# INFORMED MIXING – IMPROVING OPEN SET RECOGNITION WITH DEEP DYNAMIC DATA AUGMENTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Conventionally trained image classifiers recently excel in accuracy across diverse tasks. One practical limitation is however that they assume all potential classes to be seen during training, i.e. they can not tell ”I don't know” when encountering an unknown class. Open set recognition (OSR), which solves this problem of detecting novel classes during inference, therefore remains an open problem and is receiving increasing attention. Thereby, a crucial challenge is to learn features that are relevant for unseen categories from given data, for which these features might not be discriminative. Previous work has shown that the introduction of self-supervised contrastive learning to supervised paradigms can support diverse feature learning and thereby benefit OSR. However, the diversity in contrastive learning is commonly introduced through crafted augmentation schemes. To improve upon this aspect and ”optimize to learn” more diverse features, we propose *GradMix*, a data augmentation method that dynamically leverages gradient-based attribution maps of the model during training. The idea is to mask out the activated areas in previous epochs so that the models can pay attention to broader areas and learn to extract features beyond of what is most discriminative for every class. The resulting models are expected to learn more diverse features from the same data source and thus to improve in OSR and model generalization. Extensive experiments on open set recognition, close set classification, and out-of-distribution detection reveal that our method performs well on these tasks that can often outperform the state-of-the-art. GradMix is also beneficial for increasing robustness to common corruptions. In self-supervised learning, GradMix can increase the accuracy of downstream linear classifiers compared with baselines, indicating its benefit for model generalization. Our code is publicly released on `https://anonymous.4open.science/r/comprehensive_osr-8EEF/`

## 1 INTRODUCTION

Deep neural networks have achieved remarkable performance across various fields, particularly in object classification tasks, where they can even surpass human accuracy in certain aspects Russakovsky et al. (2015). However, traditional learning paradigms require the entire dataset to be available prior to training, and models are typically limited to recognizing only the classes present in the training set. In real-world scenarios, however, new and previously unseen classes often emerge, which may be difficult to collect in the short term or even impossible to anticipate beforehand. As a result, the challenge of recognizing novel classes during inference becomes unavoidable. This task is referred to as open set recognition (OSR), and it has garnered significant research attention in recent years Yu et al. (2017); Ge et al. (2017); Hassen & Chan (2020); Dhamija et al. (2018); Yoshihashi et al. (2019); Zhou et al. (2021); Cao et al. (2021); Miller et al. (2021); Chen et al. (2022a); Vaze et al. (2022); Xu et al. (2023); Wang et al. (2024). Throughout this paper, we will refer to known classes as ”in set” and unknown classes as ”open set”.

Existing OSR approaches leverage a range of techniques, including generative models Yu et al. (2017); Ge et al. (2017), novel learning objectives Vaze et al. (2022); Miller et al. (2021), and ensembling Wang et al. (2024). Most of these methods are fundamentally based on the principle of comprehensively modeling the in set data. Theoretical work in Wang et al. (2024) has demonstrated that OSR performance is positively correlated with feature diversity. In this paper, we propose enhancing models' ability to learn more diverse features from in set data through two key strategies. First, in conventional supervised learning paradigms, the phenomenon of class collapse, where sub-class features are suppressed Jing et al. (2021); Xue et al. (2023); Chen et al. (2022b), can hinder the learning of diverse features. By combining supervised and self-supervised learning, this issue can be mitigated, preserving a richer set of sub-class features Chen et al. (2022b). For this, we conduct concise experiments to assess OSR performance and measure feature diversity using the Intrinsic Dimension (ID) metric Ansuini et al. (2019). We compare models trained using supervised contrastive learning (SupCon) Khosla et al. (2020) with those trained using a combination of SupCon and self-supervised contrastive learning (SupCon+SimCLR) Chen et al. (2020) on CIFAR10 and TinyImageNet datasets. The results, presented in Table 1, demonstrate that the SupCon+SimCLR models learn more diverse features and achieve superior OSR performances. The experimental setup follows section 4.1. Based on these findings, we adopt this strategy for open set recognition.

| Method | Dataset | ID | AUROC (in %) |
|---|---|---|---|
| SupCon | CIFAR10 | 8.07 | 67.15 |
| | TinyImageNet | 10.04 | 72.97 |
| SupCon + SSL | CIFAR10 | **14.25** | **88.75** |
| | TinyImageNet | **17.47** | **77.48** |

Table 1: Comparison between SupCon and SupCon+SimCLR on ID and AUROC (higher ID indicates more diverse features). Higher diversity and better OSR performance coincide.

Further, we propose a novel data augmentation technique aimed at enhancing feature diversity by allowing the model to dynamically focus on previously unlearned areas within the data. Leveraging LayerGAM Jiang et al. (2021), a gradient-based method originally developed for visual explanations, we identify the regions most activated during training. We then mask these highly activated areas to generate the augmented data. This technique, which we refer to as GradMix, has shown improved OSR performance when compared to other popular mixing-based augmentation methods. Moreover, GradMix has proven advantageous for classification under common corruptions and for improving downstream classification tasks in self-supervised learning.

Our contributions are as follows:

1. We employ both supervised and self-supervised contrastive learning for open set recognition with the intuition of learning diverse features.
2. We introduce a novel data augmentation method, GradMix, designed to enable models to learn from broader areas within the data. Extensive experiments on OSR, closed set classification, and out-of-distribution detection tasks demonstrate that our method either surpasses or closely matches state-of-the-art performance across multiple datasets.
3. The proposed GradMix enhances classification performance under common corruptions and improves downstream classification tasks in self-supervised learning.

## 2 Related Work

### 2.1 Open Set Recognition

Open set recognition (OSR) involves identifying novel class samples during inference. OSR approaches generally fall into two categories: discriminative and generative models. Discriminative models Yu et al. (2017); Miller et al. (2021); Hassen & Chan (2020); Chen et al. (2022a); Vaze et al. (2022); Xu et al. (2023); Wang et al. (2024) often involve designing novel learning objectives to improve the recognition of known classes. For example, Miller et al. (2021) introduces a loss function that encourages each in set sample to be close to its class center while being as far as possible from other classes. Similarly, in Chen et al. (2022a), each in set class is separated from other classes and the learned open set space. Additionally, some methods in this category, such as Yu et al. (2017), utilize synthesized open set samples generated by techniques like GANs.

Another category of OSR approaches uses generative models to model in set data Ge et al. (2017); Cao et al. (2021). For instance, Ge et al. (2017) employs OpenMax to model in set data and detects open sets by comparing features learned from generators between testing data and in set training data. Similarly, Cao et al. (2021) utilizes a Gaussian mixture variational autoencoder to model in set data. For a comprehensive review of open set recognition methods, see Mahdavi & Carvalho (2021).

### 2.2 Mixing Data Augmentation

Mixing data augmentation methods involve combining raw samples with other images from within or outside the dataset. These methods can be broadly categorized into two types. The first category is pixel-wise mixing, initially introduced as *Mixup* Zhang et al. (2017). In this approach, augmented data is created by taking a weighted average of the original samples and their corresponding labels. An extension of this method, known as *manifold Mixup* Verma et al. (2019), extends the concept from mixing raw inputs to mixing latent representations. This extension has been shown to improve classification performance.

The second category is patch-wise mixing, where a portion of the original samples is replaced. An early method in this category is *CutOut* DeVries & Taylor (2017), which involves masking out a patch of the original samples. *CutMix* Yun et al. (2019) extends this idea by replacing patches with resized samples from the same minibatch. *SaliencyMix* Uddin et al. (2020) further refines CutMix by identifying the most representative areas in the original data using saliency detection techniques Montabone & Soto (2010). In addition to these methods, there are many other mixing data augmentation approaches. For a comprehensive review, see Lewy & Mańdziuk (2023).

## 3 Method

Our approach is centered on the concept of learning diverse features through two primary components: the incorporation of self-supervised learning within supervised paradigms and the application of gradient-based mixing augmentation. In this section, we introduce our framework for open set recognition and outline the key methodological components that drive its effectiveness. For clarity in the following discussion, we adopt the following notation: uppercase letters represent sets of scalers, while lowercase letters refer to individual samples within those sets. Vectors are indicated by bold letters, and uppercase bold letters are used to denote sets of vectors or matrices.

### 3.1 Supervised and Self-supervised Contrastive Learning

Supervised learning can sometimes lead to class collapse Jing et al. (2021), a phenomenon where the features of sub-classes become suppressed, resulting in diminished feature diversity. On the other hand, self-supervised learning has recently gained substantial attention for its ability to extract high-quality features without the need for labeled data Chen et al. (2020); He et al. (2020); Grill et al. (2020). Several studies have demonstrated that combining self-supervised learning with supervised learning helps prevent class collapse and promotes greater feature diversity, as shown in works like Xue et al. (2023); Chen et al. (2022b). Based on these findings, we integrate this approach into our open set recognition framework.

We pair SimCLR Chen et al. (2020) and supervised contrastive learning (SupCon) Khosla et al. (2020) in our method due to their shared foundational principles. SimCLR is a self-supervised contrastive learning method that learns representations by minimizing the distances between the original data and its augmented counterpart in the feature space. For a sample $\mathbf{x}_i$ and its augmented version $\mathbf{x}_j$ within a minibatch of size $2N$ (with $N$ original samples and their corresponding augmentations), their representations are denoted as $\mathbf{z}_i$ and $\mathbf{z}_j$. $\mathbf{x}_i$ and $\mathbf{x}_j$ form a positive pair, while $\mathbf{x}_i$ and all other samples in the minibatch constitute negative pairs. This learning objective is derived from *InfoNCE* Oord et al. (2018) and is described in equation 1. SimCLR is designed to maximize the mutual information between $\mathbf{z}_i$ and $\mathbf{z}_j$ Oord et al. (2018), encouraging the model to capture more of the inherent features in the sample itself. In equation 1, $\mathbb{1}_{[k \neq i]} \in 0, 1$ is an indicator function that flags negative pairs, i.e., $\mathbf{x}_i$ paired with any sample other than $\mathbf{x}_j$ in the minibatch. The function $sim(\cdot)$ measures similarity, typically using cosine function. The temperature parameter $\tau$ is a hyperparameter that controls the learning dynamics between positive and negative pairs. It is clear from equation 1 that minimizing $L_{simclr}(i)$ is equivalent to maximizing $sim(\mathbf{z_i}, \mathbf{z_j})$ while minimizing $sim(\mathbf{z_i}, \mathbf{z_k})$. The loss for the entire minibatch is the average of $L_{simclr}(i)$ across all samples.

$$L_{simclr}(i) = -\log \frac{\exp(sim(\mathbf{z_i}, \mathbf{z_j})/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(sim(\mathbf{z_i}, \mathbf{z_k})/\tau)} \tag{1}$$

SupCon extends SimCLR to a supervised fashion and has been shown to outperform cross-entropy loss in terms of model generalization for object classification tasks Khosla et al. (2020). In SupCon, positive pairs are defined as samples belonging to the same class, while negative pairs consist of samples from different classes. Here, labels are denoted by $\ell$. Similar to equation 1, the loss function for a sample $\mathbf{x}_i$ is given in equation 2. The set of all positive pairs for $\mathbf{x}_i$ is denoted by $P(i)$, where $P(i) = \{x_p | 1 \leq p \leq 2N, p \neq i, \ell_i = \ell_p\}$. As with SimCLR, the loss for the entire minibatch is the average of $L_{supcon}(i)$ across all samples.

$$L_{supcon}(i) = \frac{1}{|P(i)|} \sum_{p \in P(i)} -\log \frac{\exp(sim\,(\mathbf{z}_i \cdot \mathbf{z}_p)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{l_i \neq l_k} \exp(sim\,(\mathbf{z}_i \cdot \mathbf{z}_k)/\tau)} \tag{2}$$

We combine $L_{simclr}$ and $L_{supcon}$ linearly with weights $\theta$ and $\lambda$ respectively in a multi-task learning fashion to minimize class collapse, i.e.:

$$L_{contra} = \theta * L_{supcon} + \lambda * L_{simclr} \tag{3}$$

### 3.2 Gradient-based Mixing Augmentation

To enable the models to learn more diverse features, we propose an additional data augmentation method that encourages the models to focus on broader areas within the data. Inspired by the mixing augmentation techniques discussed in section 2.2, our approach masks out a portion of the learned areas after each epoch during training. Rather than random masking, we hope the model can pay more attention to unlearned areas in data. To address this, we propose detecting the activated areas in the input data directly using the models during training, combined with visual explanation techniques, such as GradGAM Selvaraju et al. (2017) and LayerGAM Jiang et al. (2021).

One foundational method for visual explanation is GradGAM, which visualizes the learning process by performing a weighted combination of forward activation maps. We denote the localization map of the activated areas computed by GradGAM as $\mathbf{M}_{gradcam}$. As shown in equation 4, $\mathbf{M}_{gradcam}$ is the ReLU-filtered weighted sum of the forward activation maps $\mathbf{A}^k$, where $k$ represents the index of the feature maps in the convolutional layer. As expressed in equation 5, the partial gradient of the model's output $l$ (In our context, $l$ represents the loss function used during training and $l$ is scalar) with respect to the feature map $\mathbf{A}^k$ indicates the importance of $\mathbf{A}^k$ to the final output. Practically, it is calculated as the average of the partial gradients with respect to each spatial position in $\mathbf{A}^k$, where $i$ and $j$ are the spatial indices.

$$\mathbf{M}_{gradcam} = ReLU(\sum_k \alpha_k \mathbf{A}^k) \tag{4}$$

$$a_k = \frac{\partial l}{\partial \mathbf{A}^k} = \frac{1}{|\mathbf{A}^k|} \sum_i \sum_j \frac{\partial l}{\partial \mathbf{A}^k_{i,j}} \tag{5}$$

However, the activation maps computed using GradGAM tend to be coarse and lack precision in detecting activation areas in the input. To address this, we apply an improved method called LayerGAM Jiang et al. (2021). Unlike GradGAM, LayerGAM can leverage earlier layers in CNNs to capture finer-grained activation maps. Instead of using a single weight coefficient for the entire feature map $\mathbf{A}^k$, as in equation 4 and equation 5, LayerGAM assigns individual weights to each location in $\mathbf{A}^k$, as shown in equation 6. The weighted feature maps $\mathbf{A}^k_{i,j}$ are then summed along the channel dimension to produce the final activation map $\mathbf{M}_{layercam}$, as described in equation 7.

$$\tilde{\mathbf{A}}^k_{i,j} = ReLU(a^k_{i,j})\mathbf{A}^k_{i,j} \tag{6}$$

$$\mathbf{M}_{layercam} = ReLU(\sum_k \tilde{\mathbf{A}}^k) \tag{7}$$

The activation maps computed after each epoch during training are used for data augmentation. A graphical illustration of our method is provided in figure 1. The ratio between the side lengths of the patched area and the original image, denoted as $\gamma$, follows a uniform distribution, i.e., $\gamma \sim \mathbf{U}(\gamma_{min}, \gamma_{max})$. In our work, we set $\gamma_{min}$ to 0.1 and $\gamma_{max}$ to 0.5, ensuring that the patches can neither completely cover the entire object in the original samples nor exceed the image margins. We call our method *GradMix*. GradMix is applied to the self-supervised component in equation 3,

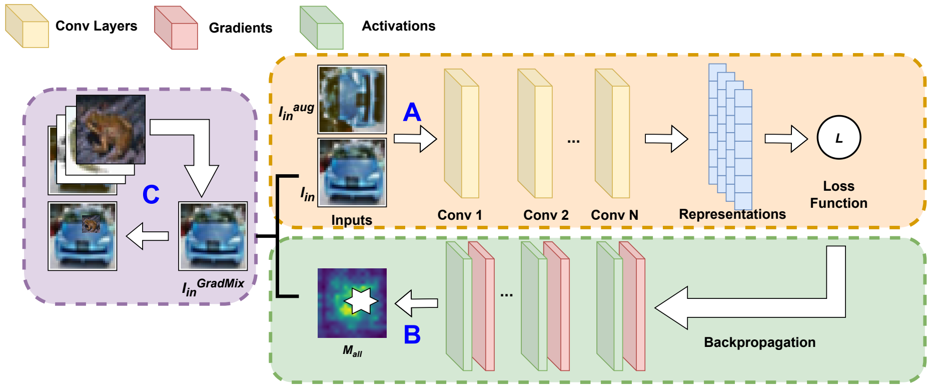


Figure 1: Graphical illustration of of GradMix. Three blocks with dashed borderlines refer to the procedures: **A**. Data is fed into the feature extractor during forward propagation; **B**: Activation maps are computed using the internal feature maps and LayerGAM method. The most activated area is selected (highlighted using white star in the graph); **C**: A random sample from the same minibatch is resized and patched on the most activated area.

and its loss values are weighted by the ratio of the patched area to the original image, which is $\gamma^2$ assuming the data is square-sized. The overall learning objective then becomes:

$$L = \theta * L_{supcon} + \lambda * (L_{simclr} + \gamma^2 * L_{simclr}^{GradMix}) \tag{8}$$

### 3.3 OSR Framework

The previous two sections have outlined our approach for diverse feature learning. In this subsection, we present our method for detecting open set samples. Given a training set $\boldsymbol{X}_{train}$, where $\boldsymbol{X}_{train} = \mathbf{X}_1, ..., \mathbf{X}_C$, with $C$ classes, and $\boldsymbol{Z}_{train} = \mathbf{Z}_1, ..., \mathbf{Z}_C$ denotes their corresponding feature representations, we compute classification scores for the test set $\boldsymbol{X}_{test}$ following algorithm 1. The classification score $s_c$ for each testing sample is the sum of the top-k similarities with its closet close set class, which will also be used for detecting open sets through comparing with pre-defined thresholds.

**Algorithm 1** Open Set Recognition Framework

```
**Input**: Feature encoder $\mathbb{E}$, $\boldsymbol{X}_{train}$, $\boldsymbol{X}_{test}$, and hyper-parameter $k$
**Output**: Set of classification scores, $G_{test}$, and predicted labels, $\hat{Y}_{test}$, for each $\mathbf{X}_{test}$ in $\boldsymbol{X}_{test}$.
**Initialize**: $\hat{Y}_{test} = \emptyset$

**for** $\mathbf{X}_{test}^i$ in $\boldsymbol{X}_{test}$ **do**
    $\mathbf{z}_{test}^i = \mathbb{E}(\mathbf{X}_{test}^i)$
    **for** $\mathbf{Z}_c$ in $\boldsymbol{Z}_{train}$ **do**
        Similarities between $\mathbf{z}_{test}$ and each element in $\mathbf{Z}_{train}$, $S_c = \mathbf{sim}(\mathbf{z}_{test}, \mathbf{Z_c})$
        Select the top $k$ values in $S_c$, $S_c^k = \mathbf{max}(S_c, k)$
        $G_{test}^i = G_{test}^i \cup \{\mathbf{sum}(S_c^k)\}$
    **end for**
    $\hat{y}_{test}^i = \mathbf{argmax}(G_{test}^i)$.
    $\hat{Y}_{test} = \hat{Y}_{test} \cup \{\hat{y}_{test}^i\}$.
    $s_c = \mathbf{max}(G_{test}^i)$
**end for**
```

# 4 Experiments

In this section, we present the experimental settings and results for open set recognition, closed set classification, out-of-distribution detection, corrupted image classification and model generalization.

## 4.1 Open Set Recognition

Following the OSR testbench that widely used in the literature Chen et al. (2022a); Vaze et al. (2022); Neal et al. (2018), we evaluate our method on six split protocols: MNIST, SVHN, CIFAR10, CIFAR+10, CIFAR+50, and TinyImageNet, which are created using the source datasets MNIST Deng (2012), SVHN Netzer et al. (2011), CIFAR10 Krizhevsky et al. (2009a), CIFAR100 Krizhevsky et al. (2009b), and TinyImageNet Deng et al. (2009). For each protocol, we repeat the experiments five times with different splits of in set and open set classes, and then average the results. In addition to GradMix, we apply the standard augmentations used in SimCLR Chen et al. (2020) for all protocols except MNIST, to avoid producing confusing samples (e.g., vertically flipping a ”7” could resemble a ”1”). Due to the same reason and the simpleness of the dataset, GradMix is not utilized to MNIST. But it can still surpass state-of-art performances. The number of training classes (denoted by $I$), the total number of testing classes (denoted by $S$), and their data sources for each protocol are provided in table 6 in appendix A.1. The complexity of each protocol is measured by the *openness* metric, $O = 1 - \sqrt{I/S}$, which describes the proportion of open set classes relative to the total number of classes. All models are trained on known classes only and evaluated on test samples from both known and unknown classes. ResNet18 He et al. (2016) is used as the feature encoder backbone in all experiments in this section, with the output feature dimension set to 128.

**Results** The area under the receiver operating characteristics curve (AUROC) is the metric for evaluating OSR performances. The receiver operating characteristics curve is plotted with the true positive rate against the false positive rate, which can be interpreted as how much the detection score histograms of open set and and in set samples are overlapped. AUROC is threshold-independent and higher AUROC represents better performance at detecting open set samples.

We compare our results with vanilla cross entropy, and the state-of-the-arts (SToAs) in literature, namely Openmax Bendale & Boult (2016), G-Openmax Ge et al. (2017), OSRCI Neal et al. (2018), C2AE Oza & Patel (2019), GRROSR Perera et al. (2020), PROSER Zhou et al. (2021), APRL Chen et al. (2022a), APRL-CS Chen et al. (2022a), OpenAUC Wang et al. (2022), ConOSR Xu et al. (2023), and MEDAF Wang et al. (2024). The results are listed table 2. Almost all methods show excellent performance on MNIST and SVHN protocols, expecially for MNIST, which is almost reaching $100\%$. Our method achieves the best performance on most complex protocols and similar performance across all other protocols. Particularly, it demonstrates clear advantage on TinyImageNet with over $1\%$ of increase. We think the reasons lie in the complexity of the dataset. The performance of contrastive learning can be increased when harder and more variant negative samples are introduced during training Shu & Lampos (2024). And complex data can provide larger room for GradMix to mine more features.

**Ablations** We perform two ablation studies: (1) evaluating OSR performance using various data augmentation mixing techniques, and (2) assessing the impact of utilizing different deep layers for computing activation maps in GradMix. Consistent with previous experiments, OSR performance is measured using the AUROC metric and each result is the average of five trials. All experiments are repeated with CIFAR10 and TinyImageNet protocols. We fix $k = 10$ in the OSR framework in each study for fair comparison.

Impact of Different Augmentation Methods We (re)-implemented and tested four augmentation methods, i.e., Mixup, CutMix, vanilla GradMix (activation maps are computed using GradGAM) and GradMix. Furthermore, we vary the hyper-parameter $\alpha$ in Mixup and CutMix. Larger $\alpha$ leads to stronger augmentations (see Zhang et al. (2017); Yun et al. (2019) for details). For vanilla GradMix and GradMix, we pick layer `conv5_2` in ResNet18 to compute activation maps.

The results, as shown in Figure 2, clearly demonstrate that models incorporating augmentations achieve superior OSR performance across both protocols. GradMix, in particular, improves AUROC by $3\%$ compared to vanilla SimCLR on both protocols. Furthermore, more advanced and stronger augmentation techniques result in greater performance gains, with GradMix consistently outperforming all other methods.

| Protocols<br>Methods | MNIST | SVHN | CIFAR10 | CIFAR+10 | CIFAR+50 | TinyImgNet |
|---|---|---|---|---|---|---|
| Cross Entropy | 97.8 | 88.6 | 67.7 | 81.6 | 80.5 | 57.7 |
| Openmax Bendale & Boult (2016) | 98.1 | 89.4 | 69.5 | 81.7 | 79.6 | 57.6 |
| G-Openmax Ge et al. (2017) | 98.4 | 89.6 | 67.5 | 82.7 | 81.9 | 58.0 |
| OSRCI Neal et al. (2018) | 98.8 | 90.1 | 69.9 | 83.8 | 82.7 | 58.6 |
| C2AE Oza & Patel (2019) | 98.9 | 92.2 | 89.5 | 95.5 | 93.7 | 74.8 |
| GRROSR Perera et al. (2020) | - | 93.5 | 80.7 | 92.8 | 92.6 | 60.8 |
| PROSER Zhou et al. (2021) | - | 94.3 | 89.1 | 96.0 | 95.3 | 69.3 |
| APRL Chen et al. (2022a) | 99.6 | 96.3 | 90.1 | 96.5 | 94.3 | 76.2 |
| APRL-CS Chen et al. (2022a) | 99.7 | 96.7 | 91.0 | 97.1 | 95.1 | 78.2 |
| OpenAUC Wang et al. (2022) | 99.4 | 95.0 | 89.2 | 95.2 | 93.6 | 75.9 |
| ConOSR Xu et al. (2023) | 99.7 | **99.1** | **94.2** | 98.1 | 97.3 | 80.9 |
| MEDAF Wang et al. (2024) | - | 95.7 | 86 | 96 | 95.5 | 80.0 |
| GradMix (Ours) | **99.8** | 94.7 | 91.33 | **98.62** | **97.64** | **81.92** |

Table 2: The area under the ROC curve (AUROC) (in %) for detecting known and unknown samples (Partial results of the baseline methods are from Chen et al. (2022a) and Xu et al. (2023)). " − " indicates there are no given results in literature. Bold numbers indicate the best results. GradMix outperforms the SoTAs in four out of six protocols and the increase is over 1% on large-scale TinyImageNet protocol.

SELECTION OF DEEP LAYERS IN GRADMIX To investigate the impact of different deep layers and their aggregation used for computing activation maps in GradMix, we evaluate the OSR performances using GradMix computed from various layers of ResNet18, specifically conv4_2, conv5_2, conv3_2 + conv4_2, conv4_2 + conv5_2, conv3_2 + conv4_2 + conv5_2. The latter three configurations represent the aggregation of layers to explore how aggregating features from multiple depths affects performance. A detailed illustration of the layer aggregation is in Appendix A.2.

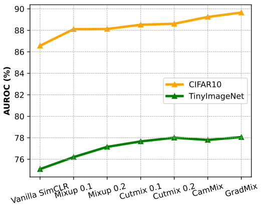


Figure 2: OSR performances of the models with different augmentation methods on CIFAR10 and TinyImageNet protocols. Clear improvements can be brought by extra data augmentations. And GradMix performs best among all augmentation methods.

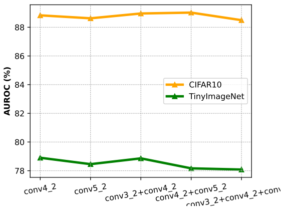


Figure 3: OSR performances of models using GradMix computed from different layers and layer aggregation in ResNet18. The results indicate that utilizing different layers, as well as aggregating multiple layers, produces no significant changes in performance.

The results are summarized in figure 3, which indicate that the utilization of different layers, as well as layer aggregation, introduces no significant performance changes, especially for higher resolution data. This suggests that the effectiveness of GradMix remains consistent regardless of the specific layers used for activation map computation. We hypothesize that this is due to the consistency of the most highly activated pixels across layers and it is invariant to linear aggregation of layers when computing the activation maps. In future work, we plan to expand the location of mixing areas to top-$k$ ($k > 1$) most activated areas within the data, with the aim of further enhancing the performance.

### 4.2 CLOSE SET CLASSIFICATION

In order to evaluate the proposed method on close set classification tasks, we train models on full CIFAR10, CIFAR100, and half TinyImageNet (the first 100 classes) datasets and test their classification accuracy as in Chen et al. (2022a); Xu et al. (2023); Wang et al. (2024). We compare the results with Cross-Entropy, ARPL, ConOSR, and MEDAF. Results are shown in table 3.

**Results** All the baselines are supervised learning methods, which are in principle better at close set classification, whereas our method employs self-supervised. For CIFAR10, there exist very confusing in set class pairs in some trails, e.g., deer and horse, which can significantly lower the close set classification accuracy. The models can learn more non-discriminative features with self-supervised learning, which can worse the problem. However, our method can still achieve significantly better performances on CIFAR100 and TinyImageNet and fair results for CIFAR10. Especially for TinyImageNet (almost 2%), we think the reasons are similar as for open set recognition discussed above that complex datasets offer greater opportunities for contrastive learning to demonstrate its effects. We evaluate how GradMix helps with generalization to common corruptions in section 4.4 and improving downstream linear classification accuracy for self-supervised learning in section 4.5. Even though the introduction of self-supervised learning can lower the close set classication accuracy, GradMix can reduce this effect.

| Method | CIFAR10 | CIFAR100 | TinyImgNet |
|---|---|---|---|
| Cross-Entropy | 94 | 71.6 | 63.7 |
| ARPL | 94.1 | 72.1 | 65.7 |
| ConOSR | 94.6 | 73 | 66.1 |
| MEDAF | 95.4 | 77 | 70.6 |
| GradMix (Ours) | 94.1 | **78** | **72.46** |

Table 3: Comparison on close set classification performances (classification accuracy in %) on CIFAR10, CIFAR100, and half TinyImageNet datasets. The results of baselines are from Chen et al. (2022a); Xu et al. (2023); Wang et al. (2024). Our method can outperform the baselines on the two larger datasets, CIFAR100 and TinyImageNet, with significant increases.

## 4.3 Out of Distribution Detection

We validate our method on out-of-distribution detection (OOD) tasks. Following the settings for OOD in Chen et al. (2022a); Wang et al. (2024), we take CIFAR10 as in distribution set and CIFAR100 as well as SVHN as out of distribution sets. vanilla cross entropy, ARPL, ARPL-CS, and MEDAF are baselines. Evaluation metrics are TNR, AUROC, DTACC, as well as AUIN/AUOUT. TNR stands for the true negative rate when the true positive rate (TPR) is $95\%$. Let TP, TN, FP, FN represent true positive, true negative, false positive and false negative respectively (same for the following text), $\mathrm{TNR} = \mathrm{TN}/(\mathrm{TP} + \mathrm{TN})$ when $\mathrm{TPR} = \mathrm{TP}/(\mathrm{FP} + \mathrm{FN})$ is $95\%$. DTACC refers to the maximum of detection accuracy across all possible thresholds. AUIN or AUOUT is the area under the precision-recall curve (AUPR) when in- or out-of-distribution samples are specified as positive respectively.

**Results** The results are given in table 4. Our method can outperform the baselines especially on the SVHN dataset. We believe the reasons lie in the model having learned many data-dependent features that overlap with the close OOD dataset CIFAR100.

| Method | In: CIFAR10 Out: CIFAR100 | | | | | In: CIFAR10 Out: SVHN | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | TNR | AUROC | DTACC | AUIN | AUOUT | TNR | AUROC | DTACC | AUIN | AUOUT |
| Cross Entropy | 31.9 | 86.3 | 79.8 | 88.4 | 82.5 | 32.1 | 90.6 | 86.4 | 88.3 | 93.6 |
| ARPL | 47.0 | 89.7 | 82.6 | 90.5 | 87.8 | 53.8 | 93.2 | 87.2 | 90.3 | 95.8 |
| APRL-CS | 48.5 | 90.3 | 83.4 | 91.1 | 88.4 | 79.1 | 96.6 | 91.6 | 94.8 | 98.0 |
| MEDAF | - | 92.5 | 85.4 | 93.2 | 91.1 | - | 99.1 | 95.3 | 98.0 | 99.6 |
| GradMix (Ours) | **85.68** | **93.22** | **96.57** | 92.5 | 90.86 | **99.94** | 98.7 | **99.81** | **98.5** | 95.3 |

Table 4: Results for out of distribution detection. Results of baselines are from Chen et al. (2022a); Wang et al. (2024). Our method can surpass the baseline and competing works in both settings.

## 4.4 Generalization to Common Corruptions

We evaluate the generalization of the classifiers trained using our method using the testbench proposed in Hendrycks & Dietterich (2018), which tests robustness of the models on common corruptions. Fifteen types of corruptions are synthesised in this testbench, namely Gaussian noise, shot noise, impulse noise, defocus blur, frosted glass blur, motion blur, zoom blur, snow, frost,

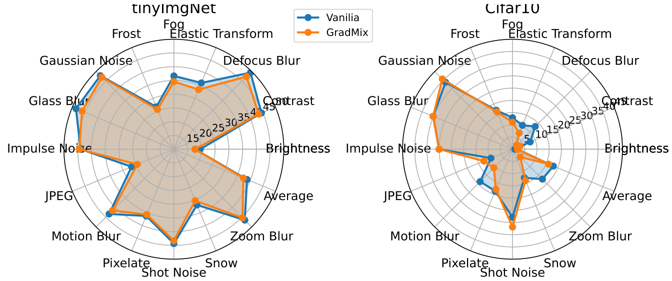


Figure 4: Classification accuracy drop of the models trained with and without GradMix on cifar10 and TinyImageNet datasets. The average accuracy drop is lower for the models with GradMix on the both datasets. And models with GradMix are more robust on most corruption types.

| Severity | Cifar10 | | TinyImgNet | |
|---|---|---|---|---|
| | SimCLR | GradMix | SimCLR | GradMix |
| 1 | 7.01 | **6.66** | 22.61 | **25.25** |
| 2 | 11.18 | **10.04** | 37.02 | **31.49** |
| 3 | 15.93 | **14.03** | 39.43 | **38.91** |
| 4 | 23.01 | **20.49** | 46.23 | **45.23** |
| 5 | 33.29 | **28.64** | 49.31 | **48.91** |
| Avg. | 18.08 | **15.97** | 38.81 | **37.36** |

Table 5: Avg. accuracy drop (lower is better) for different corruption severities w/o (left) and with GradMix (right) on Cifar10 and TinyImgNet.

fog, brightness, contrast, elastic, piexlate, JPEG compression. Each type of corruption has five levels of severity and therefore the classifiers are tested with 75 repeats. We use the accuracy drop, $D_{c,s} = A_{clean} - A_{c,s}$, as metric to measure the model robustness to the corruption type $c$ of severity $s$. We measure $D_{c,s}$ on the models trained with and without GradMix on CIFAR10 and TinyImageNet datasets. It is easy to infer that lower $D$ indicates higher robustness to the corruptions.

**Results** The results are demonstrated in figure 4 and table 5. We average $A_{c,s}$ over the five severity levels for each corruption type, i.e., $\bar{D}_c = \frac{1}{5}\sum_s D_{c,s}$, and the overall accuracy drop $\bar{D}$ is the average of $\bar{D}_c$ over the 15 corruption types. $\bar{D}_c$ of the models with GradMix are lower on both datasets. And GradMix models are more robust to most of the corruption types.

### 4.5 Model Generalization across Datasets

To validate the effectiveness of GradMix for model generalization, we apply it to self-supervised learning settings and test the downstream linear classification performances. In order to increase the experimental diversity, the models are trained with SimCLR and MoCo He et al. (2020) and different architectures of ResNet18, ResNet34, and ResNet50. For a fair comparison, the batch size for all models is 256 and the output dimensions are 128. All models are trained on TinyImageNet dataset and the settings for linear classifiers are all identical. For MoCo models, the queue size and momentum of updating key encoder are 8196 and 0.999 respectively. The baselines are SSL without extra augmentation and SSL with CutMix, which is researched in Ren et al. (2022). We record top-1 and top-5 accuracy for evaluation.

**Results** Figure 5 (Left) demonstrates the top-1 and top-5 accuracy of the self-supervised learning models trained after 200 epochs. Both top-1 and top-5 accuracy increase with the applying of extra augmentation methods and the improvements brought by GradMix are higher.

Figure 5 (Right) shows the accuracy change of the linear classifiers with different SSL model training epochs. Longer training is overall beneficial for all methods. But the improvements are higher and faster for these with GradMix. It can be concluded from the above findings that GradMix is more effective for model generalization than other baseline augmentation methods.

## 5 Analysis & Discussion

To further assess the effectiveness of GradMix, we visualize the feature activations using LayerGAM. A selection of samples is shown in figure 6 (Left), with additional examples provided in Appendix A.3. These examples are drawn from CIFAR10 and TinyImageNet datasets. It can be observed that the models trained with GradMix demonstrate larger activated areas, suggesting that GradMix enables the models to focus on a broader range of regions in the data. This broader focus likely contributes to the learning of more diverse and informative features. To quantitatively assess the activated areas in the data, we measure the number of higher-valued pixels in the activation maps, denoted as $C_{\mathbf{M}} = \sum_{i,j} \mathbb{1}_{\mathbf{M}_{i,j} > \tau}$, where $\tau$ is a predefined threshold. We vary $\tau$ from

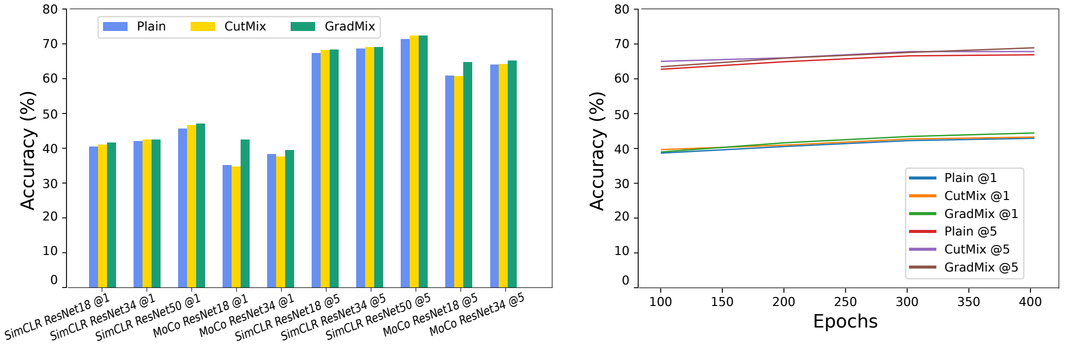
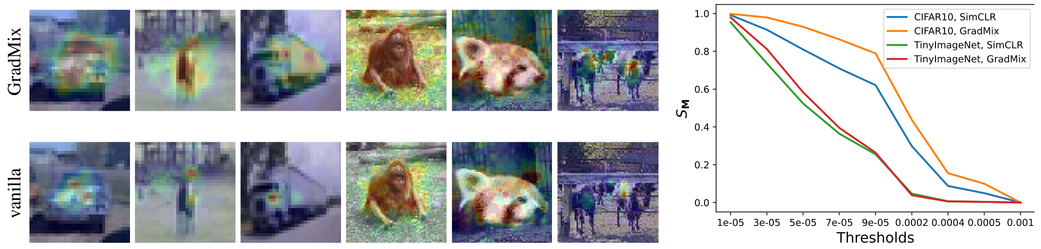


Figure 5: Top-1 and top-5 linear classification accuracy (in $\%$) on TinyImageNet. **(Left)** Features learned using self-supervised learning with different augmentation methods. The results show that the deployment of extra data augmentation can increase the downstream linear classification accuracy for self-supervised learning. GradMix can bring higher improvements than other augmentation methods, indicating its benefit for model generalization. **(Right)** Improvement over training iterations of ResNet18 trained with SimCLR and different augmentation methods. GradMix improves the accuracy more and faster than other methods.



LayerGAM visualizations

Attribution Density

Figure 6: Comparison on the attribution maps of the models trained on CIFAR10 protocol, trained with and without GradMix. **(Left)** Visualizations: (top) model trained with GradMix on CIFAR10 and TinyImgNet; (bottom) model trained without GradMix. **(Right)** Change of $S_{\mathbf{M}}$ with $\tau$. The values of GradMix models are always higher, indicating broader activated areas in data.

$10^{-5}$ to $10^{-3}$ and plot the fraction of $C_{\mathbf{M}}$ to the resolution of $\mathbf{M}$ (denoted as $S_{\mathbf{M}} = \frac{C_{\mathbf{M}}}{|\mathbf{M}|}$) in figure 6(Right). $S_M$ with GradMix is always larger than these with only vanilla SimCLR. Consistent with the visualizations, GradMix leads the models to focus on a larger portion of the data, suggesting it helps capture a broader range of features.

## 6 Conclusion

In this work, we proposed a novel approach for open set recognition, combining self-supervised learning with a gradient-based data augmentation method, guided by the idea of learn diverse features. Experimental results demonstrate that our approach surpasses most state-of-the-art methods in OSR, closed set classification, and OOD tasks. Additionally, GradMix enhances model robustness against common corruptions and boosts downstream linear classification performance in self-supervised learning, further highlighting its effectiveness for improving model generalization.

**Limitations** Recent works have reported that model sparsity can increase model robustness to adversarial attacks Timpl et al. (2022). Besides the experiments in section 4.4 on common data corruption, GradMix could be further evaluated for its effectiveness on adversarial attacks, which is beyond our current study. Furthermore, as analyzed in Chen et al. (2022c), model sparsification can allow new neural connections to grow and help the models to escape bad local minima, and hence reduce overfitting. It remains to explore if GradMix, probably other data augmentation methods in general, can achieve the same effects, which can also be the key of learning diverse attribute-related features.

## Reproducibility Statement

We used open-source data and models. We also provide detailed descriptions of our implementation in the appendix. The source code and evaluation results will be made publicly available upon acceptance.

## References

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*, pp. 1563–1572, 2016.

Alexander Cao, Yuan Luo, and Diego Klabjan. Open-set recognition with gaussian mixture variational autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6877–6884, 2021.

Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022a.

Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Ré. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning*, pp. 3090–3122. PMLR, 2022b.

Tianlong Chen, Zhenyu Zhang, Pengjun Wang, Santosh Balachandra, Haoyu Ma, Zehao Wang, and Zhangyang Wang. Sparsity winning twice: Better robust generalization from more efficient training. *arXiv preprint arXiv:2202.09844*, 2022c.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, pp. 248–255. IEEE, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018.

Zongyuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multiclass open set classification. In *British Machine Vision Conference 2017 (BMVC'17)*, London, UK, Sep. 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Mehadi Hassen and Philip K Chan. Learning a neural-network-based representation for open set recognition. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pp. 154–162. SIAM, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

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

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

Dominik Lewy and Jacek Mańdziuk. An overview of mixing augmentation methods and augmentation strategies. *Artificial Intelligence Review*, 56(3):2111–2169, 2023.

Atefeh Mahdavi and Marco Carvalho. A survey on open set recognition. In *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pp. 37–44. IEEE, 2021.

Dimity Miller, Niko Sunderhauf, Michael Milford, and Feras Dayoub. Class anchor clustering: A loss for distance-based open set recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3570–3578, 2021.

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

Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*, pp. 613–628, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y.Ng. Street view hause number dataset. `http://ufldl.stanford.edu/housenumbers`, 2011.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2307–2316, 2019.

Pramuditha Perera, Vlad I Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M Patel. Generative-discriminative feature representations for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11814–11823, 2020.

Sucheng Ren, Huiyu Wang, Zhengqi Gao, Shengfeng He, Alan Yuille, Yuyin Zhou, and Cihang Xie. A simple data mixing prior for improving self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14595–14604, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Yuxuan Shu and Vasileios Lampos. Unsupervised hard negative augmentation for contrastive learning. *arXiv preprint arXiv:2401.02594*, 2024.

Lukas Timpl, Rahim Entezari, Hanie Sedghi, Behnam Neyshabur, and Olga Saukh. Understanding the effect of sparsity on neural networks robustness. *arXiv preprint arXiv:2206.10915*, 2022.

AFM Uddin, Mst Monira, Wheemyung Shin, TaeChoong Chung, Sung-Ho Bae, et al. Saliencymix: A saliency guided data augmentation strategy for better regularization. *arXiv preprint arXiv:2006.01791*, 2020.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: a good closed-set classifier is all you need? In *International Conference on Learning Representations (ICLR'22)*, 2022.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pp. 6438–6447. PMLR, 2019.

Yu Wang, Junxian Mu, Pengfei Zhu, and Qinghua Hu. Exploring diverse representations for open set recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5731–5739, 2024.

Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. Openauc: Towards auc-oriented open-set recognition. *Advances in Neural Information Processing Systems*, 35:25033–25045, 2022.

Baile Xu, Furao Shen, and Jian Zhao. Contrastive open set recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10546–10556, 2023.

Yihao Xue, Siddharth Joshi, Eric Gan, Pin-Yu Chen, and Baharan Mirzasoleiman. Which features are learnt by contrastive learning? on the role of simplicity bias in class collapse and feature suppression. In *International Conference on Machine Learning*, pp. 38938–38970. PMLR, 2023.

Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4016–4025, 2019.

Yang Yu, Wei-Yang Qu, Nan Li, and Zimin Guo. Open category classification by adversarial sample generation. In *International Joint Conference on Artificial Intelligence (IJCAI'17)*, pp. 3357–3363, Melbourne, Australia, Aug. 2017.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2021.