Revealing the Two Sides of Data Augmentation: An Asymmetric Distillation-based Win-Win Solution for Open-Set Recognition

Yunbing Jia[1,2], **Xiaoyu Kong**[3], **Fan Tang**[2], **Yixing Gao**[1,4*], **Weiming Dong**[5], **Yi Yang** [6]
[1]School of Artificial Intelligence, Jilin University, China
[2]Institute of Computing Technology, Chinese Academy of Sciences, China
[3]Harbin Institute of Technology (Shenzhen), China
[4]Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, Ministry of Education, China
[5]MAIS, Institute of Automation, Chinese Academy of Sciences, China
[6]Didi Chuxing, China
jiayb22@mails.jlu.edu.cn, 21S151102@stu.hit.edu.cn, tfan.108@gmail.com, gaoyixing@jlu.edu.cn, weiming.dong@ia.ac.cn, yangyiian@didiglobal.com

Abstract

In this paper, we reveal the two sides of data augmentation: enhancements in closed-set recognition correlate with a significant decrease in open-set recognition. Through empirical investigation, we find that multi-sample-based augmentations would contribute to reducing feature discrimination, thereby diminishing the open-set criteria. Although knowledge distillation could impair the feature via imitation, the mixed feature with ambiguous semantics hinders the distillation. To this end, we propose an asymmetric distillation framework by feeding the teacher model extra raw data to enlarge the benefit of the teacher. Moreover, a joint mutual information loss and a selective relabel strategy are utilized to alleviate the influence of hard mixed samples. Our method successfully mitigates the decline in open-set and outperforms SOTAs by 2% $\sim$ 3% AUROC on the Tiny-ImageNet dataset, and experiments on large-scale dataset ImageNet-21K demonstrate the generalization of our method.

1 Introduction

The utilization of data augmentation (DA) strategies in training neural networks have been proven effective in expanding the training dataset [Yang *et al.*, 2022] and have become widespread in many applications [Chen *et al.*, 2021a; Xu *et al.*, 2022; Chen *et al.*, 2023; Hou *et al.*, 2024; Wang *et al.*, 2024]. As the simplest implementation, the base manipulation-based DA is the most common strategy and can be divided into two categories: single-sample-based augmentation (SSA) and multiple-sample-based augmentation (MSA). SSA creates new samples by conducting basic operations on a single sample, including rotation, flipping, blurring, or their combinations [DeVries and Taylor, 2017; Cubuk *et al.*, 2018; Hendrycks *et al.*, 2020]. Meanwhile, MSA further increases the diversity by involving more than one sample to generate the convex combination of them, i.e., cut-and-paste or addition [Zhang *et al.*, 2017; Yun *et al.*, 2019] and hence remarkably boosts the closed-set recognition ability as shown in Figure 1.

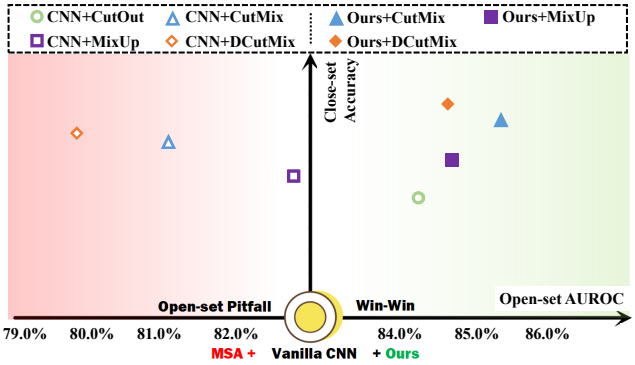

Figure 1: Illustration of the two sides of data augmentation. Despite the tremendous accuracy gain made by augmentations, multiple sample-based augmentation (MSA) tends to degrade the model's open-set performance.

Efficient and effective as MSA is, some research found it affects the performance of recognition tasks to some extent. Balestriero *et al.* [2022] demonstrate that MSA caused a drop in some classes, and Choi *et al.* [2023] argue that MSA disperses features in similar classes. However, compared with closed-set recognition, open-set recognition (OSR) is actually the biggest victim of this problem because it has no access to open-set data and hence heavily relies on the discriminative feature. As shown in Figure 1, we reveal that the significant improvement of MSA on closed-set recognition sacrifices the performance of OSR, and as closed-set recognition improves, the corresponding decline in OSR becomes more pronounced, dubbed as the two sides of MSA.

*Corresponding author

To mitigate the degradation of open-set performance caused by MSA, Roady *et al.* [2020] tempered the outputs of the models in a label-smoothing way to increase the entropies of the model's outputs. Xu *et al.* [2023] implemented the InfoNCE loss and used MixUp to enlarge the inter-class margins. However, these extra constraints alleviate the dilemma of MSA on OSR while ceiling the improvement on closed-set recognition. We argue that an ideal solution should be win-win for both closed- and open-set samples.

Based on preliminary experiments on the interplay of DA and OSR, we have two key observations: 1) *MSA performs worse than SSA on OSR since it would disperse the features*; 2) *Knowledge distillation benefits OSR but goes back to decline when MSA joins in*. Digging deeper into these observations, we found that MSA diminishes the criteria of OSR in two aspects. First, MSA degrades the magnitude of the activation of features and logits, which leads to great uncertainty in selecting unknown samples via the logits threshold. Distillation mitigates this problem somewhat by forcing the student network to mimic the activation magnitude of the teacher network. Secondly, low-discriminative features of MSA samples remain uncertain; merely distillating them still suffers OSR criteria diminution.

Motivated by the above observation and findings, we propose an asymmetric distillation framework, a win-win solution for both close-set and open-set performance. Concretely, in addition to the same MSA samples fed to the teacher and student in symmetric distillation, we introduce extra raw samples to the teacher and exert extra mutual information objectively to enlarge the teacher's benefit. The introduced objective enables the student to focus more on the class-specific features within the mixed samples. Moreover, since some hard mixed samples provide ambiguous semantic information, we filter them out by relaxedly checking the teacher's predictions and assigning them an unknown-like target to encourage the model to decrease its activation for the non-salient features of the known classes. Within this framework, the model can leverage the advantages of MSA on closed-set performance and better discriminate the novels under extra supervision.

The main contributions in this paper are as follows:

- We revealed the two sides of DA leading to the degeneration of OSR and conducted experiments to analyze how the augmented samples undermine the model.
- We introduce an asymmetric distillation framework with a cross-mutual information maximization and a two-hot label smoothing to eliminate the effect and further improve the model's open-set performance.
- We perform extensive experiments and prove the effectiveness of our proposed method on various benchmarks.

2 Reveal the Two Sides of MSA

Despite MSA achieves significant improvement in closed-set recognition, we reveal two sides of MSA on closed-set and OSR in this section. We first present two distinct observations concerning DA and OSR in Section 2.1, followed by an in-depth analysis expounded in Section 2.2. Finally, we elucidate the mechanisms through which knowledge distillation can alleviate the degradation in OSR performance induced by MSA, and we highlight inherent issues within existing symmetric distillation frameworks in Section 2.3.

2.1 Key Findings from DA and OSR Interplay

Without loss of generality, we experiment on both SSA and MSA methods on accuracy and Area Under the Receiver Operating Characteristic curve (AUROC) on Tiny-ImageNet dataset. As shown in Table 1, taking the vanilla model as the baseline, we can make the following two observations:

Observation 1) on SSA vs MSA. Both SSA and MSA exhibit efficacy in enhancing the closed-set accuracy of the model, attributed to their capabilities in expanding the dataset. Notably, MSA demonstrates superior performance in this regard, as it effectively enlarges the diversity within the training data. Nevertheless, when evaluating AUROC, SSA modestly enhances performance, while the incorporation of MSA significantly undermines OSR capabilities.

Observation 2) Distillation Benefits OSR. To verify the influence of distillation [Hinton *et al.*, 2015] on MSA, we compare the MSA-sample based distillation with the vanilla distillation framework. Following Wang *et al.* [2022], we use a non-MSA-trained network as the teacher model. As shown in Table 1, with distillation only, both accuracy and AUROC gain an improvement. Furthermore, the integration of MSA significantly enhances accuracy by a substantial margin. This implies that the dataset expanded through MSA contributes significantly to the augmentation of the model's representational capacity. Notwithstanding the attainment of more expressive and generalized features, MSA persists in yielding a decrement rather than an amelioration in performance on OSR.

2.2 MSA Diminishes the Criteria of OSR

Choi *et al.* [2023] argue that the MixUp-trained model disperses features in closed-set classes. To quantify such a phenomenon, we visualize the discrepancy among all class pairings of the vanilla model and the CutMix-trained model in Figure 2 (a). At first glance, the heatmap of the CutMix-trained model is darker than the vanilla one, which indicates the degradation of the model's activation magnitude and the lower margins among all the classes. Specifically, the drastic decrease of the gap among the similar classes 'k_2 - k_3' and 'k_2 - k_5' in Figure 2 (a) indicates that the model tends to learn an obscure boundary among these classes. In contrast, the degradation of the distinct classes such as 'k_4 - k_5' and 'k_4 - k_3' is slighter. The broken boundaries among the similar classes are vulnerable to the unknowns which have similar features with these classes. For OSR, the darker colors in the intersect regions of the known classes and the unknown classes suggest the model's degeneration of discriminating them from each other. To dig deeper into the observation, we draw a theoretical analysis in the following.

Denoting $D = \{D_k, D_{unk}\}$ as all the inputs the model may encounter during deployment, $D_{train} = \{\mathrm{x}_i, \mathrm{y}_i\}_{i=1}^{n} \subseteq D_k$ represents the training dataset and x_i and y_i are the image and the corresponding label. Given input image x_i, model's C-classes prediction $\hat{\mathrm{y}}_i$ can be obtained via $\hat{\mathrm{y}}_i = \mathrm{softmax}(\mathbf{W}\Phi_\theta(\mathrm{x}_i))$, where $\Phi_\theta(\cdot)$ is the feature extractor and

Model	Vanilla CNN		SSA						MSA				Distillation			
			+ AugMix [2020]		+ Rand. Quantization [2023]		+ CutOut [2017]		+ CutMix [2019]		+ MixUp [2017]		Vanilla [2022b]		+ CutMix [2019]	
	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC
R-101	86.72	84.03	$86.98^{+0.26}$	$84.12^{+0.09}$	$86.90^{+0.18}$	$84.14^{+0.11}$	$86.84^{+0.12}$	$84.11^{+0.08}$	$87.06^{+0.34}$	$82.62^{-1.41}$	$88.34^{+1.62}$	$83.74^{-0.29}$	$87.32^{+0.60}$	$84.65^{+0.62}$	$88.90^{+2.18}$	$84.50^{+0.47}$
R-50	86.16	83.84	$86.28^{+0.12}$	$83.92^{+0.08}$	$86.42^{+0.26}$	$84.19^{+0.32}$	$86.36^{+0.20}$	$84.26^{+0.42}$	$87.44^{+1.28}$	$83.41^{-0.43}$	$86.62^{+0.46}$	$83.13^{-0.71}$	$86.38^{+0.22}$	$84.74^{+0.90}$	$88.04^{+1.88}$	$84.10^{+0.26}$
R-18	84.28	82.84	$86.14^{+1.86}$	$84.10^{+1.26}$	$84.64^{+0.36}$	$82.93^{+0.09}$	$86.42^{+2.14}$	$84.24^{+1.44}$	$87.42^{+3.14}$	$80.99^{-1.29}$	$86.82^{+2.54}$	$82.61^{-0.23}$	$86.64^{+2.36}$	$84.24^{+2.44}$	$88.82^{+4.54}$	$82.47^{-0.37}$
V-19	82.10	80.99	$82.70^{+0.60}$	$81.69^{+0.70}$	$82.86^{+0.76}$	$81.32^{+0.33}$	$82.40^{+0.30}$	$81.17^{+0.18}$	$83.44^{+1.34}$	$75.24^{-5.75}$	$83.34^{+1.24}$	$76.81^{-4.18}$	$83.12^{+1.02}$	$81.18^{+0.19}$	$84.18^{+2.08}$	$76.72^{-4.27}$
V-16	80.90	80.83	$82.84^{+1.94}$	$81.51^{+0.68}$	$82.96^{+2.06}$	$81.86^{+1.03}$	$82.36^{+1.46}$	$81.17^{+0.34}$	$84.28^{+3.38}$	$74.98^{-5.85}$	$83.10^{+2.20}$	$75.78^{-4.95}$	$83.74^{+2.84}$	$81.62^{+0.72}$	$85.04^{+4.14}$	$78.17^{-2.66}$
V-13	80.72	80.49	$83.86^{+3.14}$	$81.70^{+1.21}$	$82.18^{+1.46}$	$81.67^{+1.18}$	$82.98^{+2.26}$	$81.45^{+2.26}$	$83.96^{+3.24}$	$74.90^{-5.59}$	$83.08^{+2.36}$	$73.00^{-7.49}$	$83.62^{+2.90}$	$81.80^{+1.31}$	$85.32^{+4.60}$	$78.87^{-1.62}$
MV2	83.20	81.31	$84.46^{+1.26}$	$81.56^{+0.25}$	$83.50^{+0.30}$	$81.52^{+0.21}$	$84.24^{+1.04}$	$81.97^{+0.66}$	$86.26^{+3.06}$	$78.82^{-2.49}$	$85.42^{+2.22}$	$78.68^{-2.63}$	$84.42^{+1.22}$	$82.36^{+1.05}$	$84.46^{+1.26}$	$82.00^{+0.69}$

Table 1: The impact of different augmentations on different models. 'R', 'V', and 'MV2' denote ResNet [2016], VGG [2014], and MobileNetV2 [2018], respectively. We report the close-set accuracy (Acc., %) and AUROC (%). The green numbers in the upper right show the improvement compared to the vanilla CNN model and the numbers in red indicate the degradation.

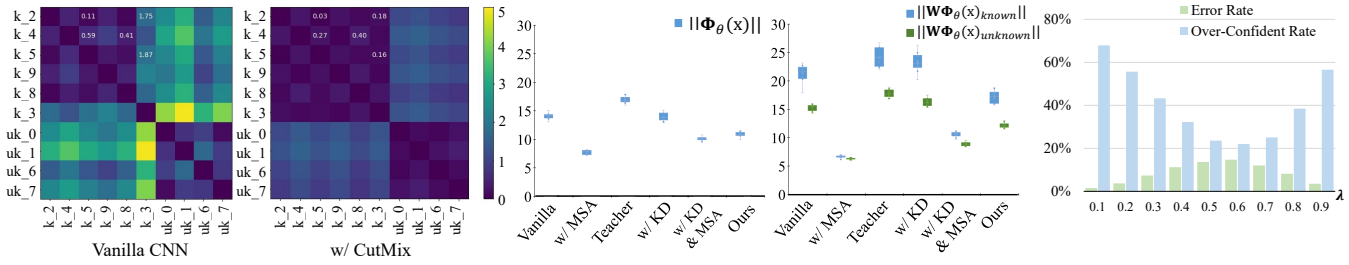

Figure 2: (a) Heatmap visualization of the distances among all the class parings on MNIST dataset. 'k' denotes the known classes and 'uk' denotes the unknown classes. The number after the underline is the ground-truth label. (b) The comparison of $||\Phi_\theta(x)||$ and $||W\,\Phi_\theta(x)||$ under different training paradigms. (c) The teacher's top-2 error rate and over-confident predictions (higher than 95%) over 10000 mixed samples under different mixing coefficients.

$\Phi_\theta(x_i) \in \mathbb{R}^D$. $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_C] \in \mathbb{R}^{C\times D}$ is the linear classification matrix and $\mathbf{W}\Phi_\theta(x_i) = [\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,C}]$ is the logits. The training is based on the cross-entropy loss $\mathcal{L}_{CE}$:

$$\begin{aligned}\mathcal{L}_{CE}(\theta, \mathbf{W}) &= -\hat{y}_{i,c} + \log(\sum_{k=1}^{C}\exp(\hat{y}_{i,k})) \\ &= -\mathbf{w}_c\,\Phi_\theta(x_i) + \log(\sum_{k=1}^{C}\exp(\mathbf{w}_k\,\Phi_\theta(x_i))).\end{aligned} \tag{1}$$

Vaze *et al.* [2022] investigated how $\mathcal{L}_{CE}$ influences OSR. The model initially embeds all the classes with a similar magnitude and gradually activates more for the known classes by increasing $||\mathbf{W}\,\Phi_\theta(x)||$ to better distinguish the unknowns. The final maximum logit score is used to provide the open-set score in their conclusion. Additionally, the model's wrong predictions during training tend to reduce $\mathbf{w}_k \cdot \Phi_\theta(x_i)\ \forall k \neq c$.

We use CutMix as an example to study the impact of MSA:

$$\begin{aligned}x_m &= \mathbf{M} \odot x_i + (\mathbf{1} - \mathbf{M}) \odot x_j, \\ y_m &= \lambda \cdot y_i + (1 - \lambda) \cdot y_j,\end{aligned} \tag{2}$$

where $\mathbf{M}$ is a mask and λ is sampled from Beta distribution $\beta(\alpha, \alpha)$. With x_m, Eq. 1 can be rewritten as:

$$\begin{aligned}\mathcal{L}_m(\theta, \mathbf{W}) = (-\lambda \cdot \mathbf{w}_{c_1}\,\Phi_\theta(x_m)) + (-(1-\lambda) \cdot \mathbf{w}_{c_2}\,\Phi_\theta(x_m)) \\ +\log(\sum_{k=1}^{C}\exp(\mathbf{w}_k\,\Phi_\theta(x_m))),\end{aligned} \tag{3}$$

where c_1 and c_2 are the ground-truth classes of x_i and x_j.

We display the comparison of $||\Phi_\theta(x)||$ and $||\mathbf{W}\,\Phi_\theta(x)||$ in Figure 2 (b) to explore how Eq. 3 influences the model's behavior. It is straightforward that the MSA-trained model suffers from a degradation of feature norm which have the direct bearing on the model's criteria of OSR. Consequently, in Figure 2 (b), $||\mathbf{W}\,\Phi_\theta(x)||$ of the MSA-trained model decreases drastically, thus harming the open-set score. The discrepancy between the known classes and the unknown classes is also reduced as can be seen in Figure 2 (b). Through the above analysis, we conclude that MSA diminishes the criteria of OSR.

2.3 Retrieve the Discrepancy by Distillation

The distillation experiments in Table 1 indicate that KD benefits OSR. However, distillation with CutMix brings a greater improvement on the model's accuracy while impairing the gain of OSR performance. We investigate how the teacher works by analyzing the distillation loss $\mathcal{L}_{Distill} = \mathcal{D}_{KL}(\hat{y}^s||\hat{y}^t)$, where the superscripts s and t denote the student and the teacher model. It encourages $\hat{y}^s$ to minimize its divergence with $\hat{y}^t$, which implicitly leads to an alignment of the magnitude of the activation. This can be concluded in Figure 2 (b) by comparing the $||\Phi_\theta(x)||$ of the KD-trained model and the teacher. In addition, the comparison of $||\mathbf{W}\,\Phi_\theta(x)||$ between the MSA-trained model and the MSA-distilled model indicates that distillation with MSA helps the model retrieve the decreased discrepancy between the known classes and the unknown classes to a certain degree. However, the MSA is still harmful to the distilled model, which suggests that the vanilla symmetric distillation framework can not mitigate this issue.

To investigate why CutMix influences the benefit of distillation, we calculate the teacher's over-confident predictions (the maximum probability is greater than 95%) and wrong predictions (the predicted class is out of c_1 and c_2) over 10000 mixed samples on Tiny-ImageNet dataset as shown in Figure 2 (c). The statistical results suggest that the teacher makes amount of unreasonable predictions on MSA samples. For example, the mixtures of the similar classes will easily be over-confidently predicted because of the redundant activation of their similar features. To solve this problem, We regularize the teacher's redundant activation by an asymmetric distillation framework with an extra mutual information supervision and a re-label mechanism in Section 3. And the wrong prediction indicates that the mixed sample does not include the discriminative features so that the model should be encouraged to decrease its activation. To achieve this, we re-label the wrong predicted samples with smoothed two-hot labels to make the model put less attention on the class-agnostic features within the mixed samples.

3 Method

3.1 Overview

The overall pipeline of the proposed asymmetric distillation framework is outlined in Figure 3. Based on the vanilla symmetric distillation in which the student and the teacher are fed with the same inputs, we introduce extra initial samples x_i and x_j to the teacher while training with the augmented input x_m to perform an asymmetric data flow between the student and the teacher. We utilize the teacher's output of x_i and x_j to exert the student's output of x_m a cross mutual information objective which forces the student to concentrate more on the class-specific features within x_m. In addition, for the confusing mixtures that are wrongly predicted by the teacher, we pick them out and re-label them with a smoothed two-hot label to decrease the student's activation of them. Achieving this can make the student less active in the class-agnostic features.

3.2 Asymmetric Distillation Framework

The asymmetric distillation framework is specially designed for the training of MSA. In our experiments, the model is randomly trained using either the original or mixed samples, with a probability of 0.5. We control different data flows with different objects for the student and the teacher during distillation to leverage the teacher's prior knowledge.

Training with the Initial Samples. We inherit the advantages of KD for the initial samples by training with the symmetric distillation framework similar to Wang *et al.* [2022]. The loss is computed by:

$$\mathcal{L}_{raw} = \mathcal{L}_{CE}(\hat{\mathrm{y}}^s, \mathrm{y}) + \mathcal{L}_{Distill}(\hat{\mathrm{y}}^s, \hat{\mathrm{y}}^t). \tag{4}$$

Asymmetric Inputs of Training MSA. The unreasonable output of the teacher emphasizes the non-salient features within the mixture. To enable the student to concentrate more on the class-specific features, we propose extra supervision to amplify the teacher's optimization of $||\Phi_\theta(\mathrm{x})||$. Concretely, we build an asymmetric distillation framework upon the vanilla symmetric distillation framework by introducing the initial samples to the teacher. The additional initial samples offer the mixed samples an extra mutual information maximization objective to amplify the teacher's impact.

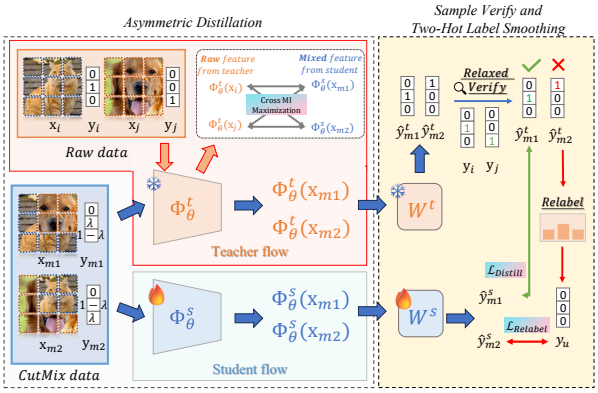

Figure 3: The proposed asymmetric distillation framework. Both the student and teacher models receive mixed data as input and perform distillation on $\Phi_\theta(x)$. Besides, the teacher model additionally accepts raw data as input to enlarge its benefit on the mixed inputs. To further decrease the student's activation of the non-discriminative features, we filter the teacher's wrong predictions of the mixed samples out and assign them a revised label to optimize.

Cross Mutual Information. Mutual Information (MI) is a fundamental measurement to quantify the relationship between random variables [Hou *et al.*, 2021; Feng *et al.*, 2023] denoting by $\mathcal{I}(v_1, v_2)$ where v_1 and v_2 are two random variables. Initially, the primary objective for MSA training can be understood as maximizing $\mathcal{I}(\Phi^s_\theta(\mathrm{x}_m), \mathrm{y}_m)$. As the mixed label y_m does not reflect the amount of the label information included in x_m well, the powerful teacher is introduced to produce the embedding $\Phi^t_\theta(\mathrm{x}_m)$ which is considered as a proper and model-friendly target of optimization. The objective of distillation can be abstracted into the maximization of $\mathcal{I}(\Phi^t_\theta(\mathrm{x}_m), \Phi^s_\theta(\mathrm{x}_m))$ which encourages the student to align its representation with the teacher.

However, since $\Phi^t_\theta(\mathrm{x}_m)$ may include class-agnostic information, especially the mixture of similar classes, we argue that merely imitating the teacher's output is not an optimal solution for OSR. An ideal objective encourages the student to maximize the discriminative features while discarding the common ones within the mixed samples. We achieve this by excluding their shared label information to amplify the teacher's impact on the class-specific features. The term $\mathcal{I}(\Phi^s_\theta(\mathrm{x}_m), \Phi^t_\theta(\mathrm{x}_m)|\mathrm{y}_j)$ rigorously quantifies the amount of information of the c_1-th class shared between $\Phi^t_\theta(\mathrm{x}_m)$ and $\Phi^s_\theta(\mathrm{x}_m)$ where '|' is an excluding operation. The maximization of this term forces the student to attend more on the characteristic features of the c_1-th class in x_m. We maximize this term for both of the two classes in x_m by a mutual information loss:

$$\begin{aligned}\mathcal{L}_{MI} = -(&\mathcal{I}(\Phi^s_\theta(\mathrm{x}_m), \Phi^t_\theta(\mathrm{x}_m)|\mathrm{y}_j) \\ &+ \mathcal{I}(\Phi^s_\theta(\mathrm{x}_m), \Phi^t_\theta(\mathrm{x}_m)|\mathrm{y}_i)).\end{aligned} \tag{5}$$

Revisiting the terms in Eq. 5, we find that excluding

the class-agnostic information of the c_1-th class in x_m, i.e. $(\Phi_\theta^t(\mathrm{x}_m)|\mathrm{y}_j)$, can be easily achieved by replacing it with $\Phi_\theta^t(\mathrm{x}_i)$ in consideration of x_i shares the same pure information of the c_1-th class with x_m. So we maximize the mutual information of x_m, x_i and x_j in a cross manner and simplify Eq. 5 to a Cross Mutual Information loss:

$$\begin{aligned}\mathcal{L}_{CMI} = -(\lambda \mathcal{I}\left(\Phi_\theta^s(\mathrm{x}_m), \Phi_\theta^t(\mathrm{x}_i)\right) \\ + (1-\lambda)\mathcal{I}(\Phi_\theta^s(\mathrm{x}_m), \Phi_\theta^t(\mathrm{x}_j)),\end{aligned} \tag{6}$$

where λ is determined by Eq. 2 to weight the contribution of x_i and x_j.

3.3 Sample Verify and Two-Hot Label Smoothing

Eq. 6 enables the model to discard the class-agnostic features in x_m. However, in Figure 2 (c), we point out that the teacher makes mistakes for some corner cases, which may be sub-optimal to the model's optimization. We revise the teacher's wrong predictions with a smoothed two-hot label to help the model learn more uncertainties within the confusing samples.

Relaxed Sample Verify. For a mixed sample x_m, we argue that it contains the non-salient parts of both the c_1-th and the c_2-th class when the teacher predicts it to the third class. We utilize a relaxed verification that checks the top-2 accuracy of $\hat{\mathrm{y}}_m^t$ to filter the teacher's wrong predictions out and assign them a revised target to optimize.

Two-Hot Label Smoothing. We aim to optimize the wrongly predicted mixed samples to decrease their activation so that the model can discard the non-discriminative features within the mixtures. The cross-entropy loss we discussed above can naturally degrade the activation for the wrong predictions by Eq. 1. In addition, we want the model to learn more uncertainties among the confusing mixtures, so we manually set a revised target for these samples. Concretely, we mix a uniform label $\bar{\mathrm{y}} \in \mathbb{R}^C$ whose elements are $1/C$ and y_m by a ratio of 0.5 to generate the target of the wrongly predicted x_m which we name it y_u. The loss is computed by:

$$\begin{aligned}\mathcal{L}_{Relabel} = &\mathbb{1}(\mathrm{argmax}(\hat{\mathrm{y}}_m^t) \neq c_1 \; and \; \mathrm{argmax}(\hat{\mathrm{y}}_m^t) \neq c_2) \\ &\mathcal{L}_{CE}(\hat{\mathrm{y}}_m^s, \mathrm{y}_u),\end{aligned} \tag{7}$$

where $\mathbb{1}(\cdot)$ is an indicator function whose value is 1 when the following expression in the brackets is true and 0 vise versa. This object encourages the model to embed the uncertain samples x_m to the origin of the feature space.

The overall loss can be denoted as:

$$\mathcal{L}_{CutMix} = \mathcal{L}_{Distill} + \mu\,\mathcal{L}_{CMI} + \eta\,\mathcal{L}_{Relabel}, \tag{8}$$

where μ and η are hyper-parameters we set to 1.0.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate the performance of our model on three benchmarks: the OSR benchmark, semantic shift benchmark, and large-scale benchmark.

- **OSR Benchmark:** within this benchmark, the method is evaluated on five datasets, including SVHN [Netzer *et al.*, 2011], CIFAR-10 [Krizhevsky *et al.*, 2009], CIFAR+10, CIFAR+50, and Tiny-ImageNet [Le and Yang, 2015]. All settings align with those of AGC.
- **Semantic Shift Benchmark:** this evaluation protocol includes three datasets: Caltech-UCSD-Birds (CUB)[Wah *et al.*, 2011], Stanford Cars[Krause *et al.*, 2013], and FGVC-Aircraft [Maji *et al.*, 2013]. The presence of specific attributes distinguishes different classes, and the difficulty of recognition is calculated based on the differences in the number of attributes. Consequently, the open-set classes of the FGVC datasets are divided into 'Easy', 'Medium', and 'Hard' levels to denote their similarities with close-set classes.
- **Large-Scale Benchmark:** Within this protocol, 200 classes from Tiny-ImageNet are used for training. Subsequently, non-overlapping 'Easy' and 'Hard' splits from Imagenet-21k are selected for evaluation, following the approach outlined by Ren *et al.* [2023].

Evaluation Metrics. In the OSR benchmark, AUROC serves as a threshold-independent metric [Davis and Goadrich, 2006]. It quantifies the probability that a positive example possesses a higher detector score or value compared to a negative example. OSCR is a metric that gauges the trade-off between accuracy and open-set detection rate by adjusting the threshold on the confidence of the predicted class.

Implementation Details. We utilize DIST [Huang *et al.*, 2022a] as the foundational distillation method. The default teacher-student pair consists of ResNet-50 [He *et al.*, 2016] and ResNet-18. The training duration spans 200 epochs, employing a batch size of 32. The initial learning rate is 0.1, subsequently reduced by a factor of 5 at the 60th, 120th, and 160th epoch. The optimization employs the SGD optimizer with a momentum of 0.9, and the weight decay is set to 5e-4.

4.2 Comparison on OSR Benchmark

To assess the recognition capability of our proposed asymmetric distillation framework in open-set scenarios, we compare it not only with traditional state-of-the-art open-set recognition methods (denoted as ARPL [Chen *et al.*, 2021b], RCSSR [Huang *et al.*, 2022a] and AGC [Vaze *et al.*, 2022], etc.) but also with method CMPKD [Wang *et al.*, 2022], which incorporates both MSA and distillation. As shown in Table 2, our method exhibits a significant improvement in open-set performance compared to traditional open-set recognition methods, achieving a gain of 0.4% on CIFAR-10 and 2.6% on TinyImageNet. In comparison to method CMPKD, our model substantially enhances the model's open-set performance while maintaining closed-set accuracy, achieving improvements of 4.9% on SVHN and 2.8% on TinyImageNet. This demonstrates that through the use of the asymmetric distillation framework and the constraints of mutual information object engineering, we have indeed succeeded in enhancing the focus of the student model on class-specific features, and finally got a win-win solution for open-set recognition.

Methods	SVHN		CIFAR-10		CIFAR+10		CIFAR+50		TinyImageNet	
	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC
CROSR (CVPR, 2019)	-	89.9	-	88.3	-	91.2	-	90.5	-	58.9
C2AE (CVPR, 2019)	-	92.2	-	89.5	-	95.5	-	93.7	-	74.8
RPL (ECCV, 2020)	-	93.4	-	82.7	-	84.2	-	83.2	-	68.8
ARPL+CS (TPAMI, 2021)	-	96.7	-	91.0	-	97.1	-	95.1	-	78.2
CSSR (TPAMI, 2022)	-	**97.9**	-	91.3	-	96.3	-	96.2	-	82.3
AGC (ICLR, 2022)	97.6	97.1	96.4	93.6	97.8	97.9	97.8	96.5	84.6	82.7
OpenMix+ (TCSVT, 2023)	-	-	95.3	86.9	96.8	93.1	96.8	92.5	58.4	75.1
CMPKD (NIPS, 2022)	97.6	92.1	96.8	84.1	97.8	95.0	97.8	91.9	86.9	82.5
Ours	**97.7**	97.2	**96.9**	**94.0**	**98.0**	**98.1**	**98.0**	**96.8**	**87.3**	**85.3**

Table 2: Comparison of AUROC (%) and close-set accuracy (Acc., %) on OSR Benchmark. The best performance values are highlighted in bold and the second best performances are underlined.

Method	CUB			SCars			FGVC-Aircraft		
	Acc.	AUROC Easy / Hard	OSCR Easy / Hard	Acc.	AUROC Easy / Hard	OSCR Easy / Hard	Acc.	AUROC Easy / Hard	OSCR Easy / Hard
ARPL (TPAMI, 2021)	85.9	83.5 / 75.5	76.0 / 69.6	96.9	94.8 / 83.6	92.8 / 82.3	91.5	87.0 / 77.7	83.3 / 74.9
AGC (ICLR, 2022)	86.2	88.3 / 79.3	79.8 / 73.1	**97.1**	94.0 / 82.2	92.2 / 81.1	**91.7**	**90.7** / 82.3	**86.8** / 79.8
Ours	**87.6**	**89.6 / 82.0**	**81.4 / 75.8**	96.9	**95.5 / 84.9**	**93.3 / 83.5**	91.1	90.1 / **83.5**	86.2 / **80.9**

Table 3: Comparison of AUROC (%), OSCR (%) and closed-set accuracy (Acc., %) on semantic shift Benchmark. Results of ARPL and AGC are from Vaze *et al.* [2022]. The best performance values are highlighted in bold.

	BackBone	Easy		Hard	
		AUROC	OSCR	AUROC	OSCR
ARPL	VGG32	51.4%	33.8%	52.4%	34.4%
AGC	VGG32	72.8%	44.1%	72.1%	43.8%
CLIP	ViT-B/32	72.9%	44.0%	72.3%	43.6%
CoOp	ViT-B/32	74.6%	54.3%	73.3%	53.3%
A^2Pt	ViT-B/32	76.6%	58.7%	74.7%	57.4%
Ours	ResNet18	77.1%	60.8%	75.7%	60.1%
Ours	ViT-B/32	**79.3%**	**64.8%**	**75.7%**	**62.8%**

Table 4: Comparison on large-scale benchmark.

Method	Backbone	Acc.	AUROC
AGC	VGG-32	84.64%	82.68%
AGC	MobileNet-V2	82.46%	80.68%
Ours	MobileNet-V2	85.70%	83.32%

Table 5: Comparison on light-weight model MobileNet-V2.

4.3 Comparison on Semantic Shift Benchmark

To further explore the discriminative ability of our model for feature extraction, we conduct experiments on the semantic shift benchmark following AGC [2022]. The results, as shown in Table 3, reveal that our method consistently outperforms the AUROC metric of the state-of-the-art method AGC by a margin of 1%$\sim$2% in both 'Easy' and 'Hard' splits while maintaining closed-set accuracy on three fine-grained datasets (CUB, SCars, and FGVC), a slightly declined by less than 1% in the 'Easy' split of the Aircraft, possibly due to the invariable backgrounds (either the sky or the runway) among the dataset. Notably, our model's performance excels in the hybrid scenario with an OSCR metric exceeding 2.7% for the SOTA method. This suggests that, by employing MSA samples through an asymmetric distillation framework, our model can discard class-agnostic representations and focus more on class-specific representations, thereby enhancing recognition performance across various scenarios, even in challenging fine-grained classification scenarios.

4.4 Comparison on Large-Scale Benchmark

To further explore the effectiveness of our method in real-world scenarios, we conducted experiments on a large-scale dataset. Specifically, we trained on TinyImageNet with only 200 classes and tested on ImageNet-21k with 2100 classes. In our evaluation, we compare our method not only with conventional OSR methods, ARPL [Chen *et al.*, 2021b] and AGC [Vaze *et al.*, 2022], but also with additional multimodal methods such as CLIP [Radford *et al.*, 2021], CoOp [Zhou *et al.*, 2022], and A^2Pt. The results, as presented in Table 4, showcase the performance of our method in both the open-set recognition metric AUROC and the hybrid recognition metric OSCR. Remarkably, our method outperforms conventional methods by approximately 5%, as well as multimodal methods by 1%. This demonstrates that even in complex real-world scenarios, the features learned through the asymmetric distillation framework remain highly discriminative. Importantly, these features are not significantly disturbed by the

Model	CutMix	Distillation	CMI	Smoothed two-hot label	Acc.	AUROC
ResNet-18					84.3%	82.8%
(a)	✓				87.4%	81.0%
(b)	✓	✓			88.8%	82.5%
(c)	✓	✓	✓		87.8%	84.8%
(d)	✓	✓	✓	✓	87.3%	85.3%

Table 6: Ablations of our proposed terms on Tiny-ImageNet.

η \ μ	0.5		1.0		2.0	
	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC
0.5	87.7%	85.1%	87.6%	85.0%	87.4%	85.0%
1.0	87.3%	84.9%	**87.3%**	**85.3%**	87.5%	85.1%
2.0	86.8%	84.9%	87.3%	84.7%	87.5%	84.9%

Table 7: Results under different hyper-parameter settings.

Method	In-distribution	Out-of-distribution	AUROC
MLS	Cifar-10	Cifar-100	87.5%
Ours			89.6%
MLS	Cifar-10	Tiny-ImageNet	88.7%
Ours			91.2%

Table 8: The evaluations on OoD detection.

Method	ChestMNIST		OCTMNIST		PneumoniaMNIST	
	AUROC	Acc.	AUROC	Acc.	AUROC	Acc.
MedMNIST	76.8%	94.7%	94.3%	74.3%	94.4%	85.4%
Ours	77.5%	94.8%	96.2%	77.0%	94.9%	90.1%

Table 9: Results on MedMNIST v2 dataset.

increase in unknown novel classes, showcasing the ability of our method to stabilize the open-set recognition performance of the model. Additionally, we replace our ResNet-18 backbone with ViT and report the results. Compared to A^2Pt and CLIP, the performance on the ImageNet-21k dataset shows the superiority of our method on the ViT backbone. And the comparison with our ResNet-18 proves that using a more powerful backbone model can reap better benefits.

4.5 Comparison on the Light-weight Model

We conduct additional experiments to assess the effectiveness of our model on lightweight networks in Table 5. In comparison to the state-of-the-art (SOTA) method AGC [Vaze *et al.*, 2022] implemented on MobileNet-V2 [Sandler *et al.*, 2018], our proposed asymmetric distillation model demonstrates superior performance. Our model exhibits improvements in both closed-set accuracy (Acc.) with a margin of +3.24% and open-set recognition AUROC with a margin of +2.64% on the TinyImageNet dataset. These results indicate that our method is not constrained by the network parameters and remains effective even in lightweight networks.

4.6 Ablation Study

In Table 6, we conduct an ablation analysis on the Tiny-ImageNet dataset to delve deeper into the effectiveness of different elements of our method. The comparison between ResNet-18 with CutMix (a) highlights the significant positive impact of multiple samples-based augmentations on improving closed-set classification (+3.1%). However, this improvement comes at the expense of a substantial reduction in the model's open-set recognition performance (-1.8%). The introduction of distillation methods partially mitigates the degradation of open-set performance (from 81.0% to 82.5%), but does not lead to improvement, less than 82.8%. However, when our Contrastive Mutual Information (CMI) objective and Smoothed Two-Hot Label method are introduced, the open-set recognition metrics AUROC of the model sequentially increase from 82.5% to 84.8% and 85.3%. Although the closed-set classification metric (Acc.) slightly decreased by 1% point and 1.5% points, respectively, there are still +3% improvements over the original ResNet. This demonstrates that our proposed CMI objective and Smoothed Two-Hot Label method significantly enhance the model's open-set recognition ability. The empirical evidence highlights the effectiveness of our additional supervision methods in facilitating the learning of class-specific features and decreasing the activation of the non-salient features of the known classes.

To validate the robustness of our method to the hyper-parameters, we test different combinations of hyper-parameters μ and η including 0.5, 1.0, and 2.0. The result in Table 7 fluctuating around 0.5% under different combinations shows that our method is insensitive to hyper-parameter settings. And the optimal result appears when μ and η are set as 1.0.

4.7 Results on Other Tasks

Our proposed method ensures the backbone model extracts features with discrimination and hence further promotes downstream task performance like OSR task. In Table 8, we evaluate our method on uncertainty-related task Out-of-Distribution (OoD) to verify our method as a general feature-strengthening tool. We equipped the maximum logit score (MLS) baseline with our method on Cifar-10/Cifar-100 and Cifar-10/Tiny-ImageNet and the improvements show the effectiveness on OoD task.

Furthermore, we focus on the fundamental and practically applicable recognition task. As a win-win solution for close- and open-set tasks, our proposed method can be regarded as an effective feature extractor enhancement strategy. Taking as an example, we verify the representation ability enhancement of our method on medical image analysis (on MedMNIST v2 dataset [Yang *et al.*, 2023]) in Table 9.

5 Conclusion

In this paper, we start by revealing the two sides of the data-mix augmentation by investigating how MSA interplays with open-set recognition. Our experiments and visualizations suggest that MSA diminishes the criteria of OSR and leads to confusion among similar classes. Based on the observations of how knowledge distillation works on OSR, we propose a win-win solution that leverages MSA to boost both the close-set and the open-set performance. The outstanding performance of our method conducted on multiple datasets demonstrates the effectiveness of our approach. It also demonstrates the potential of the known classes can help to detect novels.

Acknowledgments

This work is supported in part by the Beijing Natural Science Foundation under No. L221013, the National Natural Science Foundation of China under Grant Nos. 62102162 and 62203184, and the CCF-DiDi GAIA Collaborative Research Funds for Young Scholars.

References

[Balestriero *et al.*, 2022] Randall Balestriero, Leon Bottou, and Yann LeCun. The effects of regularization and data augmentation are class dependent. *Advances in Neural Information Processing Systems*, 35:37878–37891, 2022.

[Chen *et al.*, 2021a] Dong Chen, Fan Tang, Weiming Dong, Hanxing Yao, and Changsheng Xu. Siamcpn: Visual tracking with the siamese center-prediction network. *Computational Visual Media*, 7:253–265, 2021.

[Chen *et al.*, 2021b] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2021.

[Chen *et al.*, 2023] Dong Chen, Xingjia Pan, Fan Tang, Weiming Dong, and Changsheng Xu. Spa 2 net: Structure-preserved attention activated network for weakly supervised object localization. *IEEE Transactions on Image Processing*, 2023.

[Choi *et al.*, 2023] Hongjun Choi, Eun Som Jeon, Ankita Shukla, and Pavan Turaga. Understanding the role of mixup in knowledge distillation: An empirical study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2319–2328, 2023.

[Cubuk *et al.*, 2018] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[Davis and Goadrich, 2006] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

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

[Feng *et al.*, 2023] Runyang Feng, Yixing Gao, Xueqing Ma, Tze Ho Elden Tse, and Hyung Jin Chang. Mutual information-based temporal difference learning for human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17131–17141, 2023.

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

[Hendrycks *et al.*, 2020] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network, 2015. Accessed: 2018-12-06.

[Hou *et al.*, 2021] Xuege Hou, Yali Li, and Shengjin Wang. Disentangled representation for age-invariant face recognition: A mutual information minimization perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3692–3701, 2021.

[Hou *et al.*, 2024] Liang Hou, Qi Cao, Yige Yuan, Songtao Zhao, Chongyang Ma, Siyuan Pan, Pengfei Wan, Zhongyuan Wang, Huawei Shen, and Xueqi Cheng. Augmentation-aware self-supervision for data-efficient gan training. *Advances in Neural Information Processing Systems*, 36, 2024.

[Huang *et al.*, 2022a] Hongzhi Huang, Yu Wang, Qinghua Hu, and Ming-Ming Cheng. Class-specific semantic reconstruction for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4214–4228, 2022.

[Huang *et al.*, 2022b] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *arXiv preprint arXiv:2205.10536*, 2022.

[Krause *et al.*, 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, June 2013.

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

[Le and Yang, 2015] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[Maji *et al.*, 2013] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Ren *et al.*, 2023] Hairui Ren, Fan Tang, Xingjia Pan, Juan Cao, Weiming Dong, Zhiwen Lin, Ke Yan, and Changsheng Xu. A 2 pt: Anti-associative prompt tuning for open set visual recognition. *IEEE Transactions on Multimedia*, 2023.

[Roady *et al.*, 2020] Ryne Roady, Tyler L Hayes, and Christopher Kanan. Improved robustness to open set inputs via tempered mixup. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 186–201. Springer, 2020.

[Sandler *et al.*, 2018] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Vaze *et al.*, 2022] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: a good closed-set classifier is all you need? In *International Conference on Learning Representations*, 2022.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[Wang *et al.*, 2022] Huan Wang, Suhas Lohit, Michael Jones, and Yun Fu. What makes a "good" data augmentation in knowledge distillation – a statistical perspective. In *NeurIPS*, 2022.

[Wang *et al.*, 2024] Yue Wang, Yuke Li, James H Elder, Runmin Wu, and Huchuan Lu. Class-conditional domain adaptation for semantic segmentation. *Computational Visual Media*, pages 1–18, 2024.

[Wu *et al.*, 2023] Huimin Wu, Chenyang Lei, Xiao Sun, Peng-Shuai Wang, Qifeng Chen, Kwang-Ting Cheng, Stephen Lin, and Zhirong Wu. Randomized quantization: A generic augmentation for data agnostic self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16305–16316, 2023.

[Xu *et al.*, 2022] Yifan Xu, Huapeng Wei, Minxuan Lin, Yingying Deng, Kekai Sheng, Mengdan Zhang, Fan Tang, Weiming Dong, Feiyue Huang, and Changsheng Xu. Transformers in computational visual media: A survey. *Computational Visual Media*, 8:33–62, 2022.

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

[Yang *et al.*, 2022] Suorong Yang, Weikang Xiao, Mengcheng Zhang, Suhan Guo, Jian Zhao, and Furao Shen. Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*, 2022.

[Yang *et al.*, 2023] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

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

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

[Zhou *et al.*, 2022] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.