# TdAttenMix: Top-Down Attention Guided Mixup

**Zhiming Wang**[1], **Lin Gu**[2, 3], **Feng Lu**[1*]

[1]State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University
[2]RIKEN AIP
[3]The University of Tokyo, Japan
{zy2306418,lufeng}@buaa.edu.cn, lin.gu@riken.jp

## Abstract

CutMix is a data augmentation strategy that cuts and pastes image patches to mixup training data. Existing methods pick either random or salient areas which are often inconsistent to labels, thus misguiding the training model. By our knowledge, we integrate human gaze to guide cutmix for the first time. Since human attention is driven by both high-level recognition and low-level clues, we propose a controllable Top-down Attention Guided Module to obtain a general artificial attention which balances top-down and bottom-up attention. The proposed TdATttenMix then picks the patches and adjust the label mixing ratio that focuses on regions relevant to the current label. Experimental results demonstrate that our TdAttenMix outperforms existing state-of-the-art mixup methods across eight different benchmarks. Additionally, we introduce a new metric based on the human gaze and use this metric to investigate the issue of image-label inconsistency.

**Code** — https://github.com/morning12138/TdAttenMix

## 1 Introduction

Thanks to large amount of data, Deep Neural Networks (DNNs) have achieved significant success in recent years across a variety of applications, including recognition (Dosovitskiy et al. 2021; Zang et al. 2022; Cui et al. 2022; Tan et al. 2022; Chen, Fan, and Panda 2021), graph learning (Xia et al. 2022; Wu et al. 2023; Cheng et al. 2022), and video processing (Liu et al. 2021a; Cui et al. 2021; Liu et al. 2021b; Zhao et al. 2022). However, the data-hungry problem (Dosovitskiy et al. 2021; Touvron et al. 2021a) leads to overfitting when the training data are scarce. Therefore, a series of data augmentation techniques called mixup are proposed to alleviate this issue and enhance DNNs' generalization capabilities. Among them, CutMix (Yun et al. 2019) is an effective strategy that randomly crops a patch from the source image and pastes it into the target image. The label is then mixed by the source and target labels in proportion to the crop area ratio.

Since the randomness in CutMix (Yun et al. 2019) ignores the spatial saliency, a group of saliency-based variants (Uddin et al. 2021; Kim, Choo, and Song 2020; Walawalkar et al. 2020; Liu et al. 2022c; Dabouei et al. 2021; Chen et al. 2023) leverage the bottom-up attention as a supervisory signal. Bottom-up attention operates on raw sensory input and orients attention towards visual features of potential importance to calculate the saliency. This process only discovers *what is where* in the world (Schwinn et al. 2022), which equally looks for all salient regions in the raw sensory input. Therefore, existing saliency-based variants based on bottom-up attention are easily distracted by high saliency regions that are, in fact, irrelevant to the target label. For instance, the source image of Figure 1 is dog, but SaliencyMix (Uddin et al. 2021) become distracted by the dark rock and crops the background, including only part of the dos's ear.

Human vision entails more than just the determination of *what is where*; it involves the development of internal representations that facilitate future actions. For instance, psychological research (Buswell 1935; Yarbus 2013; Belardinelli, Herbort, and Butz 2015) found that human gaze, initially guided by bottom-up features, can be strongly influenced by the task at hand. Consequently, recent research proposes top-down mechanisms (Shi, Darrell, and Wang 2023; Schwinn et al. 2022) showing effectiveness in modeling human gaze patterns such as scanpaths (Schwinn et al. 2022) and in en-

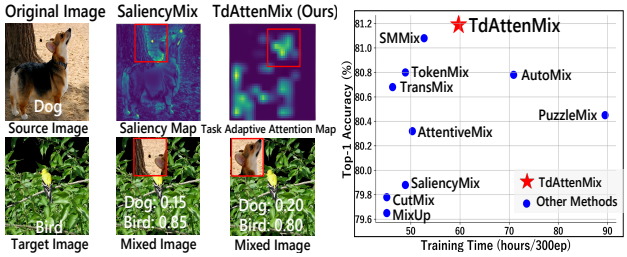


Figure 1: **Left:** SaliencyMix *vs.* TdAttenMix. Since SaliencyMix selects to crop the patch with the most salient region, it is distracted by irrelevant dark stone. Our TdAttenMix balances top-down and bottom-up attention and thus picks salient areas consistent with the dog label. **Right:** Traing time *vs.* accuracy with Deit-S on ImageNet-1k. TdAttenMix improves performance without the heavy computational overhead.

*Corresponding Author.

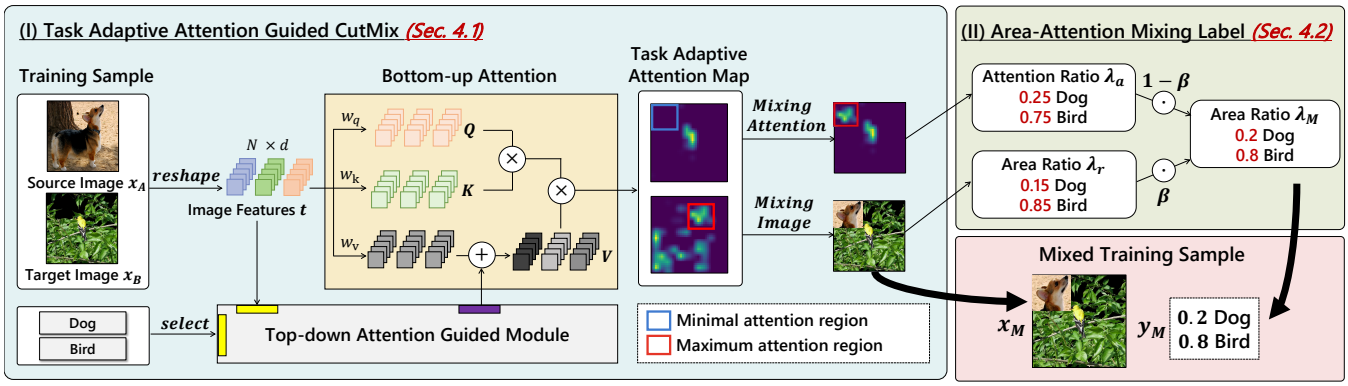


Figure 2: The framework of TdAttenMix. (1)Task Adaptive Attention Guided CutMix: compute the task adaptive attention map via manipulating the bottom-up attention using our proposed Top-down Attention Guided Module and then uses the task adaptive attention map to crop the patch. (2)Area-Attention Label Mixing: adjust label mixing based on the ratio of attention and area.

hancing downstream recognition tasks like classification and Vision Question Answering (VQA) (Shi, Darrell, and Wang 2023). For data mixing techniques, the labels of original data can be used naturally as the task at hand and the execution logic of human gaze can be modeled by the bottom-up features from the original image in conjunction with the high-level guidance of the original label.

In this paper, inspired by the task guided mechanism of human gaze, we extend the saliency-based CutMix to a general framework that balances top-down and bottom-up attention to cut and mix the training samples. Bottom-up attention learns features from the original input and looks for all salient regions while top-down attention uses characteristics of the category as the current task to adjust attention. As illustrated in Figure 1, our top-down attention mixup (TdAttenMix) crops the exemplary head region of the dog image and pastes it on the background area of the target bird image and finally obtains an image-label consistent mixed data that differs significantly from mixed data generated by SaliencyMix.

As portrayed in Figure 2, our TdAttenMix involoves two steps: Task Adaptive Attention Guided CutMix and Area-Attention Label Mixing. The first step generalizes the bottom-up attention of saliency-based CutMix to the task adaptive attention via our proposed Top-down Attention Guided Module. When mixing the image, we use Max-Min Attention Region Mixing (Chen et al. 2023) to select maximum attention region from a source image and paste it onto the region with the minimal attention score in a target image.

The second step determines the label mixing with Area-Attention Label Mixing module. Unlike the conventional approach of area-based label assignment, this module incorporates the area ratio of the mixed image along with the attention ratio of the task adaptive attention map. In the end, our TdAttenMix framework produces the mixed training sample $(x_M, y_M)$ in Figure 2.

The saliency-based CutMix variants aim to produce a sufficient amount of image-label consistent mixed data. As far as current knowledge allows, existing methods lack a quantitative approach to assess image-label inconsistency. The core of quantitative analysis lies in establishing the correct ground-truth for label assignment. Motivated by the notion that gaze mirrors human vision (Huang et al. 2020), we propose to use gaze attention on the ARISTO dataset (Liu et al. 2022b) which collects the real gaze of participants when performing fine-grained recognition tasks. This data will be used to create mixed labels, serving as the ground truth to investigate the issue of image-label inconsistency.

The contribution of this paper is three-fold:

- By our knowledge, this paper for the first time proposes a Top-down Attention Guided Module to integrate human gaze for an artificial attention that balances both top-down and bottom-up attention to crop the task-relevant patch and adjust the label mixing ratio.
- Extensive experiments demonstrate the TdAttenMix boosts the performance and achieve state-of-the-art top-1 accuracy in CIFAR100, Tiny-ImageNet, CUB-200 and ImageNet-1k. Moreover, as shown in Figure 1, our TdAttenMix can achieve state-of-the-art top-1 accuracy without the heavy computational overhead.
- We quantitatively explore the image-label inconsistency problem in image mixing. The proposed method effectively reduces the image-label inconsistency and improves the performance.

## 2 Related Work

### 2.1 CutMix and its variants

CutMix (Yun et al. 2019) randomly crops a patch from the source image and pastes it onto the corresponding location in the target image, with labels being a linear mixture of the source and target image labels proportionate to the area ratio. Since random cropping ignores the regional saliency information, researchers leverage a series of saliency-based variants based on bottom-up attention. AttentiveMix (Walawalkar et al. 2020) and SaliencyMix (Uddin et al. 2021) guide mixing patches by saliency regions in the image (based on class activation mapping or a saliency

detector(Montabone and Soto 2010)). Subsequently, PuzzleMix (Kim, Choo, and Song 2020) and Co-Mixup (Kim et al. 2020) propose combinatorial optimization strategies to find optimal mixup that maximizes the saliency information. Then AutoMix (Liu et al. 2022c) adaptively generates mixed samples based on mixing ratios and feature maps in an end-to-end manner. Inspired by the success of Vision Transformer (ViT)(Dosovitskiy et al. 2021), TokenMixup (Choi, Choi, and Kim 2022) is proposed to adaptive generate mixed images based on attention map. Moreover, concerning label assignment, recent studies have also adjusted label assignment by bottom-up attention. TransMix (Chen et al. 2022) mixes labels based on the class attention score and TokenMix (Liu et al. 2022a) assigns content-based mixes labels on mixed images. Recently, SMMix (Chen et al. 2023) motivates both image and label enhancement by the bottom-up self-attention of ViT-based model under training itself. However, these existing variants, focusing either on enhancing saliency or adjusting label assignments, are reliant on bottom-up attention, which is susceptible to being distracted by salient but label-inconsistent background areas. To relieve label inconsistency, we introduce task adaptive top-down attention into CutMix variants for the first time and propose our framework TdAttenMix.

### 2.2 Computational modeling of Attention

Computational modeling of human visual attention intersects various disciplines such as neuroscience, cognitive psychology, and computer vision. Biologically-inspired attention mechanisms can enhance the interpretability of artificial intelligence (Vuyyuru et al. 2020). The attention can be categorized into bottom-up and top-down mechanisms (Connor, Egeth, and Yantis 2004). Initially, the focus was primarily on computational modeling of bottom-up attention. Based on the Treisman's seminal work describing the feature integration theory (Treisman and Gelade 1980), current approaches assume a central role for the saliency map. Within the theory, attention shifts are generated from the saliency map using the winner-take-all algorithm (Koch and Ullman 1985). Consequently, the majority of studies have focused on improving the estimation of the saliency map (Borji and Itti 2012; Riche et al. 2013). Recently self-attention (Dosovitskiy et al. 2021) is a stimulus-driven approach that highlights all the salient objects in an image, representing a typical bottom-up attention mechanism. With the advent of increasingly large eye-tracking datasets (Liu et al. 2022b; Jiang et al. 2015), researchers have been inspired to explore task-guided top-down attention. Shi et al. (Shi, Darrell, and Wang 2023) propose a top-down modulated ViT model by mimicking the task-guided mechanism of human gaze. Shiwinn et al. (Schwinn et al. 2022) impose a biologically-inspired foveated vision constraint to neural networks to generate human-like scanpaths without training for this object. As for CutMix variants, previous saliency-based methods have utilized bottom-up attention to optimize cropping regions, whereas we explore the use of task-adaptive top-down attention to obtain a cropped region that is more consistent with the label.

## 3 Preliminary

**CutMix augmentation.** CutMix (Yun et al. 2019) is a simple data augmentation technique that combines two pairs of input and labels. $x$ and $y$ represent a training image and its corresponding label, where $x \in \mathbb{R}^{H\times W\times C}$. To create a new augmented training sample $(x_M, y_M)$, CutMix (Yun et al. 2019) utilizes a source image-label pair $(x_A, y_A)$ and a target image-label pair $(x_B, y_B)$. Mathematically, this can be expressed as follows:

$$x_M = M \odot x_A + (1 - M) \odot x_B \tag{1}$$

$$y_M = \lambda_r y_A + (1 - \lambda_r) y_B \tag{2}$$

$M \in \{0, 1\}^{H\times W}$ denotes a rectangular binary mask that indicates where to drop or keep in the two images, $\odot$ is element-wise multiplication, and $\lambda_r$ indicates the area ratio of $x_A$ in mixed image $x_M$, *i.e.*, $\lambda_r = \frac{\sum M}{HW}$.

## 4 Framework of TdAttenMix

This section formally introduces our TdAttenMix, a general image mixing framework that balances top-down and bottom-up attention to simulate the task-guided mechanism of human gaze to crop the patch and adjust the label mixing ratio. Figure 2 illustrates an overview of our proposed TdAttenMix. Details are given below.

### 4.1 Task Adaptive Attention Guided CutMix

We want to simulate the execution logic of human gaze, which is initially guided by bottom-up features and then strongly influenced by the current task.

**Bottom-up Attention.** We divide the source image $x_A$ and the target image $x_B$ into non-overlapping patches of size $P \times P$. Each image yields a total of $N = \frac{H}{P} \times \frac{W}{P}$ patches. Consequently, $x_A$ and $x_B$ are restructured as $t_A, t_B \in \mathbb{R}^{N\times(P^2C)}$, where each row corresponds to a token and $d = P^2C$. As illustrated in Figure 2, we follow SMMix (Chen et al. 2023) which obtains the attention map across all the image tokens for the bottom-up attention (Dosovitskiy et al. 2021). We obtain $Q = tw_q$, $K = tw_k$, and $V = tw_v$, where $w_q \in \mathbb{R}^{d\times d}$, $w_k \in \mathbb{R}^{d\times d}$, and $w_v \in \mathbb{R}^{d\times d}$ represent the learnable parameters of the fully-connected layers.

**Top-down Attention Guided Module.** The Top-down Attention Guided Module we propose is depicted in Figure 3. The current task at hand is the classification task. Then we extract the corresponding parameters $w_{td} \in \mathbb{R}^{d\times 1}$ from the final fully-connected layer of Vision Transformer, which is based on the current label. The parameter matrix from this layer mirrors the relationship between feature and category mapping. Thus, we can acquire the high-level guidance $V_{td}$ tied to a specific category by calculating it with the image feature $t$. The theory that top-down attention can be implemented by simply augmenting $V_{td}$ to $V$ with $K$ and $Q$ remaining constant was introduced by Shi et al. (Shi, Darrell, and Wang 2023). We ensure that the dimensionality of $V_{td}$ and $V$ is consistent through broadcasting. Furthermore, we accommodated a tunable parameter called balanced factor $\sigma$ within our framework to manage the top-down features $V_{td}$. If $\sigma = 0$, our attention map correlates with the preceding

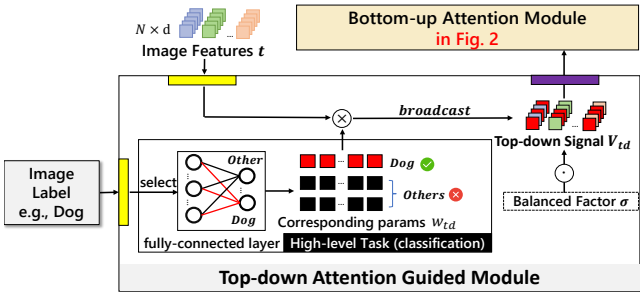


Figure 3: To simulate the top-down mechanism, we designed the Top-down Attention Guided Module by using the image label as the high-level task information to guide image feature generation, resulting in what we refer to as the "top-down signal." This top-down signal then constrains bottom-up attention to focus on regions related to the image label.

bottom-up attention utilized by SMMix (Chen et al. 2023). If $\sigma = 1$, the attention map is finalized by integrating the bottom-up features with the top-down features. As a result, we calculate the task adaptive balanced attention as follows:

$$V_{td} = \sigma \times broadcast(tw_{td}) \tag{3}$$

$$V = V + V_{td} \tag{4}$$

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}})V \tag{5}$$

Subsequently, the resulting task adaptive attention maps, $\alpha_A \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P}}$ and $\alpha_B \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P}}$ corresponding to $x_A$ and $x_B$, are obtained after reshaping operation. Now our attention map is task adaptive which focuses on the object indicated by the current task while ignoring irrelevant high saliency objects. The criterion for cropping is determined by the sum of attention scores within a given region. Then we identify the region with maximum attention scores in the source image, and the region with the minimal attention sources in the target image. Specifically, the center indices are defined as:

$$i_s, j_s = \underset{i,j}{argmax} \sum_{p,q} \alpha_A^{i+p-\lfloor \frac{h}{2} \rfloor, j+p-\lfloor \frac{w}{2} \rfloor} \tag{6}$$

$$i_t, j_t = \underset{i,j}{argmin} \sum_{p,q} \alpha_B^{i+p-\lfloor \frac{h}{2} \rfloor, j+p-\lfloor \frac{w}{2} \rfloor} \tag{7}$$

$h = \lfloor \delta \frac{H}{P} \rfloor$, $w = \lfloor \delta \frac{W}{P} \rfloor$, $\delta = Uniform(0.25, 0.75)$, $p \in \{0, 1, ..., h-1\}$, and $q \in \{0, 1, ..., w-1\}$. Then we use Max-Min Attention Region Mixing (Chen et al. 2023) which uses the maximum attention region to replace the minimal attention region to obtain the new mixed training image $x_M$ as follows:

$$x_M = x_B \tag{8}$$

$$x_M^{i_t+p-\lfloor \frac{h}{2} \rfloor, j_t+q-\lfloor \frac{w}{2} \rfloor} = x_A^{i_s+p-\lfloor \frac{h}{2} \rfloor, j_s+q-\lfloor \frac{w}{2} \rfloor} \tag{9}$$

To verify the validity of the obtained mixed image $x_M$, we examined the prediction accuracy on the mixed image $x_M$. As graphically represented in Figure 4, the prediction

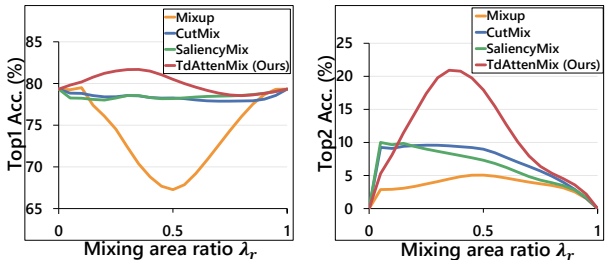


Figure 4: Top-1 accuracy of mixed data. Prediction is counted as correct if the top-1 prediction belongs to $\{y_A, y_B\}$; Top-2 accuracy is calculated by counting the top-2 predictions are equal to $\{y_A, y_B\}$. $\lambda_r$ indicates the area ratio of $x_A$ in mixed image $x_M$.

accuracy of mixed samples can be significantly improved by our method. Notably, for the top-2 accuracy, our TdAttenMix achieves 20.92% while SaliencyMix (Uddin et al. 2021) only reaches 10.00%. This demonstrates that we obtain a mixed image consistent with the labels of source and target images.

## 4.2 Area-Attention Label Mixing

To enhance the precision of the mixed label $y_M$, based on the area ratio (Eq. 2) used by CutMix (Yun et al. 2019), we adjust the area ratio using the attention scores of $\alpha_A$ and $\alpha_B$ at their respective positions within the mixed image $x_M$. More specifically, the final mixing ratio $\lambda$ are defined as follows:

$$\lambda_r = \frac{hwP^2}{HW} \tag{10}$$

$$Att_A = \sum_{p,q} \alpha_A^{i_s+p-\lfloor \frac{h}{2} \rfloor, j_s+q-\lfloor \frac{w}{2} \rfloor} \tag{11}$$

$$Att_B = \sum \alpha_B - \sum_{p,q} \alpha_B^{i_t+p-\lfloor \frac{h}{2} \rfloor, j_t+q-\lfloor \frac{w}{2} \rfloor} \tag{12}$$

$$\lambda_a = \frac{Att_A}{Att_A + Att_B} \tag{13}$$

$$\lambda = \beta \lambda_r + (1 - \beta)\lambda_a \tag{14}$$

$\lambda_r$ is the area ratio of $x_A$ in mixed image $x_M$, $Att_A$ and $Att_B$ is the sum of the task adaptive attention scores at the positions corresponding to $x_M$ in $\alpha_A$ and $\alpha_B$, $\lambda_a$ is the attention ratio of $x_A$ in mixed image $x_M$, $\beta = 0.5$, $\lambda$ is the final mixing ratio of $x_A$ in the mixed image $x_M$. The final mixed label $y_M$ is then defined as follows:

$$y_M = \lambda y_A + (1 - \lambda) y_B \tag{15}$$

Then we obtain the new mixed training sample $(x_M, y_M)$.

## 4.3 Training Objective

Our TdAttenMix framework is independent on any training model and can be used on various mainstream structures. When deployed on a ResNet-based architecture, we employ the standard classification loss and the consistency constraint losses proposed in SMMix (Chen et al. 2023). The

traditional classification loss is defined as follows, where $Y_M$ is the prediction distribution of mixed images. :

$$L_{cls} = CE(Y_M, y_M) \tag{16}$$

Then we require feature consistency constraint losses (Chen et al. 2023), which help features of the mixed images fall into a consistent space with those of the original unmixed images. The feature consistency constraint losses in our TdAttenMix is:

$$L_{con} = L_1(Y_M, \lambda Y_A + (1-\lambda)Y_B) \tag{17}$$

$Y_A$ and $Y_B$ is the prediction distributions of unmixed images $x_A$ and $x_B$. Overall, the training loss is then written as follows:

$$L_{total} = L_{cls} + L_{con} \tag{18}$$

When deployed on a ViT-based architecture, we use the same loss function like SMMix (Chen et al. 2023), which proves to be effective in learning features for mixed samples:

$$L_{fine} = \frac{1}{2}(CE(Y_A, y_A) + CE(Y_B, y_B)) \tag{19}$$

$$L_{total} = L_{cls} + L_{fine} + L_{con} \tag{20}$$

# 5 Experiments

We evaluate TdAttenMix in four aspects: 1) Evaluating image classification tasks on eight different benchmarks, 2) transferring pre-trained models to two downstream tasks, 3) Evaluating the robustness on three scenarios including occlusion and two out-of-distribution datasets. (4) In addition, we have conducted the first quantitative study on the effectiveness of saliency-based methods in reducing image-label inconsistency. Our TdAttenMix is highlighted in gray, and **bold** denotes the best results.

## 5.1 Small-scale Classification

In small-scale classification we use ResNet-18 (He et al. 2016) and ResNext-50 (Xie et al. 2017) to compare the performance. Hyperparameter settings are in the section 1 of the Supplementary. Table 1 shows small-scale classification results on CIFAR-100, Tiny-ImageNet and CUB-200. Compared to the previous SOTA methods, TdAttenMix consistently surpasses AutoMix (+0.08$\sim$ +0.84), PuzzleMix (+1.18 $\sim$ +2.97), MainfoldMix (+0.34 $\sim$ +3.35) based on various ResNet architectures. Moreover, TdAttenMix noticeably exhibits a significant gap with SaliencyMix (+2.50 $\sim$ + 4.25).

## 5.2 ImageNet Classification

Table 1 validates the performance advantage of TdAttenMix over other methods. In particular, TdAttenMix boosts the top-1 accuracy by more than +1% in ResNet-18 (He et al. 2016) and Deit-S (Touvron et al. 2021b) compared with the SaliencyMix baseline and achieves the sota result. It can be noted that TransMix, TokenMix and SMMix also exhibit good top-1 accuracy, but they are limited to ViT-special methods, causing them incompatible with all mainstream architectures (e.g., ResNet). We provide more comparisons with ViT-special methods in Section 2 of the Supplementary, and additional experiments have proven the effectiveness of our TdAttenMix. On the contrary, our TdAttenMix is an independent data argumentation method which is compatible with mainstream architectures.

## 5.3 Downstream Tasks

**Semantic segmentation.** We use ADE20k (Zhou et al. 2017) to evaluate the performance of semantic segmentation task. ADE20k is a challenging scene parsing dataset covering 150 semantic categories, with 20k, 2k, and 3k images for training, validation and testing. We evaluate DeiT backbones with UperNet (Xiao et al. 2018). As shown in Table 2, TdAttenMix improves Deit-S for +1.6% mIoU and +2.5% mAcc.

**Weakly supervised automatic segmentation (WSAS).** We compute the Jaccard similarity over the PASCAL-VOC12 benchmark (Everingham et al. 2015). The attention masks generated from TdAttenMix-DeiT-S or vanilla DeiT-S are compared with ground-truth on the benchmark. The evaluated scores can quantitatively help us to understand if TdAttenMix has a positive effect on the quality of attention map. As shown in Table 2, TdAttenMix improves Deit-S for +3.3%.

## 5.4 Robustness Analysis

**Robustness to Occlusion.** Naseer et al. (Naseer et al. 2021) studies whether ViTs perform robustly in occluded scenarios, where some of most of the image content is missing. Following (Naseer et al. 2021), we showcase the classification accuracy on ImageNet-1k validation set with three dropping settings. (1) Random Patch Dropping. (2) Salient (foreground) Patch Dropping. (3) Non-salient (background) Patch Dropping. As depicted in Figure 5, Deit-S augmented with TdAttenMix outperforms the standard Deit-S across all occlusion levels.

**Out-of-distribution Datasets.** We evaluate our TDAttenMix on two out-of-distribution datasets. (1) The ImageNet-A dataset (Hendrycks et al. 2021). The metric for assessing classifiers' robustness to adversarially filtered examples includes the top-1 accuracy, Calibration Error (CalibError) (Hendrycks et al. 2021; Kumar, Liang, and Ma 2019), and Area Under the Response Rate Accuracy Curve (AURRA) (Hendrycks et al. 2021). (2) The ImageNet-O (Hendrycks et al. 2021). The metric is the area under the precision-recall curve (AUPR) (Hendrycks et al. 2021). Table 3 indicates that TdAttenMix can have consistent performance gains over vanilla Deit-S on the out-of-distribution data.

## 5.5 Image-label Inconsistency Analysis

Previous image mixing methods did not quantitatively validate the image-label inconsistency.Motivated by the fact that gaze reflects human vision (Huang et al. 2020), we propose using the mixed label, which is based on gaze attention, as the ground-truth to validate the problem of image-label consistency. For our experiments, we utilize ARISTO dataset (Liu et al. 2022b) and the corresponding raw images. Since $\lambda$ determines the mixed label, the image-label inconsistency can be represented by the difference between the $\lambda$

| Dataset | CIFAR100 | | Tiny-ImageNet | | CUB-200 | | ImageNet-1k | |
|---|---|---|---|---|---|---|---|---|
| Network | R-18 | RX-50 | R-18 | RX-50 | R-18 | RX-50 | R-18 | Deit-S |
| Vanilla (Li et al. 2023) | 78.04 | 81.09 | 61.68 | 65.04 | 77.68 | 83.01 | 70.04 | 75.70 |
| SaliencyMix (Uddin et al. 2021) | 79.12 | 81.53 | 64.60 | 66.55 | 77.95 | 83.29 | 69.16 | 79.88 |
| MixUp (Zhang et al. 2018) | 79.12 | 82.10 | 63.86 | 66.36 | 78.39 | 84.58 | 69.98 | 79.65 |
| CutMix (Yun et al. 2019) | 78.17 | 81.67 | 65.53 | 66.47 | 78.40 | 85.68 | 68.95 | 79.78 |
| MainfoldMix (Verma et al. 2019) | 80.35 | 82.88 | 64.15 | 67.30 | 79.76 | 86.38 | 69.98 | - |
| SmoothMix (Lee et al. 2020) | 78.69 | 80.68 | 66.65 | 69.65 | - | - | - | - |
| AttentiveMix (Walawalkar et al. 2020) | 78.91 | 81.69 | 64.85 | 67.42 | - | - | 68.57 | 80.32 |
| PuzzleMix (Kim, Choo, and Song 2020) | 81.13 | 82.85 | 65.81 | 67.83 | 78.63 | 84.51 | 70.12 | 80.45 |
| Co-Mixup (Kim et al. 2020) | 81.17 | 82.91 | 65.92 | - | - | - | - | - |
| GridMix (Baek, Bang, and Shim 2021) | 78.72 | 81.11 | 65.14 | 66.53 | - | - | - | - |
| TransMix (Chen et al. 2022) | - | - | - | - | - | - | - | 80.68 |
| TokenMix (Liu et al. 2022a) | - | - | - | - | - | - | - | 80.80 |
| SMMix (Chen et al. 2023) | - | - | - | - | - | - | - | 81.08 |
| AutoMix (Liu et al. 2022c) | 82.04 | 83.64 | 67.33 | 70.72 | 79.87 | 86.56 | 70.50 | 80.78 |
| TdAttenMix (Ours) | **82.36** | **84.03** | **67.47** | **70.80** | **80.71** | **86.72** | **70.74** | **81.19** |
| Gain | +3.24 | +2.50 | +2.87 | +4.25 | +2.76 | +3.43 | +1.58 | +1.31 |

Table 1: Image classification top-1 accuracy (%) on CIFAR-100, Tiny-ImageNet, CUB-200 and ImageNet-1k. We get the performance of previous methods from the OpenMixup (Li et al. 2023) benchmark. Gain: indicates the performance improvement compared with SaliencyMix.

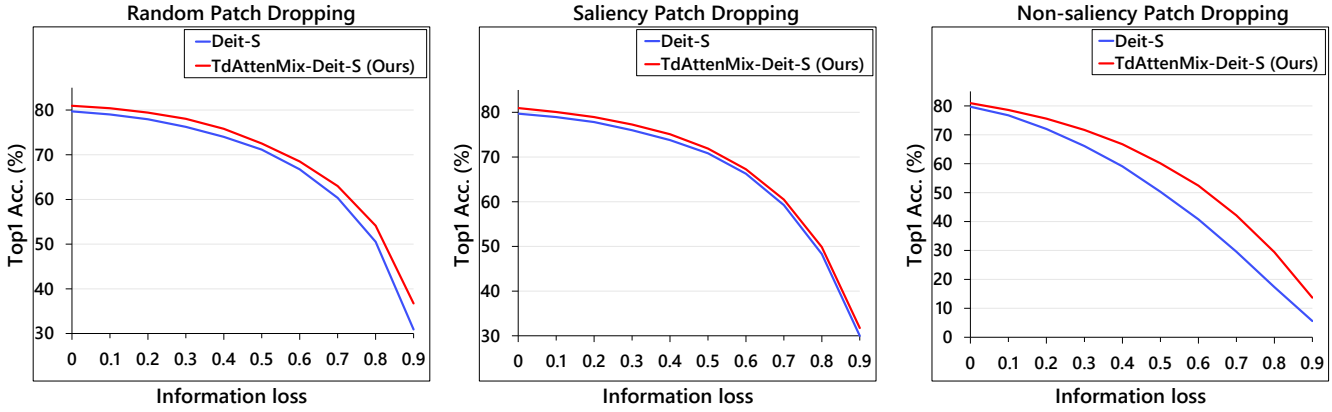


Figure 5: Robustness against occlusion. Model robustness against occlusion with different information loss ratios is studied. 3 patch dropping settings: Random Patch Dropping (left), Salient Patch Dropping (middle), and Non-Salient Patch Dropping (right) are considered.

| Models | Semantic segmentation mIoU (%) | mAcc (%) | WSAS Segmentation JI (%) |
|---|---|---|---|
| Deit-S | 31.6 | 44.4 | 29.2 |
| TdAttenMix-Deit-S | **33.2** | **46.9** | **32.5** |
| Gain | +1.6 | +2.5 | +3.3 |

Table 2: Downstream tasks. Transferring the pre-trained models to semantic segmentation task using UperNet with DeiT backbone on ADE20k dataset; Segmentation JI denotes the jaccard index for weakly supervised automatic segmentation (WSAS) on Pascal VOC.

| Models | Nat. Adversarial Example Top-1 Acc. | Calib-Error↓ | AURRA | Out-of-Dist AUPR |
|---|---|---|---|---|
| Deit-S | 19.1 | 32.0 | 23.8 | 20.9 |
| TdAttenMix-Deit-S | **22.0** | **30.4** | **29.7** | **22.0** |
| Gain | +2.9 | +1.6 | +5.9 | +1.1 |

Table 3: Model's robustness against natural adversarial examples on ImageNet-A and out-of-distribution examples on ImageNet-O.

and ground truth $\lambda_{gt}$ obtained by gaze attention for the same mixed image. So we define the metrics as:

$$Inconsistency = |\lambda_{gt} - \lambda| \tag{21}$$

$\lambda_{gt}$ is calculated based on the real human gaze, $\lambda$ is calculated based on different CutMix variants. As shown in Table 4, the inconsistency is effectively reduced for saliency-based methods. Our TdAttenMix are +7.8 higher than random based CutMix (Yun et al. 2019). The result of TdAttenMix-Bottom-up using only bottom-up attention is close to the results obtained by SaliencyMix (Uddin et al. 2021). This may be due to neither TdAttenMix-Bottom-up nor SaliencyMix has task adaptive ability, thus image-label inconsistency will be stronger than our TdAttenMix. These experimental findings strongly support the notion that

| Method | Inconsistency↓ |
|---|---|
| CutMix (Yun et al. 2019) | 26.2 |
| SaliencyMix (Uddin et al. 2021) | 18.9 |
| TdAttenMix-Bottom-up | 19.0 |
| TdAttenMix | **18.4** |
| Gain | +7.8 |

Table 4: Image-label inconsistency of different saliency-based CutMix variants. TdAttenMix-Bottom-up represents the settings of $\sigma$ to 0 to control the task adaptive balanced attention of TdAttenMix as the standard bottom-up attention. Gain: reduction of error.

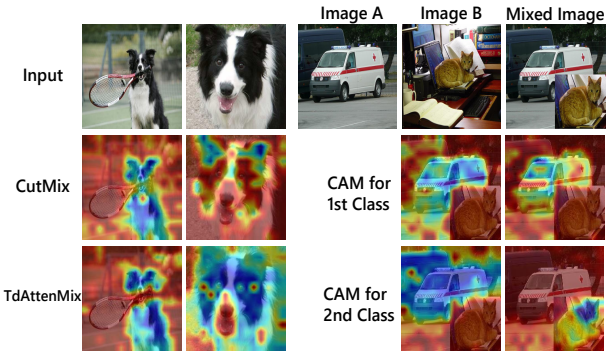


Figure 6: We show the class activation map (Selvaraju et al. 2017) of the models trained with CutMix and TdAttenMix by testing on unmixed and mixed images, respectively. Left: locate objects in the unmixed images. Right: locate objects in the mixed images.

saliency-based CutMix variants enhance training by mitigating image-label inconsistency, with top-down attention being more effective than bottom-up attention.

### 5.6 Visualization

In Figure 6, we visualize the class activation map (Selvaraju et al. 2017) of the models trained with CutMix and TdAttenMix. As shown in the left of Figure 6 that the TdAttenMix can locate object with more precision than the CutMix model in the unmixed images. Furthermore, the right of Figure 6 shows that for the mixed images, the TdAttenMix model can accurately locate objects from two different images. On the contrary, CutMix model focuses only on the class of image A. Our TdAttenMix is guided by task adaptive attention, which ensures that the information in the training data is sufficient enabling superior recognition capacity for mixed images.

### 5.7 Ablation Study

We conduct an ablation study to analyze our proposed TdAttenMix. We use ResNet-18 (He et al. 2016) as the backbone and train it on CUB-200 (Wah et al. 2011).

**Control of Task Adaptive Balanced Attention.** Our TdAttenMix balances top-down and bottom-up attention by adjusting the top-down signal $V_{td}$, enabling a shift from standard bottom-up to top-down attention.We evaluate three different task adaptive balanced attention strategies: 1) $\sigma =$ 0, 2) $\sigma = 0.5$, 3) $\sigma = 1$, 4) $\sigma = 2$, 5) $\sigma = 3$, 6) $\sigma = 4$. These strategies represent a gradual increase in task adaptive ability when bottom-up features are sufficient. This is consistent with the execution logic of human gaze, in which the top-down signal on top of the bottom-up features directs attention to achieve the best results.

| Model | $\sigma$ | Top-1 Acc.(%) |
|---|---|---|
| ResNet-18 (He et al. 2016) | 0 | 80.31 |
| | 0.5 | 80.60 |
| | **1** | **80.71** |
| | 2 | 80.29 |
| | 3 | 79.89 |
| | 4 | 79.50 |

Table 5: Control of task adaptive balanced attention. As shown in Eq. 3, the task adaptive balanced attention can be controlled by $\sigma$ which when $\sigma = 0$ represents standard bottom-up attention.

| Model | $\beta$ | Top-1 Acc.(%) |
|---|---|---|
| ResNet-18 (He et al. 2016) | 0 | 80.20 |
| | 0.3 | 80.29 |
| | **0.5** | **80.71** |
| | 0.7 | 80.19 |
| | 1 | 80.27 |
| | random | 80.29 |

Table 6: Mix ratio $\beta$ of area-attention label mixing.

**Mix ratio $\beta$ of area-attention label mixing.** $\beta$ determines the ratio of area-attention label mixing. We evaluate the performance for several values of $\beta$: 1) fixed as 0, which means only the area ratio is used to mix labels, 2) fixed as 0.3, 3) fixed as 0.5 which assigns equal weighting to the area ratio and attention ratio, 4) fixed as 0.7, 5) fixed as 1, which means only the attention ratio is used to mix labels 6) random value of $\beta$, which means $\beta$ as a random number between 0 and 1. Table 6 shows that the best results are obtained when $\beta$ is set to 0.5.

## 6 Conclusion

This paper proposes TdAttenMix, a general and effective data augmentation framework. Motivated by the superiority of human gaze, we simulate the task-guided mechanism of human gaze to modulate attention. TdAttenMix introduces a new Top-down Attention Guided Module to balance bottom-up attention for task-related regions. Extensive experiments verify the effectiveness and robustness of TdAttenMix, which significantly improves the performance on various datasets and backbones. Furthermore, we quantitatively validate that our method and saliency-based methods can efficiently reduce image-label inconsistency for the first time.

## Acknowledgements

This work was supported by Beijing Natural Science Foundation (L242019). Dr. Lin Gu was also supported by JST Moonshot R&D Grant Number JPMJMS2011 Japan.

## References


Baek, K.; Bang, D.; and Shim, H. 2021. GridMix: Strong regularization through local context mapping. *Pattern Recognition*, 109: 107594.

Belardinelli, A.; Herbort, O.; and Butz, M. V. 2015. Goal-oriented gaze strategies afforded by object interaction. *Vision research*, 106: 47–57.

Borji, A.; and Itti, L. 2012. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1): 185–207.

Buswell, G. T. 1935. How people look at pictures: a study of the psychology and perception in art.

Chen, C.-F. R.; Fan, Q.; and Panda, R. 2021. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 357–366.

Chen, J.-N.; Sun, S.; He, J.; Torr, P. H.; Yuille, A.; and Bai, S. 2022. TransMix: Attend To Mix for Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12135–12144.

Chen, M.; Lin, M.; Lin, Z.; Zhang, Y.; Chao, F.; and Ji, R. 2023. SMMix: Self-Motivated Image Mixing for Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 17260–17270.

Cheng, Z.; Liang, J.; Choi, H.; Tao, G.; Cao, Z.; Liu, D.; and Zhang, X. 2022. Physical Attack on Monocular Depth Estimation with Optimal Adversarial Patches. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 514–532. Cham: Springer Nature Switzerland.

Choi, H. K.; Choi, J.; and Kim, H. J. 2022. Tokenmixup: Efficient attention-guided token-level data augmentation for transformers. *Advances in Neural Information Processing Systems*, 35: 14224–14235.

Connor, C. E.; Egeth, H. E.; and Yantis, S. 2004. Visual attention: bottom-up versus top-down. *Current biology*, 14(19): R850–R852.

Cui, Y.; Yan, L.; Cao, Z.; and Liu, D. 2021. TF-Blender: Temporal Feature Blender for Video Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8138–8147.

Cui, Z.; Zhu, Y.; Gu, L.; Qi, G.-J.; Li, X.; Zhang, R.; Zhang, Z.; and Harada, T. 2022. Exploring Resolution and Degradation Clues as Self-supervised Signal for Low Quality Object Detection. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 473–491. Cham: Springer Nature Switzerland.

Dabouei, A.; Soleymani, S.; Taherkhani, F.; and Nasrabadi, N. M. 2021. SuperMix: Supervising the Mixing Data Augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13794–13803.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111: 98–136.

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

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15262–15271.

Huang, Y.; Cai, M.; Li, Z.; Lu, F.; and Sato, Y. 2020. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29: 7795–7806.

Jiang, M.; Huang, S.; Duan, J.; and Zhao, Q. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1072–1080.

Kim, J.; Choo, W.; Jeong, H.; and Song, H. O. 2020. Co-Mixup: Saliency Guided Joint Mixup with Supermodular Diversity. In *International Conference on Learning Representations*.

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*, 5275–5285. PMLR.

Koch, C.; and Ullman, S. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4): 219–227.

Kumar, A.; Liang, P. S.; and Ma, T. 2019. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32.

Lee, J.-H.; Zaheer, M. Z.; Astrid, M.; and Lee, S.-I. 2020. 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*, 756–757.

Li, S.; Wang, Z.; Liu, Z.; Wu, D.; Tan, C.; Jin, W.; and Li, S. Z. 2023. OpenMixup: A Comprehensive Mixup Benchmark for Visual Classification. arXiv:2209.04851.

Liu, D.; Cui, Y.; Tan, W.; and Chen, Y. 2021a. SG-Net: Spatial Granularity Network for One-Stage Video Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9816–9825.

Liu, D.; Cui, Y.; Yan, L.; Mousas, C.; Yang, B.; and Chen, Y. 2021b. Densernet: Weakly supervised visual localization using multi-scale feature aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6101–6109.

Liu, J.; Liu, B.; Zhou, H.; Li, H.; and Liu, Y. 2022a. TokenMix: Rethinking Image Mixing for Data Augmentation in

Vision Transformers. In *European Conference on Computer Vision*, 455–471.

Liu, Y.; Zhou, L.; Zhang, P.; Bai, X.; Gu, L.; Yu, X.; Zhou, J.; and Hancock, E. R. 2022b. Where to focus: Investigating hierarchical attention relationship for fine-grained visual classification. In *European Conference on Computer Vision*, 57–73. Springer.

Liu, Z.; Li, S.; Wu, D.; Liu, Z.; Chen, Z.; Wu, L.; and Li, S. Z. 2022c. AutoMix: Unveiling the Power of Mixup for Stronger Classifiers. In *European Conference on Computer Vision*, 441–458.

Montabone, S.; and Soto, A. 2010. Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *Image and Vision Computing*, 28(3): 391–402.

Naseer, M. M.; Ranasinghe, K.; Khan, S. H.; Hayat, M.; Shahbaz Khan, F.; and Yang, M.-H. 2021. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34: 23296–23308.

Riche, N.; Duvinage, M.; Mancas, M.; Gosselin, B.; and Dutoit, T. 2013. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Proceedings of the IEEE international conference on computer vision*, 1153–1160.

Schwinn, L.; Precup, D.; Eskofier, B.; and Zanca, D. 2022. Behind the Machine's Gaze: Neural Networks with Biologically-inspired Constraints Exhibit Human-like Visual Attention. *Transactions on Machine Learning Research*.

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.

Shi, B.; Darrell, T.; and Wang, X. 2023. Top-Down Visual Attention from Analysis by Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2102–2112.

Tan, C.; Gao, Z.; Wu, L.; Li, S.; and Li, S. Z. 2022. Hyperspherical Consistency Regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7244–7255.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021a. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021b. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Treisman, A. M.; and Gelade, G. 1980. A feature-integration theory of attention. *Cognitive psychology*, 12(1): 97–136.

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.

Vuyyuru, M. R.; Banburski, A.; Pant, N.; and Poggio, T. 2020. Biologically inspired mechanisms for adversarial robustness. *Advances in Neural Information Processing Systems*, 33: 2135–2146.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

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, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*.

Wu, L.; Lin, H.; Tan, C.; Gao, Z.; and Li, S. Z. 2023. Self-Supervised Learning on Graphs: Contrastive, Generative, or Predictive. *IEEE Transactions on Knowledge and Data Engineering*, 35(4): 4216–4235.

Xia, J.; Zhu, Y.; Du, Y.; and Li, S. Z. 2022. Pre-training Graph Neural Networks for Molecular Representations: Retrospect and Prospect. In *ICML 2022 2nd AI for Science Workshop*.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, 418–434.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

Yarbus, A. L. 2013. *Eye movements and vision*. Springer.

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/CVF International Conference on Computer Vision (ICCV)*.

Zang, Z.; Li, S.; Wu, D.; Wang, G.; Wang, K.; Shang, L.; Sun, B.; Li, H.; and Li, S. Z. 2022. DLME: Deep Local-Flatness Manifold Embedding. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 576–592. Cham: Springer Nature Switzerland.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhao, Z.; Wu, Z.; Zhuang, Y.; Li, B.; and Jia, J. 2022. Tracking objects as pixel-wise distributions. In *European Conference on Computer Vision*, 76–94. Springer.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 633–641.