Unlocking Better Closed-Set Alignment Based on Neural Collapse for Open-Set Recognition

Chaohua Li[1,2], **Enhao Zhang**[1,2], **Chuanxing Geng**[1,2,3], **Songcan Chen**[1,2*]

[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
[2]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
[3]Department of Computer Science, Hong Kong Baptist University
{chaohuali, zhangeh, gengchuanxing, s.chen}@nuaa.edu.cn

Abstract

In recent Open-set Recognition (OSR) community, a prevailing belief is that enhancing the discriminative boundaries of closed-set classes can improve the robustness of Deep Neural Networks (DNNs) against open data during testing. Typical studies validate this *implicitly* by empirical evidence, without a formalized understanding of *how DNNs help the closed-set features obtain more discriminative boundaries?* For this, we provide an answer from the Neural Collapse (NC) perspective: DNNs align the closed-set with a *Simplex Equiangular Tight Frame* (ETF) structure that has geometric and mathematical interpretability. Regrettably, although NC naturally occurs in DNNs, we discover that typical studies cannot guarantee the features being learned to strictly align with the ETF. Thus, we introduce a novel concept, **Fi**xed ETF **T**emplate (FiT), which holds an ideal structure associated with closed-set classes. To force class means and classifier vectors to align with FiT, we further design a **D**ual **ETF** (DEF) loss involving two components. Specifically, F-DEF loss is designed to align class means with FiT strictly, yielding optimal inter-class separability. Meanwhile, we extend a dual form to classifier vectors, termed C-DEF loss, which guides class means and classifier vectors to satisfy self-duality. Our theoretical analysis proves the validity of the proposed approach, and extensive experiments demonstrate that DEF achieves comparable or superior results with reduced computational resources on standard OSR benchmarks.

Introduction

Traditional classification tasks in computer vision typically follow the *closed-set assumption* (*i.e.,* classes absent in the training phase will not appear during testing). However, since it is impossible to provide models with complete real-world knowledge during training, they tend to misclassify unknown classes (open-set) as known ones (closed-set) during testing, leading to incorrect decisions. To address this issue, a more realistic task, Open-set Recognition (OSR), has been proposed to classify known data and recognize unknown data simultaneously (Scheirer et al. 2012; Geng, Huang, and Chen 2020).

As the research of OSR rapidly proliferates, a pivotal study was proposed by (Vaze et al. 2022). The authors found that open-set performance is positively correlated to closed-set performance. Accordingly, they claim that the essence of enhancing OSR performance lies in proficiently handling the closed-set features during training to attain more discriminative closed-set classification boundaries. This claim has sparked numerous follow-up studies based on Deep Neural Networks (DNNs) and has been effectively validated through experimental results (Cubuk et al. 2020; Zhu et al. 2023; Jiang et al. 2023; Zhang et al. 2023; Wang et al. 2024; Li et al. 2024). However, current *empirical evidence* neglects a formalized understanding of *how DNNs help closed-set features obtain more discriminative boundaries?* In this paper, we intend to compensate for this oversight from the per-

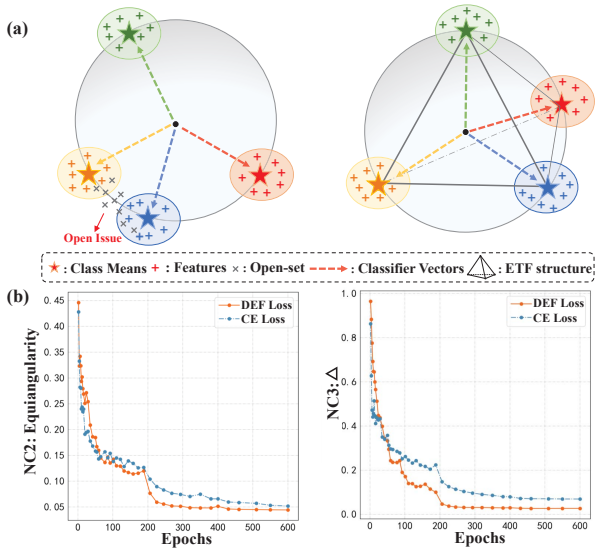

Figure 1: (a) A suboptimal ETF (*left*) leads to an uneven distribution of closed-set features, making open and closed sets inseparable during testing. An ideal ETF (*right*) holds an elegant geometric structure with equal norms and angles between each pair of classes, yielding more discriminative boundaries; (b) Results of NC2 and NC3 on TinyImageNet (*lower is better*) indicate the CE-based method (*(blue)*) fails to satisfy NC properties as effectively as our designed DEF loss (*orange*). The DEF loss forces the alignment of features and the classifiers with the ideal ETF structure.

*Corresponding author.

spective of Neural Collapse (NC), a phenomenon observed in DNNs by (Papyan, Han, and Donoho 2020). It reveals four subtle properties during the Terminal Phase of Training (TPT) (*i.e.,* an ideal stage when the model trains past zero-error), where the last-layer features collapse to their respective class means (NC1). Meanwhile, these class means collapse to the vertices of ***Simplex Equiangular Tight Frame*** (ETF) structure (NC2), and classifiers align with the same ETF structure, termed *self-duality* (NC3), to simplify the classification task (NC4). These properties indicate that the ETF describes an elegant geometric structure with equal norms and equal angles between each pair of classes, which is beneficial in many fields for making the decision regions more symmetrical and stable (Zhu et al. 2021; Peifeng et al. 2023; Zhang et al. 2024). Therefore, it provides *formalized evidence* that *the learned closed-set features existing in a Simplex Equiangular Tight Frame structure can result in better separation*, as depicted on the right of Fig. 1(a).

Unfortunately, although the NC phenomenon naturally exists in DNNs-based OSR methods, our experimental findings show these methods only partially satisfy NC property. As shown in Fig. 1(b), there is a distinct discrepancy between the CE-based method (Vaze et al. 2022) and our proposed DEF during the TPT on metrics of NC2 and NC3 (details of the metrics and additional results can be found in Appendix A). This implies a suboptimal ETF structure where class means fail to attain equal angles and the self-duality property is unsatisfied, as shown on the left of Fig. 1(a).

In response to this dilemma, we introduce a novel concept **Fi**xed ETF **T**emplate (FiT), to serve as a fixed ideal ETF structure derived from the number of class means in the closed set. It holds the optimal structure that we expect class means, features, and classifier vectors to align with. To ensure strict alignment with the proposed FiT, we design a **D**ual **ETF** loss (DEF) involving two components, F-DEF loss and C-DEF loss. Specifically, **(1)** F**-DEF loss** is designed for the feature space. It imposes class means converge to FiT, maintaining equal norms and equal angles between any pairs of class means (*i.e.,* to satisfy NC2). Notably, according to NC1, when class means align with the FiT, features will naturally align with their respective class means. **(2)** C**-DEF loss** is designed for classifier. we extend a dual form of F-DEF to classifiers. It ensures that classifiers align with the FiT in a manner consistent with the class means, thereby satisfying the *self-duality* property in NC3. Ultimately, DEF enables to achieve optimal inter-class separation and enforces more discriminative boundaries for closed set. The contributions are summarized as follows:

- From the view of Neural Collapse, we explicitly point out that learned closed-set features existing in a Simplex Equiangular Tight Frame structure can help DNNs achieve discriminative boundaries.
- Utilizing the properties of NC, we design a Dual ETF loss to ensure that the class means, features, and classifier vectors of the closed set align with the ideal ETF.
- Comprehensively experimental results indicate that DEF provides comparable or superior performance with less computational resources on standard OSR benchmarks.

Related Work

Open-set Recognition

Open set recognition was first formalized by (Scheirer et al. 2012). (Bendale and Boult 2016) later integrated deep learning into OSR and proposed OpenMax. Recently, (Vaze et al. 2022) presented a seminal paper sating that *achieving a good closed-set classifier is what OSR needs* in deep learning. We broadly categorize the recent works closely related to this paper into two categories:

Pure Closed-set Based Methods. This category focuses on training solely with pure closed-set data.. (Vaze et al. 2022) employed diverse deep learning training strategies such as longer training, better augmentations (Cubuk et al. 2020), label smoothing (Szegedy et al. 2016) as well as changing the open-set scoring rule. Subsequently, more advancing training strategies were utilized, including Vision-Language Model (Ming et al. 2022), spatial features (Liu et al. 2022), attention mechanism (Zhang et al. 2023) and multiple experts (Wang et al. 2024).

Pseudo Open-set Based Methods. The core of this category is to introduce pseudo open-set data during training to compress the regions of closed-set. It can be mainly classified into GANs-based and Augmentation-based methods. For GANs, a foundational study (Ge et al. 2017) was proposed, where a conditional GAN was trained to generate pseudo open-set data. Subsequent studies such as (Neal et al. 2018; Perera et al. 2020; Chen et al. 2021) employed various modifications to GANs to address OSR. For Augmentation, one of the major technologies employed is Mixup (Zhang et al. 2017; Verma et al. 2019), which inspired a series of studies (Zhou, Ye, and Zhan 2021; Koch, Riess, and Köhler 2023; Zhu et al. 2023; Jiang et al. 2023; Li et al. 2024).

However, despite advancements in learning closed-set classes using various techniques, existing studies lack to understand *how DNNs help closed-set features obtain more discriminative boundaries* through the formalized evidence of DNNs, which is the focus of our paper.

Neural Collapse

During the TPT, NC maintains an elegant geometric structure for class means, feature activations, and classifier vectors, achieving optimal class separability (Papyan, Han, and Donoho 2020). The NC phenomenon has applications in Representation Learning (Zhu et al. 2021; Tirer and Bruna 2022; Graf et al. 2021), Generalization and Transferability (Kothapalli 2022; Wang et al. 2023), Semi-Supervised Learning (Xiao et al. 2024) and Graph Learning (Kothapalli, Tirer, and Bruna 2024). Besides, NC has been widely applied in Imbalanced Learning to address *minority collapse* (Fang et al. 2021). They mainly concentrate on ensuring the classifiers align with the ETF structure (Yang et al. 2022; Peifeng et al. 2023; Zhang et al. 2024).

The most related works to ours are the recent study (Haas, Yolland, and Rabus 2022) and NECO (Ammar et al. 2023). (Haas, Yolland, and Rabus 2022) demonstrated substantially improved performance for open tasks by implementing L2 normalization to features. However, it only satisfies the equinormality property, ignoring other properties like

equiangular and self-duality. NECO empirically validated a novel property NC5 in the presence of OOD data. But it focuses on designing a measure score for unknown classes at the Logit aspect, without constraining the feature and classifier aspects. To the best of our knowledge, we are the first to design a method addressing class mean, feature, and classifier aspects to satisfy the properties of NC in OSR.

Preliminaries and Notations

Properties of Neural Collapse

Definition 1 (**Simplex Equiangular Tight Frame (ETF)**). *A collection of vectors $\boldsymbol{v_c} \in \mathbb{R}^d, c = 1, 2, ..., C$ and $d \geq C - 1$ is said to be a Simplex Equiangular Tight Frame if:*

$$\boldsymbol{V} = \sqrt{\frac{C}{C-1}}\boldsymbol{U}\left(\boldsymbol{I_C} - \frac{1}{C}\boldsymbol{1_C}\boldsymbol{1_C^T}\right), \tag{1}$$

where $\boldsymbol{V} = [\boldsymbol{v_1}, \cdots, \boldsymbol{v_C}] \in \mathbb{R}^{d\times C}, \boldsymbol{U} \in \mathbb{R}^{d\times C}$ allows a rotation and satisfies $\boldsymbol{U^T U} = \boldsymbol{I_C}$, $\boldsymbol{I_C}$ is the identity matrix and $\boldsymbol{1_C}$ is an all-ones vector.

Then, all vectors in the Simplex ETF obeys for $\forall i, j \in [1, C]$,

$$\|\boldsymbol{v}_i - \boldsymbol{v}_G\|_2 = \|\boldsymbol{v}_j - \boldsymbol{v}_G\|_2, \tag{2}$$

$$\boldsymbol{v_i}^T\boldsymbol{v_j} = \frac{C}{C-1}\delta_{i,j} - \frac{1}{C-1} = \begin{cases} 1, & i = j \\ -\frac{1}{C-1}, & i \neq j \end{cases}, \tag{3}$$

where $\boldsymbol{v}_G$ is the global mean. $\delta_{i,j}$ is the Kronecker delta symbol.

Consequently, the NC phenomenon can be characterized by four subtle properties (Papyan, Han, and Donoho 2020):

1. **(NC1) Variability Collapse:** the intra-class variance collapses to zero as these feature activations collapse to their respective class means:

$$\Sigma_W \to 0, \tag{4}$$

where $\boldsymbol{\Sigma}_W \triangleq \underset{i,c}{\text{Ave}}\left\{(\boldsymbol{z}_{i,c} - \boldsymbol{\mu}_c)(\boldsymbol{z}_{i,c} - \boldsymbol{\mu}_c)^\top\right\}$, $\boldsymbol{\mu}_c \triangleq \underset{i}{\text{Ave}}\{\boldsymbol{z}_{i,c}\}$, $c = 1, 2, \cdots, C$, is the mean of the $c-$th class. $\boldsymbol{z}_{i,c}$ is the feature activations of the $i-$th sample in the $c-$th class in the last layer.

2. **(NC2) Convergence to Simplex Equiangular Tight Frame (ETF):** all class means centered by the global mean will converge to having equal lengths, as well as having equal-sized angles between any pair of class means. This structure is well-studied as the Simplex ETF. *i.e.,* $\{\tilde{\boldsymbol{\mu}}_c\}_{c=1}^C$ satisfy Eq. (2) and Eq. (3), where $\tilde{\boldsymbol{\mu}}_c \triangleq (\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)/\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|_2, \boldsymbol{\mu}_G \triangleq \underset{i,c}{\text{Ave}}\{\boldsymbol{z}_{i,c}\}$.

3. **(NC3) Convergence to self-duality:** the class means and linear classifiers converge towards each other, up to rescaling:

$$\Delta = \left\| \frac{\boldsymbol{W}^\top}{\|\boldsymbol{W}\|_F} - \frac{\dot{\boldsymbol{M}}}{\|\dot{\boldsymbol{M}}\|_F} \right\|_F \to 0, \tag{5}$$

where $\dot{\boldsymbol{M}} = [\boldsymbol{\mu}_c - \boldsymbol{\mu}_G]$, $\boldsymbol{W} = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_C]$ presents the classifier vector matrix. This property is called self-duality, where the decision regions become geometrically symmetry and the class means are centered in their regions, implying the maximum separation.

4. **(NC4) Simplification to nearest class center:** the network classifiers converge to selecting the class with the nearest training class mean:

$$\arg\max_{c'} \langle \boldsymbol{\omega}_{c'}, \boldsymbol{z_t} \rangle + b_{c'} \to \arg\min_{c'} \|\boldsymbol{z_t} - \boldsymbol{\mu}_{c'}\|_2, \tag{6}$$

where b signifies a bias term and $\mathbf{z_t}$ represents the last-layer feature of a sample utilized for prediction.

Problem Statement

We denote the training set as $\mathcal{D}_t$ consisting closed-set data $\{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{i=1}^n$, where $\boldsymbol{y_i} \in \{1, ..., C\}$ is their respective label. In the training phase, we adopt a two-stage training strategy. Specifically, in the first stage, we train a feature extraction comprising an encoder network $E(\cdot)$ and a projection network $\psi(\cdot)$. Both networks are optimized by supervised contrastive learning loss and our designed F-DEF loss on features of $\psi(E(\boldsymbol{x_i}))$. In the second stage, the learned $E(\cdot)$ and $\psi(\cdot)$ are frozen and used to train a linear classifier $f(\cdot)$ on the combination of cross entropy loss and our designed C-DEF loss.

Additionally, $\mathcal{D}_{te}$ denotes the test set, containing data $\boldsymbol{t_i}$ from both closed-set and open-set classes. In the testing phase, our goal is to correctly classify each $\boldsymbol{t_i}$ to a unique closed-set class if its class is present in $\mathcal{D}_t$, or recognize it as open-set data if its class is absent during training. In the next section, we will elaborate on our primary work: designing the DEF loss to better align class means, features, and classifiers with the ETF structure.

Methodology

Fixed ETF Template

The Fixed ETF Template (FiT) is associated with the number of closed-set classes and exhibits an ideal structure which we expect class means, features and classifier vectors to align with.

Definition 2 (**Fixed ETF Template**). *Given a training set $\mathcal{D}_t$ with C closed-set classes, the Fixed ETF Template presents an ideal ETF structure for all $\boldsymbol{v}$ (it can be class means $\tilde{\boldsymbol{\mu}}$, features $\boldsymbol{z}$ or classifier vectors $\tilde{\boldsymbol{\omega}}$) maintaining equal norms and the same pair-wise angles:*

$$v_{norm} = \frac{v}{max(\| v \|_2, \epsilon)}, \tag{7}$$

where ϵ is added in the event of the zero norm.

$$\theta_{ij} = \arccos\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \arccos\left(-\frac{1}{C-1}\right) \quad \forall i, j \in [1, C], \tag{8}$$

where i, j is the index of closed-set classes.

Reflecting on our ultimate goal of obtaining discriminative boundaries, aligning class means, features, and classifier vectors with FiT, which maintains equal norms and

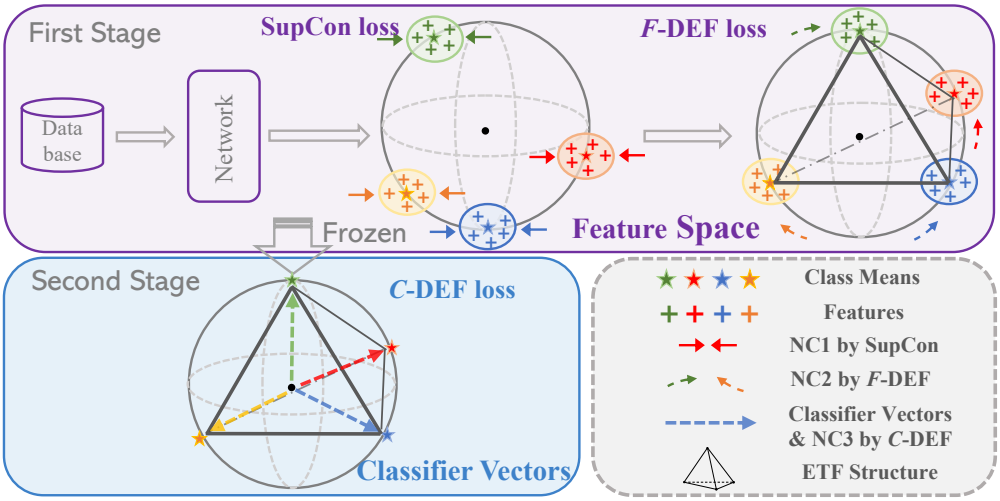

Figure 2: The pipeline of **D**ual **ETF** (DEF). In the first stage, we extract features using Supervised Contrastive Learning loss (SupCon) to minimize intra-class variance (NC1), and align the features with their corresponding class means with FiT (NC2) through F-DEF loss (*top*). In the second stage, we use the trained network to optimize a linear classifier with Cross Entropy loss combined with C-DEF loss (*below*), satisfying the self-duality property (NC3). Pseudo-code are shown in Appendix F.

an equiangular structure, will maximize and stabilize inter-class separation. In the next part, we will explain how to align these components with FiT by designing losses for class means and classifier vectors aspects.

Dual Simplex Equiangular Tight Frame Loss

According to NC1, intra-class variance will collapse to zero during the TPT phase, indicating that individual features collapse to their class mean centers. Thus, to drive NC1, we select supervised contrastive learning (Khosla et al. 2020), an effective method for feature representation that reduces intra-class variance, which has been proven effective in OSR (Xu, Shen, and Zhao 2023; Li et al. 2024). The supervised contrastive loss is formulated as:

$$\mathcal{L}_{sup} = \sum_{i} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau\right)}{\sum_{k \neq i} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_k / \tau\right)}, \quad (9)$$

where $\boldsymbol{z}_i = \psi(E(\boldsymbol{x_i}))$, $\boldsymbol{x_i} \in \mathcal{D}_t$. $P(i)$ is the set of all positive data from class i, and $|P(i)|$ is its cardinality. τ is the scalar temperature hyper-parameter.

It is important to note that, due to the assurance provided by NC1, we can align both class means $\tilde{\boldsymbol{\mu}}$ and features $\boldsymbol{z}$ with FiT in the feature space by designing a loss solely for $\tilde{\boldsymbol{\mu}}$. Next, we will detail the designed DEF loss, which consists of two components, F-DEF loss and C-DEF loss.

F-DEF Loss. In the feature space, we focus on designing a loss for $\tilde{\boldsymbol{\mu}}$ to align it with the ideal ETF structure. We can easily enforce equal norms using ℓ_2 normalization. Subsequently, our primary goal is to ensure equiangularity. We denote the loss for class means as F-DEF loss, formulated as follows:

$$\mathcal{L}_F(\tilde{\boldsymbol{\mu}}) = \mathbb{E}_{i,j \sim U(1,C)} \left[(\tilde{\boldsymbol{\mu}}_i^T \tilde{\boldsymbol{\mu}}_j - \theta_{ij})^2 \right], \quad (10)$$

where $\tilde{\boldsymbol{\mu}}_c \in \{\tilde{\boldsymbol{\mu}}_1, ..., \tilde{\boldsymbol{\mu}}_C\}$ denotes the class means of closed-set classes, θ_{ij} represent the expected angle between $\forall i, j$ and $U(1, C)$ denotes a uniform distribution.

This loss minimizes the the discrepancy between the inner products of two class means $\tilde{\boldsymbol{\mu}}_i$ and $\tilde{\boldsymbol{\mu}}_j$, and the target θ_{ij}. It pushes $\tilde{\boldsymbol{\mu}}_i$ and $\tilde{\boldsymbol{\mu}}_j$ to align with FiT. When $\tilde{\boldsymbol{\mu}}_i^T \tilde{\boldsymbol{\mu}}_j = \theta_{ij}$, $\mathcal{L}_F(\tilde{\boldsymbol{\mu}}) = 0$ indicates class means converge to the optimal ETF structure along with their respective features.

Finally, in the first stage of training (*i.e.,* training an encoder network $E(\cdot)$ and a projection network $\psi(\cdot)$), the overall loss can be described as:

$$\mathcal{L}_{feature} = \alpha \mathcal{L}_{sup} + (1 - \alpha) \mathcal{L}_F(\tilde{\boldsymbol{\mu}}), \quad (11)$$

where $\alpha \in [0, 1]$ is sampled from the Beta distribution.

C-DEF Loss. For the classifiers, we can easily extend the form of F-DEF loss to:

$$\mathcal{L}_C(\tilde{\boldsymbol{\omega}}) = \mathbb{E}_{i,j \sim U(1,C)} \left[(\tilde{\boldsymbol{\omega}}_i^T \tilde{\boldsymbol{\omega}}_j - \theta_{ij})^2 \right], \quad (12)$$

where $\tilde{\boldsymbol{\omega}}_c \in \{\tilde{\boldsymbol{\omega}}_1, ..., \tilde{\boldsymbol{\omega}}_C\}$ denotes the weights of classifier vectors.

According to NC3, if $\mathcal{L}_F$ and $\mathcal{L}_C$ converge to the optimal solution (*i.e.,* $\tilde{\boldsymbol{\mu}}_i^T \tilde{\boldsymbol{\mu}}_j = \theta_{ij} = \tilde{\boldsymbol{\omega}}_i^T \tilde{\boldsymbol{\omega}}_j$), the *self-duality* property is satisfied.

Similarly, in the second stage of training (*i.e.,* training a linear classifier $f(\cdot)$), the total loss can be described as:

$$\mathcal{L}_{classifier} = \beta \mathcal{L}_{ce} + (1 - \beta) \mathcal{L}_C(\tilde{\boldsymbol{\omega}}), \quad (13)$$

where $\beta \in [0, 1]$ is sampled from the Beta distribution. The pipeline of our DEF can be seen in Fig. 2.

Theoretical Analysis

In this subsection, we will conduct theoretical analyses from both *Closed-set Inter-class Margin* and *Open-set Misclassification Probability* perspectives.

Dataset	SVHN	CIFAR10	CIFAR+10	CIFAR+50	TinyImageNet	Average
Softmax	88.6	67.7	81.6	80.5	57.7	75.2
OpenMax	89.4	69.5	81.7	79.6	57.6	75.6
G-OpenMax	89.6	67.5	82.7	81.9	58.0	75.9
OSRCI	91.0	69.9	83.8	82.7	58.6	77.2
CPN	92.6	82.8	88.1	87.9	63.9	83.0
GFROSR	93.5	80.7	92.8	92.6	60.8	84.1
PROSER	94.3	89.1	96.0	95.3	69.3	88.8
OpenHybrid	94.7	95.0	96.2	95.5	79.3	92.1
ARPL	96.7	91.0	97.1	95.1	78.2	91.0
Class-inclusion	95.6	94.8	96.1	95.7	78.5	92.1
PMAL	96.3	94.6	96.0	94.3	81.8	92.6
MLS	97.1	93.6	97.9	96.5	83.0	93.6
ConOSR	<u>99.1</u>	94.2	98.1	97.3	80.9	93.9
DCTAU	**99.2**	<u>95.6</u>	<u>98.5</u>	<u>98.1</u>	<u>83.6</u>	<u>95.0</u>
Vanilla SupCon	98.8	93.7	97.9	97.0	79.6	93.4
DEF	<u>99.1</u>	**95.7**	**98.9**	**98.4**	**85.6**	**95.5**

Table 1: Open Set recognition results in terms of the AUROC(%). The best method is emphasized in **bold**, and the <u>underlined</u> represents the second best result.

Closed-set Inter-class Margin. Under the ETF structure, it ensures class means μ_i and μ_j have equal angles and norms, the distance between them can be described as:

$$\|\boldsymbol{\mu_i} - \boldsymbol{\mu_j}\| = \sqrt{2(\|\boldsymbol{\mu}\|^2 + \frac{1}{C-1})}, \tag{14}$$

Because C is a constant value presenting the number of closed-set classes, the closed-set inter-class margin between $\boldsymbol{\mu_i}$ and $\boldsymbol{\mu_j}$ is fixed. This minimizes inter-class overlap, resulting in better closed-set classification performance.

Open-set Misclassification Probability. We are interested in the probability that an open data falls into closed-set classification region. From Eq. (3), the symmetric geometry of the ETF structure makes it simple to find that the shortest distance L from open data to any $\boldsymbol{\mu_i}$ and $\boldsymbol{\mu_j}$ is $\frac{\|\boldsymbol{\mu_i}-\boldsymbol{\mu_j}\|}{2}$. Then we introduce the Open-set Misclassification Probability as follow:

Proposition 1. *Let $\boldsymbol{v} \sim \mathcal{N}(0, I)$ be the feature vector of an open-set data drawn from a standard normal distribution, the misclassification probability can be expressed as:*

$$P_{\text{open}} = \frac{1}{2}\left[1 - \text{erf}\left(\frac{(\|\boldsymbol{\mu}\|^2 + \frac{1}{C+1})^{-1/2}}{2}\right)\right], \tag{15}$$

where $\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\text{d}t$ represents the error function. Since C is fixed, P_{open} indicates that an open-set feature falling into erroneous classification regions is controllable. The derivation can be found in the Appendix B.

Rejecting Open-set Data

During testing, the max posterior probability is treated as the detection score, $\max_{c\in\{1,\cdots,C\}} P(y=c|t_i)$, where $t_i \in \mathcal{D}_{te}$. It can be categorized into one of the closed-set classes or recognized as an open-set data by:

$$\widehat{y} = \begin{cases} \arg\max_{c\in\{1,\cdots,C\}} P(y=c|\boldsymbol{t_i}), & \text{if } conf \geq \varepsilon, \\ \text{open-set data}, & \text{otherwise.} \end{cases} \tag{16}$$

Experiments & Results

Open-Set Recognition

Datasets. Following the protocol defined in (Neal et al. 2018) and the dataset splits specified in (Vaze et al. 2022; Li et al. 2024), we provide a summary of six standard benchmark datasets in OSR:

- **SVHN,CIFAR10.** SVHN(Netzer et al. 2011) and CIFAR10(Krizhevsky, Hinton et al. 2009) all contain 10 classes, with 6 classes randomly selected as closed set and the other 4 as open set.
- **CIFAR+10,CIFAR+50.** For these experiments, 4 classes from CIFAR-10 are selected as closed-set classes, while $10\backslash 50$ classes from CIFAR-100(Krizhevsky, Hinton et al. 2009) are used as open-set classes.
- **TinyImageNet.** TinyImageNet is a subset derived from ImageNet(Russakovsky et al. 2015) consisting of 200 classes. 20 known classes and the left 180 unknown classes are randomly sampled for evaluation.

Evaluation Metrics. Following (Chen et al. 2021; Li et al. 2024), we evaluate performance using the Area Under the Receiver Operating Characteristic (AUROC) curve (Fawcett 2006). We also present the Open Set Classification Rate (OSCR) curve (Dhamija, Günther, and Boult 2018), which integrates the evaluation for correct classifications of closed-set classes. To estimate the area under the OSCR curve, we adopt a numeric version called OpenAUC (Wang et al. 2022) in this paper. Details are shown in Appendix C.

Baselines. The baselines include Softmax Thresholding(Hendrycks and Gimpel 2016), OpenMax(Bendale and Boult 2016), GOpenMax(Ge et al. 2017), OSRCI(Neal et al. 2018), CPN(Yang et al. 2020), GFROSR(Perera et al. 2020), PROSER(Zhou, Ye, and Zhan 2021), OpenHybrid(Zhang et al. 2020), ARPL(Chen et al. 2021), Class-inclusion(Cho and Choo 2022), PMAL(Lu et al. 2022), MLS(Vaze et al.

Dataset	SVHN	CIFAR10	CIFAR+10	CIFAR+50	TinyImageNet	Average
Softmax	92.8	83.8	90.9	88.5	60.8	83.4
GCPL	93.4	84.3	91.0	88.3	59.3	83.3
RPL	93.6	85.2	91.8	89.6	53.2	82.7
ARPL	94.0	86.6	93.5	91.6	62.3	85.6
ARPL+CS	94.3	87.9	94.7	92.9	65.9	87.1
Class-inclusion	85.4	87.0	88.1	86.5	49.3	79.3
DCTAU	**96.2**	93.9	97.2	97.1	77.6	92.4
Vanilla SupCon	95.3	92.6	96.6	95.9	74.3	91.0
DEF	96.0	**94.1**	**97.7**	**97.3**	**79.8**	**93.0**

Table 2: The open set classification rate OSCR(%) results of open set recognition. The best method is emphasized in **bold**, and the underlined represents the second best result.

Dataset	SVHN	CIFAR10	CIFAR+N	TinyIN
DCTAU	5.11	4.61	2.75	5.85
DEF	**3.45**	**1.51**	**1.08**	**2.53**

Table 3: The runtime consumption (hours) of training DCTAU and our DEF for 600 epochs across four datasets. 'TinyIN' presents TinyImageNet dataset.

2022), ConOSR(Xu, Shen, and Zhao 2023) and (Li et al. 2024). The implementations can be seen in Appendix C.

Results Comparison. We provide the results of AUROC in Table 1. DEF outperforms nearly all other methods, showing notable superiority. In particular, compared to other Contrastive-based methods, our DEF surpasses the recently proposed DCTAU on four out of five datasets, with an average improvement of 0.5%. For the more challenging task TinyImageNet, DEF shows increases of **2.0%**. Results for OSCR are displayed in Table 2. Compared to the second best, DEF improves by **2.2%** on TinyImageNet, with an average increase of 0.6% across all datasets.

Some may worry that our improvements over DCTAU are modest on certain datasets, and even slightly lower on SVHN. It is important to note that DCTAU generates pseudo-unknown classes during training and requires an extra contrastive learning loss, leading to significantly higher time consumption. Therefore, as shown in Table 3, we have **considerably reduced the time cost** compared to DCTAU. Additionally, we performed comparative experiments to analyze methods based on the contrastive learning framework, as detailed in Appendix C.

Familiarity Analysis

In this group of experiments, we follow the protocol outlined in (Yoshihashi et al. 2019; Zhou, Ye, and Zhan 2021; Li et al. 2024) to illustrate the *Familiarity Hypothesis* (Dietterich and Guyer 2022), which states that the intrinsic mechanism of OSR primarily lies in detecting the absence of familiar features in the images. This implies that the detection difficulty is positively correlated with the familiarity between closed-set and open-set. We measure the performance of absence detection using macro-averaged F1-scores.

Dataset	Omniglot	MNIST-Noise	Noise
Softmax	59.5	64.1	82.9
OpenMax	68.0	72.0	82.6
CROSR	79.3	82.7	82.6
PROSER	86.2	87.4	88.2
ConSOR	95.4	98.7	98.8
DEF	**97.9**	**99.3**	**99.4**

Table 4: Familiarity Analysis on *Semantic Quality*. We report macro F1-scores.

Dataset	ImageNet (Crop)	ImageNet (Resize)	LSUN (Crop)	LSUN (Resize)
Softmax	63.9	65.3	64.2	64.7
OpenMax	66.0	68.4	65.7	66.8
CROSR	72.1	73.5	72.0	74.9
PROSER	84.9	82.4	86.7	85.6
ConOSR	89.1	84.3	91.2	88.1
DEF	**94.2**	**93.8**	**94.3**	**94.3**

Table 5: Familiarity Analysis on *Integrity Quality*. We report macro F1-scores.

Datasets. The experiments involves two settings: 1) MNIST(Lake, Salakhutdinov, and Tenenbaum 2015) serves as the closed-set for training, while Omniglot (Lake, Salakhutdinov, and Tenenbaum 2015), MNIST-Noise and Noise are used as open-set during testing. 2) CIFAR10 is used as the closed-set dataset, with open-set data sampled from ImageNet and LSUN(Yu et al. 2015). More details about the datasets are provided in Appendix D.

Results Comparison. *Semantic Quality:* Table 4 shows results for varying semantic quality of open-set data (*i.e.,* from highest to lowest semantic quality: Omniglot, MNIST-Noise, and Noise. Higher semantic quality increases detection difficulty because it contains more familiar information with MNIST.). Our method significantly outperforms others. It achieves 99.3% and 99.4% for detecting noisy open-set images, while 97.9% in Omniglot dataset for detecting se-

Figure 3: The results of θ_{ij} on four datasets during training. The θ_{ij} for these four datasets are $109°$ (*purple*), $101°$, $101°$ (*black*) and $93°$ (*blue*), respectively. The values of (i, j) on the confusion matrix represent actual angels.

Ablation	CIFAR10		TinyImageNet	
	AUROC	OSCR	AUROC	OSCR
w/o F-DEF	94.03	92.82	81.72	77.83
w/o C-DEF	95.55	94.09	84.95	79.12
Ours	**95.69**	**94.13**	**85.60**	**79.76**

Table 6: Ablation results about different employments of the designed loss on CIFAR10 and TinyImageNet.

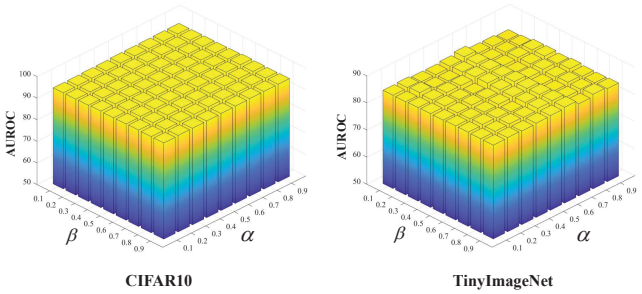

Figure 4: Sensitivity analysis about α and β for CIFAR10 and TinyImageNet datasets.

mantic open-set images. *Integrity Quality:* Table 5 presents results for varying image structures of open-set data (*i.e.,* X-Resize maintains the integrity of images, while X-Crop disrupts it, introducing more unfamiliar features relative to CIFAR10). The results show that DEF also performs best in this scenario across various datasets. In summary, our DEF can reinforce the grasp of detecting the absence of familiar features in closed-set data.

Detailed Analysis

Benefits of DEF Loss. To highlight the benefits of designed DEF loss, we conduct a series of ablation experiments by employing various loss strategies for model training. Results for the vanilla SupCon (*w/o DEF*) are shown in Table 1 and Table 2. In Table 6, the results of *w/o F-DEF* being lower than *w/o C-DEF* indicate that imposing constraints on the feature space to align it with the ETF structure is crucial for training. This provides practical evidence for a future research on *whether handling NC2 is adequate for NC?* Overall, it is obvious that the significant performance gaps between our DEF and other methods indicate the importance of aligning class means, features, and classifier vectors with the ETF structure.

Alignment to FiT. To further investigate the degree of alignment between features and FiT through DEF loss, we visualized θ_{ij} in Fig 3. Since the number of closed-set C for CIFAR+X, CIFAR10, SVHN and TinyImageNet is 4, 6, 6 and 20 respectively, the corresponding θ_{ij} values according to Eq. (8) are $109°$, $101°$, $101°$ and $93°$. Results from the confusion matrix demonstrate that our F-DEF aligns feature angles more closely with the ideal ETF structure, improving the discriminative power of classification boundaries. The visualization of alignment performance for C-DEF is provided in Appendix E.

Analysis of Hyper-parameters. (1)About α and β. As shown in Figure 4, our method is relatively insensitive to the choice of α and β. **(2)Threshold ϵ.** We carried out the analysis on ϵ, and the results are provided in Appendix E.

Conclusion

Building upon a prevailing belief in OSR that improving the discriminative boundaries of closed-set classes enhances open-set recognition, this paper advances an in-depth exploration of this belief with a formalized framework of Neural Collapse in Deep Neural Networks. Our experiments reveal that existing DNNs-based OSR methods only partially capture the properties of this natural phenomenon and finally result in falling to a suboptimal ETF structure. To address this issue, we introduce a novel concept, the Fixed ETF Template (FiT), which represents an ideal structure for closed-set classes to guide the training process. Based on this concept, we develop the Dual ETF loss (DEF) including F-DEF and C-DEF to force class means/features and classifier vectors to align strictly with the proposed FiT. We provide a theoretical analysis of how an ideal ETF structure benefits both closed-set classification and open-set recognition. Extensive experiments across various benchmarks demonstrate that our DEF achieves comparable or superior results while reducing computational resources.

Acknowledgments

This work was Supported by the National Natural Science Foundation of China (Grant No. 62376126, 62106102), the Hong Kong Scholars Program (Grant No. XJ2023035) and the Fundamental Research Funds for the Central Universities (Grant No. NS2024058).

References

Ammar, M. B.; Belkhir, N.; Popescu, S.; Manzanera, A.; and Franchi, G. 2023. NECO: NEural Collapse Based Out-of-distribution detection. *arXiv preprint arXiv:2310.06823*.

Bendale, A.; and Boult, T. E. 2016. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1563–1572.

Chen, G.; Peng, P.; Wang, X.; and Tian, Y. 2021. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 8065–8081.

Cho, W.; and Choo, J. 2022. Towards accurate open-set recognition via background-class regularization. In *European Conference on Computer Vision*, 658–674. Springer.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.

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.

Fang, C.; He, H.; Long, Q.; and Su, W. J. 2021. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43): e2103091118.

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

Ge, Z.; Demyanov, S.; Chen, Z.; and Garnavi, R. 2017. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*.

Geng, C.; Huang, S.-j.; and Chen, S. 2020. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3614–3631.

Graf, F.; Hofer, C.; Niethammer, M.; and Kwitt, R. 2021. Dissecting supervised contrastive learning. In *International Conference on Machine Learning*, 3821–3830. PMLR.

Haas, J.; Yolland, W.; and Rabus, B. 2022. Linking neural collapse and l2 normalization with improved out-of-distribution detection in deep neural networks. *arXiv preprint arXiv:2209.08378*.

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

Jiang, G.; Zhu, P.; Wang, Y.; and Hu, Q. 2023. Openmix+: Revisiting data augmentation for open set recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(11): 6777–6787.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.

Koch, T.; Riess, C.; and Köhler, T. 2023. LORD: Leveraging Open-Set Recognition with Unknown Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4386–4396.

Kothapalli, V. 2022. Neural collapse: A review on modelling principles and generalization. *arXiv preprint arXiv:2206.04041*.

Kothapalli, V.; Tirer, T.; and Bruna, J. 2024. A neural collapse perspective on feature evolution in graph neural networks. *Advances in Neural Information Processing Systems*, 36.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338.

Li, C.; Zhang, E.; Geng, C.; and Chen, S. 2024. All Beings Are Equal in Open Set Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 13446–13454.

Liu, Z.-g.; Fu, Y.-m.; Pan, Q.; and Zhang, Z.-w. 2022. Orientational distribution learning with hierarchical spatial attention for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 8757–8772.

Lu, J.; Xu, Y.; Li, H.; Cheng, Z.; and Niu, Y. 2022. Pmal: Open set recognition via robust prototype mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1872–1880.

Ming, Y.; Cai, Z.; Gu, J.; Sun, Y.; Li, W.; and Li, Y. 2022. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35: 35087–35102.

Neal, L.; Olson, M.; Fern, X.; Wong, W.-K.; and Li, F. 2018. Open set learning with counterfactual images. In *Proceedings of the European conference on computer vision (ECCV)*, 613–628.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.

Papyan, V.; Han, X.; and Donoho, D. L. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663.

Peifeng, G.; Xu, Q.; Wen, P.; Yang, Z.; Shao, H.; and Huang, Q. 2023. Feature directions matter: Long-tailed learning via rotated balanced representation. In *International Conference on Machine Learning*, 27542–27563. PMLR.

Perera, P.; Morariu, V. I.; Jain, R.; Manjunatha, V.; Wigington, C.; Ordonez, V.; and Patel, V. M. 2020. Generative-discriminative feature representations for open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11814–11823.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Scheirer, W. J.; de Rezende Rocha, A.; Sapkota, A.; and Boult, T. E. 2012. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1757–1772.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tirer, T.; and Bruna, J. 2022. Extended unconstrained features model for exploring deep neural collapse. In *International Conference on Machine Learning*, 21478–21505. PMLR.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Open-Set Recognition: a Good Closed-Set Classifier is All You Need? In *International Conference on Learning Representations*.

Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, 6438–6447. PMLR.

Wang, Y.; Mu, J.; Zhu, P.; and Hu, Q. 2024. Exploring diverse representations for open set recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5731–5739.

Wang, Z.; Luo, Y.; Zheng, L.; Huang, Z.; and Baktashmotlagh, M. 2023. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5549–5558.

Wang, Z.; Xu, Q.; Yang, Z.; He, Y.; Cao, X.; and Huang, Q. 2022. OpenAUC: Towards AUC-Oriented Open-Set Recognition. *Advances in Neural Information Processing Systems*, 35: 25033–25045.

Xiao, R.; Feng, L.; Tang, K.; Zhao, J.; Li, Y.; Chen, G.; and Wang, H. 2024. Targeted Representation Alignment for Open-World Semi-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23072–23082.

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.

Yang, H.-M.; Zhang, X.-Y.; Yin, F.; Yang, Q.; and Liu, C.-L. 2020. Convolutional prototype network for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5): 2358–2370.

Yang, Y.; Chen, S.; Li, X.; Xie, L.; Lin, Z.; and Tao, D. 2022. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? *Advances in neural information processing systems*, 35: 37991–38002.

Yoshihashi, R.; Shao, W.; Kawakami, R.; You, S.; Iida, M.; and Naemura, T. 2019. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4016–4025.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Zhang, E.; Geng, C.; Li, C.; and Chen, S. 2024. Dynamic Learnable Logit Adjustment for Long-Tailed Visual Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, H.; Li, A.; Guo, J.; and Guo, Y. 2020. Hybrid models for open set recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 102–117. Springer.

Zhang, J.; Gao, L.; Hao, B.; Huang, H.; Song, J.; and Shen, H. 2023. From global to local: Multi-scale out-of-distribution detection. *IEEE Transactions on Image Processing*.

Zhou, D.-W.; Ye, H.-J.; and Zhan, D.-C. 2021. Learning placeholders for open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Zhu, F.; Cheng, Z.; Zhang, X.-Y.; and Liu, C.-L. 2023. Openmix: Exploring outlier samples for misclassification detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12074–12083.

Zhu, Z.; Ding, T.; Zhou, J.; Li, X.; You, C.; Sulam, J.; and Qu, Q. 2021. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34: 29820–29834.