GRADMIX: IMPROVING SUPERVISED–SELF-SUPERVISED COMPATIBILITY THROUGH ATTRIBUTION-GUIDED FEATURE COVERAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised learning (SSL) has been widely applied to supervised models, either during pre-training or as an auxiliary objective, to improve representation feature coverage beyond the information directly associated with class labels. This is particularly relevant to open-set recognition (OSR), which requires models to discriminate among known classes while retaining information beyond class-discriminative labels to distinguish unknown samples. However, when applied as an auxiliary objective for OSR, SSL can lower known-class classification accuracy, leading to a generalization and discrimination dilemma. To understand this problem, we analyze the representation-learning characteristics of supervised and self-supervised learning and attribute the degradation to conflicting representation-learning signals. We then propose GradMix, a data augmentation strategy that enables the SSL objective to learn more generalizable representations, letting SSL learn a broader range of representations compatible with supervised learning, thereby expanding the solution space shared by both objectives. Experimental results on standard and domain-specific OSR benchmarks show that GradMix consistently improves over naive joint supervised–self-supervised training and achieves competitive performance against state-of-the-art OSR methods. Beyond OSR, evaluations on out-of-distribution detection, robustness to common corruptions, and linear probing of SSL representations further demonstrate its ability to improve both representation generalization and discriminative precision.

1 INTRODUCTION

Deep neural networks have achieved remarkable performance across a wide range of supervised learning tasks, including object classification Russakovsky et al. (2015) and regression Zheng et al. (2024). However, supervised models remain susceptible to overfitting and poor generalization Zhang et al. (2017a). One contributing factor is that supervised objectives prioritize information predictive of the training labels, potentially encouraging representations to rely on a limited set of class-discriminative or even spurious features Nagarajan et al. (2021); Hermann et al. (2024); Li et al. (2025). Previous studies have therefore sought to broaden the range of information captured by learned representations Yoshihashi et al. (2019); Zhang et al. (2020); Wang et al. (2024). Among these approaches, self-supervised learning (SSL) has been incorporated as an auxiliary objective because it can preserve instance-level and fine-grained information that is not directly specified by class labels Hendrycks & Gimpel (2016); Xue et al. (2023); Lee et al. (2025).

However, these methods lack of considering keeping the discrimination capabilities of the models. Because SSL is label-agnostic, it may emphasize visual evidence that is complementary to, but not compatible with, the class-discriminative features prioritized by supervised learning. Jointly optimizing the two objectives through a shared encoder can consequently produce conflicting learning signals and negative transfer, reducing classification accuracy Hu et al. (2022). This problem is particularly consequential for open-set recognition (OSR), where a model must distinguish unknown classes while retaining accurate discrimination among known classes. Although the additional information preserved by SSL may benefit unknown detection, any accompanying degradation in known-class discrimination limits this benefit. Existing approaches may steer representation learning

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

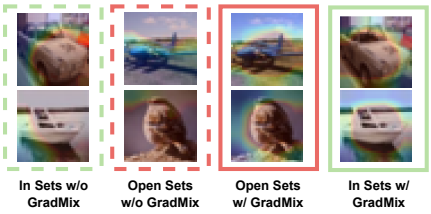

Table 1: An illustration of GradMix's effectiveness using attribution maps (green and red boxes for in-set (car, boat) and open-set (airplane, bird) samples respectively; dashed-line boxes indicate models without GradMix, while solid-line boxes represent models with GradMix). GradMix enables models to focus on broader areas in data for in-set samples while capturing object regions more effectively for open-set samples.

Reliance	Fore	Back	Full
Sup	54.2	14.4	67.6
Sup + SSL	53.4	15.2	66.3
Sup + SSL + GradMix	59.2	14.5	68.3

Table 2: Testing accuracy (in %) when the foreground (Fore) or background (Back) of the data is available. The models are trained on ImageNet100 with $\beta = 0.5$. First two rows: models trained with L_{SSL} can rely more on the background, i.e., $\mathbf{f}_{irre}$, and hence degrade the overall accuracy. The last row: the reliance on the foreground is greatly improved with *GradMix*. Details are in C in the Appendix.

toward task-relevant information using prior knowledge, such as manually specified foreground–background masks Jiang et al. (2023b). Such annotations, however, are rarely available in practical applications. To address this problem, we propose *GradMix*, an attribution-guided data augmentation strategy that regularizes the interaction between supervised and self-supervised objectives without requiring additional annotations or constraining them to learn identical representations. GradMix uses attribution maps to identify and partially suppress the image regions currently emphasized by SSL, encouraging it to explore a broader range of information in the input. Applied only to the SSL branch, this intervention allows SSL to retain complementary information while capturing additional features compatible with the supervised objective, thereby improving the balance between class discriminability and feature coverage.

We evaluate our method across a comprehensive suite of tasks, including OSR, Out-of-Distribution (OOD) detection, close-set classification, robustness under common corruptions, and linear probe on SSL representations. Our results outperform the state-of-the-art baselines and demonstrate that GradMix maintains both feature diversity and discriminative precision, which solves the discrimination-generalization trade-off problem. Crucially, we test our approach on a domain-specific fetal ultrasound brain (FUB) dataset Conti et al. (2025), where GradMix significantly outperforms pretrained Foundation Models (FMs). This highlights the applicability of our method in specialized domains where general-purpose FMs often fail.

2 Method

2.1 Problem Statement

Self-supervised learning (SSL) is often incorporated as an auxiliary objective in supervised training to preserve information beyond fulfilling supervised learning objectives Hendrycks & Dietterich (2018); Bhat et al. (2022); Lee et al. (2025), as in Equation 1. However, auxiliary SSL does not necessarily improve the supervised representation. As shown in Figure 1 and 2, increasing the contribution of the SSL objective improves unknown-class detection, as measured by AUROC, but progressively reduces closed-set classification accuracy, revealing a trade-off. This trade-off can arise because supervised and self-supervised objectives impose different relational structures on the shared representation. Supervised learning prioritizes features that distinguish the predefined classes. In contrast, SSL learns invariances from augmented views of individual samples and can preserve instance-specific and fine-grained information that is not directly required for separating the training classes. Such information may support unknown-class detection, but it is not necessarily compatible with the class-level structure favored by the supervised objective. Consequently, jointly optimizing the two objectives through a shared encoder does not automatically combine their respective benefits and can impair supervised discrimination.

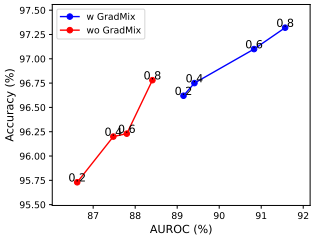

Figure 1: Pareto curves of close-set accuracy versus OSR AUROC on the CIFAR10 dataset. GradMix can improve both performances. The annotations next to the curves refer to the L_{Sup} weights, i.e., β in equation 2.

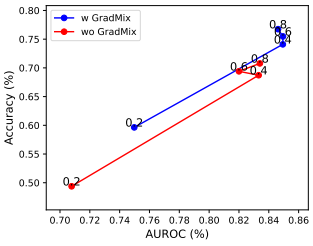

Figure 2: Pareto curves of close-set accuracy versus OSR AUROC on the Tiny-ImageNet dataset. GradMix can improve both performances. The annotations next to the curves refer to the L_{Sup} weights, i.e., β in equation 2.

Datasets	CIFAR10	ImgNet100
wo GradMix (H)	0.808	0.261
w GradMix (H)	0.848	0.517
wo GradMix (B)	0.365	0.181
w GradMix (B)	0.393	0.402

Table 3: CKA values, S_{Sup}, with the complete supervised representations. H and B stand for the head and backbone representations. Models with GradMix output representations with higher similarities to the complete supervised models.

$$L = \beta L_{Sup} + (1 - \beta) L_{SSL} \tag{1}$$

Increasing the weight of the supervised objective can reduce this interference, but it also weakens the auxiliary SSL signal and may diminish the information that benefits unknown-class detection. Similarly, forcing the two objectives to learn identical representations Yu et al. (2020); He et al. (2022) could suppress the complementary information that motivates the use of SSL. The problem is therefore not merely how to balance or align the two objectives, but how to improve their compatibility while preserving their distinct contributions. Therefore, we propose **GradMix** to address the following problem: *how can auxiliary SSL preserve information useful for supervised class discrimination while retaining information beneficial to generalization-oriented downstream tasks, such as unknown-class detection?* GradMix approaches this problem by regularizing the visual evidence explored by SSL, rather than directly aligning the two objectives or merely adjusting their relative loss weights.

2.2 GradMix

In the following text, we use $\mathbf{f}_{class}$, $\mathbf{f}_{sub}$, and $\mathbf{f}_{irre}$ to denote the class-wise, subclass-wise, and the rest irrelevant style features. Since L_{Sup} and L_{SSL} share the same multi-view input data, the $\mathbf{f}_{class}$ they capture are expected to be transferable between the two objectives in principle. However, $\mathbf{f}_{class}$ is often harder to learn than $\mathbf{f}_{sub}$ and $\mathbf{f}_{irre}$ for L_{SSL} due to the tendency of stochastic gradient descent to prioritize low-complexity signals Robinson et al. (2021); Xu et al. (2023). Then the task becomes how to make it easier for L_{SSL} to extract more $\mathbf{f}_{class}$ to align better with L_{Sup}. Since there exists no prior knowledge over the content of the data, it is infeasible to constrain the models to precisely keep the feature alignment during training with extra inductive biases. We hence take another way around that lets L_{SSL} encode more diverse $\mathbf{f}_{class}$ while keeping $\mathbf{f}_{sub}$ and $\mathbf{f}_{irre}$ during training.

Our design is inspired by erasing-based representation learning, where hiding highly activated or discriminative regions encourages a model to discover alternative visual evidence Kumar Singh & Jae Lee (2017). Specifically, we use attribution with respect to the SSL objective to identify its currently emphasized region and partially replace this region for an additional SSL view. Meanwhile, the original views are retained to avoid discarding the information already captured by SSL. We hypothesize that this asymmetric intervention encourages SSL to explore alternative evidence, some of which may encode more $\mathbf{f}_{class}$. Figure 3 provides a graphical illustration of our method. Before each forward and backward pass, the attribution maps, $\mathbf{M}_{SSL}$, of L_{SSL} are computed according to LayerCAM (see Section A in the Appendix) for one view of the inputs. The most activated areas in the data, according to $\mathbf{M}_{SSL}$, are then masked as the new inputs to the model. For masking, the ratio between the side lengths of the masked area and the original image, denoted as γ, follows a uniform distribution, i.e., $\gamma \sim \mathbf{U}(\gamma_{min}, \gamma_{max})$. In our work, we set γ_{min} to 0.1 and γ_{max} to 0.5, ensuring that the masks can neither completely cover the entire object in the original samples nor exceed the image margins. During the new forward and backward pass, compared with equation 1, L_{SSL} is the

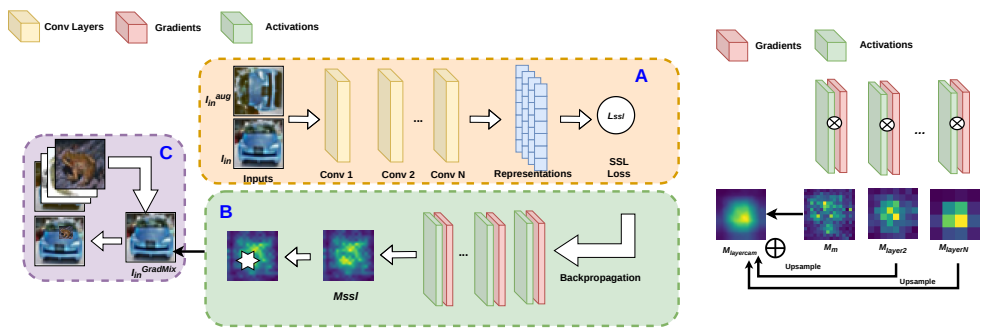

Figure 3: (**Left**) Graphical illustration of GradMix. Three blocks with dashed borderlines refer to the procedures: **A**. Data is fed into the feature extractor during forward propagation; **B**: Attribution maps are computed using the internal feature maps and the LayerCAM method. The most activated area is selected (highlighted using a white star in the graph); **C**: A random sample from the same minibatch is resized and patched on the most activated area. (**Right**) Illustration of the layer aggregation for GradMix: the attribution maps calculated using individual layers are summed to get $\mathbf{M}_{layercam}$.

weighted sum of the original L_{SSL}^{ori}, which is computed over the original two views of the inputs, and $L_{SSL}^{GradMix}$, which treats the GradMix masked data as one view, namely,

$$L = \beta L_{Sup} + (1-\beta)(L_{SSL}^{ori} + \gamma^2 \cdot L_{SSL}^{GradMix}) \tag{2}$$

Furthermore, as shown in Section 3.1, we found that aggregating attribution maps from multiple convolutional layers enhances OSR performance. Figure 3 illustrates this aggregation process. The final attribution map is the sum of these computed using individual convolutional layers, i.e., $\mathbf{M}_{layercam} = \mathbf{M}_m + \mathbf{M}_n + ... + \mathbf{M}_k$, where m, n, and k are selected indices of the layers.

To verify the principle of our design that L_{SSL} can learn richer transferable $\mathbf{f}_{class}$ with L_{Sup}, we compute the $\mathcal{C}_{rel}$ values (see Section B in the Appendix), which can reveal the transferability between L_{Sup} and L_{SSL} during training. The models with GradMix demonstrate smaller $\mathcal{C}_{rel}$ values, showing better alignment between these two objectives. Furthermore, to better demonstrate the improvement from the perspective of feature learning, we conducted a Centered Kernel Alignment (CKA) Kornblith et al. (2019) analysis. We use $S_{Sup} = CKA(\mathbf{f}, \mathbf{f}_{Sup})$ to denote the CKA similarities with the complete supervised representations. The results are in Table 3 and the representations from the GradMix models show higher similarities to the complete supervised ones. Nevertheless, the Pareto curves in Table 1 and 2 show an overall improvement for the close-set accuracy and unknown-class detection tradeoff. Models with GradMix can demonstrate larger attribution areas on the foreground in the data (see Figure 1) and higher foreground reliance (see Table 2). We believe the improved transferability between the two objectives can benefit the optimization of $\mathbf{f}_{Sup}$ itself Caruana (1997).

3 Experiments

In this section, we present the performance of GradMix for downstream tasks, including open-set recognition, close-set classification, out-of-distribution detection, corrupted image classification, and self-supervised learning model generalization. In our settings, we take supervised and self-supervised contrastive learning as L_{Sup} and L_{SSL} (see Section A in the Appendix).

3.1 Open Set Recognition

Settings Following the OSR testbench widely used in the literature Chen et al. (2022); Vaze et al. (2022); Neal et al. (2018), we evaluate our method on six open and close set 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). In each protocol, five different splits are applied, and all final results are the average of these five trials.

	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. (2022)	99.6	96.3	90.1	96.5	94.3	76.2
APRL-CS Chen et al. (2022)	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
OpenMix+ Jiang et al. (2023a)	–	–	86.9	93.1	92.5	75.1
CMI Jia et al. (2024)	–	–	94.0	98.1	96.8	85.3
Attentive CutMix Walawalkar et al. (2020)	78.75	79.91	74.32	84.51	89.65	64.52
GradMix (Ours)	**99.8**	94.7	91.33	**98.62**	**97.64**	81.92

Table 4: 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. (2022) and Xu et al. (2023)). " – " indicates there are no given results in the literature. Bold numbers indicate the best results. GradMix outperforms the SoTAs in four out of six protocols, and the increase is over 1% on the large-scale TinyImageNet protocol. OpenMix+, CMI, and Attentive CutMix are the methods that are closely related to our work; they are therefore grouped.

ResNet18 He et al. (2016) is used as the feature encoder backbone in all experiments of this section, with the output feature dimension set to 128. Furthermore, we evaluate the method on a domain-specific fetal ultrasound brain (FUB) dataset Conti et al. (2025) as well, for which the backbone is VGG16 Russakovsky et al. (2015). We use the area under the receiver operating characteristic curve (AUROC) to evaluate the OSR performance. AUROC is threshold-independent, and a higher AUROC represents better performance at detecting open-set samples. The complete experimental settings are detailed in Appendix E.

Results For the standard testbench, we compare our results with vanilla cross entropy classifier, as well as the state-of-the-art and popular OSR methods, 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. (2022), APRL-CS Chen et al. (2022), OpenAUC Wang et al. (2022), ConOSR Xu et al. (2023), MEDAF Wang et al. (2024), OpenMix+ Jiang et al. (2023a), and CMI Jia et al. (2024). Additionally, we test Attentive CutMix with its settings in the original paper Walawalkar et al. (2020). The results are presented in Table 4. Our method achieves advanced performance on most protocols, especially complex protocols, such as CIFAR+50 and TinyImageNet. 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). For the FUB datasets, the baselines are the same OSR framework but with the representations extracted using DINO Caron et al. (2021). The results are in Table 5. Although our method applied a much smaller model, it still greatly outperforms the DINO representations.

Ablations We perform two ablation studies: (1) comparing OSR performance when applying different mixing data augmentation techniques, and (2) assessing the impact of utilizing different deep layers and layer aggregation for computing attribution maps in GradMix. Consistent with previous experiments, OSR performance is evaluated using AUROC and each result is the average of five trials. All experiments are repeated with CIFAR10 and TinyImageNet protocols. We fix $k = 10$ (see Section D in Appendix) in each study for fair comparison.

A. IMPACT OF DIFFERENT AUGMENTATION METHODS We have (re)-implemented and tested three augmentation methods i.e., Mixup Zhang et al. (2017b), CutMix Yun et al. (2019), and GradMix,

as well as the models without extra augmentation. Furthermore, we vary the hyper-parameter α in Mixup and CutMix. Larger α leads to stronger augmentations (see Zhang et al. (2017b); Yun et al. (2019) for details).

The results, as shown in Figure 4 (left), clearly demonstrate that the models incorporating augmentations achieve superior OSR performance across both protocols. GradMix, in particular, improves AUROC by over 3% compared to without extra augmentation on both protocols. Furthermore, more advanced and stronger augmentation techniques can result in greater performance gains, with GradMix consistently outperforming all other methods.

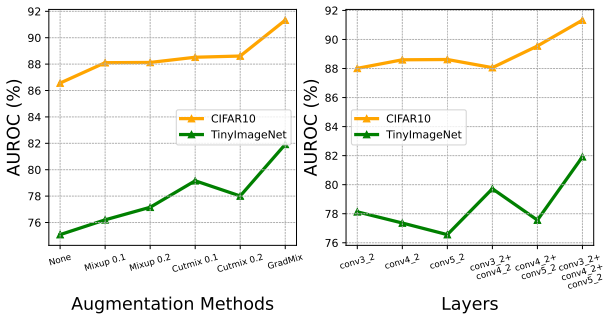

Figure 4: **Left:** OSR performances of the models with different augmentation methods on CIFAR10 and TinyImageNet protocols. Clear improvements can be brought by extra data augmentations. And GradMix performs best among all augmentation methods. **Right:** OSR performances of models with GradMix computed using different layers and layer aggregation in ResNet18. The results indicate that different layers can produce non-negligible changes in performance. And layer aggregation can in general provide better performance than single layers.

B. SELECTION OF THE DEEP LAYERS IN GRADMIX In order to investigate the impact of different deep layers and their aggregations for computing attribution maps in GradMix, we evaluate the OSR performances of the models with GradMix computed using various layers of ResNet18, specifically conv3_2, conv4_2, conv5_2, conv3_2+conv4_2, conv4_2+conv5_2, conv3_2+conv4_2+conv5_2. The latter three configurations represent layer aggregations.

The results are summarized in Figure 4 (right), which indicate that the utilization of different layers, as well as layer aggregation, introduces non-negligible performance changes, especially for higher resolution data, i.e., TinyImageNet. For both testing protocols, layer aggregation brings obvious improvements to OSR performance. For CIFAR10, GradMix with early layers imposes better performance whereas it is the inverse case for TinyImageNet. We think the reason lies in that the features are more diverse for more complex data and attribution maps from early layers can better preserve these features. Based on the above results, we adopt layer aggregation in our method.

3.2 CLOSE SET CLASSIFICATION

Settings 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. (2022); Xu et al. (2023); Wang et al. (2024). We compare the results with cross-entropy, ARPL, ConOSR, and MEDAF. Besides, we also compare GradMix with Mixup and CutMix under the same settings. Results are shown in Table 6.

Results All the baselines are pure supervised learning methods, which are in principle better for close-set classification. However, our method can still achieve significantly better performances on CIFAR100 and TinyImageNet and fair results for CIFAR10. As mentioned in 2.2, SSL with GradMix and supervised learning can promote with each other for learning $\mathbf{f}_{class}$. Besides, compared with Mixup and CutMix, GradMix can still show higher accuracy on almost all datasets, showing the higher generalization performance of SSL with GradMix.

Features	AUROC
DINO ViT-S/8	83.99
DINO ViT-S/16	70.96
DINO ViT-B/8	75.91
DINO ViT-B/16	78.52
Sup+SSL	91.39
Sup+SSL+GradMix	94.32

Table 5: OSR performances on the FUB dataset. The baselines are variant versions of DINO features. Our method can outperform the general-purpose FMs.

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
Mixup	92.52	76.16	71.22
CutMix	95.3	77.76	70.9
GradMix (Ours)	94.1	**78**	**72.46**

Table 6: 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. (2022); Xu et al. (2023); Wang et al. (2024) and Mixup, as well as CutMix trained under the same Sup+SSL settings. Our method can outperform the baselines on the two larger datasets, CIFAR100 and TinyImageNet, with significant increases.

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)	**50.3**	91.63	**92.15**	92.5	90.86	**80.5**	96.7	**96.82**	96.9	95.3

Table 7: GradMix performs comparably or better than previous works in out of distribution detection. Results of baselines are from Chen et al. (2022); Wang et al. (2024).

3.3 Out-of-Distribution Detection

Settings We validate our method on out-of-distribution detection (OOD) tasks. Following the settings for OOD in Chen et al. (2022); Wang et al. (2024), we take CIFAR10 as in distribution set and CIFAR100 and SVHN as out of distribution sets. vanilla cross entropy, ARPL, ARPL-CS, and MEDAF are baselines. Evaluation metrics are TNR, AUROC, DTACC, and AUIN/AUOUT. Details on complete settings are in Appendix G.

Results The results are given in Table 7. GradMix can outperform the baselines, especially on SVHN, indicating that the model has learned many data-dependent features that overlap with the close OOD dataset CIFAR100.

3.4 Generalization to Common Corruptions

Settings We evaluate the robustness of the classifiers trained using our method on the data with common corruptions proposed in Hendrycks & Dietterich (2018), in which 15 types of corruptions are synthesized, namely Gaussian noise, shot noise, impulse noise, defocus blur, frosted glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic, piexlate, and JPEG compression. Each of the corruption has five levels of severity, and therefore the classifiers are tested with 75 repeats for all corruptions. We use accuracy drop, $D_{c,s} = A_{clean} - A_{c,s}$, as a metric to quantify the model's robustness to the corruption type c of severity s, and A_{clean} denotes the testing accuracy on data without corruptions. We record $D_{c,s}$ on the models trained with and without GradMix on the full CIFAR10 and TinyImageNet datasets. Lower $D_{c,s}$ indicates higher robustness to the corruption (see Appendix H for more details).

Results The results are given in Figure 5 and Table 8. In Figure 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 compare each $\bar{D}_c$ with and without GradMix. For most corruption types, $\bar{D}_c$s are smaller in the models with GradMix. In Table 8, we

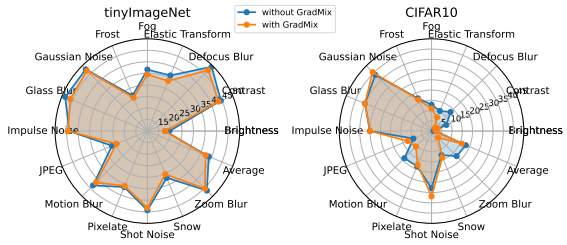

Figure 5: Classification accuracy drop of the models trained with and without GradMix on CIFAR10 and TinyImgNet. The average accuracy drop $\bar{D}_c$s are lower for the models with GradMix for most corruption types on both datasets.

	CIFAR10		TinyImageNet	
Severity	w/o GradMix	GradMix	w/o GradMix	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 8: Avg. accuracy drop $\bar{D}_s$ for different corruption severities of the models without (left) and with GradMix (right) on CIFAR10 and TinyImageNet datasets. Better results are in bold. For almost all severity levels, GradMix model shows lower $\bar{D}_s$ and is therefore more robust.

average $A_{c,s}$ over the all corruption types for each severity level, i.e., $\bar{D}_s = \frac{1}{15}\sum_c D_{c,s}$. For almost all severity levels, $\bar{D}_s$ in models with GradMix are lower, indicating that GradMix classifiers are more robust to common corruptions.

3.5 Model Generalization across Datasets

Settings To validate the effectiveness of GradMix for model generalization, we apply it to self-supervised learning and test the downstream linear classification performances (linear probe). In order to increase the experimental diversity, the models are trained with SimCLR and MoCo v1 He et al. (2020) as well as multiple network architectures of ResNet18, ResNet34, and ResNet50. Furthermore, in order to evaluate the effectiveness of GradMix for high-resolution data, we conduct the linear probe on the models trained on ImageNet100 Rebuffi et al. (2017). We record top-1 and top-5 accuracy. Detailed settings are given in Appendix J.

Results The results are shown in Figure 6 and Table 9 for TinyImageNet and ImageNet100 datasets respectively. Both top-1 and top-5 accuracy increase with the application of extra augmentation methods, and the improvements brought by GradMix are consistently higher. Qualitative visualizations of GradMix attribution maps from LayerCam on ImageNet100 are in K.

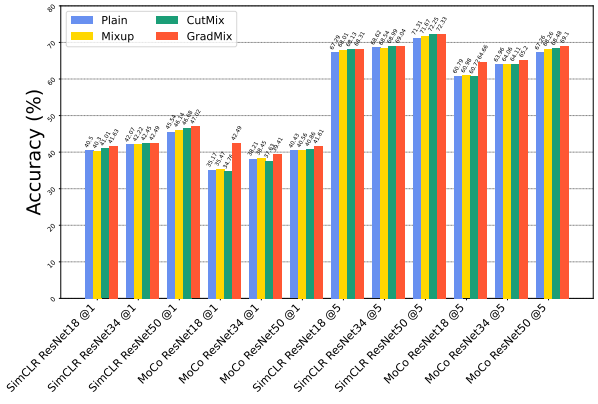

Figure 6: Top-1 and top-5 linear classification accuracy (in %) on TinyImageNet with features learned using self-supervised learning with different augmentation methods. GradMix indicates benefits for model generalization.

	SimCLR		MoCo v1	
	Acc@1	Acc@5	Acc@1	Acc@5
-	48.1	75.78	40.31	66.54
+ Mixup	49.32	77.43	42.53	67.1
+ CutMix	50.61	78.67	44.87	69.21
+ GradMix	**54.8**	**81.36**	**47.75**	**72.37**

Table 9: Results from the linear probe on SSL models trained on ImageNet100, showing that additional mixing data augmentations can lead to performance gains, with GradMix consistently achieving the best results.

4 Related Work

Open Set Recognition OSR methods 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. (2022); Vaze et al. (2022); Xu et al. (2023); Wang et al. (2024) can either focus on designing learning objectives that enhance feature learning for known classes, such as Miller et al. (2021), or leverage synthesized open set samples generated by generative models like GANs, such as Yu et al. (2017). Another category utilizes generative models to model the close-sets Ge et al. (2017); Cao et al. (2021). For a more thorough review of open set recognition methods, see Mahdavi & Carvalho (2021).

Mixing Data Augmentation Mixing data augmentation combines raw samples with other images from within or outside the dataset. These methods can be broadly categorized into two types: pixel-wise mixing, in which augmented data is created by taking a weighted average of the original samples and their corresponding labels, such as *Mixup* Zhang et al. (2017b), and patch-wise mixing, where a portion of the original samples is replaced, such as *CutOut* DeVries & Taylor (2017) and *CutMix* Yun et al. (2019). Besides, there are numerous other mixing data augmentation approaches. Lewy & Mańdziuk (2023) presents a comprehensive review.

5 Conclusion & Outlook

This work investigated the tension that arises when self-supervised learning is introduced as an auxiliary objective for supervised representation learning. Although SSL preserves information beneficial to generalization, increasing its contribution can impair the class discrimination required by the supervised task. We propose *GradMix*, an attribution-guided augmentation that partially replaces the visual evidence currently emphasized by SSL while retaining the original views and leaving the supervised branch unchanged. Rather than directly aligning gradients or forcing the two objectives to learn identical representations, GradMix regulates the feature-acquisition process of SSL. Across the evaluated settings, GradMix improves the discrimination–generalization trade-off, with benefits for closed-set classification, open-set recognition, out-of-distribution detection, corruption robustness, and SSL linear probing. The directional-curvature, representation-similarity, and Pareto curves suggest that these improvements arise from a more compatible interaction between the supervised and self-supervised objectives.

Several extensions are worth exploring. GradMix is studied here using supervised and self-supervised contrastive objectives, but its underlying principle may also apply to non-contrastive SSL methods and other combinations of primary and auxiliary learning objectives. Extending the method to vision transformers and pretrained foundation models could further reveal how attribution-guided feature exploration behaves across architectures and pretraining paradigms. Another promising direction is to characterize how the alternative evidence encouraged by GradMix contributes to different downstream tasks, including fine-grained recognition, domain adaptation, and continual learning. Such analysis could also inform adaptive strategies that select the attribution layers, replacement regions, and augmentation strengths according to the data or training stage.

6 Use of LLMs

We applied LLMs for polishing the language, implementing methods and creating visualizations, looking for literature, and aiding internal peer review.

REFERENCES

Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, pp. 1563–1572, 2016.

Prashant Shivaram Bhat, Bahram Zonooz, and Elahe Arani. Task agnostic representation consolidation: a self-supervised based continual learning approach. In *Conference on Lifelong Learning Agents*, pp. 390–405. PMLR, 2022.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE TPAMI*, 2022.

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.

Edoardo Conti, Riccardo Rosati, Lorenzo Federici, Adriano Mancini, and Maria Chiara Fiorentin. Challenging dinov3 foundation model under low inter-class variability: A case study on fetal brain ultrasound. *arXiv preprint arXiv:2511.01915*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. IEEE, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Sign. Process. Letters*, 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.

Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. 2022.

Zongyuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. In *BMVC*, London, UK, Sep. 2017.

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 *CVPR*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pp. 9729–9738, 2020.

Yun He, Xue Feng, Cheng Cheng, Geng Ji, Yunsong Guo, and James Caverlee. Metabalance: improving multi-task recommendations via adapting gradient magnitudes of auxiliary tasks. In *Proceedings of the ACM web conference 2022*, pp. 2205–2215, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2018.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Katherine Hermann, Hossein Mobahi, Thomas Fel, and Michael Mozer. On the foundations of shortcut learning. In *International Conference on Learning Representations*, volume 2024, pp. 43832–43868, 2024.

Ziniu Hu, Zhe Zhao, Xinyang Yi, Tiansheng Yao, Lichan Hong, Yizhou Sun, and Ed Chi. Improving multi-task generalization via regularizing spurious correlation. *Advances in Neural Information Processing Systems*, 35:11450–11466, 2022.

Yunbing Jia, Xiaoyu Kong, Fan Tang, Yixing Gao, Weiming Dong, and Yi Yang. Revealing the two sides of data augmentation: An asymmetric distillation-based win-win solution for open-set recognition. *arXiv preprint arXiv:2404.19527*, 2024.

Guosong Jiang, Pengfei Zhu, Yu Wang, and Qinghua Hu. Openmix+: Revisiting data augmentation for open set recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(11): 6777–6787, 2023a.

Guosong Jiang, Pengfei Zhu, Yu Wang, and Qinghua Hu. Openmix+: Revisiting data augmentation for open set recognition. *IEEE transactions on circuits and systems for video technology*, 33(11): 6777–6787, 2023b.

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

Dayal Singh Kalra, Jean-Christophe Gagnon-Audet, Andrey Gromov, Ishita Mediratta, Kelvin Niu, Alexander H Miller, and Michael Shvartsman. A scalable measure of loss landscape curvature for analyzing the training dynamics of llms. *arXiv preprint arXiv:2601.16979*, 2026.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

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.

Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Chungpa Lee, Jeongheon Oh, Kibok Lee, and Jy-yong Sohn. A theoretical framework for preventing class collapse in supervised contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 487–495. PMLR, 2025.

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

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim GJ Rudner, C Bayan Bruss, and Andrew Gordon Wilson. Out-of-distribution detection methods answer the wrong questions. *arXiv preprint arXiv:2507.01831*, 2025.

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.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *ECCV*, 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.

Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *CVPR*, 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 *CVPR*, pp. 11814–11823, 2020.

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. *Advances in neural information processing systems*, 34:18420–18432, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, pp. 2001–2010, 2017.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

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. *IJCV*, 115:211–252, 2015.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE TPAMI*, 35(7):1757–1772, 2012.

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 *ICCV*, pp. 618–626, 2017.

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

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: a good closed-set classifier is all you need? In *ICLR*, 2022.

Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. *arXiv preprint arXiv:2003.13048*, 2020.

Yu Wang, Junxian Mu, Pengfei Zhu, and Qinghua Hu. Exploring diverse representations for open set recognition. In *AAAI*, 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. *NeurIPS*, 35:25033–25045, 2022.

Baile Xu, Furao Shen, and Jian Zhao. Contrastive open set recognition. In *AAAI*, 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 *CVPR*, pp. 4016–4025, 2019.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in neural information processing systems*, pp. 5824–5836, 2020.

Yang Yu, Wei-Yang Qu, Nan Li, and Zimin Guo. Open category classification by adversarial sample generation. In *IJCAI*, 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 *CVPR*, pp. 6023–6032, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017a.

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

Song Zhang, Gehui Shen, Jinsong Huang, and Zhi-Hong Deng. Self-supervised learning aided class-incremental lifelong learning. *arXiv preprint arXiv:2006.05882*, 2020.

Jiyang Zheng, Yu Yao, Bo Han, Dadong Wang, and Tongliang Liu. Enhancing contrastive learning for ordinal regression via ordinal content preserved data augmentation. In *International Conference on Learning Representations*, volume 2024, pp. 49585–49601, 2024.

Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *CVPR*, pp. 4401–4410, 2021.