

# BackMix: Regularizing Open Set Recognition by Removing Underlying Fore-Background Priors

Yu Wang, *Member, IEEE*, Junxian Mu, Hongzhi Huang, Qilong Wang, *Member, IEEE*,  
Pengfei Zhu, Qinghua Hu, *Senior Member, IEEE*

**Abstract**—Open set recognition (OSR) requires models to classify known samples while detecting unknown samples for real-world applications. Existing studies show impressive progress using unknown samples from auxiliary datasets to regularize OSR models, but they have proved to be sensitive to selecting such known outliers. In this paper, we discuss the aforementioned problem from a new perspective: Can we regularize OSR models without elaborately selecting auxiliary known outliers? We first empirically and theoretically explore the role of foregrounds and backgrounds in open set recognition and disclose that: 1) backgrounds that correlate with foregrounds would mislead the model and cause failures when encounters ‘partially’ known images; 2) Backgrounds unrelated to foregrounds can serve as auxiliary known outliers and provide regularization via global average pooling. Based on the above insights, we propose a new method, Background Mix (BackMix), that mixes the foreground of an image with different backgrounds to remove the underlying fore-background priors. Specifically, BackMix first estimates the foreground with class activation maps (CAMs), then randomly replaces image patches with backgrounds from other images to obtain mixed images for training. With backgrounds de-correlated from foregrounds, the open set recognition performance is significantly improved. The proposed method is quite simple to implement, requires no extra operation for inferences, and can be seamlessly integrated into almost all of the existing frameworks. The code is released on <https://github.com/Vanixxz/BackMix>.

**Index Terms**—Classification, open set recognition, unknown detection, fore-background priors, spurious correlation.

## I. INTRODUCTION

CONVENTIONAL artificial intelligence models primarily tackle visual tasks within closed-set situations, where classes remain consistent throughout both training and testing phases [1]. However, in real-world applications, unknown classes may arise during test, challenging the reliability of the closed-set assumption and leading to an open set scenario [2]–[5]. Open set recognition (OSR) is a task that aims to handle such challenging scenarios, in which the models are required

This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0116500, in part by the National Natural Science Foundation of China under Grants 62476195, U23B2049, 62436002, 62222608, and 62266035, in part by Tianjin Natural Science Funds for Distinguished Young Scholar under Grant 23JCJQJC00270, in part by Tianjin Young Scientific and Technological Talents Project under grant QN20230305, and in part by Tianjin Science and Technology Plan Project under grants 24YDTPJC00150 and 24JCYBJC00950.

Yu Wang, Junxian Mu, Hongzhi Huang, Qilong Wang, Pengfei Zhu, and Qinghua Hu are with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China, and also with the Haihe Laboratory of Information Technology Application Innovation (Haihe Lab of ITAI), Tianjin, China.

Corresponding authors: Pengfei Zhu (zhupengfei@tju.edu.cn) and Qinghua Hu (huqinghua@tju.edu.cn)

to recognize unknown images while accurately classifying known classes.

Existing OSR methods can be mainly categorized into three groups: discriminative methods that design open set oriented classification strategies [6]–[11], generative methods that generate input distribution or pseudo-unknown samples to explicitly reserve a certain space for unknown classes [12]–[18], and auxiliary data-based methods that utilize additional samples from manually selected datasets as available outliers [19]–[22]. The significant difference between the first two methods and auxiliary data-based methods lies in whether using available outliers to regularize the model’s performance on out-of-distribution data. Many studies show that proper unknown samples can enable models to recognize unknowns better and adapt to open environments with less complexity [23], [24].

Although auxiliary data-based methods have impressive performance in unknown detection, we find that they are quite sensitive to the selection of out-of-distribution data on different tasks (See Section III-B2). Such a problem raises a straightforward question: **Can we regularize open set recognition models without elaborately selecting auxiliary known outliers?**

In pursuit of this goal, we look into the mechanisms of image recognition. In natural images, foregrounds typically encompass the distinctive regions correlated with categories, while cognitive studies suggest that humans utilize backgrounds as contextual cues during object recognition [25]–[27]. For example, cars are typically found on roads but rarely in water, while fish predominantly inhabit the water and are seldom observed on roads. Likewise, image classifiers have been shown to effectively utilize and derive benefit from such underlying priors in object identification [28]–[31]. Even more, they can achieve notable performance when foregrounds are masked out [31].

To rely on these priors, one has to assume that the classes learned during training will only appear in matching backgrounds during test. However, in OSR, partial distributions of images may change due to the appearance of unknown classes, in which there are two typical cases: **1) varying foreground**, *i.e.*, unknown classes arise in known backgrounds. In this case, the model is likely to predict an unknown sample to a known class according to its known background; **2) varying background**, *i.e.*, known classes arise in rare backgrounds. In this case, the model may mistakenly recognize the foreground due to the unseen information brought by the background [32]. When encountering the above two cases of varying foregrounds or backgrounds, the model may fail to

identify test samples due to the ‘partially’ known information. Unfortunately, the impact of fore-background priors has been neglected in existing OSR methods, which poses significant challenges to them in tackling the aforementioned cases.

**In this paper, we delve deeply into the role of fore-background priors in OSR and properly use such information to serve as an effective regularization of unknown classes without auxiliary data.** Firstly, we vary the background distribution during both the training and test phases and empirically find that: 1) The model trained on raw images fails to generalize when known objects appear in unseen backgrounds due to the disruption of the learned fore-background correlations; 2) The model becomes more robust to the presence of unknown backgrounds when the foreground class is de-correlated from the seen backgrounds during training. Subsequently, we provide a theoretical analysis to explain such findings. We show that backgrounds can serve as known outliers and provide extra regularization via global average pooling (GAP), which is equivalent to using manually selected outliers more flexibly and robustly.

Based on the above insights, we propose a new method, Background Mix (BackMix), which mixes the foreground of an image with different backgrounds to remove the underlying priors. To avoid precise segmentation annotations and additional costs, we use class activation maps (CAMs) [33] to estimate the foreground region roughly. Then, two random images serve as the target image (TI) and the background image (BI) for mixing, respectively. The processed input is obtained by replacing random patches of TI with background patches of BI. Extensive experiments demonstrate that BackMix enhances both closed-set and open set performance under various evaluation metrics when applied to existing OSR methods or compared to data augmentation techniques. The proposed method is quite simple to implement, does not require additional operation to make inferences, and can be seamlessly integrated into almost all existing OSR frameworks.

In general, our work has the following contributions:

- 1) We thoroughly discuss the role of fore-background priors and demonstrate that the fore-background priors can mislead the model in OSR. This issue can be resolved by releasing the correlation between foreground and background during training.
- 2) We provide insights into the regularization effect of class-unrelated backgrounds, which can enhance open set performance by serving as outliers. Moreover, the internal regularization mechanism is as effective as well-designed auxiliary data-based methods.
- 3) We propose BackMix that involves rough foreground estimation using CAMs and mixing up backgrounds from different images to release the inherent correlation.
- 4) BackMix is simple to implement and can be seamlessly integrated into other methods. Experimental results show that BackMix significantly improves conventional and state-of-the-art OSR methods by up to 23.6% on the AUROC, even enhancing the plain baseline to outperform advanced methods.

The remainder of this paper is organized as follows. Section II briefly reviews studies related to this work. Section III deeply explores and analyzes the role of backgrounds in OSR. Based on the analysis, we propose and elaborate on a new method in Section IV. Experimental implementation, metrics and results are provided in Section V. Finally, conclusions and future work are drawn in Section VI.

## II. RELATED WORK

In this section, we review the literature that relates to our work, mainly including open set recognition methods, data augmentation methods, and researches on spurious correlations.

### A. Open Set Recognition

Open set recognition aims to detect unknown classes while maintaining accuracy in classifying known classes. In this regard, a similar task out-of-distribution (OOD) detection also attempts to address such a problem. We review related methods and divide them into the following three groups.

**Discriminative methods.** Bendale and Boulton [6] addressed the limitations of SoftMax in OSR by introducing OpenMax, which calibrates classification scores using extreme value theory. Liang *et al.* [7] proposed to combine temperature scaling and input preprocessing for improving detection performance without retraining the model. Perera *et al.* [8] developed GDFR, enhancing feature quality with a self-supervised auxiliary task. Liu *et al.* [9] demonstrated that energy scores are more effective than SoftMax scores in distinguishing unknown samples and can be flexibly used as a score function. Zhou *et al.* [10] employed class and data placeholders to reserve space for unknown classes and adjust overconfident predictions. Xu *et al.* [11] used supervised contrastive learning to boost the model’s ability to extract robust representations. These methods enhance discriminative power by reinforcing feature learning or implementing tailored classification strategies for open scenarios. However, without specific adaptations for open space, their performance remains limited.

**Generative methods.** To constrain the boundary between known and unknown, Neal *et al.* [12] generates images that are close to known classes in latent space. Lee *et al.* [13] used a generative classifier and adopted the score function based on the Mahalanobis distance. Oza and Patel [14] utilized an auto-encoder (AE) as the classifier, identifying unknown samples via reconstruction error. Chen *et al.* [15] proposed RPL, a distance-based method using reciprocal points, later refined to ARPL [16] with adversarial constraints to limit known class space. Yang *et al.* [17] embedded prototypes of known classes in feature space and replaced SoftMax with a prototype model to exclude unknowns. To represent known classes without devouring, Huang *et al.* [18] developed plugged class-specific AEs at the top of the backbone to generate manifolds for known classes. These methods use generated samples or distributions to model classes and enlarge the discrepancy between known and unknown samples, while the generative modules introduce computational cost and instable performance.

**Auxiliary data-based methods.** Hendrycks *et al.* [19] introduced outlier exposure (OE), training models to assign uniform probabilities to outliers from auxiliary datasets. Dhamija *et al.* [20] reduced the intensity of global features in outlier images. Perera and Patel [21] developed global negative filters using a large dataset to decrease activation for unknown samples. Recognizing the limitations of available auxiliary data and potential overfitting, Kong and Ramanan [22] proposed training with both real outliers and generated samples. Cen *et al.* [23] examined the effectiveness of OE, suggesting the inclusion of unknown samples in training for few-shot unified OSR tasks. While auxiliary datasets aid in modeling open space with reduced complexity, they introduce biases tied to outlier distributions, making model performance highly sensitive to the choice of auxiliary data.

Despite significant progress in OSR, current methods primarily focus on modeling known classes or unknown space, overlooking the impact of image backgrounds. We argue that joint modeling of foregrounds and backgrounds may hinder performance. This insight motivates our new approach and offers valuable perspectives for future OSR studies.

### B. Data Augmentation

To mitigate overfitting and enhance the generalization of the model, data augmentation techniques have been extensively employed in various tasks. Current approaches can be broadly categorized into masking-based and mixing-based methods.

**Masking-based methods.** DeVries and Taylor [34] introduced Cutout to remove random image regions for occlusion-invariant training, while Singh and Lee [35] used random patch hiding to encourage learning from whole objects. To avoid excessive masking, Chen *et al.* [36] proposed GridMask, which applies grid-pattern masking. These techniques mask image sections while preserving primary objects, reducing the original correlations in training images.

**Mixing-based methods.** Zhang *et al.* [37] introduced Mixup, combining inputs and labels as linear mixtures of two images. Yun *et al.* [38] developed Cutmix, which swaps random regions within a batch to create new inputs. Zhou *et al.* [39] mixed images across source domains to generate new styles, enhancing training diversity. We notice these operations can help mix image backgrounds, and this should lead to a positive effect on open set recognition. However, experiments in Section V-B show that the improvement in closed-set performance is accompanied by degradation in open set performance—likely due to unintended label mixing during processing, which interferes with predictions in OSR settings.

Data augmentation techniques are generally designed with a closed-set assumption, and their effectiveness in OSR remains largely unexplored. Furthermore, many are heuristic, offering limited reliability across diverse and challenging open scenarios. In this work, we provide both empirical and theoretical insights to establish a feasible and robust approach for OSR.

### C. Spurious Correlations

Spurious correlations occur when models depend on irrelevant or secondary features instead of class-related ones,



Fig. 1. Illustration of operations applied on the generated dataset. Raw, FG, and BG represent the original image, the foreground of the image, and the background of the image, respectively. The star (\*) on Raw and BG denotes that another source image is randomly sampled from the dataset.

risking inaccurate predictions if not properly addressed [40], we categorize existing methods into two classes:

**Representations enhancement methods.** Srivastava *et al.* [41] used human annotations to capture unmeasured confounders and mitigate distribution shifts. Creager *et al.* [42] enabled the model to learn invariant features by incorporating environment inference tasks. Yao *et al.* [43] proposed to improve out-of-distribution robustness by augmenting the data with a mixup-based method to learn invariant predictors. These methods aim to enhance the model’s ability to capture essential representations by optimizing data or features.

**Debiasing optimization methods.** Du *et al.* [44] proposed to down weight examples with high feature-label bias, reducing the model’s reliance on such shortcuts. Liu *et al.* [45] reduced spurious correlations by correcting logits to balance group accuracy and minimize bias. Asgari *et al.* [46] masked learned dominant features, encouraging the model to explore and rely on alternative, unbiased features. These methods use specific debiasing optimization objectives to eliminate spurious correlations, making the model more robust.

Recently, Ming *et al.* [47] analyzed the impact of spurious correlations in OOD detection tasks, the proposed BackMix takes a simple and effective way to address this issue in OSR tasks directly. By mixing diverse backgrounds with foregrounds, BackMix effectively mitigates misleading fore-background priors, improving both closed-set classification accuracy and unknown detection performance without requiring auxiliary data or model component.

## III. EXPLORING FORE-BACKGROUND PRIORS IN OPEN SET RECOGNITION

In this section, we first explore and discuss the effect of fore-background priors in open set recognition. Based on the analysis, we then provide insights into how class-unrelated backgrounds regularize open set classifiers.

### A. Fore-Background Priors in Training Set Misleads Open Set Classifiers

In OSR settings, the model may be faced with partially known test samples. According to the maximum entropy principle [48], it is clear that models are supposed to not rely on background information. Training with raw images that have class-related backgrounds, *i.e.*, some objects appearing only in certain backgrounds, inadvertently introduces prior correlation

TABLE I

PERFORMANCE UNDER VARIOUS TRAINING AND TEST SETTINGS. THE GAIN OR LOSS VALUES ARE CALCULATED COMPARED TO THE SETTING I WHICH USES RAW IMAGES FOR TRAINING. THE IMAGES USED DURING THE TEST PHASE ARE REPRESENTED AS ‘KNOWN TEST DATA / UNKNOWN TEST DATA’.

Setting	Training Data	Raw / Raw		FG+BG* / Raw		FG only / FG only		Raw / Raw (iNaturalist)	
		Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC
I	Raw	80.1	67.2	69.3	59.0	72.5	64.7	80.1	74.9
II	FG only	17.9 <sub>62.2</sub> ↓	52.7 <sub>14.5</sub> ↓	30.2 <sub>39.1</sub> ↓	53.0 <sub>6.0</sub> ↓	92.4 <sub>19.8</sub> ↑	82.0 <sub>17.3</sub> ↑	17.9 <sub>62.2</sub> ↓	50.9 <sub>24.0</sub> ↓
III	FG+Raw*	73.6 <sub>6.5</sub> ↓	68.3 <sub>1.1</sub> ↑	85.8 <sub>16.5</sub> ↑	83.3 <sub>24.3</sub> ↑	89.3 <sub>16.7</sub> ↑	75.4 <sub>10.7</sub> ↑	73.6 <sub>6.5</sub> ↓	85.4 <sub>10.5</sub> ↑
IV	FG+BG*	76.4 <sub>3.7</sub> ↓	69.5 <sub>2.3</sub> ↑	87.3 <sub>18.0</sub> ↑	83.6 <sub>24.6</sub> ↑	91.5 <sub>18.9</sub> ↑	81.0 <sub>16.3</sub> ↑	76.4 <sub>3.7</sub> ↓	85.6 <sub>10.7</sub> ↑

to the model. These priors may be helpful under the closed-set assumption, where the test sample distribution matches the training ones. However, foreground-background correlations become a hindrance in OSR, potentially misleading classifiers.

To verify the above ideas, we first conducted an experiment to observe the performance trend with varied backgrounds during training and the test. Following Deng *et al.* [49], we used samples from the COCO dataset [50], which contains pixel-level foreground annotations. We chose 12 classes as the known and 6 classes as the unknown during open set evaluation. With segmentation labels, we generated three variants as follows (See Fig. 1). *FG only*: the original background is erased for the image; *FG+Raw\**: the original background is erased and replaced with another random raw image in the dataset; *FG+BG\**: the original background is erased and replaced with a random background of another image.

To observe the impact of fore-background priors, we trained models on the original dataset and the above three variants, respectively. During test, four different settings were considered to evaluate the performance of each model comprehensively:

- 1) Using raw images for known data and unknown data;
- 2) Using raw images for unknown data while replacing backgrounds of known data with an unknown image randomly to break the foreground-background correlations;
- 3) Using images that have been removed backgrounds for both known and unknown data;
- 4) Using raw images for known data while unknown data are from an out-of-distribution dataset iNaturalist [51], which has little semantic overlap with known classes.

We used the classification accuracy and the Area Under Receiver Operating Characteristic curve (AUROC) to evaluate closed-set and open set performance, respectively. The maximum SoftMax probability [52] served as the score function to reject unknown samples. Results in Table I show that:

**Training priors have a negative impact once the correlation breaks down.** As the closed-set performance of learning Raw (Setting I) drops 11% and 8% when using the FG+BG\* and FG only images as test known samples, the model becomes uncertain about classification without backgrounds and gets worse when given unrelated backgrounds.

**Simply removing backgrounds is not a practical solution.** As learning FG only (Setting II) never considers the existence of backgrounds, its performance seriously degrades when faced with images that have a background. Results suggest that replacing the background of images with pure grey may not be a robust solution for releasing correlations in practice.

**Releasing fore-background correlations enhances open set performance.** Learning *FG+Raw\** (Setting III) and *FG+BG\** (Setting IV) consistently enhance open set performance, which indicates the correlations between foregrounds and backgrounds are hindrances to OSR. Prior experiments showed that using constant backgrounds (pure grey) prevents misleading classifiers but compromises robustness in practice. Therefore, using class-unrelated backgrounds is a feasible way to release the correlations.

**Avoiding multiple objects appearing in the foreground improves closed-set performance.** Comparing learning *FG+Raw\** and *FG+BG\**, *FG+BG\** shows better accuracy as it avoids having multiple objects appear in a single image. Therefore, if segmentation is challenging, using raw images to refill the backgrounds exchanges open set performance improvement with a slight closed-set performance drop.

**Connections to existing opinions.** Previous study on image background shows that models with better classification performance often rely less on backgrounds [31]. Meanwhile, Vaze *et al.* [53] suggested that a good closed-set classifier inherently enhances open set performance. Our main finding bridges both conclusions, suggesting that good classifiers emphasize foregrounds and are more robust against unknowns.

### B. Exploiting Class-Unrelated Backgrounds for Open Set Classifier Regularization

In this section, we first outline two desirable properties that a reliable classifier should have for OSR. Then, we verify that regularizing classifiers with class-unrelated backgrounds shares similarities with OE but with fewer limitations.

1) *The main theory*: Based on the analysis of foreground-background correlations in OSR, we suppose that models should focus on the foreground objects rather than being misled by backgrounds. To accurately and robustly identify unknowns, with  $\mathbf{z}_f$  and  $\mathbf{z}_b$  representing the features of foreground and background, a reliable OSR model  $\mathcal{W}$  is expected to possess the following two desired properties:

**Property 1.** *The information of foreground and background in any image  $\mathbf{x}$  are independent of each other to model  $\mathcal{W}$ . That is, random variables  $\mathbf{z}_f$  and  $\mathbf{z}_b$  are independent.*

**Property 2.** *For a certain model  $\mathcal{W}$ , image backgrounds are independent of its prediction on image categories. That is, background feature  $\mathbf{z}_b$  and image category  $y$  are independent random variables.*

Possessing Property 1 is mostly beneficial for classification, except for certain special cases, *e.g.*, a person may wear less on the beach. As we have discussed in the previous section, it is difficult for most models to possess the Property 2, but this is quite significant in the open scenarios.

In modern DNNs, global average pooling (GAP) is widely used [54], which is designed to replace intensive fully-connected layers, thereby minimizing overfitting. GAP processes the feature map  $\mathcal{Z} \in \mathbb{R}^{H \times W \times C}$  and reduces its spatial dimensions by averaging values to generate a global representation  $\mathbf{z}_g = \text{GAP}(\mathcal{Z}) = \frac{1}{|\mathcal{Z}|} \sum_{(i,j) \in \mathcal{Z}} \mathbf{z}_{ij}$ . We subsequently explain how DNNs depress background features and regularize networks robust to outliers. Firstly, the global representation  $\mathbf{z}_g$  can be decomposed into a linear combination of foreground global features  $\mathbf{z}_f$  and background global features  $\mathbf{z}_b$ :

$$\begin{aligned} \mathbf{z}_g &= \frac{1}{HW} \sum_{(i,j) \in P} \mathbf{z}_{ij} \\ &= \frac{|P_f|}{HW} \times \frac{1}{|P_f|} \sum_{(i,j) \in P_f} \mathbf{z}_{ij} + \frac{|P_b|}{HW} \times \frac{1}{|P_b|} \sum_{(i,j) \in P_b} \mathbf{z}_{ij} \\ &= \lambda \times \text{GAP}(\mathcal{Z}_f) + (1 - \lambda) \times \text{GAP}(\mathcal{Z}_b) \\ &= \lambda \times \mathbf{z}_f + (1 - \lambda) \times \mathbf{z}_b, \end{aligned} \quad (1)$$

where  $P_f$  and  $P_b$  are the pixel sets of the foreground or background, respectively,  $\lambda$  is the proportion of foreground pixels, and  $\mathcal{Z}_*$  is a subset of  $\mathcal{Z}$  with elements picked by  $P_*$ , *i.e.*,  $\mathcal{Z}_* = \{\mathbf{z}_{ij} \mid (i, j) \in P_*\}$ .

Modern DNN-based models are trained with cross-entropy loss. With such a criterion, the global representation  $\mathbf{z}_g$  is trained to minimize the conditional entropy over class label  $H(y \mid \mathbf{z}_g)$ , which relates to the lower bound of the final cross-entropy loss. This objective is also equivalent to maximizing the mutual information between the features and the class label, as  $H(y \mid \mathbf{z}_g) = H(y) - I(y; \mathbf{z}_g)$ , where the entropy of  $y$  is a constant value. We next decompose the maximum mutual information objective with the following theorem. The proofs for two theorems are in the supplemental material.

**Theorem 1.** *For model  $\mathcal{W}$  with the given properties, the mutual information maximization objective decomposes to  $I(y; \mathbf{z}_g) = I(y; \mathbf{z}_f) - I(y; \mathbf{z}_f \mid \mathbf{z}_g)$ , where maximizing  $I(y; \mathbf{z}_f)$  is the classification objective and minimizing  $I(y; \mathbf{z}_f \mid \mathbf{z}_g) \geq 0$  is a regularization term.*

The regularization term  $-I(y; \mathbf{z}_f \mid \mathbf{z}_g)$  is to some extent unclear. To interpret how it functions, we next provide an analysis of its optimal solution.

**Theorem 2.** *The regularization term is optimized to zero if 1) **constant value solution:**  $\mathbf{z}_b$  is a constant value, or 2) **orthogonal subspace solution:**  $\mathbf{z}_f$  and  $\mathbf{z}_b$  are from different feature subspaces, *i.e.*,  $\mathbf{z}_g$  can be equivalently represented by the concatenation of  $\mathbf{z}_f$  and  $\mathbf{z}_b$ .*

**Constant value solution.** As ReLU activation maps all negative raw inputs to zero, the only feasible constant value is zero. Otherwise, raw inputs before GAP must stay constant, which is improbable. Our theorems demonstrate that the GAP regularizer suppresses background feature intensity to

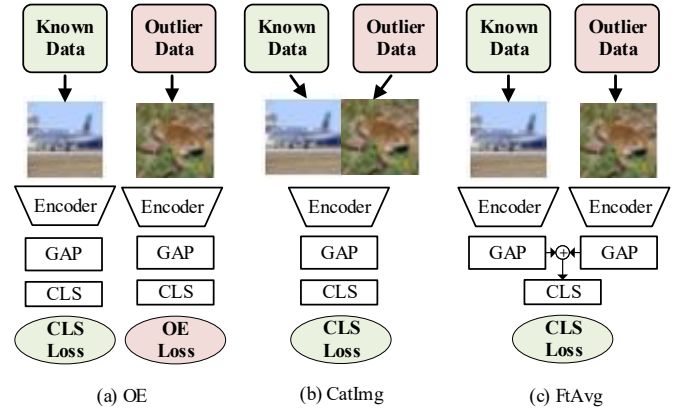


Fig. 2. Illustration of three different models compared in the OE experiment. (a) The traditional OE method trains known samples and outliers separately. (b) CatImg concatenates outliers to known samples, serving as constructed image backgrounds. (c) FtAvg inputs known samples and outliers to the backbone separately and restricts them from interacting under simulated GAP.

make their global feature zero, implicitly treating backgrounds as auxiliary outliers, which links backgrounds and available outliers in OE [19]. During training, OE optimizes an auxiliary task that samples from unknown classes should have minimum Maximum SoftMax Probability (MSP), *i.e.*, predicting uniform probability for known classes on unknown samples. The training objective is defined as:

$$\mathcal{L} = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_{in}} (H(y; \mathbf{x})) + \alpha \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{oe}} (H(u; \mathbf{x})), \quad (2)$$

where  $\mathcal{D}_{in}$  represents the distribution for known classes,  $\mathcal{D}_{oe}$  is the distribution for known outliers and  $u$  is the uniform class distribution, and  $\alpha$  is the weight for the auxiliary objective. As is pointed out in [20], such a regularization objective makes the deep feature from the penultimate layer zero for outliers.

**The orthogonal subspace solution.** Under this solution, the model encodes foregrounds and backgrounds as separate feature sets, which leads the model to learn discriminative features for distinguishing between foregrounds and backgrounds. Fore-background discrimination improves model robustness as well as enriches the hidden features, and more diverse features can potentially improve closed-set performance.

2) *Verification experiments:* To verify our theory and quantitatively evaluate the power of GAP regularizer, we conducted experiments following OE [19], which is one of the most representative auxiliary data-based methods and has outstanding performance. CIFAR10 [55] was used as the known dataset, while LSUN-Fix and ImageNet-Fix [56] were adopted as unknown datasets. LSUN-Fix and ImageNet-Fix contain randomly sampled and resized images from LSUN [57] and ImageNet [58], respectively. Auxiliary outliers were from TinyImage [19] or CIFAR100 [55]. We compared the performance of the following settings depicted in Fig. 2:

- **Plain:** the closed-set classifier baseline [52], where the auxiliary dataset is not utilized.
- **OE:** the OE baseline [19], where the model is regularized to predict a uniform distribution probability across known classes for auxiliary outliers.

TABLE II  
RESULTS FOR OUTLIER EXPOSURE EXPERIMENTS. THE GAIN OR LOSS VALUES ARE CALCULATED COMPARED TO THE PLAIN BASELINE.

Outlier	Method	Accuracy	Test Unknown LSUN-Fix				Test Unknown IMGN-Fix			
			AUROC	TNR@95	FT-Auc	FT-Cos	AUROC	TNR@95	FT-Auc	FT-Cos
-	Plain	95.8	89.5	45.8	66.5	43.0	89.7	48.1	66.5	43.6
TinyImage	OE	95.40.4↓	98.48.9↑	96.450.6↑	98.632.1↑	19.024.0↓	97.78.0↑	88.640.5↑	97.731.2↑	18.225.4↓
	CatImg	95.70.1↓	98.79.2↑	93.647.8↑	98.932.4↑	27.515.5↓	97.67.9↑	85.837.7↑	97.931.4↑	26.916.7↓
	FtAvg	95.90.1↑	98.79.2↑	94.448.6↑	98.932.4↑	25.217.8↓	97.78.0↑	81.833.7↑	97.931.4↑	25.418.2↓
CIFAR100	OE	96.10.3↑	97.68.1↑	88.642.8↑	97.430.9↑	15.427.6↓	97.37.6↑	86.938.8↑	97.230.7↑	16.327.3↓
	CatImg	95.80.0-	97.47.9↑	85.840.0↑	97.430.9↑	27.515.5↓	97.17.4↑	83.935.8↑	97.130.6↑	26.517.1↓
	FtAvg	95.70.1↓	96.97.4↑	81.836.0↑	96.930.4↑	25.717.3↓	96.97.2↑	81.833.7↑	96.830.3↑	25.318.3↓

- **CatImg**: an adaptation of GAP regularizer to OE problem. Each in-distribution image (shaped  $32 \times 32 \times 3$ ) is concatenated with a random known outlier image on the height dimension to form the  $64 \times 32 \times 3$  input.
- **FtAvg**: prevents the foreground and backgrounds from meeting early in the encoder to verify that the regularization functions via GAP. Known class features (serve as  $\mathbf{z}_f$ ) and outlier features (serve as  $\mathbf{z}_b$ ) are mixed by GAP, which follows Eq. (1) with  $\lambda = 0.5$ . Finally, the classification head inputs the mixed feature and is trained with the original known class label.

Besides the mentioned accuracy and AUROC, we evaluated models with three additional metrics: **TNR@95**, the rate of correctly rejecting unknown samples given a 95% true positive rate. As AUROC nears one, TNR@95 is more distinguishable in comparison. **FT-Auc**, the AUROC uses the Euclidean norm of global features as the score function, confirming that the GAP regularizer discovers a constant zero value solution, which suppresses the feature magnitude of unknown samples. **FT-Cos**: the average cosine similarity between each pair of known and unknown samples. If the model finds an orthogonal subspace solution, FT-Cos approaches zero.

**GAP regularizers prefer constant zero solution.** The performance on FT-Auc in Table II shows that the GAP regularizer can achieve even better performance simply with the activation intensity of the global feature. FT-Cos reduction implies that two solutions in Theorem 2 work together. OE also suppresses and orthogonalizes outlier features, while it shows significantly lower FT-Cos values than GAP regularizers. Thus, we suggest using constant zero activation intensity instead of classification head scores for GAP regularizers.

**Class-unrelated backgrounds regularize classifier via GAP.** The performance of CatImg and FtAvg shows no significant difference, suggesting GAP effectively regularizes the classifier using background patches. We suppose that local spatial pooling in the encoder has a similar function, explaining a slight overall improvement of CatImg.

**GAP regularizers show comparable performance without auxiliary optimization tasks.** Table II reveals that GAP regularizers attain comparable or superior AUROC relative to OE and exhibit robust regularization for known outliers. Meanwhile, OE necessitates manual outlier data specification and a custom loss function, while the GAP regularizer autonomously detects and regulates class-unrelated patches, *i.e.*,

TABLE III  
RESULTS FOR OUTLIER EXPOSURE WITH DIFFERENT DATASETS SERVED AS AUXILIARY OUTLIERS UNDER THE OE SETTINGS. THE GAIN OR LOSS VALUES ARE CALCULATED COMPARED TO THE BASELINE WITHOUT AUXILIARY OUTLIER.

Known	Outlier	Accuracy	TNR@95	AUROC
CIFAR10	-	95.8	59.7	87.9
	DTD	76.119.7↓	31.428.3↓	77.410.5↓
	LSUN-Fix	81.114.7↓	33.426.3↓	77.310.6↓
	Flower102	79.716.1↓	34.824.9↓	77.610.3↓
	TinyImage	95.40.4↓	76.817.1↑	92.64.7↑
CIFAR100	-	74.5	35.3	74.7
	DTD	46.727.8↓	15.719.6↓	58.716.0↓
	LSUN-Fix	50.224.3↓	21.413.9↓	66.38.4↓
	Flower102	49.724.8↓	20.714.6↓	64.510.2↓
	TinyImage	75.91.4↑	39.94.6↑	78.13.4↑

well-assumed backgrounds.

**OE is sensitive to selected outliers.** We additionally used three datasets as auxiliary outliers for training under the OE setting, including DTD [59], LSUN-Fix [56], and Flower102 [60]. DTD is a texture dataset with limited semantic information, LSUN-Fix has minimal overlap with the unknown test dataset, and Flower102 comprises abundant floral data. CIFAR10 or CIFAR100 is used as the known dataset, while the other is the unknown dataset. Besides the mentioned metrics, we used threshold-independent AUPR to evaluate performance based on precision and recall. Results in Table III show that the performance of OE varies significantly depending on the chosen outliers. OE shows impressive performance only when the auxiliary dataset has abundant semantic information. Moreover, the low accuracy in most settings indicates the limitations of OE in improving closed-set performance.

Based on the theoretical analysis and experimental validation, we find that GAP regularizers can use unknown regions within a single image for regularization. Moreover, GAP regularizers exhibit comparable performance to OE and prevent the model from overfitting to selected outliers, thereby ensuring performance stability and reducing the cost.

#### IV. BACKGROUND MIX FOR OPEN SET RECOGNITION

Under the common OSR setting, only known class samples and their labels are provided. Neither segmentation annotation nor auxiliary outlier data is available during training. We now

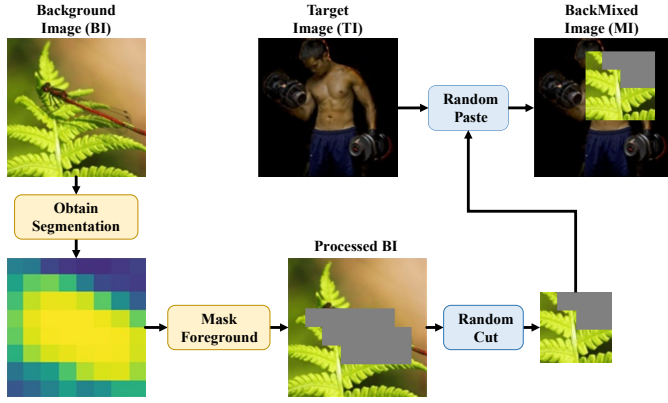


Fig. 3. Illustration of BackMix. BackMix first estimates and masks the foreground of the background image, then randomly cuts patches and pastes them on the target image to obtain the mixed image as the training sample.

propose a solution to apply our findings in practical open set recognition tasks.

#### A. The Proposed Method

We generally follow the cut-and-paste operation in Cutmix. Instead of treating both images equally as Cutmix does, we assign specific roles for the two images to be mixed. One image is designated as the target image (TI), and its label is used for training. The other image is identified as the background image (BI), whose foreground has been masked. Our main idea is to remove multiple foreground instances as well as multiple training objectives.

To mask out the foreground in BI without available annotations, we use Class Activation Maps (CAMs) for approximate estimation. Zhou *et al.* [33] employed GAP to generate CAMs, providing an estimation of visual cues for specific categories. The design of CAM leverages the fact that regions relevant to the model’s predictions have higher activation values compared to the background. We modify the classification head from GAP-Linear-SoftMax to Conv.1×1-GAP-SoftMax. Despite their computational equivalence, the latter architecture directly produces CAMs via the 1×1 convolution. We also apply a pixel-wise SoftMax across channels to ensure values remain within the [0, 1] range. The channel corresponding to the ground-truth category is then used to estimate the foreground regions.

Formally, suppose  $(x_T, y_T)$  and  $(x_B, y_B)$  are TI and BI, respectively, we generate the mixed sample  $(\tilde{x}, \tilde{y})$  by

$$\tilde{x} = \mathbf{M}_b \odot (1 - \mathbf{C}_B) \odot x_B + (1 - \mathbf{M}_b) \odot x_T, \quad \tilde{y} = y_T, \quad (3)$$

where  $\mathbf{M}_b \in \{0, 1\}^{H \times W}$  denotes a binary mask of the region where  $x_T$  cuts out its image patches and pastes the patches from  $x_B$ ,  $\mathbf{C}_B \in \{0, 1\}^{H \times W}$  denotes the foreground mask, and  $\odot$  denotes the element-wise multiplication. Rather than directly using the raw soft estimation from CAM, we sharpen the activation map to generate  $\mathbf{C}_B$ , where the highest  $k$  values are set to one and the rest set to zero.

The framework of BackMix is illustrated in Fig. 3 and is detailed in Algorithm 1. At the very beginning of training, we

---

#### Algorithm 1 Pseudo-code for BackMix

---

**Require:**  $\mathcal{D}_{(x,y)}$ : Training Set,  $k$ : mask ratio,  $s$ : cut size.

**Ensure:** A model  $\mathcal{W}$  that can well handle OSR tasks.

$\theta \leftarrow$  initialize parameters of model  $\mathcal{W}$ .

$\mathbf{C} \leftarrow$  initialize a CAM bank as soft foreground estimation.

**repeat**

**repeat**

    Pick a TI  $(x_T, y_T)$  and a BI  $(x_B, y_B)$  within a batch.

    Compute  $\mathbf{C}_B$  with  $k$  and  $\mathbf{C}$ .

    Compute mix region  $\mathbf{M}$  with  $s$ .

    Generate sample  $\tilde{x} = \mathbf{M} \odot (1 - \mathbf{C}_B) \odot x_B + (1 - \mathbf{M}) \odot x_T$ .

**until** all images have been selected as TI

  Obtain  $\hat{\mathbf{C}}$  with respect to  $\tilde{x}$ .

  Update segmentation  $\mathbf{C} \leftarrow \beta \cdot \hat{\mathbf{C}} + (1 - \beta) \cdot \mathbf{C}$ .

$g_\theta \leftarrow \nabla_{\theta} \log(\mathcal{W}_\theta(\tilde{\mathbf{X}})) + \log(1 - \mathcal{W}_\theta(\tilde{\mathbf{X}}))$ .

$\theta \leftarrow \theta - \eta g_\theta$ .

**until** convergence

---

initialize a bank  $\mathbf{C}$  to store the rough foreground estimation generated with CAMs. In the consequent training steps, we randomly pick an image as TI  $(x_T, y_T)$  and another image as BI  $(x_B, y_B)$  within a batch, and then compute the foreground mask  $\mathbf{C}_B$  with  $\mathbf{C}$  and the mask ratio  $k$ . Then, we mask out the foreground of  $x_B$  and mix it with  $x_T$  to generate  $\tilde{x}$ , and the value of mix region  $\mathbf{M}$  depending on cut size  $s$ . After that, we update stored mask  $\mathbf{C}$  with the CAM  $\hat{\mathbf{C}}$  generated by the network. Finally, the mixing process terminates until all images have been served as TI, and the mixed images  $\tilde{\mathbf{X}}$  are used to train the network and update the parameter.

During the initial training steps, CAM may not provide an accurate foreground estimation as the network has yet to learn effective representations. To tackle this, we use a uniform distribution to initialize the mask bank  $\mathbf{C}$  and update these masks using an exponential moving average as follows:

$$\mathbf{C}_t = \beta \cdot \hat{\mathbf{C}}_t + (1 - \beta) \cdot \mathbf{C}_{t-1}, \quad (4)$$

where  $\mathbf{C}_{t-1}$  and  $\mathbf{C}_t$  represent the estimated masks obtained in the  $t-1$ -th and  $t$ -th training step, respectively.  $\hat{\mathbf{C}}_t$  denotes the mask calculated only using CAM in  $t$ -th step and  $\beta$  is the exponential decay rate.

As we train the model with only the label of TI, avoiding the effect of foreground objects in the pasted patches is quite important. Unlike the random cut size used in Cutmix, we cut a square region of a fixed size. With a reasonable cut size, *e.g.*, half of the image width or height, TI is guaranteed to dominate the label even if unexpected objects are pasted, which accounts for at most 25% of the mixed image. One concern is that the masked regions might not provide background information. However, this actually falls into the Cutout if the cut region is fully masked. Such a situation also helps regularize the classifier. Visualizations in Section V-C4 demonstrate that BackMix can effectively estimate the foreground of BI and avoid obscuring the main object in TI.

During the test phase, samples with scores below the threshold are deemed unknown and are rejected, while the remaining samples are classified as known and the model outputs corresponding predictions. The threshold is set to ensure

TABLE IV  
AUROC SCORE COMPARISON OF DIFFERENT OSR METHODS IN UNKNOWN DETECTION TASKS.

Method	SVHN	CIFAR10	CIFAR+10	CIFAR+50	Tiny-IN
OSRCI [12]	91.0	69.9	83.8	82.7	58.6
CROSR [61]	89.9	88.3	91.2	90.5	58.9
C2AE [14]	92.2	89.5	95.5	93.7	74.8
CGDL [62]	93.5	90.3	95.9	95.0	76.2
GDFR [8]	93.5	83.1	91.5	91.3	64.7
PROSER [10]	94.3	89.1	96.0	95.3	69.3
Plain*	88.6	67.7	81.6	80.5	57.7
+ BackMix	97.0 <sub>8.4</sub> ↑	91.3 <sub>23.6</sub> ↑	91.9 <sub>10.3</sub> ↑	91.6 <sub>11.1</sub> ↑	80.4 <sub>22.7</sub> ↑
ARPL [16]	95.3	89.8	91.3	90.8	76.0
+ BackMix	96.4 <sub>1.1</sub> ↑	91.0 <sub>1.2</sub> ↑	93.4 <sub>2.1</sub> ↑	92.3 <sub>1.5</sub> ↑	76.3 <sub>0.3</sub> ↑
CSSR [18]	96.7	90.7	91.5	90.9	80.6
+ BackMix	97.1 <sub>1.0</sub> ↑	94.2 <sub>3.5</sub> ↑	96.4 <sub>4.9</sub> ↑	95.7 <sub>4.8</sub> ↑	83.1 <sub>2.5</sub> ↑

that 95% of the known samples are correctly classified. As BackMix can be considered as a data augmentation technique, it is applicable for any score function of the original method. For example, when applying it to the SoftMax baseline, the MSP is used as the score function.

### B. Discussion

**Is it necessary to segment foregrounds precisely?** Generating foreground masks via CAM may not yield precise segmentation, because CAM only highlights the discriminative regions predicted by the model. If the discriminative regions are pasted to TI, the model can be confused, as the true label changes from  $y_B$  to  $y_T$ . Masking out these regions can preserve the well-learned knowledge for class  $y_B$ . In cases where image backgrounds are mistakenly estimated as foregrounds with high probability, a strategy that masks out only a certain fraction ( $k$ ) of the foregrounds would be beneficial. As long as the real foregrounds have higher confidence, the backgrounds remain and are pasted to TI. After several training steps, these mistakes can be corrected as they are randomly pasted onto all samples. Consequently, our framework can boost the quality of foreground segmentation by itself during training.

**Is it necessary to mix the labels?** Although we mix two different images, it is not necessary to mix two image labels. The most discriminative regions for BI are masked out and are supposed to make few differences. Note that it is not necessary to paste the BI to the background regions of TI carefully. Pasting BI on foreground regions of TI functions like Cutout, which simulates occlusions for training images. In Section V, we empirically show that mixing labels of two images could limit the OSR performance.

## V. EXPERIMENTS

**Implementation details.** We set the fixed cut size to  $0.5 \times$  the height and width of the image. The exponential decay rate  $\beta$  for segmentation masks was set to 0.1. We set  $k = 0.25$ , i.e., 25% pixels with the highest segmentation scores are masked out for BI. We trained WideResNet40-4 [63] on small-scale datasets (e.g. CIFAR10) and ResNet18 [64] on ImageNet30 and Tiny-ImageNet, setting the batch size to 128

TABLE V  
OPEN SET RECOGNITION PERFORMANCE WITH CIFAR10 AS KNOWN AND VARIOUS DATASETS AS UNKNOWN.

Method	IMGN-C	IMGN-R	LSUN-C	LSUN-R
CROSR [61]	72.1	73.5	72.0	74.9
GDFR [8]	75.7	79.2	75.1	80.5
C2AE [14]	83.7	82.6	78.3	80.1
CGDL [62]	84.0	83.2	80.6	81.2
PROSER [10]	84.9	82.4	86.7	85.6
ConOSR [11]	89.1	84.3	91.2	88.1
Plain*	63.9	65.3	64.2	64.7
+ BackMix	92.6 <sub>28.7</sub> ↑	90.4 <sub>25.1</sub> ↑	92.6 <sub>28.4</sub> ↑	93.3 <sub>28.6</sub> ↑
ARPL [16]	80.6	82.5	85.3	82.7
+ BackMix	92.3 <sub>11.7</sub> ↑	91.3 <sub>8.8</sub> ↑	92.9 <sub>7.6</sub> ↑	94.2 <sub>11.5</sub> ↑
CSSR [18]	88.3	89.5	92.2	90.4
+ BackMix	93.7 <sub>5.4</sub> ↑	93.0 <sub>3.5</sub> ↑	94.7 <sub>2.5</sub> ↑	94.8 <sub>4.4</sub> ↑

and the learning rate to 0.1, with a cosine annealing learning rate scheduler and an SGD optimizer.

### A. Comparison with OSR Methods

To verify the effectiveness of the proposed method, we applied BackMix to the baseline strategy MSP (Plain\*) [52] and also integrated it to state-of-the-art methods ARPL [16] and CSSR [18]. Notice that to ensure a fair comparison, only simple data augmentation (e.g. RandomHorizontalFlip and RandomCrop) used in the original method was retained, and all experiments were conducted under the original settings.

1) *Unknown detection:* For the OSR task, we first followed the setting from [12] to conduct the unknown detection experiments. Five standard benchmarks were adopted in the experiments, including CIFAR10 [55], SVHN [65], CIFAR+10, CIFAR+50 and Tiny-ImageNet [66].

For ten-class datasets CIFAR10 and SVHN, we randomly selected six classes as known classes to appear during training, and the remaining four classes as unknown classes for testing. For CIFAR+N datasets, we selected four non-animal classes from CIFAR10 as known classes, while using  $N$  animal classes from CIFAR100 as unknown classes and set  $N=10$  and  $N=50$  to test performance at the different scenarios. For the large-scale and more challenging dataset Tiny-ImageNet (Tiny-IN), we selected 20 classes as known and the remaining 180 classes as unknown. Following the standard evaluation protocol, we adopted the threshold-independent metric AUROC, and all the reported results were the average of five trials.

**BackMix significantly enhances the open set performance of various methods and surpasses the state-of-the-art by removing the fore-background priors.** We compared the proposed method with classic and advanced OSR methods. The performance values of other methods were from [8], [10], [14], [18], [61], [62] or reproduced with the official code under our settings. Results in Table IV show that BackMix significantly improves the performance of the plain SoftMax with little cost, making it exceed many complex methods, especially in the CIFAR10 (+23.6%) and Tiny-ImageNet (+22.7%) experiments. In addition, by applying our method to state-of-the-art methods, ARPL, and CSSR, they



TABLE VI  
COMPARISON FOR DISTINGUISHING KNOWN DATASET CIFAR10 FROM NEAR OOD DATASET CIFAR100 AND FAR OOD DATASET SVHN.

Method	In:CIFAR10 / Out:CIFAR100				In:CIFAR10 / Out:SVHN			
	DTACC	AUROC	AUIN	AUOUT	DTACC	AUROC	AUIN	AUOUT
GCPL [17]	80.2	86.4	86.6	84.1	86.1	91.3	86.6	94.8
RPL [15]	80.6	87.1	88.8	83.8	87.1	92.0	89.6	95.1
CSI [56]	84.4	91.6	92.5	90.0	92.8	97.9	96.2	99.0
OpenGAN [22]	84.2	89.7	87.7	89.6	92.1	95.9	93.4	97.1
Plain*	79.8	86.3	88.4	82.5	86.4	90.6	88.3	93.6
+ BackMix	84.9 <sub>5.1</sub> ↑	91.3 <sub>5.0</sub> ↑	93.0 <sub>4.6</sub> ↑	88.1 <sub>5.6</sub> ↑	88.5 <sub>2.1</sub> ↑	94.1 <sub>3.5</sub> ↑	93.5 <sub>5.2</sub> ↑	97.5 <sub>3.9</sub> ↑
ARPL [16]	80.8	88.2	90.4	84.4	82.8	90.5	84.6	95.3
+ BackMix	84.0 <sub>3.2</sub> ↑	91.1 <sub>2.9</sub> ↑	92.1 <sub>1.7</sub> ↑	89.0 <sub>4.6</sub> ↑	94.9 <sub>12.1</sub> ↑	98.5 <sub>8.0</sub> ↑	97.6 <sub>13.0</sub> ↑	99.1 <sub>3.8</sub> ↑
CSSR [18]	83.1	90.3	91.3	87.8	94.1	98.1	97.1	98.2
+ BackMix	86.3 <sub>3.2</sub> ↑	93.0 <sub>2.7</sub> ↑	93.7 <sub>2.4</sub> ↑	91.7 <sub>3.9</sub> ↑	96.4 <sub>2.3</sub> ↑	99.2 <sub>1.1</sub> ↑	98.4 <sub>1.3</sub> ↑	99.6 <sub>1.4</sub> ↑

obtained consistent performance improvement in all settings and achieved new state-of-the-art performance. Therefore, we demonstrated the effectiveness of the proposed method design and its applicability.

2) *Open set classification*: We then followed a common experimental setup from [61] to test the performance when introducing unknown samples from other datasets. The models were trained on CIFAR10, while in the test phase, the samples from ImageNet [67] and LSUN [57] were used as unknowns. These two datasets are cropped or resized so that their image size can remain the same as known samples, which form ImageNet-Crop (IMGN-C), ImageNet-Resize (IMGN-R), LSUN-Crop (LSUN-C), and LSUN-Resize (LSUN-R). For a fair comparison, we used the release version of the four datasets from [7]. The performance was evaluated by macro-averaged F1-scores in 11 classes (including 10 known classes and 1 unknown class).

**BackMix establishes impressive capability in classifying known samples and distinguishing unknown samples.** We reported results in Table V and the values other than ours are taken from [10], [17], [56], [62] or reproduced using the official code under our settings. We can observe that BackMix makes the plain baseline exceed many recent complex methods and further improves the CSSR [18] to outperform existing state-of-the-art OSR methods by a significant margin.

3) *Out-of-Distribution detection*: Considering the unknown classes from different datasets, we also followed the setting of [16] to carry out the out-of-distribution detection experiment. For the out-of-distribution (OOD) detection task, in addition to the AUROC and AUPR, we used the DTACC following [16], which calculates the maximum known or unknown classification accuracy a model can achieve over all possible decision thresholds. The metric AUPR becomes AUIN (or AUOUT) if known (or unknown) samples are specified as the positive class.

**BackMix improves performance on detecting unknown samples from datasets that are either slightly or significantly different from the known dataset.** Results in Table VI indicate that applying BackMix on different methods significantly improves their performance of OOD detection tasks and reached higher results compared with state-of-the-art methods. Concretely, BackMix enhanced ARPL by up

TABLE VII  
COMPARISON FOR CLOSED-SET CLASSIFICATION AND OPEN SET DETECTION PERFORMANCE ON SPLIT IMAGENET30.

Augmentation	Accuracy	AUROC	AUIN	AUOUT
Plain*	94.9	89.9	86.4	92.9
Cutout [34]	95.7 <sub>0.8</sub> ↑	90.4 <sub>0.5</sub> ↑	86.7 <sub>0.3</sub> ↑	92.9 <sub>0.0</sub> -
Mixup [37]	95.9 <sub>1.0</sub> ↑	89.8 <sub>0.1</sub> ↓	80.8 <sub>5.6</sub> ↓	94.0 <sub>1.1</sub> ↑
Cutmix [38]	96.8 <sub>1.9</sub> ↑	89.7 <sub>0.2</sub> ↓	76.7 <sub>9.7</sub> ↓	93.5 <sub>0.6</sub> ↑
BackMix	97.2 <sub>2.3</sub> ↑	91.3 <sub>1.4</sub> ↑	87.7 <sub>1.3</sub> ↑	94.1 <sub>1.2</sub> ↑

to 13.0% on AUIN and CSSR by up to 3.9% on AUOUT. Note that the improvement brought by BackMix on CSSR is less obvious compared to other baselines, which is probably because CSSR already had outstanding performance. Besides, we applied BackMix on the plain baseline with various post-processing methods as score functions in the supplemental material. BackMix consistently improves performance on both prediction-based and feature-based score functions.

### B. Comparison with Data Augmentation Techniques

Since the BackMix can be seen as a data augmentation technique, we compared it with three commonly used strong techniques—Cutout [34], Mixup [37], and Cutmix [38]—using the plain SoftMax strategy. As discussed in Section II-B, these techniques inherently mix backgrounds.

To evaluate the closed-set and open set performance, we adopted several metrics including closed-set accuracy, AUROC, AUIN (or AUOUT), and DTACC as used in previous experiments. All the above metrics are threshold-free. Following the standard setting in V-A1, we conducted experiments on split ImageNet30 [68], which contains 30 classes with 1300 training images and 100 test images per class, and the image resolution is 224×224. We took the first 10 classes as known classes and the rest 20 as unknown classes during test.

**BackMix is more suitable for the OSR task compared to other data augmentation techniques.** Results in Table VII indicate that BackMix obtains a significant improvement in both closed-set and open set scenarios. The compared methods enhanced the closed-set accuracy but obtained limited improvement or even performance drop on open set metrics, which infers that mixing labels may help improve closed-set

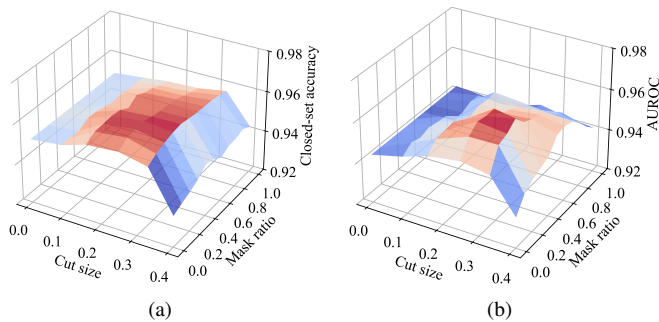


Fig. 4. Parameter Analysis for BackMix, where (a) presents the closed-set classification accuracy, (b) presents the AUROC with varying cut size  $s$  and mask ratio  $k$ . The values of open set metric AUROC are averaged on the six unknown datasets.

performance but can limit open set performance. The proposed BackMix used hard labels and avoided multiple objects in foregrounds, keeping both closed and open set performance.

### C. Further Analysis

1) *Closed-set classification performance*: For the OSR task, accurate classification of known classes is also very important. Therefore, we compared the closed-set performance of different methods on CIFAR10. To test the effectiveness of BackMix in improving classification performance, we applied it to three baselines: plain SoftMax, ARPL, and CSSR.

Results in Table VIII show that applying BackMix on different baselines obtains consistent improvement in the closed-set performance, and previous experiments show that BackMix improves the open set performance concurrently, which indicates BackMix is suitable for the OSR task.

2) *Is BackMix sensitive to hyperparameters*: We performed parameter analysis on the cut size  $s$  and mask ratio  $k$  here. Cut size  $s$  represents the ratio of the cutting area to the whole image area, e.g., a cut box sized  $0.5 \times 0.5$  corresponds to  $s = 0.25$ . In this experiment, ten classes from CIFAR10 were used as known, while six datasets, including ImageNet-Crop, ImageNet-Resize, ImageNet-Fix, LSUN-Crop, LSUN-Resize, and LSUN-Fix, were treated as unknown. We used the preprocessed version in [7], [56]. We varied the cut size  $s$  from 0 to 0.4 and the mask ratio  $k$  from 0 to 1. It should be noted that BackMix converts to Cutout with  $k = 1$ , where BI

TABLE VIII  
CLOSED-SET CLASSIFICATION ACCURACY PERFORMANCE COMPARISON ON THE CIFAR10 DATASET.

Method	Accuracy
CROSR [61]	94.0
CGDL [62]	91.2
GCPL [17]	93.3
Plain*	94.0
+ BackMix	95.1 <sub>1.1</sub> ↑
ARPL [16]	92.7
+ BackMix	93.2 <sub>0.5</sub> ↑
CSSR [18]	94.2
+ BackMix	95.6 <sub>1.4</sub> ↑

TABLE IX  
OOD DETECTION PERFORMANCE OF BACKMIX IN THE FINETUNING STAGE. WE USED CIFAR10 AS THE IN-DISTRIBUTION DATASET AND CIFAR100 AS THE OOD DATASET.

Method	1-shot		4-shot		16-shot	
	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC
CoOp [69]	89.8	91.6	90.6	91.5	91.2	91.1
+BackMix	90.7 <sub>0.9</sub> ↑	92.1 <sub>0.5</sub> ↑	91.3 <sub>0.7</sub> ↑	92.1 <sub>0.6</sub> ↑	91.7 <sub>0.5</sub> ↑	91.6 <sub>0.5</sub> ↑
LoCoOp [70]	89.6	91.2	89.8	91.4	91.4	90.4
+BackMix	90.7 <sub>1.1</sub> ↑	91.6 <sub>0.4</sub> ↑	90.9 <sub>1.1</sub> ↑	91.9 <sub>0.5</sub> ↑	91.7 <sub>0.3</sub> ↑	91.0 <sub>0.6</sub> ↑

is processed to be pure grey, and with  $k = 0$ , we do not mask anything in BI at all.

**BackMix maintains stable performance and is not sensitive to hyperparameters except for extreme value cases.** We plotted closed-set accuracy and AUROC results in Fig. 4. Proper values (0.2-0.33) for cut size as well as mask ratio boost both closed and open set performance. For the mask ratio  $k$ , masking out a proper fraction of possible foregrounds does improve open set performance. It is also interesting to find that though the closed-set performance for  $k = 1$  (Cutout) raises, the open set performance drops. Therefore, we emphasize that mixing backgrounds is important for OSR while avoiding multiple foregrounds is essential for preserving closed-set performance. This meets the conclusion of the empirical study in III-A. Additionally, we also demonstrated the robustness of BackMix in background selection in the supplemental material.

3) *BackMix on large-scale pretrained models*: **BackMix further improves pretrained model performance in the few-shot finetuning stage.** We used CIFAR10 as the known dataset and CIFAR100 as the unknown dataset to evaluate the performance of model on the few-shot OOD detection task. BackMix was applied to the few-shot finetuning methods CoOp [69] and LoCoOp [70] based on large-scale pretrained model CLIP [71]. Results in Table IX show that BackMix leverages the powerful representational capacity of the large-scale pretrained model and further enhances both the closed-set and open set performance of the model even with only 1 sample from each known class.

4) *Visualizations*: **BackMix estimates the foreground regions accurately without additional annotations.** We provide examples of the estimated foreground masks from experi-

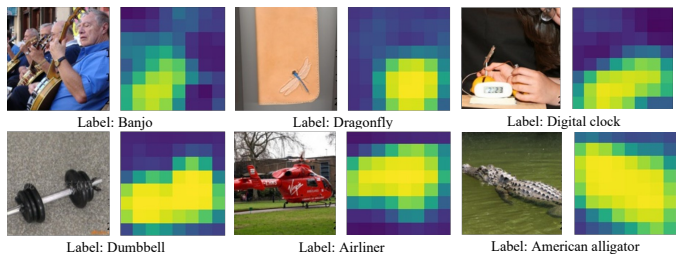


Fig. 5. Examples of the estimated foreground masks, and labels have been annotated below the corresponding image. The rough segmentation using CAM can effectively estimate the foreground.

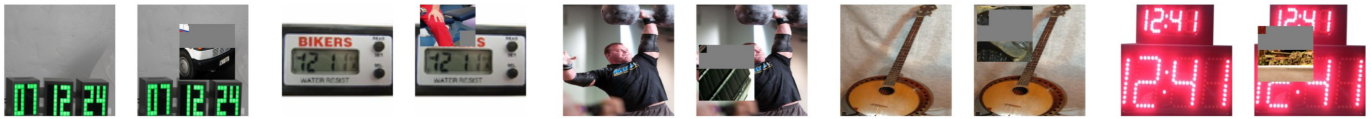


Fig. 6. Examples of the BackMix processed images. The pasted background patches contain almost no foreground objects from another image. By setting a reasonable cut size, we can ensure that the processed training samples retain sufficient information about the classification object in the target image.

ments on ImageNet30. As shown in Fig. 5, the estimations are accurate enough to mask out most of the foreground regions, verifying the effectiveness and feasibility of the BackMix.

**BackMix prevents the performance degradation caused by foreground occlusion.** To better illustrate the effectiveness of our method in image processing, we provide some processed samples in Fig. 6. The cut portions from the BI do not contain significant foreground objects, which avoids misleading the model after being pasted to TI. Due to the setting of proper mixing ratio, our method preserves sufficient classification information for the model to make judgments, regardless of whether the majority or minority of the region of interest is present. Therefore, it avoids situations where key areas are occluded and cannot be accurately classified.

## VI. CONCLUSIONS

In this paper, we discuss open set recognition from a new perspective of fore-background priors. We explore the role of fore-background priors and provide insights into how classifiers model the backgrounds in OSR. Empirical and theoretical analyses show that the underlying fore-background priors have negative impacts on OSR performance while removing these priors can enhance the performance. More importantly, class-unrelated backgrounds serve as auxiliary known outliers and provide extra regularization via global average pooling. Inspired by the insights, we design a new BackMix method that mixes the foreground of an image with the backgrounds of different images. The proposed method is simple to implement, requires no segmentation annotations or priors, and can seamlessly integrate into the learning process of almost all of the existing methods. Extensive experiments show that our method can significantly enhance state-of-the-art open set recognition methods and show clear advantages over existing data augmentation methods.

## REFERENCES

- [1] Z.-H. Zhou, “Open-environment machine learning,” *Nat. Sci. Rev.*, vol. 9, no. 8, pp. 1–11, 2022.
- [2] F. Angiulli, R. Ben-Eliyahu-Zohary, and L. Palopoli, “Outlier detection for simple default theories,” *Artif. Intell.*, vol. 174, no. 15, pp. 1247–1253, 2010.
- [3] C. Basich, J. Svegliato, K. H. Wray, S. Witwicki, J. Biswas, and S. Zilberstein, “Competence-aware systems,” *Artif. Intell.*, p. 103844, 2022.
- [4] N. Lu, G. Zhang, and J. Lu, “Concept drift detection via competence models,” *Artif. Intell.*, vol. 209, pp. 11–28, 2014.
- [5] M. Hanheide, M. Göbelbecker, G. S. Horn, A. Pronobis, K. Sjö, A. Aydemir, P. Jensfelt, C. Gretton, R. Dearden, M. Janicek, H. Zender, G.-J. Kruijff, N. Hawes, and J. L. Wyatt, “Robot task planning and explanation in open and uncertain worlds,” *Artif. Intell.*, vol. 247, pp. 119–150, 2017.
- [6] A. Bendale and T. E. Boulton, “Towards open set deep networks,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 1563–1572.
- [7] S. Liang, Y. Li, and R. Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks,” in *Int. Conf. Learn. Represent.*, 2018.
- [8] P. Perera, V. I. Morariu, R. Jain, V. Manjunatha, C. Wigginton, V. Ordonez, and V. M. Patel, “Generative-discriminative feature representations for open-set recognition,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 11 814–11 823.
- [9] W. Liu, X. Wang, J. D. Owens, and Y. Li, “Energy-based out-of-distribution detection,” in *Adv. Neural Inform. Process. Syst.*, 2020, pp. 21 464–21 475.
- [10] D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, “Learning placeholders for open-set recognition,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 4401–4410.
- [11] B. Xu, F. Shen, and J. Zhao, “Contrastive open set recognition,” in *AAAI Conf. on Artif. Intell.*, vol. 37, no. 9, 2023, pp. 10 546–10 556.
- [12] L. Neal, M. L. Olson, X. Z. Fern, W.-K. Wong, and F. Li, “Open set learning with counterfactual images,” in *Eur. Conf. Comput. Vis.*, 2018, pp. 620–635.
- [13] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” in *Adv. Neural Inform. Process. Syst.*, vol. 31, 2018, pp. 7167–7177.
- [14] P. Oza and V. M. Patel, “C2ae: Class conditioned auto-encoder for open-set recognition,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 2307–2316.
- [15] G. Chen, L. Qiao, Y. Shi, P. Peng, J. Li, T. Huang, S. Pu, and Y. Tian, “Learning open set network with discriminative reciprocal points,” in *Eur. Conf. Comput. Vis.*, 2020, pp. 507–522.
- [16] G. Chen, P. Peng, X. Wang, and Y. Tian, “Adversarial reciprocal points learning for open set recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8065–8081, 2022.
- [17] H.-M. Yang, X.-Y. Zhang, F. Yin, Q. Yang, and C.-L. Liu, “Convolutional prototype network for open set recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2358–2370, 2022.
- [18] H. Huang, Y. Wang, Q. Hu, and M.-M. Cheng, “Class-specific semantic reconstruction for open set recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4214–4228, 2023.
- [19] D. Hendrycks, M. Mazeika, and T. Dietterich, “Deep anomaly detection with outlier exposure,” in *Int. Conf. Learn. Represent.*, 2019.
- [20] A. R. Dhamija, M. Günther, and T. E. Boulton, “Reducing network agnostophobia,” in *Adv. Neural Inform. Process. Syst.*, vol. 31, 2018, pp. 9157–9168.
- [21] P. Perera and V. M. Patel, “Deep transfer learning for multiple class novelty detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 11 536–11 544.
- [22] S. Kong and D. Ramanan, “Opengan: Open-set recognition via open data generation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–10, 2022.
- [23] J. Cen, D. Luan, S. Zhang, Y. Pei, Y. Zhang, D. Zhao, S. Shen, and Q. Chen, “The devil is in the wrongly-classified samples: Towards unified open-set recognition,” in *Int. Conf. Learn. Represent.*, 2023, pp. 1–12.
- [24] Y. Bai, Z. Han, B. Cao, X. Jiang, Q. Hu, and C. Zhang, “Id-like prompt learning for few-shot out-of-distribution detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.
- [25] F. P. De Lange, M. Heilbron, and P. Kok, “How do expectations shape perception?” *Trends Cogniti. Sci.*, vol. 22, no. 9, pp. 764–779, 2018.
- [26] A. Oliva and A. Torralba, “The role of context in object recognition,” *Trends Cogniti. Sci.*, vol. 11, no. 12, pp. 520–527, 2007.
- [27] C. Ranganath and M. Ritchey, “Two cortical systems for memory-guided behaviour,” *Nat. Rev. Neurosci.*, vol. 13, no. 10, pp. 713–726, 2012.
- [28] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. B. Tenenbaum, and B. Katz, “Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models,” in *Adv. Neural Inform. Process. Syst.*, 2019.
- [29] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, “Distributionally robust neural networks for group shifts: On the importance of regular-

- ization for worst-case generalization,” in *Int. Conf. Learn. Represent.*, 2020.
- [30] R. Shetty, B. Schiele, and M. Fritz, “Not using the car to see the sidewalk – quantifying and controlling the effects of context in classification and segmentation,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8210–8218.
- [31] K. Y. Xiao, L. Engstrom, A. Ilyas, and A. Madry, “Noise or signal: The role of image backgrounds in object recognition,” in *Int. Conf. Learn. Represent.*, 2021.
- [32] A. Rosenfeld, R. Zemel, and J. K. Tsotsos, “The elephant in the room,” 2018. [Online]. Available: <https://arxiv.org/abs/1808.03305>
- [33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2921–2929.
- [34] T. DeVries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout,” *arXiv preprint arXiv:1708.04552*, 2017.
- [35] K. K. Singh and Y. J. Lee, “Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.
- [36] P. Chen, S. Liu, H. Zhao, and J. Jia, “Gridmask data augmentation,” *arXiv preprint arXiv:2001.04086*, 2020.
- [37] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” in *Int. Conf. Learn. Represent.*, 2017.
- [38] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, “Cutmix: Regularization strategy to train strong classifiers with localizable features,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 6023–6032.
- [39] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, “Domain generalization with mixstyle,” in *Int. Conf. Learn. Represent.*, 2021.
- [40] W. Ye, G. Zheng, X. Cao, Y. Ma, X. Hu, and A. Zhang, “Spurious correlations in machine learning: A survey,” *arXiv preprint arXiv:2402.12715*, 2024.
- [41] M. Srivastava, T. Hashimoto, and P. Liang, “Robustness to spurious correlations via human annotations,” in *Int. Conf. Mach. Learn.*, 2020, pp. 9109–9119.
- [42] E. Creager, J.-H. Jacobsen, and R. Zemel, “Environment inference for invariant learning,” in *Int. Conf. Mach. Learn.*, 2021, pp. 2189–2200.
- [43] H. Yao, Y. Wang, S. Li, L. Zhang, W. Liang, J. Zou, and C. Finn, “Improving out-of-distribution robustness via selective augmentation,” in *Int. Conf. Mach. Learn.*, 2022, pp. 25 407–25 437.
- [44] Y. Du, J. Yan, Y. Chen, J. Liu, S. Zhao, Q. She, H. Wu, H. Wang, and B. Qin, “Less learn shortcut: analyzing and mitigating learning of spurious feature-label correlation,” in *Int. Joint Conf. Artif. Intell.*, 2023, pp. 5039–5048.
- [45] S. Liu, X. Zhang, N. Sekhar, Y. Wu, P. Singhal, and C. Fernandez-Granda, “Avoiding spurious correlations via logit correction,” in *Int. Conf. Learn. Represent.*, 2023.
- [46] S. Asgari, A. Khani, F. Khani, A. Gholami, L. Tran, A. Mahdavi Amiri, and G. Hamarneh, “Masktune: Mitigating spurious correlations by forcing to explore,” in *Adv. Neural Inform. Process. Syst.*, vol. 35, 2022, pp. 23 284–23 296.
- [47] Y. Ming, H. Yin, and Y. Li, “On the impact of spurious correlation for out-of-distribution detection,” in *AAAI conf. on artificial intelligence*, vol. 36, no. 9, 2022, pp. 10 051–10 059.
- [48] E. T. Jaynes, “Information theory and statistical mechanics,” *Phys. Rev.*, vol. 106, pp. 620–630, 1957.
- [49] W. Deng and L. Zheng, “Are labels always necessary for classifier accuracy evaluation?” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 15 069–15 078.
- [50] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
- [51] G. V. Horn, O. M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. J. Belongie, “The inaturalist species classification and detection dataset,” *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8769–8778, 2018.
- [52] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” in *Int. Conf. Learn. Represent.*, 2017, pp. 1–12.
- [53] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, “Open-set recognition: A good closed-set classifier is all you need,” in *Int. Conf. Learn. Represent.*, 2022.
- [54] M. Lin, Q. Chen, and S. Yan, “Network in network,” 2013. [Online]. Available: <https://arxiv.org/abs/1312.4400>
- [55] A. Krizhevsky, “Learning multiple layers of features from tiny images.” *Tech Report*, 2009.
- [56] J. Tack, S. Mo, J. Jeong, and J. Shin, “Csi: Novelty detection via contrastive learning on distributionally shifted instances,” in *Adv. Neural Inform. Process. Syst.*, vol. 33, 2020, pp. 11 839–11 852.
- [57] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao, “Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop,” *arXiv preprint arXiv:1506.03365*, 2015.
- [58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.
- [59] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi, “Describing textures in the wild,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014.
- [60] M.-E. Nilsback and A. Zisserman, “Automated flower classification over a large number of classes,” *Indian Conf. Comput. Vis., Graph. Imag. Process.*, pp. 722–729, 2008.
- [61] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Nae-mura, “Classification-reconstruction learning for open-set recognition,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 4016–4025.
- [62] X. Sun, Z. Yang, C. Zhang, K.-V. Ling, and G. Peng, “Conditional gaussian distribution learning for open set recognition,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 13 480–13 489.
- [63] S. Zagoruyko and N. Komodakis, “Wide residual networks,” in *Brit. Mach. Vis. Conf.*, 2016, pp. 1–15.
- [64] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.
- [65] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” in *Adv. Neural Inform. Process. Syst.*, 2011, pp. 1–9.
- [66] H. Pouransari and S. Ghili, “Tiny imagenet visual recognition challenge,” *CS 231N*, 2014.
- [67] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
- [68] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, “Using self-supervised learning can improve model robustness and uncertainty,” in *Adv. Neural Inform. Process. Syst.*, vol. 32, 2019, pp. 15 663–15 674.
- [69] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [70] A. Miyai, Q. Yu, G. Irie, and K. Aizawa, “Locoop: few-shot out-of-distribution detection via prompt learning,” in *Adv. Neural Inform. Process. Syst.*, 2023, pp. 76 298–76 310.
- [71] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.



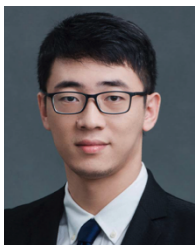
**Yu Wang** received the B.S. degree in communication engineering, the M.S. degree in software engineering, and the Ph.D. degree in computer applications and techniques from Tianjin University in 2013 and 2016, and 2020, respectively. He is currently an assistant professor at Tianjin University. His research is focused on data mining and machine learning, especially multi-granularity learning in the open and dynamic environment for computer vision and industrial applications.



**Junxian Mu** received the B.S. degree in computer science and technology from Dalian University of Technology in 2023, and is currently pursuing her M.S. degree in Tianjin University. Her research focuses on open set recognition and unknown detection in computer vision.



**Hongzhi Huang** received the B.S. degree in computer science and technology from Tianjin University in 2020, and is pursuing his M.S. degree in Tianjin University. His research interests are open set recognition and out-of-distribution detection in computer vision.



**Qilong Wang** received the PhD degree from the School of Information and Communication Engineering, the Dalian University of Technology, China, in 2018. He currently is an associate professor of Tianjin University. His research interests include visual understanding and deep learning, particularly deep models with high-order statistical modeling and self-attention mechanism. He is served as an Area Chair of CVPR 2024.



**Pengfei Zhu** received the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong, China, in 2015. He received his B. S. and M. S. from Harbin Institute of Technology, Harbin, China in 2009 and 2011, respectively. Now he is an associate professor with the College of Intelligence and Computing, Tianjin University. His interests are focused on machine learning and computer vision.



**Qinghua Hu** received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008, respectively. After that he joined Department of Computing, The Hong Kong Polytechnical University as a postdoctoral fellow. He became a full professor with Tianjin University in 2012, and now is a Chair Professor at College of Intelligence and Computing. His research interest is focused on uncertainty modeling, multi-modality learning, incremental learning and continual learning these years, funded by National

Natural Science Foundation of China and The National Key Research and Development Program of China. He has published more than 300 peer-reviewed papers in IEEE TKDE, IEEE TPAMI, IEEE TNNLS, etc.

# Supplemental Material: BackMix: Regularizing Open Set Recognition by Removing Underlying Fore-Background Priors

Yu Wang, *Member, IEEE*, Junxian Mu, Hongzhi Huang, Qilong Wang, *Member, IEEE*, Pengfei Zhu, Qinghua Hu, *Senior Member, IEEE*

## I. THEORETICAL PROOFS

**Theorem 1.** *For model  $\mathcal{W}$  with the given properties, the mutual information maximization objective decomposes to  $I(y; \mathbf{z}_g) = I(y; \mathbf{z}_f) - I(y; \mathbf{z}_f | \mathbf{z}_g)$ , where maximizing  $I(y; \mathbf{z}_f)$  is the classification objective and minimizing  $I(y; \mathbf{z}_f | \mathbf{z}_g) \geq 0$  is a regularization term.*

*Proof.* We obtain the proof with data processing inequality. First of all, we have the following equations

$$\begin{aligned} I(y; \mathbf{z}_g, \mathbf{z}_f) &= I(y; \mathbf{z}_g | \mathbf{z}_f) + I(y; \mathbf{z}_f) \\ &= I(y; \mathbf{z}_f | \mathbf{z}_g) + I(y; \mathbf{z}_g). \end{aligned} \quad (1)$$

Given foreground feature  $\mathbf{z}_f$ , global feature  $\mathbf{z}_g$  depends only on background feature  $\mathbf{z}_b$  (Property 1). Also because background feature  $\mathbf{z}_b$  is independent from class label  $y$  (Property 2), we have  $I(y; \mathbf{z}_g | \mathbf{z}_f) = 0$ . Therefore, Eq. (1) can be re-organized to  $I(y; \mathbf{z}_g) = I(y; \mathbf{z}_f) - I(y; \mathbf{z}_f | \mathbf{z}_g)$ .  $\square$

**Theorem 2.** *The regularization term is optimized to zero if 1) constant value solution:  $\mathbf{z}_b$  is a constant value, or 2) orthogonal subspace solution:  $\mathbf{z}_f$  and  $\mathbf{z}_b$  are from different feature subspaces, i.e.,  $\mathbf{z}_g$  can be equivalently represented by the concatenation of  $\mathbf{z}_f$  and  $\mathbf{z}_b$ .*

*Proof.* (a) Let  $\mathbf{z}_b$  be the constant, which is no longer a random variable. Thus,  $I(y; \mathbf{z}_g) = I(y; \mathbf{z}_f + \mathbf{c}) = I(y; \mathbf{z}_f)$ , i.e., the extra regularization term is zero. (b) If  $\mathbf{z}_f$  and  $\mathbf{z}_b$  are from different feature subspaces, the probability density satisfies  $p(\mathbf{z}_g) = p(\mathbf{z}_f)p(\mathbf{z}_b)$ , then

$$\begin{aligned} p(\mathbf{z}_f | \mathbf{z}_g)p(y | \mathbf{z}_g) &= \frac{p(\mathbf{z}_f, \mathbf{z}_g)p(y, \mathbf{z}_g)}{p(\mathbf{z}_g)p(\mathbf{z}_g)} \\ &= \frac{[p(\mathbf{z}_g | \mathbf{z}_f)p(\mathbf{z}_f)]p(y, \mathbf{z}_f, \mathbf{z}_b)}{p(\mathbf{z}_f)p(\mathbf{z}_b)p(\mathbf{z}_g)} \end{aligned}$$

This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0116500, in part by the National Natural Science Foundation of China under Grants 62476195, U23B2049, 62436002, 62222608, and 62266035, in part by Tianjin Natural Science Funds for Distinguished Young Scholar under Grant 23JCJQJC00270, in part by Tianjin Young Scientific and Technological Talents Project under grant QN20230305, and in part by Tianjin Science and Technology Plan Project under grants 24YDTPJC00150 and 24JCYBJC00950.

Yu Wang, Junxian Mu, Hongzhi Huang, Qilong Wang, Pengfei Zhu, and Qinghua Hu are with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China, and also with the Haihe Laboratory of Information Technology Application Innovation (Haihe Lab of ITAI), Tianjin, China.

$$\begin{aligned} &= \frac{p(y, \mathbf{z}_f)p(\mathbf{z}_b)}{p(\mathbf{z}_g)}, \quad (2) \\ p(y, \mathbf{z}_f | \mathbf{z}_g) &= \frac{p(y, \mathbf{z}_f, \mathbf{z}_g)}{p(\mathbf{z}_g)} = \frac{p(y, \mathbf{z}_g | \mathbf{z}_f)p(\mathbf{z}_f)}{p(\mathbf{z}_g)} \\ &= \frac{p(y | \mathbf{z}_f)p(\mathbf{z}_b)p(\mathbf{z}_f)}{p(\mathbf{z}_g)} = \frac{p(y, \mathbf{z}_f)p(\mathbf{z}_b)}{p(\mathbf{z}_g)}. \quad (3) \end{aligned}$$

As Eq. (2) equals to Eq. (3), we have  $p(y, \mathbf{z}_f | \mathbf{z}_g) = p(\mathbf{z}_f | \mathbf{z}_g)p(y | \mathbf{z}_g)$ , implying random variables  $y$  and  $\mathbf{z}_f$  are independent conditioned on  $\mathbf{z}_g$ . Therefore, we have  $I(y; \mathbf{z}_f | \mathbf{z}_g) = 0$ .  $\square$

## II. VERIFICATION EXPERIMENTS DETAILS

### A. Experimental Setups for Synthesized Dataset

**Dataset construction.** We chose 18 classes from COCO [1] in total. Specifically, 12 classes were selected as known classes: airplane, bicycle, bird, boat, bottle, bus, car, dog, horse, TV, motorcycle, and person, while 6 classes were selected to serve as unknown classes in the test: sheep, cow, elephant, bench, toilet, and cat. For each class, we cropped at most 2,000 instances according to annotated bounding boxes. To preserve spaces for backgrounds, the crop boxes were given a margin of 100 pixels to the object bounding boxes. When cropping an instance from an image, we cropped the corresponding segmentation mask simultaneously to operate on foregrounds and backgrounds during training.

**Classifier Training.** For each dataset variant, we trained a ResNet18 [2] with a learning rate of 0.1, batch size of 128, and a cosine annealing learn rate scheduler. We applied a SGD optimizer with Nesterov momentum of 0.9 and weight decay of  $5e-4$  is adopted for training. And standard data augmentations are used, i.e., first resize the short side of the image to 256 then apply RandomHorizontalFlip and RandomCrop.

### B. Experiments for Outlier Exposure

**Setups.** In section III-B2 of main paper, we followed the experimental setups as well as the training framework in OE [3]. The known outlier dataset, TinyImages, is directly inherited from [3], which is obtained from the 80 Million Tiny Images [4] with images that appear in CIFAR10 removed. We adopted WideResNet40-4 [5] in these experiments with learn rate 0.1, batch size 128 and a cosine annealing learn rate scheduler. We applied a SGD optimizer with Nesterov

TABLE I  
COMPARISON OF USING DIFFERENT BACKGROUND SELECTION STRATEGIES IN THE TASK OF UNKNOWN DETECTION.

Methods	CIFAR10		CIFAR+50		Tiny-ImageNet	
	AUROC	Accuracy	AUROC	Accuracy	AUROC	Accuracy
Different	90.9	97.1	91.0	97.3	79.7	78.5
Random	91.3	97.5	91.6	97.9	80.4	79.2

momentum of 0.9 and weight decay of  $5e-4$  was adopted for training.

### III. SUPPLEMENTAL EXPERIMENTS

#### A. The Selection of Available Backgrounds

To investigate the impact of background selection within datasets on model performance, we tested the effects of using only images from different known classes (Different) as background regularization and randomly using any image within the batch (Random) as regularization on the CIFAR10 and Tiny-ImageNet datasets under the unknown detection settings.

Results in Table I indicate that using only images from different classes as background images for mixing has a relatively small impact on the model performance. The slight performance drop in the Different setting may be due to not mixing target images with the same label background images that are selected randomly within a batch. This also suggests that BackMix establishes the robustness in the selection of backgrounds.

#### B. BackMix with Different Score Functions

As highlighted in [10], using different score functions may offer alternative perspectives in handling spurious correlations. We extended our analysis by testing the proposed method with the baseline MSP (Plain\*) [6], prediction-based methods ODIN [9], Energy [8], and feature-based method Mahalanobis Distance [7]. We used CIFAR10 as the in-distribution (InD) dataset, with CIFAR100, SVHN, LSUN-Crop, and ImageNet-Crop as the out-of-distribution (OOD) datasets. We recorded

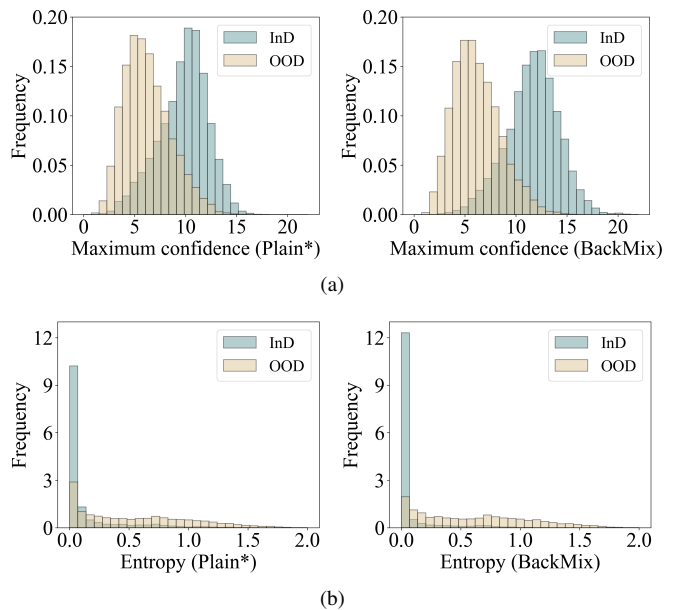


Fig. 1. Comparison of the (a) maximum confidence score and the (b) prediction probability entropy output by Plain\* and BackMix models on the in-distribution dataset CIFAR10 and the OOD dataset CIFAR100.

the DTACC and AUROC values of the baseline model under different score functions, with and without the BackMix.

**BackMix enables the model to learn the foreground object of known samples and reduces the uncertainty of predictions, further enhancing the effectiveness of prediction-based and feature-based score functions.** Results in Table II show that BackMix establishes stable performance improvement with different score functions. Fig. 1 illustrates the highest confidence of the baseline method (Plain\*) and BackMix output, along with the entropy of the prediction probabilities normalized by SoftMax. This indicates that BackMix can reduce the uncertainty of the model’s predictions, and thus improve the model’s discriminative ability for known and unknown samples. Consequently, it can further enhance the performance of prediction-based score functions.

Moreover, we observed that the Mahalanobis distance [7], which measures the distance of test samples from known

TABLE II  
DISTINGUISHING IN-DISTRIBUTION DATASET CIFAR10 FROM OOD DATASETS CIFAR100, SVHN, LSUN-CROP, AND IMAGENET-CROP WITH DIFFERENT SCORE FUNCTIONS UNDER VARIOUS METRICS.

Method	CIFAR100		SVHN		LSUN-Crop		Imagnet-Crop	
	DTACC	AUROC	DTACC	AUROC	DTACC	AUROC	DTACC	AUROC
Plain* [6]	79.8	86.3	86.4	90.6	87.9	93.7	88.6	94.5
+ BackMix	84.9 <sub>5.1</sub> ↑	91.3 <sub>5.0</sub> ↑	88.5 <sub>2.1</sub> ↑	94.1 <sub>3.5</sub> ↑	92.2 <sub>4.3</sub> ↑	96.9 <sub>3.2</sub> ↑	93.4 <sub>4.8</sub> ↑	97.8 <sub>3.3</sub> ↑
Mahalanobis [7]	79.4	86.1	95.4	98.7	98.9	99.7	98.9	99.9
+ BackMix	81.7 <sub>2.3</sub> ↑	88.1 <sub>2.0</sub> ↑	95.9 <sub>0.5</sub> ↑	99.3 <sub>0.6</sub> ↑	98.9 <sub>0.0</sub> –	99.7 <sub>0.0</sub> –	99.1 <sub>0.2</sub> ↑	99.9 <sub>0.0</sub> –
Energy [8]	80.6	86.6	86.7	90.4	88.3	94.7	88.9	95.1
+ BackMix	85.6 <sub>5.0</sub> ↑	92.2 <sub>5.6</sub> ↑	89.3 <sub>2.6</sub> ↑	94.8 <sub>3.4</sub> ↑	92.2 <sub>3.9</sub> ↑	97.2 <sub>2.5</sub> ↑	93.4 <sub>4.5</sub> ↑	98.1 <sub>3.0</sub> ↑
ODIN [9]	80.4	87.9	88.0	95.5	94.3	98.5	96.0	99.3
+ BackMix	85.0 <sub>4.6</sub> ↑	92.1 <sub>4.2</sub> ↑	95.2 <sub>7.2</sub> ↑	99.0 <sub>3.5</sub> ↑	96.4 <sub>2.1</sub> ↑	99.3 <sub>0.8</sub> ↑	97.7 <sub>1.7</sub> ↑	99.7 <sub>0.4</sub> ↑

TABLE III  
THE TRANSFERABILITY OF BACKBONES PRETRAINED USING BACKMIX AND OTHER DATA AUGMENTATION METHODS ON MULTIPLE VISUAL DOWNSTREAM TASKS WITH DIFFERENT METHODS.

Augmentation	Object Detection		Image Captioning	
	SSD [11] (mAP)	Faster-RCNN [12] (mAP)	NIC [13] (BLEU-1)	NIC [13] (BLEU-4)
Plain*	76.7	75.6	61.4	22.9
Mixup [14]	76.6 <sub>0.1</sub> ↓	73.9 <sub>1.7</sub> ↓	61.6 <sub>0.2</sub> ↑	23.2 <sub>0.3</sub> ↑
Cutout [15]	76.8 <sub>0.1</sub> ↑	75.0 <sub>0.6</sub> ↓	63.0 <sub>1.6</sub> ↑	24.0 <sub>1.1</sub> ↑
Cutmix [16]	77.6 <sub>0.9</sub> ↑	76.7 <sub>1.1</sub> ↑	64.2 <sub>2.8</sub> ↑	24.9 <sub>2.0</sub> ↑
BackMix	77.9 <sub>1.2</sub> ↑	77.1 <sub>1.5</sub> ↑	68.5 <sub>7.1</sub> ↑	25.6 <sub>2.7</sub> ↑

TABLE IV  
COMPARISON OF VARIOUS IMAGE CAPTIONING METRICS FOR BACKBONES PRETRAINED USING BACKMIX AND OTHER DATA AUGMENTATION METHODS ON THE COCO DATASET.

Augmentation	BLEU-2	BLEU-3	METEOR	ROUGE-L	CIDEr
Plain*	43.8	31.4	22.8	44.7	71.2
Mixup [14]	44.1 <sub>0.3</sub> ↑	31.6 <sub>0.2</sub> ↑	22.9 <sub>0.1</sub> ↑	47.9 <sub>3.2</sub> ↑	72.2 <sub>1.0</sub> ↑
Cutout [15]	45.3 <sub>1.5</sub> ↑	32.6 <sub>1.2</sub> ↑	22.6 <sub>0.2</sub> ↓	48.2 <sub>3.5</sub> ↑	74.1 <sub>2.9</sub> ↑
Cutmix [16]	46.3 <sub>2.5</sub> ↑	33.6 <sub>2.2</sub> ↑	23.1 <sub>0.3</sub> ↑	49.0 <sub>4.3</sub> ↑	77.6 <sub>6.4</sub> ↑
BackMix	50.6 <sub>6.8</sub> ↑	36.1 <sub>4.7</sub> ↑	23.1 <sub>0.3</sub> ↑	50.2 <sub>5.5</sub> ↑	80.6 <sub>9.4</sub> ↑

samples in feature space, performs well on far OOD datasets such as SVHN, but diminishes effectiveness in the near OOD dataset CIFAR100, where similar feature distributions limit its ability to separate samples. BackMix can also effectively enhance the Mahalanobis distance method, verifying its ability to help the model learn foreground features.

### C. The Transferability of BackMix

Following the Cutmix [16] setup, we applied BackMix on ResNet50 [2] for data augmentation during the ImageNet1K classification pretraining phase and then finetuned the trained model on the downstream visual tasks of object detection and image captioning.

According to [16], we finetuned the pretrained models on the object detection using SSD [11] and Faster-RCNN [12] algorithms. The Pascal VOC 2007 and 2012 [17] trainval was used as training data and we evaluated the model on the VOC 2007 test data using mean Average Precision (mAP) as the evaluation metric, which measures the average precision at different recall levels across all classes. We conducted image captioning experiments on the COCO dataset [1] using NIC [13] with the model pretrained with various data augmentation methods. The Bilingual Evaluation Understudy (BLEU) -n is used to evaluate the generated captions by comparing n-gram matches, where “n” represents the length of contiguous word sequences. We presented results in Table III, where the results of methods other than ours are taken from [16]. Compared to other methods, BackMix delivers greater performance improvements across different algorithms and downstream tasks.

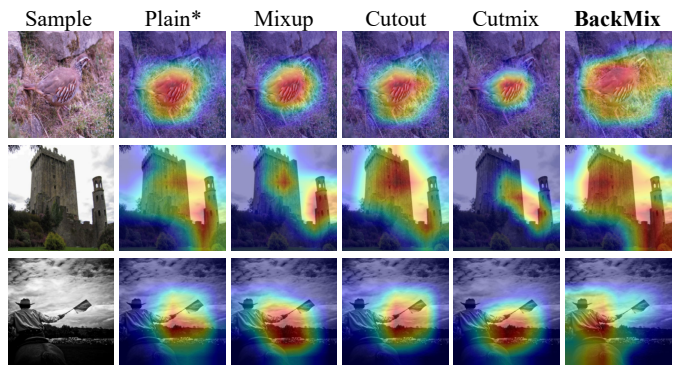


Fig. 2. Comparison of Grad-CAM results on the ImageNet1K dataset using different data augmentation methods.

We provided more metrics comparison on image captioning task in Table IV. METEOR evaluates the quality of natural language generation tasks. ROUGE-L is a metric used to evaluate the quality of generated text by measuring the longest common subsequence between the generated and reference text. CIDEr evaluates image captions by comparing how much the generated caption matches human-written reference captions. Results demonstrate that BackMix delivers significant performance improvements over the baseline methods across various metrics.

Figure 2 shows the Grad-CAM [18] of models trained using baseline strategy (Plain\*) and different data augmentation techniques on test images from the ImageNet1K dataset. Compared to other methods, BackMix exhibits more comprehensive attention and can accurately focus on the main body of the corresponding label class in scenes with multiple objects. As mentioned in representation learning methods [19], [20], a backbone with better feature extraction capabilities can perform better on downstream tasks. Therefore, the more noticeable performance gains of BackMix in various visual downstream tasks may be attributed to its enhancement of the model’s representation capabilities.

Additionally, we found that data augmentation methods with soft targets, like Mixup and Cutmix, tend to focus on limited regions, while methods with one-hot labels have relatively more comprehensive focus regions according to the main object. It is suggested that soft targets might enable the model to perform closed-set classification using a small number of highly discriminative features. However, it could lead to confusion when the model encounters unknown samples similar to known classes in open scenarios, as it may not be able to use comprehensive information for unknown detection [21].

### D. BackMix on Large-scale Pretrained Models

1) *Implementation description:* As a data augmentation method for OSR or OOD detection tasks, BackMix applies to various model architectures and can further enhance the OSR performance of large-scale pretrained models. Incorporating BackMix into both vision-only and vision-language pretrained models can be achieved as follows:



TABLE V

THE CLOSED-SET AND OPEN SET PERFORMANCE COMPARISON OF APPLYING BACKMIX TO MULTIPLE FINETUNING METHODS ON PRETRAINED MODELS FOR THE OUT-OF-DISTRIBUTION DETECTION TASK. CIFAR10 WAS USED AS THE IN-DISTRIBUTION DATASET, WHILE CIFAR100 AND SVHN WERE USED AS OUT-OF-DISTRIBUTION DATASETS.

Setting	Method	Accuracy	CIFAR100			SVHN		
			AUROC	OSCR	Macro-F1	AUROC	OSCR	Macro-F1
1-shot	CoOp	89.8	91.6	85.0	74.4	98.8	89.5	86.3
	+ BackMix	90.7 <sub>0.9</sub> ↑	92.1 <sub>0.5</sub> ↑	85.8 <sub>0.8</sub> ↑	75.4 <sub>1.0</sub> ↑	99.0 <sub>0.2</sub> ↑	90.3 <sub>0.8</sub> ↑	87.4 <sub>1.1</sub> ↑
	LoCoOp	89.6	91.2	84.6	74.2	98.5	89.0	85.1
	+ BackMix	90.7 <sub>1.1</sub> ↑	91.6 <sub>0.4</sub> ↑	85.2 <sub>0.6</sub> ↑	76.0 <sub>1.8</sub> ↑	99.0 <sub>0.5</sub> ↑	90.1 <sub>1.1</sub> ↑	88.0 <sub>2.9</sub> ↑
4-shot	CoOp	90.6	91.5	85.3	74.4	99.0	90.1	87.3
	+ BackMix	91.3 <sub>0.7</sub> ↑	92.1 <sub>0.6</sub> ↑	86.3 <sub>1.0</sub> ↑	76.0 <sub>1.6</sub> ↑	99.1 <sub>0.1</sub> ↑	90.8 <sub>0.7</sub> ↑	88.3 <sub>1.0</sub> ↑
	LoCoOp	89.8	91.4	84.6	73.8	98.6	89.2	84.8
	+ BackMix	90.9 <sub>1.1</sub> ↑	91.9 <sub>0.5</sub> ↑	85.7 <sub>1.1</sub> ↑	76.1 <sub>2.3</sub> ↑	99.0 <sub>0.4</sub> ↑	90.4 <sub>1.2</sub> ↑	87.9 <sub>3.1</sub> ↑
16-shot	CoOp	91.2	91.1	85.6	74.5	98.6	90.7	85.8
	+ BackMix	91.7 <sub>0.5</sub> ↑	91.6 <sub>0.5</sub> ↑	86.3 <sub>0.7</sub> ↑	75.8 <sub>1.3</sub> ↑	99.0 <sub>0.4</sub> ↑	91.3 <sub>0.6</sub> ↑	87.7 <sub>1.9</sub> ↑
	LoCoOp	91.4	90.4	85.5	71.9	93.7	87.6	79.8
	+ BackMix	91.7 <sub>0.3</sub> ↑	91.0 <sub>0.6</sub> ↑	86.3 <sub>0.8</sub> ↑	73.8 <sub>1.9</sub> ↑	96.1 <sub>2.4</sub> ↑	89.7 <sub>2.1</sub> ↑	82.5 <sub>2.7</sub> ↑
Full-data	CoOp	93.2	91.1	87.2	73.5	98.8	92.6	88.3
	+ BackMix	94.0 <sub>0.8</sub> ↑	92.7 <sub>1.6</sub> ↑	89.0 <sub>1.8</sub> ↑	77.7 <sub>4.2</sub> ↑	98.7 <sub>0.1</sub> ↓	93.4 <sub>0.8</sub> ↑	88.7 <sub>0.4</sub> ↑
	LoCoOp	94.3	93.1	89.6	77.8	95.3	91.2	84.1
	+ BackMix	94.6 <sub>0.3</sub> ↑	93.3 <sub>0.2</sub> ↑	90.0 <sub>0.4</sub> ↑	78.7 <sub>0.9</sub> ↑	97.3 <sub>2.0</sub> ↑	93.3 <sub>2.1</sub> ↑	87.0 <sub>2.9</sub> ↑
	VPT	96.2	95.0	92.4	82.5	97.5	94.4	84.6
	+ BackMix	96.5 <sub>0.3</sub> ↑	95.5 <sub>0.5</sub> ↑	93.0 <sub>0.6</sub> ↑	83.8 <sub>1.3</sub> ↑	98.3 <sub>0.8</sub> ↑	95.0 <sub>0.6</sub> ↑	85.6 <sub>1.0</sub> ↑

- **Vision-only model.** The method of obtaining attention maps in the main text applies to other more complex models in the ResNet [2] series. Considering the difference in architecture, we adopted a similar way like Grad-CAM [18] of obtaining activation maps for models using vision transformer (ViT) [22] as the backbone. Grad-CAM identifies influential input regions by backpropagating gradients from the prediction layer, generating a heatmap highlighting areas crucial to the prediction.
- **Vision-language model.** Due to the unique ability to integrate and process both visual and textual information, pre-trained vision-language models (VLMs) have established superior performance on classification tasks. CLIP [23] is one of the most representative VLMs, which constrains the model only output high similarity on matched text and image features. Similar to the vision-only model, we capture attention weights contributing to model predictions by backpropagating class-specific gradients through the attention layers. The final attention map created by accumulating and normalizing these contributions shows the regions most influential to the predictions.

Thus, BackMix can be flexibly applied to mainstream large-scale pretrained models and corresponding finetuning methods.

2) *Experiments:* To validate the effectiveness of our approach on large-scale pretrained models, we conducted experiments on the pure ViT-based finetuning method VPT [24], as well as the CLIP-based finetuning methods CoOp [25] and LoCoOp [26]. VPT introduces prompts within visual models to adapt them without altering backbone weights.

CoOp finetunes CLIP by using learnable prompts and LoCoOp further uses local features for OOD regularization.

We evaluated the CoOp and LoCoOp across 1-shot, 4-shot, 16-shot and full-data settings. Considering that VPT requires training a classification head, we evaluated its performance in the full-data setting. CIFAR10 served as the InD dataset, and CIFAR100 and SVHN as OOD datasets. We adopted the ViT-B/16 model trained by [23] as the backbone for all methods and input both original and processed data for BackMix.

**BackMix consistently improves the performance of pre-trained model across data scales, proving its strength in the finetuning stage.** Results in Table V indicate that BackMix can further enhance the pretrained model’s closed-set classification and unknown detection capabilities across different scales of training data, even in the scenario where only one training sample from each class is available. Benefiting from the powerful feature extraction capabilities of the pretrained models, BackMix can more accurately distinguish between the background and foreground of images, thereby achieving superior performance.

### E. BackMix on Mitigating Spurious Correlations

Following the settings in [10], we combined various OOD detection methods with BackMix and test these methods on the Waterbirds [27] dataset with different spurious correlation  $r$  to evaluate the performance of BackMix on mitigating impacts of spurious correlation. As proposed in [10], the correlation  $r$

TABLE VI  
COMPARISON OF THE BASELINE AND DIFFERENT SCORE FUNCTIONS COMBINED WITH BACKMIX ON THE WATERBIRDS DATASET WITH VARYING DEGREES OF SPURIOUS FORE-BACKGROUND CORRELATION (CORR.).

Method	Corr.	Spurious OOD		SVHN		iSUN		LSUN	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Plain*		55.7	90.2	12.7	96.6	19.0	95.2	16.1	94.3
+ BackMix		55.3 <sub>0.4↓</sub>	90.8 <sub>0.6↑</sub>	12.3 <sub>0.4↓</sub>	97.1 <sub>0.5↑</sub>	18.3 <sub>0.7↓</sub>	95.5 <sub>0.3↑</sub>	15.8 <sub>0.3↓</sub>	94.8 <sub>0.5↑</sub>
Mahalanobis	$r = 0.5$	56.1	86.3	0.5	100.0	0.6	99.9	0.6	100.0
+ BackMix		55.7 <sub>0.4↓</sub>	87.0 <sub>0.7↑</sub>	0.2 <sub>0.3↓</sub>	100.0 <sub>0.0-</sub>	0.2 <sub>0.4↓</sub>	100.0 <sub>0.1↑</sub>	0.3 <sub>0.3↓</sub>	100.0 <sub>0.0-</sub>
Energy		54.9	90.5	6.5	98.9	10.2	97.5	12.6	97.2
+ BackMix		54.6 <sub>0.3↓</sub>	91.0 <sub>0.5↑</sub>	6.4 <sub>0.1↓</sub>	99.0 <sub>0.1↑</sub>	10.0 <sub>0.2↓</sub>	97.5 <sub>0.0-</sub>	12.3 <sub>0.3↓</sub>	97.4 <sub>0.2↑</sub>
ODIN		54.3	90.5	6.1	99.1	10.1	97.5	12.7	96.9
+ BackMix		54.3 <sub>0.0-</sub>	90.7 <sub>0.2↑</sub>	6.0 <sub>0.1↓</sub>	99.3 <sub>0.2↑</sub>	10.1 <sub>0.0-</sub>	98.0 <sub>0.5↑</sub>	12.4 <sub>0.3↓</sub>	97.3 <sub>0.4↑</sub>
Plain*		73.0	81.3	37.4	94.8	46.7	91.6	46.8	91.6
+ BackMix		71.5 <sub>1.5↓</sub>	85.1 <sub>3.8↑</sub>	35.8 <sub>1.6↓</sub>	95.8 <sub>1.0↑</sub>	43.3 <sub>3.4↓</sub>	92.7 <sub>1.1↑</sub>	44.0 <sub>2.8↓</sub>	93.1 <sub>1.5↑</sub>
Mahalanobis	$r = 0.7$	71.8	82.9	0.5	99.9	0.9	99.8	1.1	99.8
+ BackMix		71.3 <sub>0.5↓</sub>	83.6 <sub>0.7↑</sub>	0.2 <sub>0.3↓</sub>	100.0 <sub>0.1↑</sub>	0.6 <sub>0.3↓</sub>	99.9 <sub>0.1↑</sub>	0.7 <sub>0.4↓</sub>	100.0 <sub>0.2↑</sub>
Energy		72.5	84.6	36.9	95.4	41.9	91.9	40.4	92.4
+ BackMix		71.3 <sub>1.2↓</sub>	86.1 <sub>1.5↑</sub>	35.3 <sub>1.6↓</sub>	96.1 <sub>0.7↑</sub>	39.5 <sub>2.4↓</sub>	93.3 <sub>1.4↑</sub>	38.6 <sub>1.8↓</sub>	93.7 <sub>1.3↑</sub>
ODIN		73.6	81.3	36.4	95.5	42.1	92.0	40.9	92.3
+ BackMix		71.9 <sub>1.7↓</sub>	83.8 <sub>2.5↑</sub>	35.3 <sub>1.1↓</sub>	96.2 <sub>0.7↑</sub>	39.8 <sub>2.3↓</sub>	93.1 <sub>1.1↑</sub>	39.1 <sub>1.8↓</sub>	93.5 <sub>1.2↑</sub>
Plain*		86.1	74.4	44.7	92.3	51.4	89.6	49.5	90.1
+ BackMix		81.5 <sub>4.6↓</sub>	80.3 <sub>5.9↑</sub>	42.6 <sub>2.1↓</sub>	93.2 <sub>0.9↑</sub>	49.5 <sub>1.9↓</sub>	91.0 <sub>1.4↑</sub>	46.2 <sub>3.2↓</sub>	91.9 <sub>1.8↑</sub>
Mahalanobis	$r = 0.9$	79.5	76.3	0.9	99.8	1.1	99.6	1.9	99.5
+ BackMix		76.2 <sub>3.3↓</sub>	80.5 <sub>4.2↑</sub>	0.4 <sub>0.5↓</sub>	100.0 <sub>0.2↑</sub>	0.7 <sub>0.4↓</sub>	99.9 <sub>0.3↑</sub>	0.8 <sub>1.1↓</sub>	99.9 <sub>0.4↑</sub>
Energy		84.5	75.2	44.3	92.5	50.9	90.0	49.7	89.8
+ BackMix		81.7 <sub>2.8↓</sub>	80.6 <sub>5.4↑</sub>	42.4 <sub>1.9↓</sub>	93.8 <sub>1.3↑</sub>	49.4 <sub>1.5↓</sub>	91.1 <sub>1.1↑</sub>	46.6 <sub>3.1↓</sub>	91.5 <sub>1.7↑</sub>
ODIN		84.8	74.8	44.6	92.6	50.8	90.0	49.2	90.4
+ BackMix		81.6 <sub>3.2↓</sub>	79.9 <sub>5.1↑</sub>	42.5 <sub>2.1↓</sub>	94.1 <sub>1.5↑</sub>	49.8 <sub>1.0↓</sub>	91.2 <sub>1.2↑</sub>	46.0 <sub>3.2↓</sub>	91.6 <sub>1.2↑</sub>

is defined as:

$$\begin{aligned}
 r &= P(e = \text{water} | y = \text{waterbirds}) \\
 &= P(e = \text{land} | y = \text{landbirds}), \quad (4)
 \end{aligned}$$

where  $e$  denotes the label of environment and  $y$  denotes the label of foreground object. For this dataset,  $r=0.5$  indicates that classes appear uniformly across different backgrounds, reflecting a low level of spurious correlation, while  $r=0.9$  means that classes almost exclusively appear in strongly correlated backgrounds, indicating a high level of spurious correlation. We use a subset of images of land and water from the Places dataset as the ‘Spurious OOD’ dataset, while using SVHN, iSUN, and LSUN as non-spurious datasets. FPR95 and AUROC are adopted as evaluation metrics. FPR95 measures the false positive rate when 95% of the known samples are correctly accepted.

#### BackMix can mitigate impacts of spurious correlations.

When  $r=0.7$  and  $r=0.9$ , spurious correlations exist in the data. Results in Table VI indicates that BackMix significantly reduces the proportion of misclassified unknown samples with spurious correlations and improves the overall performance of the model. Particularly for Spurious OOD, BackMix significantly enhances the model’s detection capability. This

indicates that the fore-background decoupling of BackMix can effectively mitigate impacts of inherent associations in the dataset. Moreover, BackMix can be integrated with various OOD detection methods to achieve better performance.

**Mitigating impacts of spurious correlations between foreground and background can improve the model’s performance on OSR and OOD detection tasks.** When  $r=0.5$ , the model performs significantly better than when  $r=0.9$ . By removing foreground-background correlations, the model is better able to focus on the primary classification subject, thus avoiding misclassification caused by unknown data with spurious correlations. This conclusion also validates the effectiveness of BackMix in removing fore-background correlations in the training set for OSR and OOD detection tasks.

#### F. More visualizations

We present more images processed with BackMix in Fig. 3. With the progressively optimized foreground segmentation, the model effectively masks foreground objects in the BI, successfully avoiding the introduction of multiple object classes in a single image after being processed by BackMix.



Fig. 3. Samples processed with BackMix on the ImageNet30 dataset.

TABLE VII  
COMPARISON OF THE KEY CHARACTERISTICS OF REPRESENTATIVE STUDIES IN SPURIOUS CORRELATIONS AND BACKMIX.

Study	Impacts of Fore-Background Correlation on OSR / OOD Detection	Address Fore-Background Correlations	Applicability
Spurious OOD [10]	1. Analyze impacts on elaborately constructed datasets 2. Feature-based score function mitigates negative impacts	✗	Limited experimental scenarios
Existing Methods [28]–[31]	✗	Focus on foreground feature mainly	Tasks with spurious correlation priors using additional annotations
<b>Ours</b>	1. Analyze impacts on common datasets 2. Analyze impacts of multiple fore-background processing cases 3. Analyze background regularization	1. Remove fore-background priors with no additional information 2. Use foreground to predict 3. Use background as available outliers	✓

#### IV. FURTHER DISCUSSION

##### A. Connection between BackMix and Existing Methods

1) *Connections to OSR methods:* Most existing OSR methods jointly model the foregrounds and backgrounds, thereby suffering from failing to identify the test samples that are partially known, *i.e.*, varying foregrounds or backgrounds. OE methods select an auxiliary dataset that serves as a regularizer to alleviate such a problem, but the dataset requires careful design, which is seldom feasible in real-world applications. Theoretically, we find that the proposed BackMix that removes fore-background priors and uses a GAP regularizer works similarly to OE but does not require any auxiliary data.

2) *Connections to spurious correlation mitigation methods:* To better illustrate the contributions of BackMix relative to existing methods, we compare key characteristics of the work by Ming *et al.* [10] (Spurious OOD), some representative studies in spurious correlations [28]–[31] and BackMix in Table VII. It evaluates whether each study explores impacts of fore-background correlations on OSR / OOD detection tasks, the way of addressing fore-background correlations, and the applicability to general scenarios. Unlike existing works, our study deeply explore impacts of fore-background priors and impacts of different types of foregrounds and backgrounds in OSR / OOD detection tasks, respectively. More importantly, we proposed a new method BackMix that uses backgrounds as outlier regularizations with no additional information. We proved that the simple BackMix is equivalent to the OE [3] method which elaborately uses auxiliary outliers.






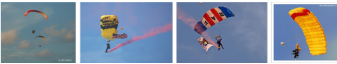

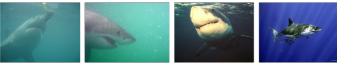


**Impacts of fore-background correlations on OSR / OOD detection.** Ming *et al.* [10] (Row 1) conducted experiments on elaborately constructed datasets with spurious correlations

and found that fore-background correlations negatively impact model performance. They experimentally concluded that feature-based score function [7] in OOD detection can mitigate negative impacts. In contrast, we tested models directly on standard datasets and analyzed how fore-background correlations affect model performance in general cases. Furthermore, we explored ways to enhance model performance while mitigating fore-background correlations, including cases where there are multiple foregrounds and no background in the image.. Based on theoretical and experimental analysis, we proposed a strategy using backgrounds as available outliers for regularization.

**Address fore-background correlations.** Existing methods [28]–[31] (Row 2) mitigate impacts of fore-background correlations on the performance by enabling the model learn foreground more accurately and comprehensively. However, most of these methods do not directly address negative impacts of fore-background correlations on OSR / OOD detection. Based on our experiments, we found that using background for regularization can improve the performance of model in OSR / OOD detection by treating backgrounds as available outliers and mitigating negative impacts of fore-background correlations.

**Applicability.** Ming *et al.* [10] found that Mahalanobis method [7] is relatively effective on data with spurious correlations and far OOD data. Results in Table II indicate that Mahalanobis method [7] still suffer due to the inherent prior correlations between foreground and background features, leading to poor performance in handling complex problems such as near OOD data (CIFAR100). Other methods on addressing spurious correlations require additional information [28], [29], [31],

TABLE VIII  
INFORMATION OF 10 CLASSES WITH STRONG FORE-BACKGROUND CORRELATIONS SELECTED FROM THE IMAGENET1K.

Foreground	Background	Images
Hare	Grassland	
Hot pot	Cooker	
Taxicab	City road	
Cardoon	Leaves	
Dog sled	Snow	
Parachute	Sky	
Container ship	Sea surface	
Great white shark	Ocean	
Entertainment center Indoor		
Lizard	Sand	

model components [31], and training objectives [28], [30], [31]. Compared with the aforementioned methods, BackMix can serve as a general data augmentation method, seamlessly integrating with existing methods to improve OSR / OOD detection performance, thus offering greater applicability.

Thus, BackMix provides a simpler and more effective method, with significant advantages in its easy deployment and specific alignment with OSR, enhancing robustness against both fore-background correlations and OOD samples.

### B. BackMix with Outlier Exposure

We consider applying BackMix on Outlier Exposure settings, where TIs are taken from known classes, and BIs are taken from auxiliary outliers. The results show that BackMix improves the plain classifier but is significantly inferior to OE methods and the two GAP regularizers. The gap lies in that BackMix uses only a small patch of backgrounds. While both OE methods and our GAP regularizers use the full outlier image. We infer that the foreground semantics in outliers may be of greater help than plain backgrounds. In other words, the richer information background images provide, the better the regularization of a classifier. In future work, it is still worth exploring and fully exploiting the regularization power hidden in image backgrounds.

### C. Limitations

As the BackMix uses CAM to determine the background regions and then applies background regularization to mitigate

TABLE IX  
COMPARISON OF THE OPEN SET RECOGNITION PERFORMANCE OF THE MODEL ON THE CONSTRUCTED STRONG FORE-BACKGROUND CORRELATION DATASET.

Method	4-shot		16-shot		Full-data	
	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC
Plain*	53.2	64.6	67.2	65.9	96.8	90.0
+ BackMix	53.2 <sub>0.0</sub> ↓	65.2 <sub>0.6</sub> ↑	67.6 <sub>0.4</sub> ↑	66.8 <sub>0.9</sub> ↑	98.0 <sub>1.2</sub> ↑	91.6 <sub>1.6</sub> ↑

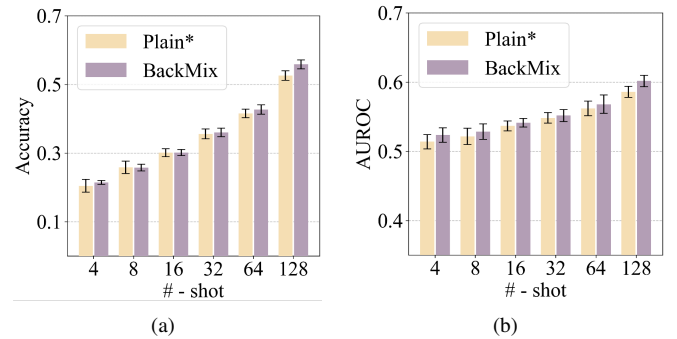


Fig. 4. Comparison of the (a) closed-set accuracy and (b) AUROC performance of the Plain\* and the BackMix with varying numbers of the training samples from each known class. We conducted experiments under each setting with five different random seeds and reported the average results.

the influence of fore-background priors in the training set, the performance improvement of BackMix is limited when each foreground only appears in specific background without overlapping, and the background regions extracted by CAM are inaccurate in few-shot learning from scratch.

**Each foreground only appears in a specific background without overlapping.** When each class in a dataset appears only in a fixed background, the performance improvement from BackMix is limited. In such cases, the training set’s fore-background priors also apply during test.

To simulate this rare scenario, we selected 10 classes from the ImageNet dataset, each with a fixed, non-overlapping background as shown in Table VIII. Five classes were designated as known, and the remaining five as unknown during test. Table IX presents the closed-set classification accuracy and AUROC performance under few-shot and full data settings.

With strong fore-background correlations and no overlap, backgrounds can effectively aid classification. In few-shot settings, the model can learn the background from limited data. As the sample size increases, the improvements become more significant. However, due to the distinct fore-background differences in the dataset, the performance of the baseline model is already impressive, and improvements brought by BackMix are relatively limited. Note that BackMix only provides limited performance gains and will not affect the capability of the original method in this case.

Practically, known and unknown classes are likely to appear in similar scenes or belong to classes without strong fore-background correlations. Therefore, BackMix is effective in real open scenarios.

**Few-shot learning from scratch.** Since BackMix relies on

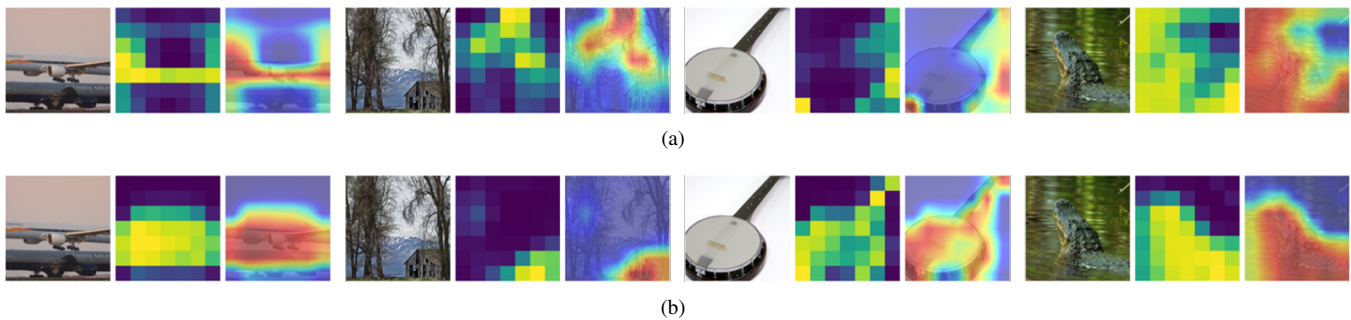


Fig. 5. Foreground segmentation masks learned by the model on the ImageNet30 dataset. Brighter areas indicate higher activation intensity, (a) under the 4-shot training setting, (b) under the 64-shot training setting.

class activation maps (CAMs) to segment and mask the foreground, limited model capacity can cause multiple foregrounds to appear in a single image. As we have verified in experiments (See Section III in the main paper) when multiple foreground objects appear in the image and one-hot labels are used, the performance is weaker than those with only one foreground object.

Thus, in scenarios with few samples and no pretraining, it is difficult for the model to capture foreground regions and it may even use the image background for classification. We used the CIFAR10 dataset as the InD dataset, varying the number of training samples per class from 4 to 128, to test the model’s closed-set classification accuracy and unknown detection capability. Figure 4 shows that BackMix can only bring a slight or even no improvement when the number of training samples is minimal. As the number of samples increases, the gains of BackMix in both closed-set and open-set performance become more pronounced. In Figure 5, we show the foreground segmentation of some images learned by the model when there are 4 training samples and 64 training samples per class, respectively. When the number of samples is too small, the learned segmentation of foreground objects is inaccurate, which results in multiple classified foregrounds in processed images and further confuses the model.

This limitation can be addressed by finetuning a pretrained model that has strong representation capacity with a small amount of downstream task data.

## V. DETAILS OF THE DATASETS

- **COCO [1]**. The COCO dataset provides 80 classes, more than 330,000 images, 200,000 tags, and more than 150, ten thousand people. It includes 91 objects, 328,000 images, and 250,000 labels and is widely used in target detection and segmentation. With its precise segmentation, we use it in Section III-A of the main paper for verifying unbalanced backgrounds may mislead the open set classifier. In Section III-C of the supplemental material, we use it for testing the transferability of models trained with different data augmentation techniques on the image captioning task.
- **iNaturalist [32]**. iNaturalist is a natural image dataset from the real world. It includes 8142 fine-grained classes, covering plants, birds, insects, and other natural organisms. The sample distribution is different from the com-

mon data set used for classification, showing an extreme long-tailed distribution. Take advantage of its real-world features and use it in Section III-A of the main paper as an unknown class in the test phase.

- **CIFAR10 [33]**. It includes 60,000 RGB images with size of  $32 \times 32$ , among which 50,000 are divided into train set and the remaining 10,000 make up test set. This small-scale dataset includes ten common classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck, and each class has 6,000 images. It is used in Section III-B2, Section V-A, Section V-B1, Section V-B2, Section V-C1, Section V-C2 of the main paper and used in Section III-A, Section III-B, Section III-D, Section IV-C of the supplemental material.
- **LSUN-Crop/-Resize/-Fix**. The large-scale Scene Understanding dataset [34] includes images from ten different scenes, *e.g.* the kitchen, living room, and bedroom. It only has a test set of 10,000 images and can be reconstructed by cropping and resizing to LSUN-Crop and LSUN-Resize. LSUN-Fix is constructed by combining random sampling and resizing. They are regarded as unknown sets in Section III-B2, Section V-A2, Section V-B2 of the main paper and Section III-B, Section III-E of the supplemental material.
- **ImageNet-Crop/-Resize/-Fix**. The ImageNet-Crop and ImageNet-Resize both have 10000 RGB images with the size of  $32 \times 32$ . Liang *et al.* [9] constructed them by cropping or downsampling the images from a subset of ImageNet [35]. The ImageNet-Fix is from [36], which is constructed by randomly sampling and resizing the images in ImageNet. We use them in Section III-B2, Section V-A2, Section V-B2 of the main paper and Section III-B of the supplemental material.
- **TinyImage [4]**. It consists of nearly 80 million tiny RGB images of size  $32 \times 32$  collected from the web. Images in TinyImage are labeled with one of the 75,062 non-abstract nouns and are mainly used for image classification tasks. It is used as the outlier after removing its overlap with CIFAR10 in Section III-B2.
- **CIFAR100 [33]**. CIFAR100 is a hierarchical dataset, which consists of RGB images size of  $32 \times 32$ . Its 60,000 images are divided into 100 classes, which are divided into 20 super classes, each with 500 training images and 100 test images. We use it as outlier in Section III-B2

and seen it as near OOD of CIFAR10 for OOD detection in Section V-A3 of the main paper and Section III-B, Section III-D, Section IV-C of the supplemental material.

- **DTD [37]**. The Describable Textures Dataset is a texture dataset consisting of 5,640 images across 47 classes, with each class containing 120 images. The dataset serves as a comprehensive resource for texture classification and recognition algorithms. We use it as an auxiliary outlier dataset that has limited semantic information in Section III-B2.
- **Flower102 [38]**. The Oxford 102 Flower dataset comprises 102 classes of flowers. Each class contains between 40 and 258 images. We use it as an auxiliary outlier dataset that has limited semantic information in Section III-B2.
- **SVHN [39]**. The Street View House Number Dataset is derived from Google Street View house number and has images of digits 0-9. Its training set contains 73,257 digit images, and its test set contains 26,032 digit images. We tested the unknown detection capability of the model in this dataset in Section V-A1. It is used as the far OOD dataset of CIFAR10 in Section V-A3 in main paper and used as OOD dataset in Section III-B, Section III-D, and Section III-E of the supplemental material.
- **CIFAR+10**. The fixed openness of the model is also restricted due to the fixed data of known and unknown classes in Section V-A1 experiments on CIFAR10. Following the protocol in [40], we select four classes in CIFAR10 as known classes and select 10 classes from CIFAR100 as unknown classes. To avoid possible overlap of classes, we only select non-animal classes in CIFAR10 while only selecting from animal classes in CIFAR100.
- **CIFAR+50**. Like CIFAR+10, CIFAR+50 uses 50 animal classes from CIFAR100 as unknown classes. As the number of unknown classes further increases, the openness correspondingly increases and becomes more challenging in Section V-A1 experiments in main paper. We also use this setting in Section III-A of the supplemental material.
- **Tiny-ImageNet [41]**. Tiny ImageNet contains 100,000 RGB images of size 64×64. It is divided into 200 classes, each with 500 training images, 50 validation images, and 50 test images. Open set tasks also pose a greater challenge because of the richer classes available. In Section V-A1 of the main paper and Section III-A of the supplemental material, we randomly select 20 classes as known classes and the remaining 180 classes as unknown classes.
- **ImageNet30 [42]**. ImageNet30 is composed of 30 low overlapped class images selected from ImageNet [35]. Each class has 1300 training images and 100 test images. In our Section V-C1, Section V-C2 experiments, we take the first 10 classes as known classes and the last 20 as unknown classes in the dictionary order.
- **ImageNet1K [35]**. The ImageNet1K dataset comprises 1000 classes with a total of approximately 1.2 million labeled images. It serves as a standard benchmark for image classification tasks and has been crucial in the development and evaluation of deep learning models. In

Section III-C of the supplemental material, we use all training samples for training the visual backbone.

- **Pascal VOC [17]**. The Visual Object Classes dataset is a significant benchmark in computer vision, primarily used for object detection and image classification. It includes a variety of object classes, such as animals, vehicles, and household items, and provides detailed annotations like bounding boxes and class labels. In Section III-C of the supplemental material, we use this dataset for testing the performance of model on the object detection task.
- **WaterBirds [27]**. The Waterbirds dataset is a synthetic dataset that commonly used to study the problem of spurious correlations. It uses bird images and corresponding segmentation annotations from the CUB [43] dataset, combined with background images from Places365 [44] dataset, to create two classes: water birds and land birds. The CUB dataset contains 200 classes of birds and the Places365 dataset composes of 434 scene classes. It introduces a correlation between the type of bird and the background (water or land), which can mislead models to rely on the background for classification instead of focusing on the bird itself. In Section III-E of the supplemental material, we use this dataset for testing the performance of model against the spurious correlations in the training stage.
- **iSUN [45]**. The iSUN dataset contains a rich variety of natural scene images. In Section III-E of the supplemental material, we use this dataset as the OOD dataset.

## REFERENCES

- [1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.
- [3] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *Int. Conf. Learn. Represent.*, 2019.
- [4] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, 2008.
- [5] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Brit. Mach. Vis. Conf.*, 2016, pp. 1–15.
- [6] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Int. Conf. Learn. Represent.*, 2017, pp. 1–12.
- [7] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Adv. Neural Inform. Process. Syst.*, vol. 31, 2018, pp. 7167–7177.
- [8] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Adv. Neural Inform. Process. Syst.*, 2020, pp. 21464–21475.
- [9] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *Int. Conf. Learn. Represent.*, 2018.
- [10] Y. Ming, H. Yin, and Y. Li, "On the impact of spurious correlation for out-of-distribution detection," in *AAAI conf. on artificial intelligence*, vol. 36, no. 9, 2022, pp. 10051–10059.
- [11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Eur. Conf. Comput. Vis.* Springer, 2016, pp. 21–37.
- [12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2016.
- [13] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3156–3164.

- [14] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” in *Int. Conf. Learn. Represent.*, 2017.
- [15] T. DeVries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout,” *arXiv preprint arXiv:1708.04552*, 2017.
- [16] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, “Cutmix: Regularization strategy to train strong classifiers with localizable features,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 6023–6032.
- [17] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2010.
- [18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
- [19] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 9650–9660.
- [20] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, “Dinov2: Learning robust visual features without supervision,” *Transactions on Machine Learning Research Journal*, pp. 1–31, 2024.
- [21] W. Moon, J. Park, H. S. Seong, C.-H. Cho, and J.-P. Heo, “Difficulty-aware simulator for open set recognition,” in *Eur. Conf. Comput. Vis.*, 2022, pp. 365–381.
- [22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Int. Conf. Learn. Represent.*, 2021.
- [23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [24] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, “Visual prompt tuning,” in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 709–727.
- [25] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [26] A. Miyai, Q. Yu, G. Irie, and K. Aizawa, “Locoop: few-shot out-of-distribution detection via prompt learning,” in *Adv. Neural Inform. Process. Syst.*, 2023, pp. 76 298–76 310.
- [27] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, “Distributionally robust neural networks,” in *Int. Conf. Learn. Represent.*, 2019.
- [28] E. Creager, J.-H. Jacobsen, and R. Zemel, “Environment inference for invariant learning,” in *Int. Conf. Mach. Learn.*, 2021, pp. 2189–2200.
- [29] H. Yao, Y. Wang, S. Li, L. Zhang, W. Liang, J. Zou, and C. Finn, “Improving out-of-distribution robustness via selective augmentation,” in *Int. Conf. Mach. Learn.*, 2022, pp. 25 407–25 437.
- [30] S. Asgari, A. Khani, F. Khani, A. Gholami, L. Tran, A. Mahdavi Amiri, and G. Hamarneh, “Masktune: Mitigating spurious correlations by forcing to explore,” in *Adv. Neural Inform. Process. Syst.*, vol. 35, 2022, pp. 23 284–23 296.
- [31] S. Liu, X. Zhang, N. Sekhar, Y. Wu, P. Singhal, and C. Fernandez-Granda, “Avoiding spurious correlations via logit correction,” in *Int. Conf. Learn. Represent.*, 2023.
- [32] G. V. Horn, O. M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. J. Belongie, “The inaturalist species classification and detection dataset,” *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8769–8778, 2018.
- [33] A. Krizhevsky, “Learning multiple layers of features from tiny images,” *Tech Report*, 2009.
- [34] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao, “Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop,” *arXiv preprint arXiv:1506.03365*, 2015.
- [35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.
- [36] J. Tack, S. Mo, J. Jeong, and J. Shin, “Csi: Novelty detection via contrastive learning on distributionally shifted instances,” in *Adv. Neural Inform. Process. Syst.*, vol. 33, 2020, pp. 11 839–11 852.
- [37] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi, “Describing textures in the wild,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014.
- [38] M.-E. Nilsback and A. Zisserman, “Automated flower classification over a large number of classes,” *Indian Conf. Comput. Vis., Graph. Imag. Process.*, pp. 722–729, 2008.
- [39] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” in *Adv. Neural Inform. Process. Syst.*, 2011, pp. 1–9.
- [40] L. Neal, M. L. Olson, X. Z. Fern, W.-K. Wong, and F. Li, “Open set learning with counterfactual images,” in *Eur. Conf. Comput. Vis.*, 2018, pp. 620–635.
- [41] H. Pouransari and S. Ghili, “Tiny imagenet visual recognition challenge,” *CS 231N*, 2014.
- [42] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, “Using self-supervised learning can improve model robustness and uncertainty,” in *Adv. Neural Inform. Process. Syst.*, vol. 32, 2019, pp. 15 663–15 674.
- [43] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” 2011.
- [44] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, 2017.
- [45] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao, “Turkergaze: Crowdsourcing saliency with webcam based eye tracking,” *arXiv preprint arXiv:1504.06755*, 2015.