Appendix

Anonymous submission

A Theoretical Analysis

Gradient Derivation of $\mathcal{L}^k$

$$\mathcal{L}^k = -\frac{\boldsymbol{z}_p}{\tau}\cdot\boldsymbol{z}_i - \log\Big[\sum_{k\neq i}\exp\left(\boldsymbol{z}_i\cdot\boldsymbol{z}_k/\tau\right) + \sum_{\boldsymbol{x}_{\boldsymbol{u}}^{\boldsymbol{i}}\in U(i)}\exp(\boldsymbol{z}_i\cdot\boldsymbol{u}_i/\tau)\Big], \tag{1}$$

where we denote $C_k = \exp\left(\boldsymbol{z}_i\cdot\boldsymbol{z}_k/\tau\right)$ and $O_i = \exp\left(\boldsymbol{z}_i\cdot\boldsymbol{u}_i/\tau\right)$ and both terms are scalars.

The gradient for $\mathcal{L}^{sup}$ with respect to $\boldsymbol{z}_i$ can be depicted as:

$$\frac{\partial\mathcal{L}^{sup}}{\partial\boldsymbol{z}_i} = -\frac{1}{\tau}\left(\boldsymbol{z}_p + \frac{\sum_{k\neq i}C_k\cdot\boldsymbol{z}_k}{\sum_{k\neq i}C_k}\right), \tag{2}$$

Then the gradient for $\mathcal{L}^k$ with respect to $\boldsymbol{z}_i$ can be depicted as:

$$\begin{aligned}\frac{\partial\mathcal{L}^k}{\partial\boldsymbol{z}_i} &= -\frac{1}{\tau}\left(\boldsymbol{z}_p + \frac{1}{S}\sum_{k\neq i}C_k\cdot\boldsymbol{z}_k + \frac{1}{S}\sum_{\boldsymbol{x}_{\boldsymbol{u}}^{\boldsymbol{i}}\in U(i)}O_i\cdot\boldsymbol{u}_i\right)\\ &= -\frac{1}{\tau}\left(\boldsymbol{z}_p + G_{NK} + G_{TAU}\right),\end{aligned} \tag{3}$$

where $S = \sum_{k\neq i}C_k + \sum_{\boldsymbol{x}_{\boldsymbol{u}}^{\boldsymbol{i}}\in U(i)}O_i$. The components in $(\cdot)$ can be divided into three parts involved in the gradient update process: the representation of positively labelled data, the gradient of negatively labelled known data and the gradient of TAU data.

Hard Negative Mining

Analysis regarding *Fairly Contrast* has already been presented in the paper. Here, we will discuss *Hard Negative Mining* across three situations.

Situation 1. When negatively labelled known data is hard and TAU is easy, we have $\boldsymbol{z_i}\cdot\boldsymbol{z_k}\approx 1$ and $\boldsymbol{z_i}\cdot\boldsymbol{u_i}\approx 0$, $C_k = e^{1/\tau}$, $O_i = 1$. Therefore,

$$\frac{\partial\mathcal{L}^{sup}}{\partial\boldsymbol{z}_i} = -\frac{1}{\tau}\left(\boldsymbol{z}_p + \frac{\sum_{k\neq i}e^{1/\tau}\cdot\boldsymbol{z}_k}{\sum_{k\neq i}e^{1/\tau}}\right), \tag{4}$$

$$\frac{\partial\mathcal{L}^k}{\partial\boldsymbol{z}_i} = -\frac{1}{\tau}\left(\boldsymbol{z}_p + \frac{1}{S_1}\sum_{k\neq i}e^{1/\tau}\cdot\boldsymbol{z}_k + \frac{1}{S_1}\sum_{\boldsymbol{x}_{\boldsymbol{u}}^{\boldsymbol{i}}\in U(i)}\boldsymbol{u}_i\right), \tag{5}$$

where $S_1 \approx \sum_{k\neq i}e^{1/\tau}$. In Eq. (5), the weights of $\boldsymbol{z}_i$ is $e^{1/\tau}$ larger than $\boldsymbol{u}_i$', which enhances the contribution of hard negatives $\boldsymbol{z}_i$ to the gradient update process. Besides, compared with Eq. (4), the additional term $\sum_{\boldsymbol{x}_{\boldsymbol{u}}^{\boldsymbol{i}}\in U(i)}\boldsymbol{u}_i$ implies that $\mathcal{L}^k$ can take into account the contribution of pseudo-unknowns to gradient updates, which greatly benefits the encoder.

Situation 2. When negatively labelled known data is easy and TAU is hard, we have $\boldsymbol{z_i}\cdot\boldsymbol{z_k}\approx 0$ and $\boldsymbol{z_i}\cdot\boldsymbol{u_i}\approx 1$, $C_k = 1$, $O_i = e^{1/\tau}$. Therefore,

$$\frac{\partial\mathcal{L}^k}{\partial\boldsymbol{z}_i} = -\frac{1}{\tau}\left(\boldsymbol{z}_p + \frac{1}{S_2}\sum_{k\neq i}\boldsymbol{z}_k + \frac{1}{S_2}\sum_{\boldsymbol{x}_{\boldsymbol{u}}^{\boldsymbol{i}}\in U(i)}e^{1/\tau}\cdot\boldsymbol{u}_i\right), \tag{6}$$

where $S_2 \approx \sum_{\boldsymbol{x}_{\boldsymbol{u}}^{\boldsymbol{i}}\in U(i)}e^{1/\tau}$. It especially enhances the contribution of hard negatives $\boldsymbol{u}_i$ to the gradient updates, which means the inner ability of hard negative mining of ours.

Situation 3. When negatively labelled known data and TAU both are hard, we have $\boldsymbol{z_i}\cdot\boldsymbol{z_k}\approx 1$ and $\boldsymbol{z_i}\cdot\boldsymbol{u_i}\approx 1$, $C_k = O_i = e^{1/\tau}$. Therefore,

$$\frac{\partial\mathcal{L}^k}{\partial\boldsymbol{z}_i} = -\frac{1}{\tau}\left(\boldsymbol{z}_p + \frac{1}{S_3}\sum_{k\neq i}e^{1/\tau}\cdot\boldsymbol{z}_k + \frac{1}{S_3}\sum_{\boldsymbol{x}_{\boldsymbol{u}}^{\boldsymbol{i}}\in U(i)}e^{1/\tau}\cdot\boldsymbol{u}_i\right), \tag{7}$$

where $S_3 = \sum_{k\neq i}e^{1/\tau} + \sum_{\boldsymbol{x}_{\boldsymbol{u}}^{\boldsymbol{i}}\in U(i)}e^{1/\tau}$. In this situation, the contribution of the hard negatives irrespective of negatively labelled known data or TAU data is enhanced.

B More About Experiments

Implementation Details

All experiments are conducted on a Nvidia RTX 3090 GPU. The parameters of networks are optimized by the Adam optimizer with weight decay = 0.0001 on all datasets except for the TinyImageNet optimized by the SGDM optimizer with momentum = 0.9. The learning rates, strategy of learning rate decay and warm for contrastive learning and classifier training are all followed with (Xu, Shen, and Zhao 2023). The batch size of loaded training data is 128 by default. In the data augmentation part, the hyper-parameters of RandAugment are fixed as $N = 1, M = 5$.

Metrics Details

Area Under the Receiver Operating Characteristic (AUROC) curve, which is widely used in OSR is a threshold-independent metric that measures the probability of a positive example being assigned a higher detection score than a negative example(Fawcett 2006).

Additionally, the goal of OSR task is not limited to only detecting unknown data from test instances that AUROC can evaluate, it also involves considering the accuracy of known classes. Thus, a suitable metric for evaluating correct classifications of known classes, Open Set Classification Rate (OSCR) (Dhamija, Günther, and Boult 2018), has been widely used in many recent OSR researches. OSCR sets the correct classification rate (CCR) in relation to the false positive rate (FPR), where δ is a probability threshold. Therefore,

$$CCR(\delta) = \frac{|\{x \in \mathcal{D}_{tr} \wedge \arg max_k P(k|\boldsymbol{x}) = \hat{k} \wedge P(\hat{k}|x) \geq \delta\}|}{|\mathcal{D}_{tr|}}, \quad (8)$$

$$FPR(\delta) = \frac{|\{x \in \mathcal{D}_U \wedge \arg max_k P(k|\boldsymbol{x}) \wedge P(\hat{k}|x) \geq \delta\}|}{|\mathcal{D}_{U|}}. \quad (9)$$

C Experiments for Out-of-Distribution Detection

Details about Datasets

For X-Crop datasets, the original images are cropped into 32*32. And for X-Resize datasets, the original images are resized into 32*32.

Further Elaboration

The results of these two experiments are completely in accordance with the *Familiarity Hypothesis*(Dietterich and Guyer 2022) that states instead of detecting the novel features presented in an image, the existing model succeeds in OSR task primarily by detecting the absence of familiar features in the image for recognizing the unknowns.

The difficulty of OOD detection depends on the degree of unfamiliarity (*i.e.,* distinction in semantic information within the features) between OOD and ID data. The greater the distinction between ID and OOD data, the more unfamiliar the features of OOD become, leading to easier detection of OOD data.

According to this, we support it through these two experiments from different aspects to modify the *familiarity* between ID and OOD data:

Semantic Quality. In ID:MNIST setting, we vary the semantic quality of OOD data. The performance gap between noisy images (MNIST and MNIST-Noise) and semantic OOD images (Omniglot) can be interpreted as the following. The features of OOD data from the Omniglot dataset possess clear semantic information, which contains much familiar semantic information with MNIST's. Therefore, they can be regarded as the high semantic quality OOD data, which makes detection become more challenging on this. However, OOD data from the MNIST-Noise and Noise datasets exhibit lower semantic quality than Omniglot since the semantic information of the features has been blurred, leading to more unfamiliar features with MNIST.

Image Structure. In ID:CIFAR10 setting, we vary the image structure of OOD data. The performance gap between X-Resize and X-Crop OOD data can be interpreted as the following. X-Resize are full original images resized into 32*32 pixels, which maintains the global semantic structure of images. Thus, X-Resize OOD data contains more familiar features with CIFAR10's, which leads to the detection becoming more challenging on these datasets. Meanwhile, the process of Crop disrupts the global semantic structure of images, which makes more unfamiliar features with CIFAR10. So the performance of X-Crop is more than X-Resize in terms of F1 Scores.

D Ablation of Hyper-parameters

Threshold ϵ

In this subsection, we explore the impact of different threshold percent ϵ on macro F1-Scores, which is sensitive to the value of the hyper-parameter ϵ, by adjusting it employed in the experiments for Out-of-Distribution Detection. We vary the value of ϵ from 1 to 15, and show the macro F1-Scores of two settings and the accuracy of ID and OOD data, respectively.

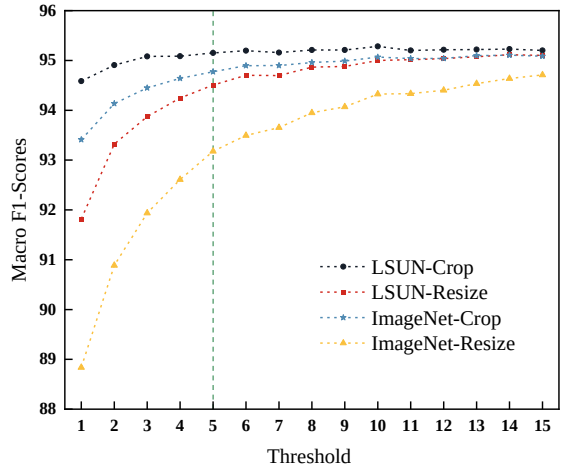
The results of these groups of experiments are shown in Fig. 1. From Fig. 1(a) and (b) (ID:CIFAR10 Setting), we can observe that with ϵ increased, the macro F1-Scores of four OOD datasets are increasing simultaneously, even the hardest dataset, ImageNet-Resize, can also achieve around 95% ultimately. The trend begins to stabilize when ϵ is set to 5. The accuracy of four OOD datasets. The pattern of the accuracy of OOD data changing with ϵ matches that of the macro F1-Scores. Meanwhile, in our framework, the impact of accuracy on ID data is not significant, consistently staying around 94.8%.

In Fig. 1(c) and (d) (ID:MNIST Setting), the hardest OOD dataset is Omniglot and its macro F1-Scores increases with ϵ increased and is steady when ϵ is 5. The macro F1-Scores of MNIST-Noise is also stable when ϵ is 5. Especially, the macro F1-Scores of Noise data reaches almost 100% right from the beginning and consistently maintained it. The accuracy of these keeps the same trend as that of macro F1-Scores.

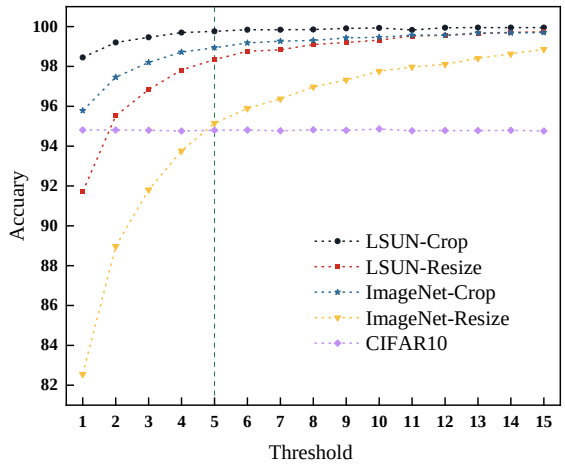
From Fig. 1, considering both macro F1-Scores and accuracy together, we set the value of threshold percentile ϵ as 5 in these experiments.

Visualization

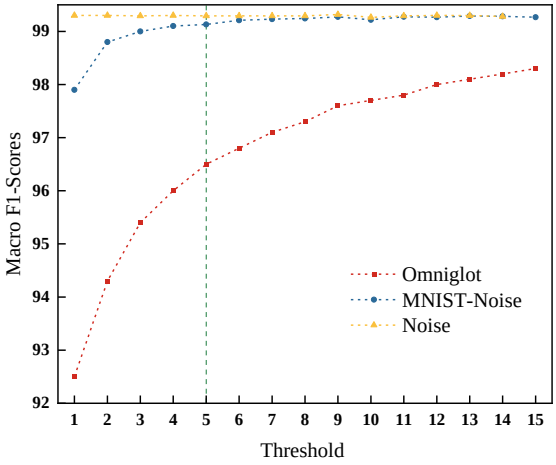
We compare the results of experiments including K+1 stereotype, DCTAU (w/o DC) and complete DCATU (K+K) for Open Set Recognition on MNIST dataset. The results shown in Fig. 2 validates the effectiveness of K+K pattern and our DCTAU from the visualization.



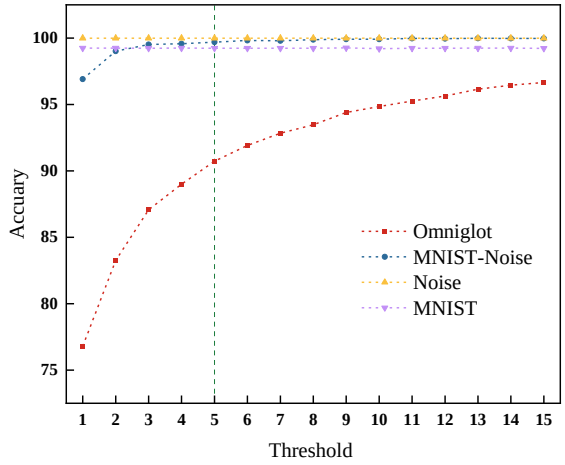
(a) Macro F1-Scores in ID: CIFAR10 Setting



(b) Accuracy in ID: CIFAR10 Setting

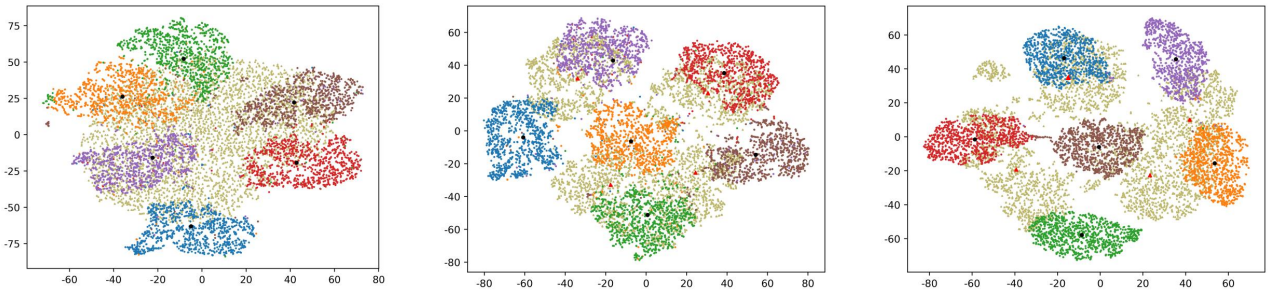


(c) Macro F1-Scores in ID: MNIST Setting



(d) Accuracy in ID: MNIST Setting

Figure 1: Ablation of Hyper-parameters ϵ.



(a) (b) (c)

Figure 2: Visualizations of (a) K+1 stereotype, (b) DCTAU (w/o DC) and DCATU (K+K). Unknown clasees are the blackish green point. The black circles are the centers of known classes and the red triangles are the centers of unknown classes.

References

Dhamija, A. R.; Günther, M.; and Boult, T. 2018. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31.

Dietterich, T. G.; and Guyer, A. 2022. The familiarity hypothesis: Explaining the behavior of deep open set methods. *Pattern Recognition*, 132: 108931.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern recognition letters*, 27(8): 861–874.

Xu, B.; Shen, F.; and Zhao, J. 2023. Contrastive Open Set Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10546–10556.