Saliency Grafting: Innocuous Attribution-Guided Mixup with Calibrated Label Mixing

Joonhyung Park[1], **June Yong Yang**[1], **Jinwoo Shin**[1], **Sung Ju Hwang**[1,2], **Eunho Yang** [1,2]
[1]Korea Advanced Institute of Science and Technology (KAIST)
[2]AITRICS
{deepjoon, laoconeth, jinwoos, sjhwang82, eunhoy}@kaist.ac.kr

Abstract

The Mixup scheme suggests mixing a pair of samples to create an augmented training sample and has gained considerable attention recently for improving the generalizability of neural networks. A straightforward and widely used extension of Mixup is to combine with regional dropout-like methods: removing random patches from a sample and replacing it with the features from another sample. Albeit their simplicity and effectiveness, these methods are prone to create harmful samples due to their randomness. To address this issue, 'maximum saliency' strategies were recently proposed: they select only the most informative features to prevent such a phenomenon. However, they now suffer from lack of sample diversification as they always deterministically select regions with maximum saliency, injecting bias into the augmented data. In this paper, we present a novel, yet simple Mixup-variant that captures the best of both worlds. Our idea is two-fold. By stochastically sampling the features and 'grafting' them onto another sample, our method effectively generates diverse yet meaningful samples. Its second ingredient is to produce the label of the grafted sample by mixing the labels in a saliency-calibrated fashion, which rectifies supervision misguidance introduced by the random sampling procedure. Our experiments under CIFAR, Tiny-ImageNet, and ImageNet datasets show that our scheme outperforms the current state-of-the-art augmentation strategies not only in terms of classification accuracy, but is also superior in coping under stress conditions such as data scarcity and object occlusion.

1 Introduction

Modern deep neural networks (DNNs) have achieved unprecedented success in various computer vision tasks, *e.g.*, image classification (He et al. 2016a), generation (Brock, Donahue, and Simonyan 2019) and segmentation (He et al. 2017). However, due to their over-parameterized nature, DNNs require an immense amount of training data to generalize well for test data. Otherwise, DNNs are predisposed to memorize the training samples and exhibit lackluster performance on the unseen data - in other words, incur overfitting.

Acquiring a sufficient amount of data for a given task is not always possible as it consumes valuable manpower and budget. One common approach to combat such data scarcity is *data augmentation*, which aims to enlarge the effective size of a dataset by producing virtual samples from the training data through means such as injecting noise (Amodei et al. 2016) or cropping out regions (DeVries and Taylor 2017). Datasets diversified with these augmented samples are shown to effectively improve the generalization performance of the trained model. Furthermore, data augmentation is proven to be effective not only for promoting generalization but also in boosting the robustness of a model (Hendrycks et al. 2019) and acquiring visual representations without human supervision (Chen et al. 2020; He et al. 2020).

To this end, conventional augmentation methods have focused on creating new images by transforming a given image using means such as flipping, resizing, and more. However, a recently proposed augmentation method called Mixup (Zhang et al. 2017) proposed the idea of crafting a new sample out of a pair of samples by taking a convex combination of them. Inspired by this pioneering work, Yun et al. (2019) proposed CutMix, a progeny of Mixup and Cutout (DeVries and Taylor 2017), which crops a random region of an image and pasting it on another. These methods are able to generate a wider variety of samples while effectively compensating for the loss of information caused by actions such as cropping. However, such context-agnostic nature of these methods gives way to creating samples that are potentially harmful. Since the images are combined randomly without considering their contexts and labels, incorrect augmentation is destined to occur (see Figure 1(d)). For instance, an object can be cropped out and replaced by a different kind of object from another image, or the background part of the image can be pasted on top of an existing object. Even worse, their labels are naively mixed according only to their mixing proportions, disregarding any information transfer or loss caused by the data mixing. The harmfulness of semantically unaware label mixing was previously reported in (Guo, Mao, and Zhang 2019). This mismatch in data and its supervision signal yields harmful samples.

To address this problem, saliency-guided augmentation methods have been recently proposed (Walawalkar et al. 2020; Kim, Choo, and Song 2020; Uddin et al. 2021). These approaches allegedly refrain from generating harmful samples by preserving the region of maximum saliency based on the saliency maps of the image. Attentive Cutmix (Walawalkar et al. 2020) preserves the maximum

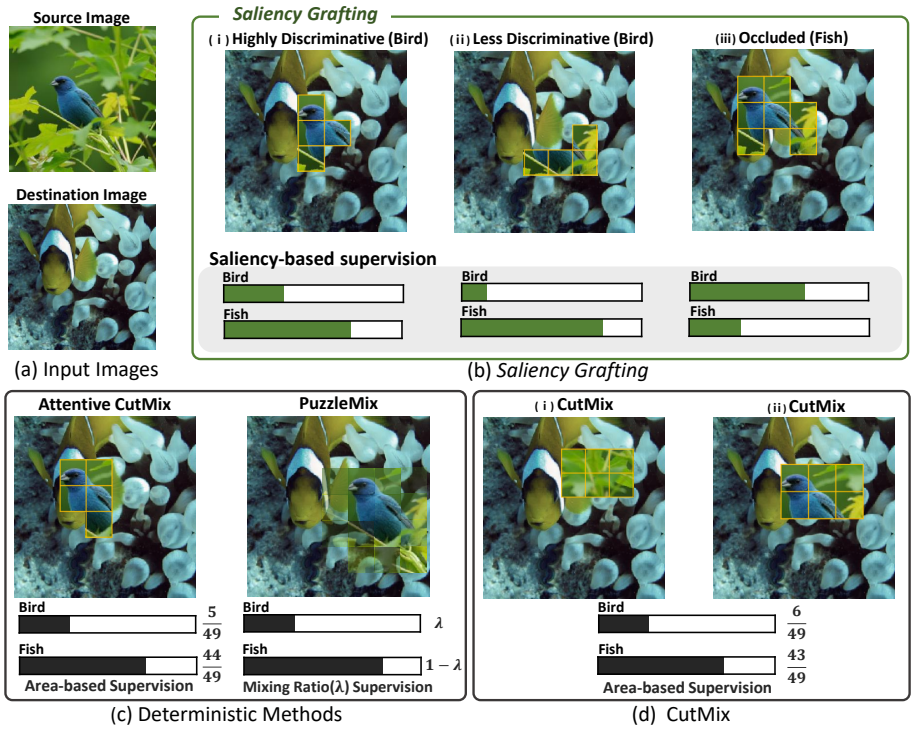

Figure 1: Comparison of augmented samples generated by mixup-based augmentations. (a) Source and destination images to be used in augmentation. (b) *Saliency Grafting* produces diverse samples, including samples that do not contain the maximum saliency region. For all kinds of diverse samples, their labels are correctly rectified. (c) Deterministic saliency-based methods produce semantically plausible labels, but lack diversity since the maximum saliency region is always included. (d) CutMix generates diverse samples but produces misleading labels.

saliency regions of the donor image by locating the k-most salient patches of it and merging them on top of the acceptor image. SaliencyMix (Uddin et al. 2021) constructs a bounding box around the maximum saliency region and crops this box to place it on top of the acceptor image. PuzzleMix (Kim, Choo, and Song 2020) tries to salvage the most salient regions of each image by mixing one to another and solving an optimal transport problem and region-wise mixup ratio to maximize the saliency of the created sample. However, these precautionary measures sacrifice sample diversity - which is the advantage of previous CutMix-based methods. Unlike CutMix that teaches the model to attend to the whole object by probabilistically choosing diverse regions of the image, maximum saliency methods lose this feature as the most discriminative region is always included in the resulting image, biasing the model to depend on such regions. Moreover, they still overlook making appropriate supervision to describe the augmented image properly, and use semantically inaccurate labels determined by the mixing ratio or the size of the pasted region, which can easily mislead the network (see Figure 1(c)).

To solve the drawbacks present in contemporary augmentation methods, we propose *Saliency Grafting*, a novel data augmentation method that can generate *diverse and innocuous* augmented data (see Figure 1(b)). Instead of blindingly selecting the maximum saliency region, our method scales and thresholds the saliency map to grant all salient regions equal chance. The selected regions are then imposed with Bernoulli distribution and sampled to generate stochastic patches. These patches are then 'grafted' on top of another image. Moreover, to compensate for the side effects of grafting such as label mismatch, we propose a novel label mixing strategy: saliency-guided label mixing. By mixing the labels of the two images according to their *saliency* instead of their area, potential bad apples are effectively neutralized.

Our contribution is threefold:

- We discuss the potential weaknesses of current Mixup-based augmentation strategies and present a novel data augmentation strategy that can generate diverse yet meaningful data through saliency-based sampling.
- We present a novel label mixing method to calibrate the generated label to match the information contained in the newly generated data.
- Through extensive experiments, we show that models trained with our method outperform others - even under data corruption or data scarcity.

2 Related Works

Data augmentation Image data augmentation played a formidable role in breakthroughs of deep learning based

computer vision (LeCun et al. 1998; Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2015). Recently, regional dropout methods such as Cutout (DeVries and Taylor 2017), Dropblock (Ghiasi, Lin, and Le 2018) and Random Erasing (Zhong et al. 2020) were proposed to promote generalization by removing selected regions of an image or a feature map to diversify the model's focus. However, the removed regions are bound to suffer from information loss. The recently proposed Mixup (Zhang et al. 2017) and its variants (Verma et al. 2019; Guo, Mao, and Zhang 2019), shifted the augmentation paradigm by not only transforming a sample but using a pair of samples to create a new augmented sample via convex combination. Although successful on multiple domains, Mixup is met with lost opportunities when applied to images as it cannot exploit their spatial locality. To remedy this issue, Cutmix (Yun et al. 2019), a method combining Cutout and Mixup, was proposed. By cropping out a region then filling it with a patch of another image, Cutmix executes regional dropout with less information loss. However, in Cutmix, a new problem arises as the random cut-and-paste strategy incurs semantic information loss and label mismatch. To fix this issue, methods exploiting maximum saliency regions were proposed. Attentive Cutmix (Walawalkar et al. 2020) selects the top-k regions to cut and paste to another image. SaliencyMix (Uddin et al. 2021) creates a bounding box around the maximum saliency region, and pastes the box on another image. Puzzlemix (Kim, Choo, and Song 2020) takes a slightly different approach, where it selects maximum saliency regions of the two images and solves a transportation problem to maximize the saliency of the mixed image. However, since the maximum saliency region is always pertained, the model is deprived of the opportunities to learn from challenging but beneficial samples present in CutMix.

Saliency methods In neuroscience literature, Koch and Ullman (1987) first proposed saliency maps as a means for understanding the attention patterns of the human visual cortex. As contemporary CNNs bear close resemblance to the visual cortex, it is plausible to adapt this tool to observe the inner workings of CNNs. These saliency techniques inspired by human attention are divided into two groups: bottom-up (backward) and top-down (forward) (Katsuki and Constantinidis 2014). For backward methods, saliency is determined in a class-discriminative fashion. Starting from the output of the network, the saliency signal is back-propagated starting from the label logit and attributed to the regions of the input image. (Simonyan, Vedaldi, and Zisserman 2013; Zhou et al. 2016; Selvaraju et al. 2017) utilize the backpropagated gradients to construct saliency maps. Methods such as (Montavon, Samek, and Müller 2018; Nam et al. 2020) proposed to backpropagate saliency scores with carefully designed backpropagation rules that preserve the total saliency score across a selected layer. On the other hand, forward saliency techniques start from the input layer and accumulate the detected signals up the network. The accumulated signals are then extracted at a higher convolutional layer (often the last convolutional layer) to obtain a saliency map. Unlike backward approaches, forward methods are class-agnostic as the convolutional layers extract features from all possible objects inside an image to support the last classifier. These maps are used in a variety of fields such as classification (Oquab et al. 2015) and transfer learning (Zagoruyko and Komodakis 2017).

3 Preliminaries

We first clarify the notations used throughout the section by describing a general form of Mixup-based augmentation procedures. Let $f_\theta(\cdot)$ be a Convolutional Neural Network (CNN) parametrized by θ. For a given batch B of input data $\{x_1, \ldots, x_m\} \in \mathcal{X}^{|B|}$ and the corresponding labels $\{y_1, \ldots, y_m\} \in \mathcal{Y}^{|B|}$, a mixed image $\tilde{x}$ is generated by the augmentation function $\phi(\cdot)$ and the corresponding label $\tilde{y}$ is created through the label mixing function $\psi(\cdot)$: $\tilde{x} = \phi(x_i, x_j)$ and $\tilde{y} = \psi(y_i, y_j)$ for data index i and its random permutation j.

Then, Mixup-based augmentation methods define their own $\phi(\cdot)$ as a pixel-wise convex combination of two randomly selected pair, as follows:

$$\phi(x_i, x_j) = M_\lambda \odot h(x_i) + (\mathbf{1} - M_\lambda) \odot h(x_j) \quad (1)$$

where M_λ is a mixing matrix controlled by a mixing ratio λ, $\odot$ is the element-wise Hadamard product, and $h(\cdot)$ is some pre-processing function.

The vanilla (input) Mixup defines the augmentation function ϕ as $\phi(x_i, x_j) = \lambda x_i + (1 - \lambda)x_j$. Manifold Mixup uses similar function $\phi(x_i, x_j) = \lambda h(x_i) + (1 - \lambda)h(x_j)$ but with the latent features. In CutMix, the augmentation function ϕ is defined as $\mathbf{1}_i^{\text{Rect}} \odot x_i + (1 - \mathbf{1}_i^{\text{Rect}}) \odot x_j$. This method randomly cuts a rectangular region $\mathbf{1}_i^{\text{Rect}}$ from the source image x_i with area proportional to λ and pastes it onto the destination image x_j. PuzzleMix, recent saliency-based Mixup variant, employs the augmentation function $\phi(x_i, x_j) = Z^* \odot \Pi_i^{*T} x_i + (\mathbf{1} - Z^*) \odot \Pi_j^{*T} x_j$. This method exploits the image transportation plan Π and region-wise mask matrix Z to maximize the saliency of the mixed image. Note that unlike the vanilla Mixup, Z is discretized region-wise mixing matrix that satisfies $\lambda = \frac{1}{n} \sum_s \sum_t Z_{st}$ for given mixing ratio λ. To find the optimal transportation plan Π^* and region-wise mask Z^* for the maximum saliency, PuzzleMix solves additional optimization problems in an alternating fashion, per *each* iteration.

Although it is a simpler scalar function, the label function $\psi(\cdot)$ is also defined in a similar form to the augmentation function $\phi(\cdot)$:

$$\psi(y_i, y_j) = \rho y_i + (1 - \rho) y_j \quad (2)$$

where ρ is a label mixing coefficient determined by the sample pair $(x_i, y_i), (x_j, y_j)$ and the mixing ratio λ from ϕ. However, in all methods mentioned above, this ρ simply depends on λ, disregarding the contents of sample pair x_i and x_j: $\tilde{y} = \lambda y_i + (1 - \lambda) y_j$.

4 Saliency Grafting

We now describe our simple approach, *Saliency Grafting*, that creates diverse and innocuous Mixup augmentation based on the content of instances being merged. Two key

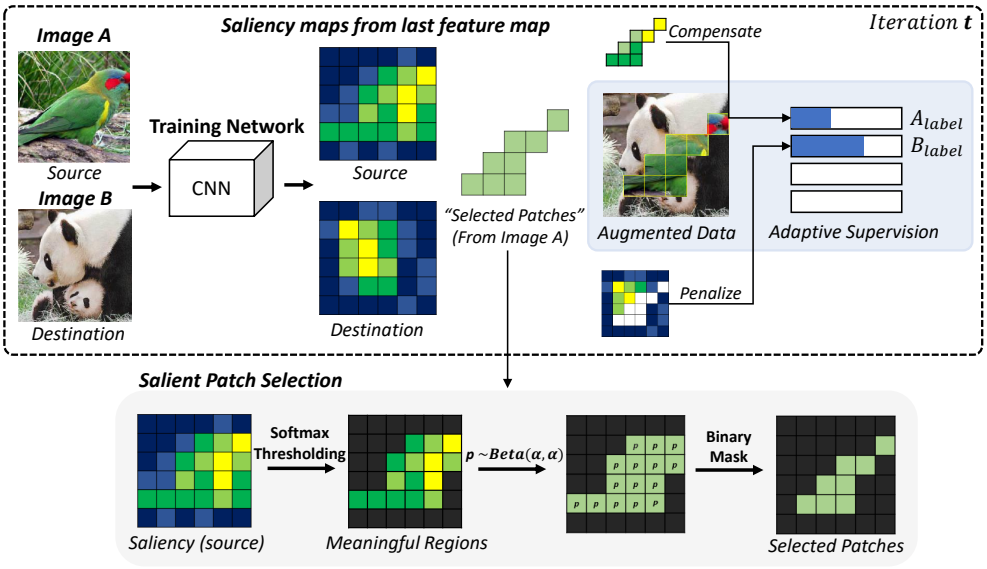

Figure 2: Overview of *Saliency Grafting*. The source and destination images are drawn from the mini-batch and fed forward through the training network, producing their respective forward saliency maps. The source saliency map is thresholded and sampled using a region-wise i.i.d. Bernoulli distribution. Patches of the source image that corresponds to the resulting image are grafted to the destination image.

innovations in *Saliency Grafting* are stochastic patch selection (Section 4.1) and label mixing (Section 4.2), both of which utilize the saliency information at the core. Last but not least, another important element of *Saliency Grafting* is choosing a saliency map generation method (Section 4.3) for the above two main components while keeping the learning cost to a minimum. The overall procedure is described in Figure 2. Now we discuss the details of each component in the subsequent subsections.

4.1 Stochastic Salient Patch Selection

The stochastic patch selection of *Saliency Grafting* aims to choose regions that can create diverse and meaningful instances. The key question here is how to select regions to be grafted, given a saliency matrix S_i for the source image x_i (whose element S_{st} indicates the saliency for a region (s,t) of x_i). As in recent studies (Kim, Choo, and Song 2020; Walawalkar et al. 2020), if only regions with high intensity of S_{st} are always selected, then these regions - which are already easy to judge by the model - are continuously augmented in the iterative training procedure. As a result, the model is repeatedly exposed to the same grafting patch, which would iteratively amplify the model's attention on the selected regions and deprive the opportunity to learn how to attend to other parts and structures of the object.

In order to eliminate this *selection bias*, the patch selection of *Saliency Grafting* consists of two steps: i) softmax thresholding and ii) stochastic sampling.

Softmax thresholding To neutralize the selection bias due to the intensity of saliency, we normalize the saliency map by applying the softmax function and then binarize the map with some threshold σ:

$$S'_{st}(\boldsymbol{x};T) = \frac{\exp\left(S_{st}(\boldsymbol{x})/T\right)}{\sum_h^H \sum_w^W \exp\left(S_{hw}(\boldsymbol{x})/T\right)}, \tag{3}$$

$$S''_{st}(\boldsymbol{x};T) = \begin{cases} 1, & \text{if } S'_{st}(\boldsymbol{x};T) > \sigma \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

given the temperature hyperparameter T to control the sharpness of the normalized saliency map. Here, threshold σ has a variety of options, but we adopt the threshold as $\sigma_{\text{mean}} = \frac{1}{HW}\sum_h^H \sum_w^W S'_{hw}$, using the mean value of the normalized saliency map.

Stochastic sampling Although the selection bias is significantly mitigated by thresholding, the high intensity regions are never *removed*, as the softmax function preserves the order of the regions. To address this issue, we stochastically sample the grafting regions based on the binarized saliency map produced above. The final mixing matrix M_i is constructed by taking the Hadamard product of S''_i and a region-wise i.i.d. random Bernoulli matrix of same dimensions $P \sim \mathbf{Bern}(p_B)$: $M_i = P \odot S''_i$. Here, the batch-wise sampling probability p_B is drawn from a Beta distribution $p_B \sim Beta(\alpha, \alpha)$. The final augmentation function ϕ for *Saliency Grafting* is $M_i \odot x_i + (1 - M_i) \odot x_j$.

4.2 Calibrated Label Mixing Based On Saliency

In addition to the method of grafting diverse and innocuous augmentations described in the previous section, attaching an appropriate label for supervision to the generated data is also the core of *Saliency Grafting*. Although extreme, to

highlight the drawbacks of the existing label mixing strategy used in all baselines, suppose that source image x_i is combined with destination image x_j, both of which have saliency concentrated in some small regions. Suppose further that this region of x_i is selected and grafted to the region where the original class of destination x_j is concentrated. Then, most of the information of class y_i is retained while most of the information on class y_j is lost. However, if the label is determined in proportion to the mixing rate or the size of the area used, as all the baselines do, the generated label will be close to class y_j since most areas of it originally came from the destination image x_j.

To tackle this issue, we propose a novel label mixing procedure that can adaptively mix the labels again based on saliency maps. Regarding the destination image x_j receiving the graft, the ground truth label y_j is penalized according to the degree of occlusion. Specifically, the importance of the destination image $I(S_j, 1 - M_i)$[1] given the mixing matrix M_i is calibrated using the saliency values of the remaining part not occluded by the source image, $I(S_j, 1 - M_i) = \frac{\|S_j \odot (\mathbf{1} - M_i)\|_2}{\|S_j\|_2}$. On the other hand, with regard to the source image x_i giving the graft, the corresponding label y_i is compensated in proportion to the importance of the selected region: $I(S_i, M_i) = \frac{\|S_i \odot M_i\|_2}{\|S_i\|_2}$.

The final label mixing ratio is computed based on the relative importance of x_i and x_j, so that their coefficients sum to 1 to define the calibrated label mixing function ψ.

$$\psi(y_i, y_j) = \lambda(S_i, S_j, M_i) y_i + \big(1 - \lambda(S_i, S_j, M_i)\big) y_j \quad (5)$$

$$\text{where} \quad \lambda(S_i, S_j, M_i) = \frac{I(S_i, M_i)}{I(S_i, M_i) + I(S_j, 1 - M_i)}$$

4.3 Saliency Map Generation

Technically, *Saliency Grafting* can be combined with various saliency generation methods without the dependence on a specific method. However, the caveat here is that the performance of *Saliency Grafting* is, by design, highly affected by the quality of the saliency map, or how accurately the saliency map corresponds to the ground truth label. From this point of view, the forward saliency methods, which incur less false negatives, may support *Salient Grafting* more stably than the backward methods (see Section 2 for forward and backward saliency methods). We also provide the performance comparison in Appendix A. This is because the backward methods are likely to break down and exclude true salient regions when the model fails to predict the true label, whereas the forward methods preserve all the feature maps inside the saliency map, i.e., they act like a class-agnostic saliency detector (Mahendran and Vedaldi 2016).

In an environment where there is no separate pre-trained model, another advantage of using forward saliency is gained: Saliency maps can be naturally constructed based on the terms already calculated in the learning process. In this environment, since the generated maps can be noisy in the early phases of training, we employ warmup epochs without data augmentation.

We now describe the choice of generating the saliency maps to guide our augmentation process. We adopt the channel-collapsed absolute feature map of the model as our saliency map, mainly due to its simplicity: $S^{(l)} = \sum_{c=1}^{C} \left|A_c^{(l)}\right|$ where $A \in \mathbb{R}^{C \times H \times W}$ is the feature map at the l-th layer. Albeit it is possible to extract saliency maps from any designated layer in the network, we extract the maps from the last convolutional layer as it generally conveys the high-level spatial information (Bengio, Courville, and Vincent 2013). In practice, we randomly select the up/down-sampling scale of saliency maps per each mini-batch.

5 Experiments

We conduct a collection of experiments to test *Saliency Grafting* against other baselines. First, we test the prediction performance on standard image classification datasets. Next, to confirm our claim that *Saliency Grafting* can safely boost the diversity of augmented data, we design and conduct experiments to assess the sample diversity of each augmentation method. We also conduct multiple stress tests to measure the enhancement in generalization capability. Finally, we perform an ablation study to investigate the contribution of each sub-component of *Saliency Grafting*. Note that we train the models with both original and augmented images.

5.1 Classification Tasks

CIFAR-100 We evaluate our method *Saliency Grafting* on CIFAR-100 dataset (Krizhevsky 2009) using two neural networks: PyramidNet-200 with widening factor $\tilde{\alpha} = 240$ (Han, Kim, and Kim 2017) and WRN28-10 (Zagoruyko and Komodakis 2016). For the PyramidNet-200, we follow the experimental setting of Yun et al. (2019), which trains PyramidNet-200 for 300 epochs. The baselines results on PyramidNet-200 are as reported in Yun et al. (2019). For WRN28-10, the network is trained for 400 epochs as following studies (Kim, Choo, and Song 2020; Verma et al. 2019). In this experiment, we reproduce other baselines following the original setting of each paper. Detailed settings are provided in Appendix B. As shown in Table 1 and Table 2, *Saliency Grafting* exhibits significant improvements for both architectures compared to other baselines. Furthermore, when used together with Shakedrop regularization (Yamada et al. 2019), *Saliency Grafting* achieves additional enhancement - **13.05%** Top-1 error.

Tiny-ImageNet We evaluate our method on another benchmark dataset - Tiny-ImageNet (Chrabaszcz, Loshchilov, and Hutter 2017). We train ResNet-18 (He et al. 2016b) for 600 epochs and report the converged error rates of the last 10 epochs, following one of Tiny-ImageNet experimental settings in (Kim, Choo, and Song 2020). Other data augmentation methods are evaluated using their author-released code and hyperparameters. Detailed experimental settings are described in Appendix B. The obtained results are shown in Table 3. In line with the CIFAR-100 experiments, *Saliency Grafting* consistently exhibits the best performance on this benchmark dataset.

[1] We use ℓ_2 norm to define the importance I in the sense that the overall saliency is simply the same as the sum of saliency in each region, but similar importance can be obtained with other norms.

PyramidNet-200 ($\tilde{\alpha} = 240$) (# params: 26.8 M)	Top-1 Error (%)	Top-5 Error (%)
Vanilla	16.45	3.69
Cutout	16.53	3.65
DropBlock	15.73	3.26
Mixup ($\alpha = 1.0$)	15.63	3.99
Manifold Mixup ($\alpha = 1.0$)	16.14	4.07
ShakeDrop	15.08	2.72
Cutout + Mixup	15.46	3.42
Cutout + Manifold Mixup	15.09	3.35
CutMix	14.47	2.97
CutMix + ShakeDrop	13.81	2.29
Attentive CutMix (N = 6)	15.24 $\pm$0.09	3.46 $\pm$0.06
SaliencyMix	14.74 $\pm$0.17	3.07 $\pm$0.04
PuzzleMix	14.78	3.08
Saliency Grafting	**13.94** $\pm$0.11	**2.79** $\pm$0.09
Saliency Grafting + ShakeDrop	**13.05** $\pm$0.06	**2.18** $\pm$0.03

Table 1: Error rates on CIFAR-100 for PyramidNet-200 in comparison to recent regularization methods. The averaged best error rates are reported.

WRN28-10 (# params: 36.5 M)	Top-1 Error (%)	Top-5 Error (%)
Vanilla	20.74 $\pm$0.06	5.70 $\pm$0.03
Mixup	17.59 $\pm$0.07	5.18 $\pm$0.14
Manifold Mixup†	18.04 $\pm$0.08	-
CutMix	17.47 $\pm$0.24	4.80 $\pm$0.42
AugMix	19.19 $\pm$0.04	4.36 $\pm$0.02
SaliencyMix (w/ dropout)†	16.23 $\pm$0.08	-
PuzzleMix	16.00 $\pm$0.03	3.84 $\pm$0.04
Saliency Grafting	**15.32** $\pm$0.13	**3.54** $\pm$0.03

Table 2: Error rates on CIFAR-100 for WRN28-10 in comparison to data augmentation methods. The averaged best error rates with standard errors are reported. † indicates the reported result in the original paper.

ImageNet For the ImageNet (Russakovsky et al. 2015) experiment, we train ResNet-50 for 100 epochs. We follow the training protocol in Wong, Rice, and Kolter (2020), which includes cyclic learning rate, regularization on batch normalization layers, and mixed-precision training. This protocol also gradually resizes images during training, beginning with larger batches of smaller images and moving on to smaller batches of larger images later (*image-resizing policy*). The baselines results are as reported in (Kim, Choo, and Song 2020). As shown in Table 4, *Saliency Grafting* achieves again the best performance in both Top-1/Top-5 error rates. We confirm that ours can bring further performance improvement without image-resizing scheduling.

Additional experiments Due to the space constraint, two experiments are deferred to Appendix A. The first experiment shows that *Saliency Grafting* is useful for **speech** dataset beyond the image classification task, and the second experiment (**weakly supervised object localization**) implies that the final model learned through *Saliency Grafting* contains more useful saliency information.

ResNet-18 (# params: 11.3 M)	Top-1 Error (%)	Top-5 Error (%)
Vanilla	38.54 $\pm$0.15	18.53 $\pm$0.11
Mixup	37.37 $\pm$0.14	18.09 $\pm$0.11
CutMix	35.76 $\pm$0.15	15.82 $\pm$0.20
SaliencyMix	36.61 $\pm$0.13	16.31 $\pm$0.27
PuzzleMix	35.79$\pm$0.17	16.31 $\pm$0.15
Saliency Grafting	**35.16** $\pm$0.12	**15.02**$\pm$0.07

Table 3: Error rates on Tiny-ImageNet for ResNet-18 in comparison to data augmentations. The converged error rates with standard errors are reported.

ResNet-50 (# params: 25.6M)	Top-1 Error (%)	Top-5 Error (%)
Vanilla	24.31	7.34
Mixup	22.99	6.48
Manifold Mixup	23.15	6.50
CutMix	22.92	6.55
AugMix	23.25	6.70
PuzzleMix	22.49	6.24
Saliency Grafting	22.35	**6.19**
Saliency Grafting ***(w/o image-resizing)***	**22.26**	6.29

Table 4: Comparison of state-of-the-art data augmentation methods on ImageNet dataset.

5.2 Sample Diversity

Generating data *k*-times We design an experiment to compare *Saliency Grafting* and other augmentations in terms of sample diversity. For every iteration, each method trains the network by generating additional augmented data k times from the mini-batch; each method tries to diversify the mini-batch by producing k independent augmented batches with its own randomness. To ensure sufficient diversity, the mixing ratio λ is newly sampled for each augmented data. While varying k from 1 to 6, we evaluate whether each method can obtain the performance gain due to sample diversity. We train the WRN28-10 for 200 epochs and use 20% of the CIFAR-100 dataset to better confirm the diversity effect of the augmented data. In Figure 3, the performance of *Saliency Grafting* consistently improves as k increases, whereas PuzzleMix, one of the representative maximum saliency strategies, does not show performance gain when k increases. We believe this is the direct evidence that generating multiple augmented instances by sampling the random mixing ratio is insufficient to ensure sample diversity in the case of maximum saliency approaches. However, as our method exploits temperature-scaled thresholding with stochastic sampling, the model easily attends to the entire object as k increases. Hence, the sample diversity can be guaranteed innocuity.

5.3 Stress Testing

Data scarcity The situation where data augmentation is most needed is when data is scarce. In this condition, it is important to improve the generalization performance by increasing the data volume while preventing overfitting. To

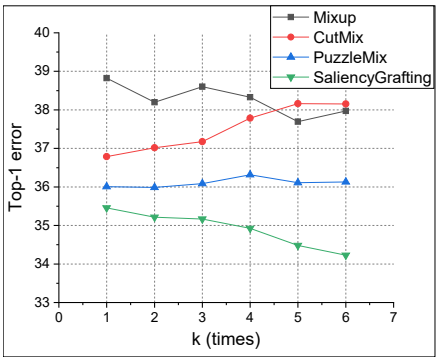

Figure 3: Comparing sample diversity by generating k-times augmented data from a mini-batch. The experiment was repeated five times and averaged best error rates are reported.

this end, we test our method against data scarcity by reducing the number of data per class to 50%, 20%, and 10%, with the WRN28-10 on the CIFAR-100. In Table 5, our method exhibits the best performance in every condition. These results are in line with the fact that, as explored by Rolnick et al. (2017), corrupted labels degrade the performance when data is scarce. As our method uses calibrated label to reduce the mismatch between data and labels, the generalization performance can be enhanced in data scarcity conditions.

# of data per class	50 (10 %)	100 (20 %)	250 (50 %)
Vanilla	59.96 $\pm$0.50	44.29 $\pm$0.51	31.19 $\pm$0.20
Mixup	51.29 $\pm$0.27	38.80 $\pm$0.51	27.11 $\pm$0.10
CutMix	52.90 $\pm$0.18	38.76 $\pm$0.15	26.58 $\pm$0.10
PuzzleMix	54.69 $\pm$0.54	38.66 $\pm$0.16	26.69 $\pm$0.05
Saliency Grafting	**51.01** $\pm$0.31	**37.35** $\pm$0.21	**25.56** $\pm$0.24

Table 5: Top-1 errors on the CIFAR-100 with reduced number of data per class. The experiment was repeated 3 times.

Partial occlusion To expose the bias induced by previous saliency augmentations, we conduct an 'occlusion experiment', where we remove the top-k salient regions from the images then evaluate. Table 6 shows that as the occluded area gets larger, *Saliency Grafting* scores the highest as stochastic sampling removes the bias.

Method	**Top-1 Error (%)**		
(ResNet-18)	$k = 0\%$	$k = 12.5\%$	$k = 25\%$
SaliencyMix	36.61	44.73	55.89
PuzzleMix	35.79	50.91	66.23
SaliencyGrafting	**35.16**	**42.98**	**52.19**

Table 6: Top-1 error rates on TinyImageNet with top-$k\%$ salient regions removed.

5.4 Ablation Study

Stochastic selection VS deterministic selection We argued that the deterministic region selection process of existing saliency methods leads to performance degradation. This was partly shown in Table 1, 3, and 4 where such methods perform worse than CutMix. Here, we directly study the contribution of stochastic selection. We measure the accuracy on CIFAR-100 with two architectures where the deterministic top-k selection of Attentive CutMix (Walawalkar et al. 2020) is replaced by our stochastic selection. For fair comparison, the softmax temperature T is adjusted to satisfy $\mathbb{E}_i[\sum_s \sum_t M_{i,st}] = k$. Results show that stochastic selection indeed outperforms deterministic selection (Table 7).

Effect of threshold σ Here, we conduct an experiment where we vary the saliency threshold σ. Figure 4 shows that as we lower σ below the normalized saliency mean σ_{mean}, non-salient regions are introduced, and the performance degenerates. At $\sigma = 0$, SG becomes saliency-agnostic (which is near-equivalent to the CutMix strategy), and the performance of SG converges to the vicinity of CutMix.

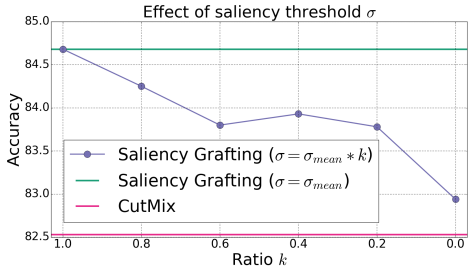

Figure 4: Effect of σ on CIFAR-100 with WRN28-10.

Label mixing strategies In Section 4.2, we discussed the pitfalls of naive area-based label mixing and proposed saliency-based label mixing as a solution. Here, we compare the two strategies. We experiment on CIFAR-100 with two architectures and replace the mixing strategy of *Saliency Grafting* with area-based mixing. Results in Table 7 confirm that saliency-based mixing outperforms area-based mixing.

Method	**WRN28-10**	**PyramidNet-200**
Saliency Grafting (SG)	Top-1 Error (%)	Top-1 Error (%)
Deterministic + area labels	16.34	14.63
Stochastic + area labels	15.67	14.14
Stochastic + saliency labels	**15.32**	**13.94**

Table 7: Ablation study of our method *Saliency Grafting*.

6 Conclusion

We have presented *Saliency Grafting*, a data augmentation method that generates diverse saliency-guided samples via stochastic sampling and neutralizing any induced data-label mismatch with saliency-based label mixing. Through extensive experiments, we have shown that models equipped with *Saliency Grafting* outperform existing mixup-based data augmentation techniques under both normal and extreme conditions while using less computational resources.

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)) and National Research Foundation of Korea (NRF) grants (2019R1C1C1009192). This work was also partly supported by KAIST-NAVER Hypercreative AI Center.

References

Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, 173–182.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large scale gan training for high fidelity natural image synthesis. In *ICLR*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Chrabaszcz, P.; Loshchilov, I.; and Hutter, F. 2017. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*.

DeVries, T.; and Taylor, G. W. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv:1708.04552*.

Ghiasi, G.; Lin, T.-Y.; and Le, Q. V. 2018. Dropblock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems*, 10727–10737.

Guo, H.; Mao, Y.; and Zhang, R. 2019. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3714–3722.

Han, D.; Kim, J.; and Kim, J. 2017. Deep pyramidal residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5927–5935.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *International Conference on Learning Representations*.

Katsuki, F.; and Constantinidis, C. 2014. Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist*, 20(5): 509–521.

Kim, J.-H.; Choo, W.; and Song, H. O. 2020. Puzzle Mix: Exploiting Saliency and Local Statistics for Optimal Mixup. In *International Conference on Machine Learning (ICML)*.

Koch, C.; and Ullman, S. 1987. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, 115–141. Springer.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, University of Tront*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Mahendran, A.; and Vedaldi, A. 2016. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3): 233–255.

Montavon, G.; Samek, W.; and Müller, K.-R. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73: 1–15.

Nam, W.-J.; Gur, S.; Choi, J.; Wolf, L.; and Lee, S.-W. 2020. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks. *AAAI*.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 685–694.

Rolnick, D.; Veit, A.; Belongie, S.; and Shavit, N. 2017. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR*.

Uddin, A. F. M. S.; Monira, M. S.; Shin, W.; Chung, T.; and Bae, S.-H. 2021. SaliencyMix: A Saliency Guided Data Augmentation Strategy for Better Regularization. In *International Conference on Learning Representations*.

Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, 6438–6447. PMLR.

Walawalkar, D.; Shen, Z.; Liu, Z.; and Savvides, M. 2020. 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)*, 3642–3646. IEEE.

Wong, E.; Rice, L.; and Kolter, J. Z. 2020. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*.

Yamada, Y.; Iwamura, M.; Akiba, T.; and Kise, K. 2019. Shakedrop regularization for deep residual learning. *IEEE Access*, 7: 186126–186136.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, 6023–6032.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zagoruyko, S.; and Komodakis, N. 2017. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random Erasing Data Augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.