carry the most discriminative information for classification.



Figure 5: Two images of two different classes. Small parts of the images are occluded.

Figure 5 illustrates this phenomenon. The two images belong to different classes, yet they are visually similar. Even with only a small part of the image occluded, it becomes significantly more difficult to differentiate the two breeds—especially for non-experts. This suggests that critical class-specific features, such as the eyes, play a key role in classification. When CutMix occludes or replaces these key regions, the model may struggle to learn effective discriminative features.

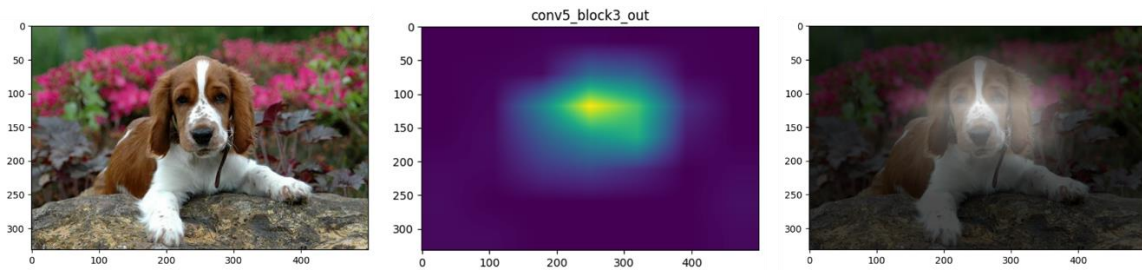


Figure 6: Class Activation Map of a Stanford Dogs dataset sample.

Figure 6 presents another example from the Stanford Dogs dataset. We visualized the features that the *No Aug* model focuses on during classification using a Class Activation Map (CAM). The results indicate that the model primarily attends to the dog’s face—particularly the eyes and ears. This suggests that the model relies heavily on a very limited region of the image for its classification decision.

In this context, the application of CutMix may introduce significant challenges. Since CutMix replaces a portion of an image with a patch from another class, it may obscure or remove critical visual features such as the eyes or ears—regions that are essential for distinguishing between fine-grained dog breeds. As a result, the model may fail to learn the subtle differences necessary for accurate classification. This suggests that for datasets like Stanford Dogs, where intra-class similarity is high and discriminative features are localized, CutMix may hinder rather than help the learning process.

4.2. Effectiveness of CutMix in Food-101 dataset.

In the case of the Food-101 dataset, the discriminative features of each class are not limited to a specific localized region but are instead observed repeatedly across various parts of the image. As a result, even when a portion of an image is replaced via CutMix, there is a high likelihood that key features necessary for classification remain intact. This suggests that CutMix can be effective in preserving or even enhancing generalization performance in this dataset. This is further supported by the visualization in Figure 7.



Figure 7: Carbonara image samples with regions replaced via CutMix

Even when portions of a carbonara image are replaced using the CutMix method, key visual features representative of the carbonara class often remain intact. A comparable example can be found with pizza. Even if seven out of eight slices are removed from a whole pizza, the remaining single slice is usually sufficient to identify it as pizza. This suggests that the visual cues defining the "pizza" class are not confined to a small, localized region but are instead distributed and repeated throughout the image.

Consequently, region-based augmentation methods such as CutMix can guide the model to learn more robust and generalized representations in datasets like Food-101. This helps prevent overfitting and contributes to improved performance.

4.3. Common Characteristics of Datasets Where CutMix Is Effective

In the original CutMix paper, its effectiveness was demonstrated on datasets such as CIFAR-10, CIFAR-100, and ImageNet. These datasets, like Food-101, share a common property: the visual features necessary for class discrimination are not confined to a small, localized region of the image but are instead distributed more broadly.

For instance, in classes such as "car" or "airplane," the class-defining features are often observed across

the entire shape or outline. As a result, replacing a portion of an image with content from another class does not necessarily eliminate the critical cues needed for correct classification. Moreover, CutMix may help prevent overfitting to spurious or overly specific features—for example, if a model learns to overly rely on wheels to identify a "car," it may misclassify occluded cars or objects like motorcycles. By introducing diverse training views through CutMix, the model can learn more robust and generalizable representations.

5. Conclusion

In this study, we investigated the effects of different data augmentation techniques, especially CutMix, on classification performance using two distinct datasets: Stanford Dogs and Food-101. Our results revealed contrasting outcomes; while CutMix failed to mitigate overfitting and improve generalization on the Stanford Dogs dataset, it demonstrated significant effectiveness on the Food-101 dataset by substantially reducing overfitting and enhancing validation accuracy.

We further analyzed these differences and identified a key factor: the spatial distribution of class-discriminative features within the images. Datasets like Food-101 and commonly used benchmarks such as CIFAR and ImageNet exhibit features spread across multiple regions of the image, allowing CutMix to preserve sufficient class information even when parts of the image are replaced. In contrast, Stanford Dogs images often rely on subtle, localized features critical for class distinction, making CutMix ineffective.

This finding highlights the importance of considering dataset characteristics when selecting augmentation strategies. Applying augmentation methods without such considerations may not always yield performance gains and can sometimes impede learning.