Understanding Open-Set Recognition by Jacobian Norm and Inter-Class Separation

Jaewoo Park[a], Hojin Park[a], Eunju Jeong[b], Andrew Beng Jin Teoh[a]

[a]*Electrical and Electronic Engineering Department, Yonsei University, Seoul, South Korea*
[b]*Woowa Brothers Corp., Seoul, South Korea*

Abstract

The findings on open-set recognition (OSR) show that models trained on classification datasets are capable of detecting unknown classes not encountered during the training process. Specifically, after training, the learned representations of known classes dissociate from the representations of the unknown class, facilitating OSR. In this paper, we investigate this emergent phenomenon by examining the relationship between the Jacobian norm of representations and the inter/intra-class learning dynamics. We provide a theoretical analysis, demonstrating that intra-class learning reduces the Jacobian norm for known class samples, while inter-class learning increases the Jacobian norm for unknown samples, even in the absence of direct exposure to any unknown sample. Overall, the discrepancy in the Jacobian norm between the known and unknown classes enables OSR. Based on this insight, which highlights the pivotal role of inter-class learning, we devise a marginal one-vs-rest (m-OvR) loss function that promotes strong inter-class separation. To further improve OSR performance, we integrate the m-OvR loss with additional strategies that maximize the Jacobian norm disparity. We present comprehensive experimental results that support our theoretical observations and demonstrate the efficacy of our proposed OSR approach.

Keywords:
Open-Set Recognition, Representation Learning, Metric-Learning, Object Classification.

1. Introduction

In recent years, deep neural network (DNN) based models have demonstrated remarkable success in *closed-set recognition*, where the train and test sets share the same categorical classes to classify. In practical environments, however, a deployed model can encounter instances of class categories *unknown* during its training. Detecting these unknown class instances is crucial in safety-critical applications such as autonomous driving and cybersecurity. A solution to this is *open-set recognition* (OSR), where a classifier trained over K known classes can classify them and reject unknown class instances in the test stage [1].

A predominant approach in DNN-based OSR is to train a discriminative model over known classes with a metric-learning loss, and derive a score (or decision) function that captures the difference between the known and unknown in terms of their representations. For the score function to work effectively, the unknown class must be dissociated from the known class in the representation space. Interestingly, [1] along with subsequent works [2, 3, 4] observed that training *over known classes alone* results in this separation; the model separates the unknown class from the known classes even though the model did not utilize any unknown class instance during its training.

However, the underlying mechanism of this phenomenon has rarely been explored in the context of representation learning.

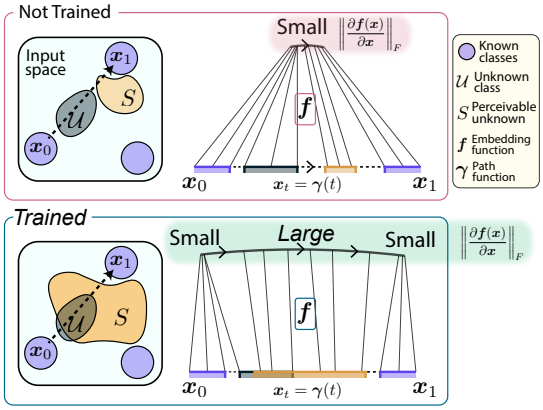

Figure 1: During the closed-set metric learning, the model learns only over the known classes C_k, but the learning also changes the representation of *unknown class*. We ask why. We discover that the intra-class learning diminishes the Jacobian norm of known class representations, while the inter-class learning increases the Jacobian norm of the unknown. The resulting disparity in Jacobian norm separates the unknown from the known.

This work aims to analyze this phenomenon, namely, how the closed-set metric learning separates the unknown class from the known classes in the representation space.

To this end, we analyze the Jacobian norm of representation $\|\frac{\partial f(x)}{\partial x}\|_F$, which is the Frobenius norm of the Jacobian matrix. We discover that inter-class separation learning within known classes plays a crucial role in OSR, as it alters the representations of *unknown* class instances without direct exposure to them. Specifically, inter-class learning elevates the Jacobian

Email addresses: julypraise@yonsei.ac.kr (Jaewoo Park), 2014142100@yonsei.ac.kr (Hojin Park), eunju.jeong@woowahan.com (Eunju Jeong), bjteoh@yonsei.ac.kr (Andrew Beng Jin Teoh)

norm of the unknown, whereas intra-class learning diminishes the Jacobian norm of the known. This resulting disparity between the known and unknown in terms of Jacobian norm leads to a differentiation between their respective representations.

We provide comprehensive theoretical validation for our hypothesis, which is further reinforced by a wealth of empirical evidence. Additionally, inspired by the integral role of inter-class learning in segregating unknown class instances, we develop a marginal one-vs-rest (m-OvR) loss function designed to foster substantial inter-class separation. Furthermore, we incorporate the model loss with auxiliary techniques to enhance the Jacobian norm disparity, ultimately strengthening the distinction between known and unknown classes.

The contributions of our works are summarized as follows:

1. We theoretically show that the closed-set metric learning separates the representations of unknown class from those of the known classes by making their Jacobian norm different. In particular, we discover that the inter-class learning is the key factor in this process as it alters the unknown class instances' representations without directly accessing them.
2. We empirically validate our theory, observing that the Jacobian norm difference between the known and unknown classes is strongly correlated to the unknown class detection performance.
3. Based on the integral role of inter-class learning for the unknown class segregation, we devise a marginal one-vs-rest (m-OvR) loss that can induce strong inter-class separation within the known classes. We further integrate the model loss with auxiliary techniques that can enhance the unknown class segregation via the Jacobian norm difference.

We highlight that our primary objective is not to advance the state-of-the-art in the field. Rather, our foremost contribution lies in providing a theoretical elucidation of how a model gains awareness of the unknown through closed-set metric learning. Additional contributions encompass the empirical validation of our theoretical framework, as well as an examination of prevalent deep learning methodologies within the context of our proposed theory.

To the best of our knowledge, this is the inaugural study to investigate open set recognition (OSR) representations in relation to their Jacobian norm.

2. Related Works

2.1. Theoretical/empirical works on OSR.

Recent theoretical works [5, 6] tackle OSR with theoretical guarantees on the performance but with specific distributional modeling assumptions (e.g. Gaussian mixture). [7] conduct theoretical studies in a more general setting by extending the classical closed-set PAC framework [8] to open-set environments, deriving analytical bounds of the generalization error in the context of OSR. [9] relates OSR to transfer learning and interprets the unknown class samples as covariate shifts. This enables the substitution of theoretical bounds derived in the transfer learning setting [10] to open-set environments. On the other hand, [11] observes that for a model trained only with known class samples, the magnitudes of representation vectors tend to exhibit relatively larger values over the known class than over the unknown ones. [12] empirically proved that the standard discriminative models detect unknown classes mainly based on their unfamiliar features rather than based on the novelty of unknown category.

2.2. OSR Methods

For a general, broad survey on OSR models, the readers are recommended to [4, 13]. Here, we focus on reviewing state-of-the-art OSR models, mainly focusing on discriminative ones.

The basic baseline model [14, 1, 15] trained by softmax cross-entropy loss is known to perform both closed-set classification and unknown class detection reasonably effectively. To enhance its unknown detection mechanism, OpenMax [16] applied probabilistic modification on the softmax activation based on extreme value theory. DOC [17] replaced the softmax cross entropy with the one-vs-rest logistic regression, finding its effectiveness on invalid topic rejection in natural language. RPL [18] proposed to maximize inter-class separation in the form of reciprocal, followed by a variant [19] that utilizes synthetic, adversarially generated unknown class. CPN [20] learns embedding metrics by modeling each known class as a group of multiple prototypes. PROSER [21] leverages latent mixup samples [22, 23] as a generated unknown class and places their representations near the known class representations. [24], on the other hand, proposed a collection of multiple one-vs-rest networks to mitigate the over-confidence and poor generalization issue, and utilizes a collective decision score for effective OSR.

Recently, [15] demonstrated that the basic SCE baseline could outperform all other OSR baselines if the SCE model is trained with strong data augmentation and utilizes state-of-the-art optimization techniques. On the other hand, [25] showed that a prior on well separated discriminative embedding is still critical for effective open-set recognition.

2.3. Jacobian Norm in Deep Discriminative Models

Within the domain of discriminative learning, though not explicitly in the context of OSR, the Jacobian of the representation function has been examined in closed-set settings within various contexts. [26, 27] demonstrate that the explicit minimization of the Frobenius norm of the Jacobian of classification prediction output, specifically the softmax output and logit, promotes a smoothness prior on it, subsequently enhancing the generalization of recognition in closed-set scenarios. Nevertheless, the explicit computation of the Jacobian demands substantial computational resources. To address this, [28] has introduced an efficient method of computing the Jacobian norm via its random projection, serving as an unbiased estimator of the raw Jacobian norm.

[27, 29, 30] have noted that the smoothness prior, as enforced by Jacobian norm penalization, reduces the sensitivity of the

network output to minute input perturbations, thereby making the network robust against adversarial examples. On a theoretical level, [31] has identified a close connection between weight decay and the Jacobian norm, establishing that under ideal conditions, a gradient update with weight decay equates to penalizing the Frobenius norm of the Jacobian matrix of representation.

However, all preceding studies on Jacobian analysis have been confined to the context of closed-set learning. To the best of our knowledge, our research represents the first instance of analyzing the Jacobian norm within an open-set scenario, providing a rigorous examination of its relationship to the unknown class.

3. Theory: Understanding the Separation of Unknown Class Representations via Jacobian Norm

We theoretically demonstrate that training a discriminative model over known classes separate the representations of known classes from those of the unknown class by decreasing Jacobian norm over the known classes while increasing the Jacobian norm over the unknown class (Cor. 5 in Sec. 3.2). The limitation of Jacobian norm theory is given in Sec. 3.3. Our observation is summarized in Sec. 3.4 with its depiction in Fig. 2.

3.1. Problem Setup and Notation

During closed-set metric learning, the representation embedding function $\boldsymbol{f} : \mathbb{R}^d \rightarrow \mathbb{R}^{d_z}$ of a discriminative model is trained to minimize *intra-class distances* $\mathcal{D}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{w}_y)$ and maximize *inter-class distances* $\mathcal{D}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}'))$ for known class samples $\boldsymbol{x}$ and $\boldsymbol{x}'$ paired with different class labels y and y' ($y \neq y'$). The prototype vector $\boldsymbol{w}_y \in \mathbb{R}^{d_z}$ is a proxy for the y-th known class C_y, and is formulated as a learnable parameter. The known set $\mathcal{K} = \cup_{k=1}^{K} C_k$ consists of K disjoint disconnected known classes C_k. The train samples $\boldsymbol{x}$ and $\boldsymbol{x}'$ are sampled from the known set, while the labels y and y' from the corresponding label space $\mathcal{Y}_{\mathcal{K}} = \{1, \ldots, K\}$. The open set $\mathcal{O} := \mathcal{X} \backslash \mathcal{K}$ is the complement of the known set in the (bounded) global space $\mathcal{X} = [-1, 1]^d \subseteq \mathbb{R}^d$. The unknown class $\mathcal{U}$ we consider is a *proper* subset of the open set $\mathcal{O}$. Since our task is not to discriminate within the unknown class, we treat the unknown class as a single class, although it may consist of a diverse type of object.

During training, the model has no access to the unknown class $\mathcal{U}$, and is trained only with the K number of known classes to discriminate them. After training, the OSR model should not only discriminate each class in the known set but also need to differentiate the unknown from the known. Hence, the unknown class should be separated from all known classes in the representations space such that $\boldsymbol{f}(\mathcal{K}) \cap \boldsymbol{f}(\mathcal{K}) = \varnothing$.

3.2. Derivation of the Theory

We prove our theoretical claims by observing how the embedding function $\boldsymbol{f}$ changes on a class interpolating path (i.e., a path $\boldsymbol{\gamma}:[0, 1] \rightarrow \mathcal{X}$ that interpolates two different known classes C_i and C_j by traversing t from 0 to 1 with $\boldsymbol{x}_0 \in C_i$ and $\boldsymbol{x}_1 \in C_j$ as depicted in Fig. 1). The detailed assumptions and full proofs to the theoretical statements are given in Appendix A and Appendix B, respectively.

Firstly, we show that, during the closed-set supervision, the intra-class distance minimization minimizes the length of the projected path over the known class:

Proposition 1. *Minimizing intra-class distances* $\mathcal{D}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{w}_k)$ *to* 0 *for all* $\boldsymbol{x} \in C_k$ *minimizes the length of the projected path* $\boldsymbol{f}(\boldsymbol{\gamma}([0, 1]) \cap C_k)$ *for an arbitrary path* $\boldsymbol{\gamma}$ *from* C_k.

On the other hand, the inter-class distance maximization is presumed to increase the length of any linear path between the known classes C_i and C_j in the representation space by Assumption 2b. In summary, intra-class distance minimization reduces the projected path length, while the inter-class distance maximization increases the projected path length.

Now, the increasing/decreasing trend of the projected path length due to the metric learning is transferred to the Jacobian norm $\|\frac{d\boldsymbol{f}(\boldsymbol{\gamma}(t))}{dt}\|_2$ via the path length equation

$$\text{length}(\boldsymbol{f} \circ \boldsymbol{\gamma}) = \int_0^1 \left\| \frac{d\boldsymbol{f}(\boldsymbol{\gamma}(t))}{dt} \right\|_2 dt. \tag{1}$$

Accordingly, we expect that intra-class distance minimization minimizes the Jacobian norm over the known class intersecting path. In contrast, inter-class distance maximization increases the Jacobian norm over the open set intersecting path. This description, however, is constrained to the local paths. The following theorem assures that this phenomenon is extendible from the local path to the global region. In other words, the closed-set metric learning minimizes the Jacobian norm over the known classes and increases the Jacobian norm over the open set $\mathcal{O}$.

Theorem 2. *Let* C_i, C_j, *and* C_k *be different known classes.*

(a) Minimizing intra-class distances $\mathcal{D}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{w}_k)$ *for all* $\boldsymbol{x} \in C_k$ *minimizes* $\|\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}}\|_F$ *over* C_k.

(b) Maximizing inter-class distances $\mathcal{D}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}'))$ *for all* $\boldsymbol{x} \in C_i$ *and* $\boldsymbol{x}' \in C_j$ *strictly increases* $\int_{\mathcal{O}} \|\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}}\|_F \, d\boldsymbol{x}$.

Theorem 2b indicates that the length of the projected path can be accessed from the global integral of the Jacobian norm. Thereby, we find that the strictly increasing trend of Jacobian norm integral is positively correlated to the strictly increasing trend of the projected inter-class path length. Based on our overall observations, we deduct the below corollaries:

Corollary 3. *Minimizing the intra-class distances minimizes the Jacobian norm* $\|\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}}\|_F$ *over the known classes* $\mathcal{K}$.

Corollary 4. *Maximizing the inter-class distances strictly increases strictly increases*

$$\text{Vol}(S) \textit{ and/or } \mathbb{E}_{\boldsymbol{x} \sim S}[\|\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\|_F] \tag{2}$$

where S *is the support of Jacobian norm*

$$S := \{\boldsymbol{x} \in \mathcal{O} : \|\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}}\|_F > 0\}, \tag{3}$$

whose Jacobian norm is greater than 0*, and* $\mathrm{Vol}(S)$ *is the volume of* S*. Hence, if* $S \cap \mathcal{U} \neq \varnothing$*, then the inter-class maximization enlarges the volume* $\mathrm{Vol}(\mathcal{U} \cap S)$ *and/or increases the Jacobian norm of unknown class samples* $\boldsymbol{x} \in \mathcal{U} \cap S$*.*

Hence, maximizing the inter-class distances between the known classes access to the unknown class samples indirectly via the region S of high Jacobian norm, and increases the Jacobian norm of unknown class representations.

Overall, by metric learning, the model increases the expected Jacobian norm difference between the known and unknown

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{U}}[\|\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}\|_F] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{K}}[\|\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}\|_F]. \tag{4}$$

The increased *Jacobian norm difference* then separates the known classes from the unknown class in the representation space:

Corollary 5. *The inter/intra-class learning separates the unknown class from known classes in the representation space by inducing the Jacobian norm difference between the known and unknown.*

3.3. Limitation of the Theory on Jacobian Norm

We highlight that the Jacobian norm characteristic **is only one of the many explanatory factors** that demystifies how closed-set metric learning derives OSR; our analysis does not fully characterize all connections between closed-set metric learning and OSR. One apparent phenomenon our theory does not explain is that known and unknown representations can be separated in the metric space with having the same Jacobian norm value. Moreover, our theory is limited in characterizing the support set S. As the support set does not include the whole part of the open set, there would be some unknown class that is not included in the support. In this case, the Jacobian norm difference indicated in Eq. (4) would not be explanatory.

3.4. Summary of Theory

Training a discriminative model over the known classes reduces the Jacobian norm over known class samples, while increasing the volume of region of high Jacobian norm in the open set. Due to the increased volume of high Jcaobian norm region, the unknown class samples likely fall into this region, and thus involve high Jacobian norm values. Overall, the embedding representations of known classes are separated from those of unknown class because the Jacobian norms of known classes are low while the Jacobian norms of unknown class are high. Our theoretical finding is summarized in Fig. 2.

4. Empirical Verification of the Theory

In this section, we empirically verify the theory developed in Sec. 3 in multiple aspects.

4.1. Experiment Setup

We empirically analyze the relationship between the Jacobian norm difference and the unknown class detection to evidence our theoretical analysis. To this end, we train our proposed model as described in Sec. 5 and 6, and evaluate over the standard OSR benchmark datasets [14]. To compute the degree of separation between known and unknown, we use the detection score provided in Sec. 5.3 and evaluate the area under the receiver-operating-characteristic curve (AUC) metric [32]. The discriminative (cluster) quality of known class representations is measured in Davies-Bouldin Index (DBI) [33], which measures the ratio of intra-class distance to inter-class distance. All experiments are conducted with one 12GB GPU RTX 2080-ti. Due to resource limitations, empirical observations are made on standard OSR datasets rather than recently proposed high-resolution OSR datasets [15].

Datasets. For the empirical analysis, we test on the standard OSR datasets as described in Protocol A of Sec. 6.1. Each dataset consists the K number of known classes and 1 unknown class, overall constituting $K+1$ semantic classes. The unknown class can be constituted by a diverse set of semantic classes, but is regarded as a single chunk. The known classes must have no semantic overlap with the unknown class.

4.2. Empirical Observations

Jacobian norm before and after training. Fig. 3 demonstrate that the gradient norm separates the representations only after training. Fig. 4 displays the gradient norm over the linearly interpolated data samples $\boldsymbol{x}_t$ for $t \in [0, 1]$ between two different class samples $\boldsymbol{x}_0 \in C_i$ and $\boldsymbol{x}_1 \in C_j$. It shows that the interpolated samples inside the open region have a larger gradient norm than those in the known classes. These empirical observations support our theory.

In practice, however, the inter/intra-class distance optimizations conflict; thus, the overall gradient norm increases for both the known and unknown.

Moreover, on some datasets (SVHN and TinyImageNet), the inter-class separation may not be substantial due to innate data characteristics such as small inter-class data variance. Accordingly, based on Theorem 2b, the weak inter-class separation induces relatively smaller difference in Jacobian norm between the known and unknown, resulting a larger overlap between them.

The dynamics of the Jacobian norm during training. Fig. 5 shows the dynamics of different quantities during training. The intra/inter-class distance optimization increases the quality of cluster separation measured by DBI. Accordingly, the linear projected path length between different known classes in the representation space increases (Fig. 5b). As a result, the model increases both the Jacobian norm difference (Fig. 5c) and the degree of separation between known and unknown classes (Fig. 5d) as claimed by the theory.

Although the global trend has a simple correspondence between these metrics, a more careful look at the graphs of Fig. 5 shows that the metrics involve different phases during training. Specifically, the intra/inter ratio is stable at the early stage of

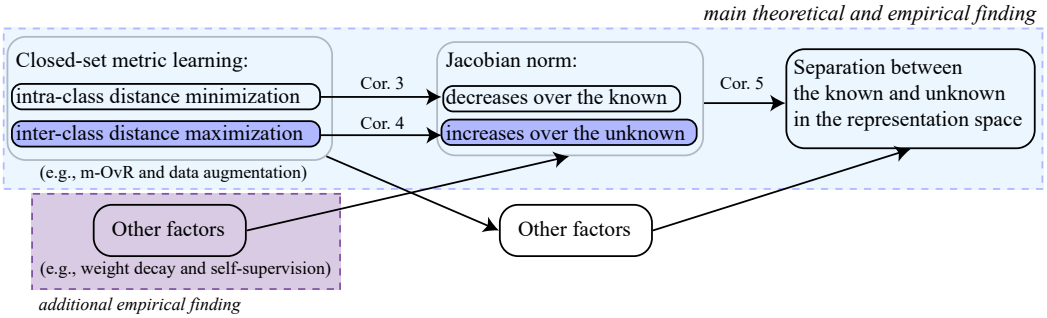

Figure 2: The summary of our theory on how a model becomes aware of the unknown by the closed-set metric learning over the known classes.

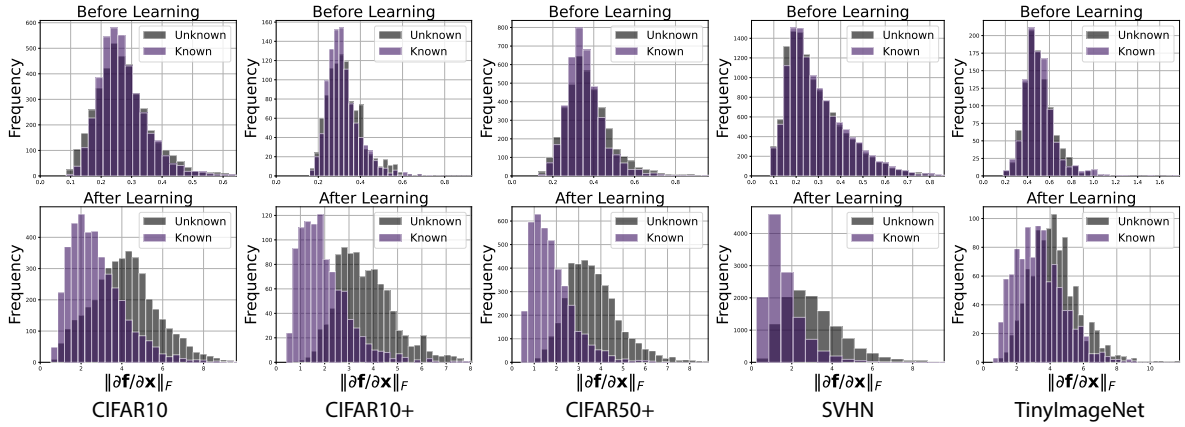

Figure 3: The distribution of Jacobian norms of representations **before and after training**. Although the model is trained only on the known class data, the model learns to increase the Jacobian norm of unknown class representation, while lowering the Jacobian norm of the known class representation.

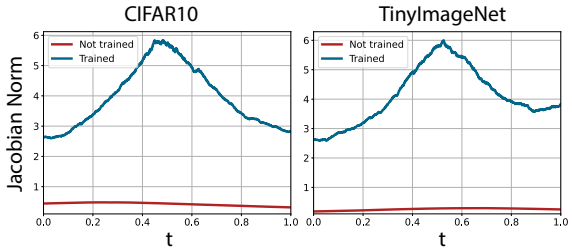

Figure 4: Given known class samples $\boldsymbol{x}_0 \in C_i$ and $\boldsymbol{x}_1 \in C_2$ from two different classes C_i and C_j, we **linearly interpolate** between $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ by $\boldsymbol{x}_t := (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1$. Then, we measure the Jacobian norm of the representation $\boldsymbol{f}(\boldsymbol{x}_t)$. When $t \approx 0.5$, the interpolated sample x_t passes through the open set, where unknown class samples arise.

training. On the other hand, the inter-class distance is still increasing even at a later stage. The Jacobian norm difference rises more gradually, and the rate of increase becomes large at the last stage. The separation between the known and unknown also increases largely at the early stage but continues to improve even later in training. These observations show that the known and unknown class representations are separated as the model makes their Jacobian norm different. Still, the Jacobian norm is not the only factor contributing to their separation.

The correlation between Jacobian norm and discriminative metrics. For each dataset, we measure the following three metrics during different training iterations: the discriminatory quality of known class representations (DBI), the unknown class detection performance (AUC), and the averaged Jacobian norm difference between known and unknown classes.

Fig. 6(1st row) shows that the degree of separation between the known and unknown strongly correlates to the Jacobian norm difference. This observation evidences our theoretical claim that the closed-set metric learning separates the unknown by increasing their Jacobian norm difference during training. There is, however, nonlinearity between these two metrics, showing that the Jacobian norm difference is not the only factor contributing to the separation of unknown class representation.

Fig. 6(2nd row) shows a similar correlation trend between the intra/inter-class distance ratio (DBI) and the Jacobian norm difference. However, the nonlinearity between them is severe. The plot indicates that the Jacobian norm difference abruptly increases at a later stage of training where the intra/inter ratio is already small and stable.

The relation between the Jacobian norm difference and the number of discriminative classes. Theorem 2 states that the inter-class distance maximization between a single pair of inter classes (C_i, C_j) can cause to increase in the Jacobian norm difference. Therefore, we hypothesize that a larger number of inter-class pairs would improve the Jacobian norm difference, contributing to better separation between known and unknown class representations. The results are given in Fig. 7 supports the hypothesis by showing that the Jacobian norm difference tends to become larger with a larger number K of known

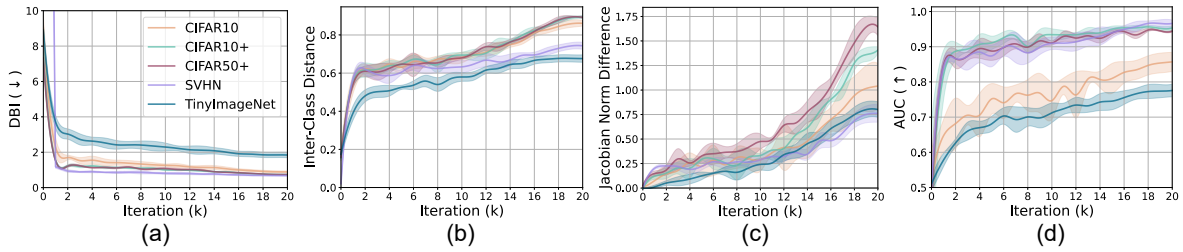

Figure 5: Several metrics are measured while a discriminative model (ours) is trained. (a) The discriminative quality of known class representations is measured in DBI. (b) The averaged inter-class distances between known classes. (c) The Jacobian norm difference between the known and unknown classes. (d) The degree of separation between known and unknown class representations. **All metrics are improved as the discriminative model learns**.

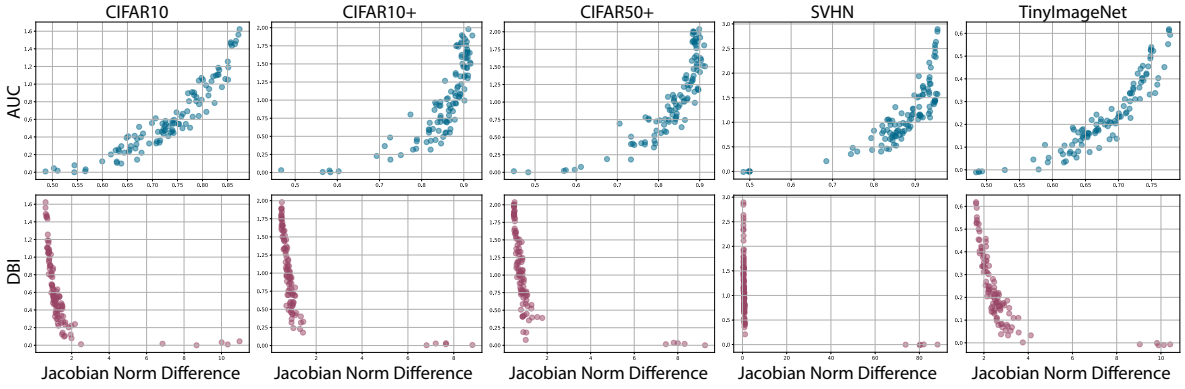

Figure 6: We measure the detection performance (in AUC), the discriminative quality of known classes (in DBI), and the averaged Jacobian norm difference **for a single model during different training iterations, indicating a strong correlation between these metrics**.

classes. We note that the exceptions may occur as some known classes are more similar to the unknown class examples; adding to the train data a known class that is similar to the unknown may slightly reduce the Jacobian norm difference.

5. Method

We develop an effective OSR method based on our theoretical finding given in Fig. 2. Firstly, we devise a margin-based one-vs-rest that can induce powerful inter-class separation between different known classes. Then, we integrate the loss term with other regularizers that enhance the separation of the unknown via the Jacobian norm difference. Finally, for the unknown class detection in the inference stage, we utilize the sample-wise loss function as it is aware of both the Jacobian norm difference and proximity to the known class prototypes.

5.1. Training: marginal One-vs-Rest Loss (m-OvR)

Our analysis indicates that the powerful inter-class separation is the key to separate the known from the unknown in the Jacobian norm and therefore in the representation space. Motivated upon this theory, we devise a marginal one-vs-rest (m-OvR) loss that induces powerful inter-class separation by preventing the collapse between inter-class prototypes $\boldsymbol{w}_k$ and effective inter-class gradients. The m-OvR loss is given by

$$\mathcal{L}(\boldsymbol{x}, y) = -\sum_{k=1}^{K} 1\{y = k\} \log p(k|\boldsymbol{x}) + 1\{y = k\} \log (1 - p(k|\boldsymbol{x})) \tag{5}$$

where $(\boldsymbol{x}, y)$ is a labeled sample, and $1\{\cdot\}$ is an indicator function. The class probability $p(k|\boldsymbol{x})$ is given by $\sigma(T\, s_k)$ where σ is the sigmoid activation, s_k is the cosine similarity between the representation $\boldsymbol{f}(\boldsymbol{x})$ and the k-th class proxy prototype $\boldsymbol{w}_k$, and T is a scale term to calibrate the sigmoid probability.

During training, the bare minimization of the loss in Eq. (5) involves a harmful behavior; particularly, minimizing the loss in Eq. (5) collapses the inter-class prototypes as observed by below proposition:

Proposition 6. *The minimum OvR loss collapses all prototypes* $\boldsymbol{w}_k = \boldsymbol{w}_{k'}$ *except* $\boldsymbol{w}_y$.

This inter-class collapse weakens the inter-class separation. We mitigate this situation by inserting a margin in the similarity computation; namely, during the training of the OvR metric-learning loss, the similarity is computed by

$$s_k = \cos(\arccos(\boldsymbol{w}_k \cdot \boldsymbol{f}(\boldsymbol{x})) + m) \tag{6}$$

where $m > 0$ is the margin. The margin ensures an angular gap of degree $2m$ between inter-class prototypes, thus preventing their collapse:

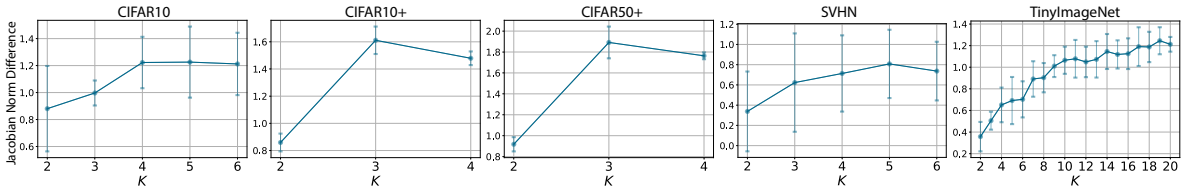

Figure 7: Increasing the number K of known classes increases the Jacobian norm difference between the known and unknown classes.

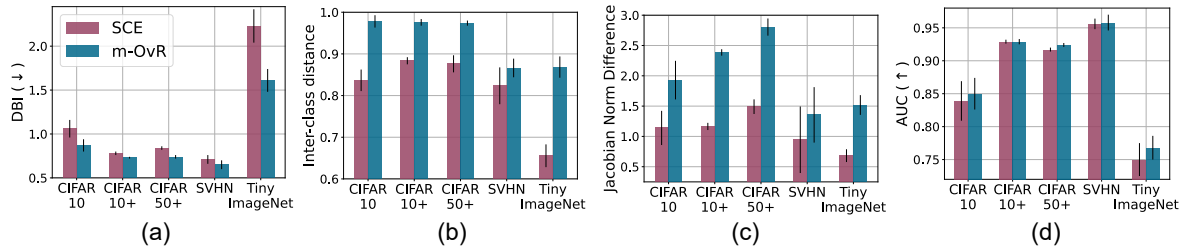

Figure 8: Comparison of the loss functions (SCE and m-OvR) with respect to (a) the discriminative quality in DBI, (b) the average of the pairwise distance between class-wise mean features, (c) Jacobian norm difference, and (d) unknown class detection performance in AUC.

Proposition 7. *For the nonzero margin $m > 0$, however, the angle gap can be assured between different prototypes $\measuredangle(\boldsymbol{w}_{k_1}, \boldsymbol{w}_{k_2}) \geq 2m$.*

In addition, the proposed m-OvR induces more powerful inter-class separation than the standard softmax cross-entropy (SCE) loss:

Proposition 8. *Assume $s_y > 0$. Then, the inter-class gradient for the m-OvR $\frac{\partial s_k^{m\text{-}OvR}}{\partial \theta}$ is greater than that for the SCE $\frac{\partial s_k^{SCE}}{\partial \theta}$.*

Therefore, m-OvR is more effective at increasing the Jacobian norm difference and, hence, the unknown class detection performance accordingly.

The empirical observations given in Fig. 8 indicate the effectiveness of m-OvR compared to the SCE loss in terms of Jacobian norm difference, discriminative quality of known class representations, and the unknown class detection performance based on the detector in Sec. 5.3.

5.2. Training: Subsidiary Techniques to Improve OSR

Using the Jacobian norm principle from Sec. 3, we explain how the standard techniques (weight decay, auxiliary self-supervision, and data augmentation) improve the separation between known and unknown class representations, thereby improving the OSR performance. Our final model is combined with these techniques.

Data Augmentation. The training data is usually limited. Hence, directly applying metric learning to the raw data without augmentation results in suboptimal inter-class separation and intra-class compactness. The Jacobian norm difference between known and unknown class representations would be negligible in this case. Applying data augmentation resolves this issue by expanding the training set size based on the prior human knowledge of the data. Furthermore, the improved Jacobian norm difference by data augmentation enhances the unknown class detection (Fig. 10).

Weight Decay. Based on [34], the embedding similarity s_k is optimized based on the gradient

$$\partial s_k / \partial \widehat{\boldsymbol{f}} = (\boldsymbol{w}_k - s_k \boldsymbol{f}) \cdot \|\widehat{\boldsymbol{f}}\|_2^{-1}. \tag{7}$$

Thus, the small norm $\|\widehat{\boldsymbol{f}}\|_2$ of the (unnormalized) representation can incite stronger inter-class separation. The weight decay decreases this norm by decreasing the values of the network parameters in $\widehat{\boldsymbol{f}}$ [31]. Based on our theory, the enhanced inter-class separation results in higher Jacobian norm values of the unknown class representation, resulting in better separation between the known and unknown in the representation space. The experimental results in Fig. 11 precisely verify this theoretical observation.

Auxiliary Self-Supervision. To improve the unknown class detection performance, several works [35, 36] employ an auxiliary supervision task to predict the degree of rotation (either 0, 90, 180, or 270) on the rotated images. This extra discriminative task poses additional inter-class separation learning on the model. Based on our observations in Sec. 3 and 4 and Fig. 7, posing additional inter-class separation increases the Jacobian norm of the unknown, thereby improving the separation between the known and unknown class representations (Fig. 12). We note, however, that the auxiliary self-supervision should be accompanied with care; predicting rotation in a standard manner may collapse the original class prototypes w_k as the rotation prediction head regards the original classes as a single 0-degree class. Hence, we add the auxiliary self-supervision loss $\mathcal{L}_{self}$ with a small coefficient $\lambda_{self} = 0.1$.

Our final metric-learning objective is to minimize the combined loss $\mathcal{L} + \lambda_{self}\mathcal{L}_{self}$ with data augmentation and weight decay.

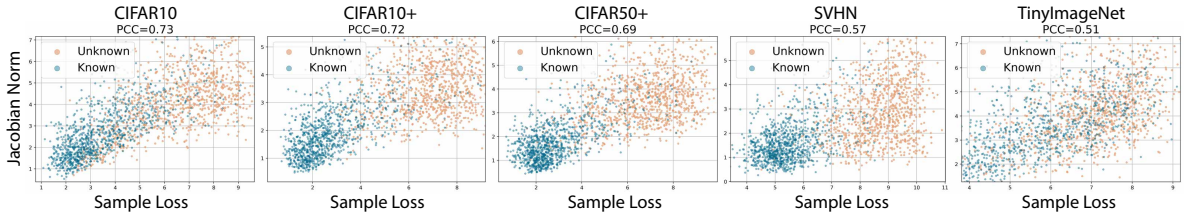

Figure 9: The correlation between sample-wise losses and the Jacobian norm of the corresponding representations.

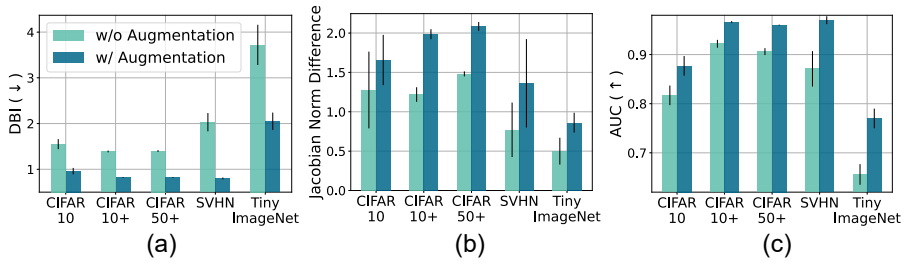

Figure 10: The effect of data augmentation with respect to (a) the discriminative quality in DBI, (b) Jacobian norm difference, and (c) unknown class detection performance in AUC.

5.3. Inference: Unknown Class Detection by the Sample-Wise Loss Function

To effectively detect unknown class samples during the inference stage, we utilize the sample-wise loss function. Based on our theoretical finding, the loss function is aware of both the Jacobian norm difference and the closeness to the known class prototype:

$$\mathcal{L}(\boldsymbol{x}) \text{ low/high} \Longleftrightarrow \mathbb{E}_{\boldsymbol{x}_u \sim \mathcal{U}} \left[\left\| \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}(\boldsymbol{x}_u) \right\|_2 \right] - \left\| \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}(\boldsymbol{x}) \right\|_2 \text{ low/high} \tag{8}$$

$$\text{and } \min_k \mathcal{D}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{w}_k) \text{ low/high}$$

Hence, the loss function (1) differentiates the the known class representations in the low Jacobian norm region from the unknown class representations residing in the region of high Jacobian norm, and (2) separates the known class close to the prototypes w_k from the unknown class instances. The positive correlation indicated in Fig. 9 vindicates the property of loss function described in Eq. (8).

6. Experiments for Comparison

The experiment section is outlined as follows: (1) We compare our method with other baseline OSR models for the unknown class detection task under two different widely-used protocols, Protocol A [14] and Protocol B [39]. (2) We conduct a careful ablation study of our method, analyzing each component in terms of the unknown class detection performance and the Jacobian norm. (3) We visualize and analyze the Jacobian norm of representation with respect to the metric distances in the representation space. To this end, we compare our proposed model with a baseline model trained with the bare SCE loss.

Our proposed model is trained with the m-OvR loss in all experiments below. Unless specified, we always include weight decay, data augmentation, and auxiliary self-supervision in our model. The default model hyperparameters are as follows: the scale term $T = 32$, margin $m = 0.5$, the auxiliary self-supervision coefficient 0.1, and the weight decay $1e{-}3$.

We consider three backbones to extract the representation: WRN-16-4 [41], VGG [14], and ResNet-18. For WRN-16-4 and VGG, our model is trained by SGD with 20k training iterations unless specified otherwise. Its learning rate is regulated under a cosine scheduler, initiating from 0.1 and decaying to $1e{-}5$. The batch size is 128. In the case of the ResNet-18 backbone, on the other hand, the model is trained for 200 epochs under the SGD optimizer with a momentum of 0.9 and a learning rate of 0.06 that decays to 0 by the cosine learning scheduler.

In all experiments, the model is trained only with known classes so that the model never sees any unknown class sample during training.

6.1. Performance Comparison - Protocol A

Datasets-Protocol A. In this protocol [14], we use five different OSR datasets to compare different OSR methods in terms of the closed-set classification accuracy and unknown class detection performance.

Our method is evaluated for unknown class detection performance (AUC) and closed-set accuracy (ACC). The protocol used in [14] is adopted with the following benchmark datasets:

- CIFAR10 and SVHN: Among the total ten classes, K=6 classes are chosen as the known ones, regarding the rest as a single unknown class. CIFAR10 [42] consists of generic object images while SVHN [43] of street view numbers.
- CIFAR10+ and CIFAR50+: To make CIFAR10 more challenging, CIFAR10+ and CIFAR50+ are considered,

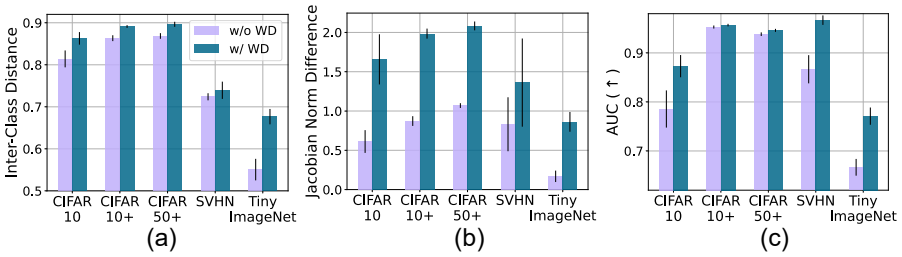

Figure 11: The effect of weight decay (WD) with respect to (a) averaged inter-class distance, (b) Jacobian norm difference, and (c) unknown class detection performance in AUC.

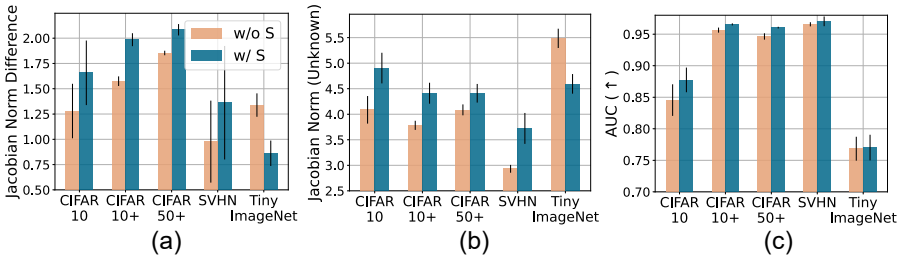

Figure 12: The effect of auxiliary self-supervision (S) with respect to (a) Jacobian norm difference, (b) Jacobian norm of unknown class, and (c) unknown class detection performance in AUC.

Table 1: Unknown class detection performance (in AUC) and closed-set accuracy (ACC) for OSR where the unknown class is derived from the same distribution. The results are the averages from 5 random splits. * indicates that the values are taken from the references. 'Arch.' denotes the backbone network used.

Method	Arch.	CIFAR10		CIFAR10+		CIFAR50+		SVHN		TinyImageNet	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
SCE* [1] (ICLR'17)	VGG	-	67.7	-	81.6	-	80.5	-	88.6	-	57.7
OpenMax* [16] (CVPR'16)	VGG	80.1	69.5	-	81.7	-	79.6	94.7	89.4	-	57.6
RPL* [18] (ECCV'20)	VGG	-	82.7	-	84.2	-	83.2	-	93.4	-	68.8
PROSER* [21] (CVPR'21)	VGG	92.6	89.1	-	96.0	-	95.3	96.4	94.3	52.1	69.3
CPN* [20] (TPAMI'22)	VGG	92.9	82.8	-	88.1	-	87.9	96.4	92.7	-	63.9
ODL* [37] (TPAMI'22)	VGG	-	88.5	-	91.1	-	90.6	-	95.4	-	74.6
Ours (combined)	VGG	**96.4**	**89.5**	**96.3**	**96.2**	**96.5**	**95.7**	**97.4**	**95.7**	**78.2**	**75.3**
SCE (ICLR'17)	WRN	92.1	76.5	94.0	84.7	94.0	83.8	97.0	92.1	65.8	66.1
N-SCE	WRN	93.7	76.8	93.7	85.3	93.7	84.4	97.1	91.8	64.5	66.4
DOC [17] (EMNLP'17)	WRN	91.6	78.0	93.9	88.2	93.9	88.1	97.0	93.5	59.6	65.5
RPL (ECCV'20)	WRN	94.7	82.0	95.9	91.1	96.1	90.9	97.5	93.4	71.4	70.4
CPN (TPAMI'22)	WRN	91.2	76.2	93.7	84.6	93.7	83.2	96.7	92.3	59.7	64.6
Ours (combined)	WRN	**97.0**	**89.0**	**97.7**	**96.6**	**97.6**	**96.0**	**98.0**	**97.0**	**79.1**	**77.0**
CSSR* [38] (TPAMI'22)	ResNet-18	-	91.3	-	96.3	-	96.2	-	97.9	-	82.3
GoodOSR [15] (ICLR'22)	ResNet-18	**96.3**	91.4	**97.4**	96.0	**97.4**	94.2	96.6	97.5	83.1	82.2
Ours	ResNet-18	96.0	**92.9**	97.3	**98.0**	97.3	**96.5**	**96.8**	**98.2**	**83.2**	**82.9**

in which K=4 known classes are selected from CIFAR10 while 10 (or 50) classes from CIFAR100 [42] constitute a single unknown class.

- TinyImageNet: In TinyImagenet (TIN) [44] with more diverse categories, K=20 classes constitutes the known, while the other 180 remaining ones form a single unknown class.

Results-Protocol A. The comparison results are given in Table 1, which indicate that our proposed methodology is effective for OSR across different backbone architectures, including VGG, WRN, and ResNet-18.

A significant attribute of our methodology lies in the employment of our margin-based loss, m-OvR, which not only optimizes intra-class compactness but also ensures inter-class separation by circumventing inter-class collapse, as detailed in Prop. 7. This aspect renders our work as an improvement over the prevailing techniques such as RPL, CPN, and OvRN-CD, which predominantly focus on the inter-class aspects alone. Furthermore, our methodology incorporates carefully chosen subsidiary techniques, including weight decay, representation unit-normalization, and self-supervision through rotation prediction, which can efficaciously enhance OSR.

Table 2: OSR performance under macro-averaged F1-score. * indicates that the values are taken from the references. 'Arch.' states the backbone network used. The weight decay is applied in default.

Method	Arch.	Param.	ImageNet-crop	ImageNet-resize	LSUN-crop	LSUN-resize	Avg.
SCE* [1] (ICLR'17)	VGG	1.1M	63.9	65.3	64.2	64.7	64.5
OpenMax* [16] (CVPR'16)	VGG	1.1M	66.0	68.4	65.7	66.8	66.7
CROSR* [39] (CVPR'19)	VGG	1.1M	72.1	73.5	72.0	74.9	73.1
GFROSR* [40] (CVPR'20)	VGG	1.1M	75.7	79.2	75.1	80.5	77.6
PROSER* [21] (CVPR'21)	VGG	1.1M	**84.9**	82.4	**86.7**	85.6	84.9
OvRN-CD* [24] (TNNLS'22)	VGG	1.1M	83.5	82.5	84.6	83.9	83.6
Ours	VGG	1.1M	84.2	**88.4**	85.1	**88.1**	**86.5**
SCE	WRN-16-4	2.7M	79.1	79.2	80.3	80.8	79.9
m-OvR	WRN-16-4	2.7M	80.5	79.8	79.2	81.2	80.2
SCE + A	WRN-16-4	2.7M	84.5	88.5	87.0	88.8	87.2
m-OvR + A	WRN-16-4	2.7M	87.4	89.0	89.0	90.0	88.9
SCE + A + S	WRN-16-4	2.7M	84.6	87.5	87.5	87.5	86.8
m-OvR + A + S	WRN-16-4	2.7M	**89.1**	**90.5**	**90.4**	**90.9**	**90.2**

Table 3: Ablation of our proposed model by analyzing its training components: the m-OvR loss, unit-normalization (N) of representations, the margin m in the similarity computation, weight decay (W), data augmentation (A), and auxiliary self-supervision (S). Models are evaluated in terms of unknown class detection performance (AUC), closed-set accuracy (ACC), and detection accuracy (DetACC). SCE substitutes in the absence of m-OvR.

	Model	OvR	N	m	W	A	S	CIFAR10 AUC/ACC/DetACC	CIFAR10+ AUC/ACC/DetACC	CIFAR50+ AUC/ACC/DetACC	SVHN AUC/ACC/DetACC	TIN AUC/ACC/DetACC	Avg. AUC/ACC/DetACC
	Baseline							76.5/92.1/70.8	84.7/94.1/78.9	83.9/94.1/78.1	92.1/97.1/86.1	66.1/65.8/62.8	80.7/88.6/75.3
	m-OvR	✓	✓	✓				79.3/91.6/72.8	90.1/93.9/82.6	89.8/94.1/82.1	93.8/97.1/87.6	66.2/62.3/62.7	83.9/87.8/77.6
one out	a				✓	✓	✓	85.0/96.1/78.1	90.4/96.8/84.1	89.3/96.9/82.6	94.4/97.7/88.7	74.7/77.8/70.0	86.7/93.1/80.7
	b		✓		✓	✓	✓	83.9/95.9/77.0	92.9/97.0/85.8	91.7/97.1/85.0	95.6/97.7/88.6	75.0/77.0/69.8	87.8/92.9/81.2
	c	✓	✓	✓		✓	✓	79.2/93.1/63.7	95.7/95.6/89.0	94.5/95.5/86.8	95.5/97.1/88.3	66.1/67.0/62.8	86.2/89.7/78.1
	d	✓	✓	✓	✓		✓	81.7/95.1/77.5	92.2/95.7/86.5	90.6/95.7/85.1	87.1/97.6/88.4	65.6/66.4/63.8	83.4/90.1/80.3
w/o A+S	e				✓			81.6/94.1/75.7	87.7/95.6/81.3	87.1/95.8/80.7	94.4/97.7/89.3	70.1/67.8/65.4	84.3/90.2/78.5
	f		✓		✓			79.9/93.4/73.8	86.8/95.0/80.7	85.0/94.9/78.7	92.7/97.3/87.0	67.3/67.0/63.6	82.3/89.6/76.7
	g	✓	✓	✓	✓			80.3/93.8/74.1	89.6/95.2/82.8	88.6/95.0/81.4	92.8/97.4/86.3	68.9/65.4/64.7	84.0/89.4/77.9
w/o S	h				✓	✓		83.2/95.9/76.5	91.1/97.1/84.6	90.6/97.0/84.1	94.1/97.5/87.7	75.6/76.5/69.9	86.9/92.8/80.5
	i		✓		✓	✓		84.6/96.1/77.0	95.6/97.2/88.8	94.6/97.2/86.2	96.6/97.9/90.5	76.8/77.8/70.8	89.6/93.2/82.7
	j	✓	✓	✓	✓	✓		84.6/96.1/77.0	95.6/97.2/88.8	94.6/97.2/86.2	96.6/97.9/90.5	76.8/77.8/70.8	89.6/93.2/82.7
on m	k	✓	✓		✓	✓	✓	85.0/96.4/78.6	92.9/97.2/86.3	92.4/97.2/85.4	95.8/98.0/89.9	76.8/79.2/71.3	88.6/93.6/82.3
	Ours	✓	✓	✓	✓	✓	✓	87.7/97.2/80.0	96.6/97.8/89.6	96.0/97.6/88.1	97.0/98.0/91.5	77.0/79.1/71.0	**90.9/93.9/84.0**

We note that our approach, even without the use of complex training tricks but solely utilizing the m-OvR loss, is comparable to the state-of-the-art GoodOSR. The pivotal differentiation lies in that GoodOSR boosts the OSR performance by excessive hyperparameter tuning and various cutting-edge training tricks, while ours is simply based on the loss function design.

6.2. Performance Comparison - Protocol B

Datasets-Protocol B. In this experiment, the model is trained over K known classes and classifies K+1 where the K + 1-th class is the unknown class. The protocol given in [39] is adopted. For benchmarking, we use CIFAR10 classes as the known with K=10. The unknown class is either ImageNet [45] or LSUN [46] that comprises scenery images. They are resized or cropped, constituting ImageNet-crop, ImageNet-resize, LSUN-crop, or LSUN-resize. Following the convention given in [21, 39], we choose the threshold τ for the inference score in Sec. 5.3 so that 10% of the validation set is detected as unknown class samples. The performance is evaluated using macro-averaged F1-score [47].

Results-Protocol B. The result in Table 2 shows that our proposed method outperforms all other baselines in the average performance. Under the WRN-16-4 architecture, m-OvR shows superiority over SCE, significantly more effective than SCE when applied with augmentation (A) and self-supervision (S). This is mainly due to the large Jacobian norm difference derived from the highly discriminative representations of the m-OvR (as observed in Fig. 8) triggers a strong separation between the known and unknown class representations.

6.3. Ablation Study

Ablation on Training Components. Each component in our model is more carefully evaluated in this experiment. For this purpose, we use the standard metrics used in OSR; namely, AUC for the unknown class detection performance, the closed-set accuracy (ACC), and detection accuracy (DetACC) [2]. The second block in the row shows that the m-OvR loss outperforms the SCE loss by a large margin, even when there is no data augmentation (A), weight decay (W), and self-supervision (S). The representation embedding normalization (N) improves the performance by preventing trivial increase of the Jacobian norm. The third block in a row ('one out') of Table 3 along with the model-j compares each component by removing one of them out, verifying the effectiveness of each in the entire model. When the standard data augmentation is available (i.e. the fifth block), m-OvR effectively utilizes the data, thus more effectively separating the known from the unknown than SCE. Finally, the sixth block analyzes the margin, which improves the

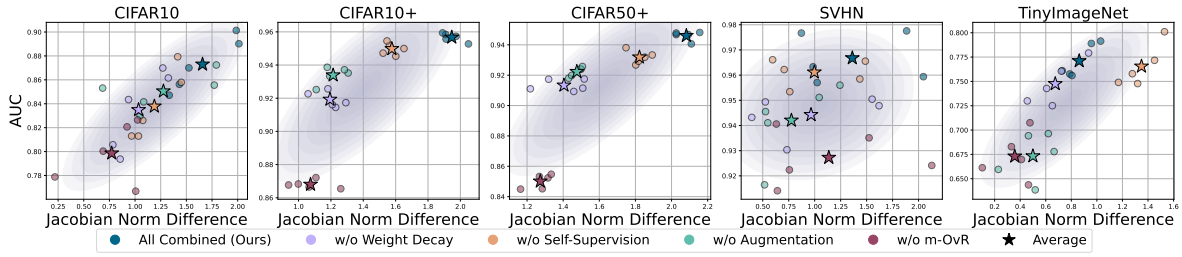

Figure 13: The plot of Jacobian norm difference versus the unknown class detection performance (AUC). Each point corresponds to a distinct model trained over a different known set, following the protocol of [14]. Different colors indicate different methods. The plot shows the positive correlation, and the all combined model has the largest Jacobian norm differnce.

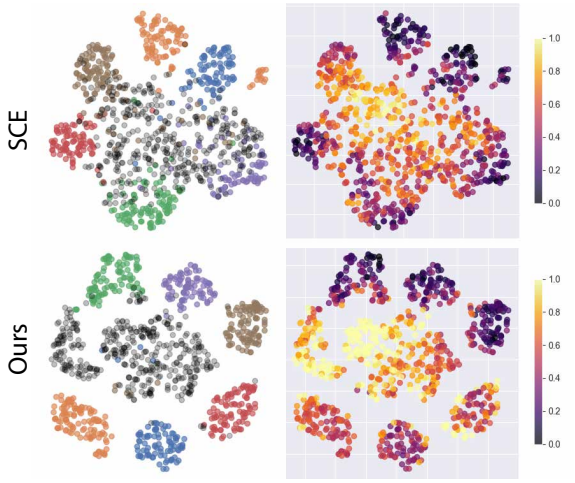

Figure 14: The 2-dimensional t-SNE [48] visualization of $\boldsymbol{f}(\boldsymbol{x})$ trained on MNIST under the protocol of [14]. In the left column, the black color denotes the unknown class. The temperature in the heat map (right column) indicates the (min-max normalized) Jacobian norm $\|\partial \boldsymbol{f}/\partial \boldsymbol{x}\|_F$. The figure shows that the larger the Jacobian norm difference between the known and unknown (i.e., the color contrast in the right column figures), the better the separation between the known and unknown.

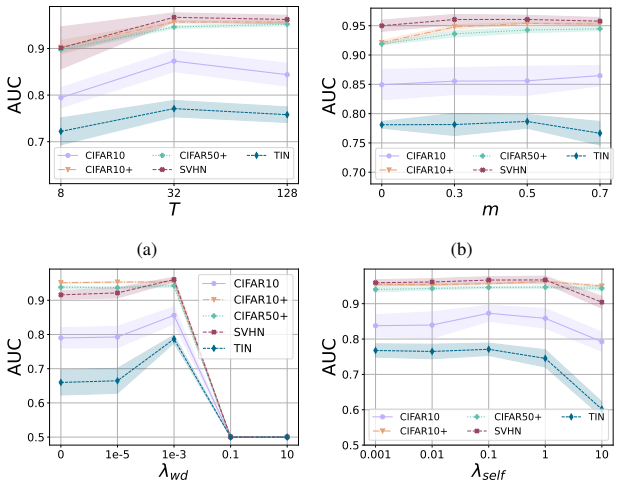

Figure 15: Unknown class detection performance (AUC) versus (a) the scale T, (b) the margin m, (c) the coefficient of the weight decay, and (d) the coefficient of the auxiliary self-supervision loss.

effectiveness of the loss-based unknown class detector by resolving the prototype misalignment issue.

Ablation with Jacobian Norm Difference The scatter plot for each fixed dataset in Fig. 13 shows that the degree of separation between the known and unknown class representations positively correlates to the Jacobian norm difference. The correlations in CIFAR10 and TinyImageNet are strong, while CIFAR10+ and CIFAR50+ exhibit some degree of nonlinearity. In SVHN, on the other hand, the correlation is comparatively weak due to the performance saturation. Moreover, this proves that the large Jacobian norm difference is not the only factor that captures distance separation between the known and unknown, as already remarked by Sec. 3.3).

Ablation on Model Hyperparameters. We analyze the hyperparameters of our overall model. Fig. 15 shows that the unknown class detection performance is robust for a sufficiently large scale term T, and the margin m should not be too large.

On the other hand, if the weight decay coefficient λ_{wd} is overly large, then it collapses the embedding to a constant (i.e., zero vector). At the same time, overly small λ_{wd} has no impact as a regularizer. Finally, we remark that selecting a proper coefficient for the weight decay is not tricky by observing the train loss dynamic during the early stage.

As already remarked in Sec. 5.2, the rotation-based self-supervision auxiliary loss contributes positively only when its coefficient λ_{self} is small (i.e., smaller than 1). The unknown class detection performance is robust for the small values of λ_{self}.

6.4. Visual Analysis of the Jacobian Norm of Representation

In the 2-dimensional visualization of Fig. 14 obtained by applying t-SNE on the embedding representations of data samples, the known classes exhibit small Jacobian norm values while the unknown samples have larger Jacobian norm values. Moreover, the degree of distance-wise separation becomes high when the Jacobian norm contrast between the known and unknown classes is more vivid.

7. Conclusion

We have demonstrated that closed-set metric learning distinguishes the unknown from the known by causing their representations' Jacobian norm values to differ. Crucially, inter-class

learning serves as the primary factor in this process, as it modifies the unknown class samples' representations without directly accessing them. Recognizing the significant role of inter-class learning in OSR, we developed a marginal one-vs-rest loss function designed to promote robust inter-class separation. By integrating this loss with other techniques that amplify the Jacobian norm disparity between known and unknown classes, we have successfully showcased the efficacy of our method on standard OSR benchmarks.

Acknowledgment. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NO. NRF-2022R1A2C1010710).

References

[1] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, arXiv preprint arXiv:1610.02136 (2016).

[2] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, Advances in neural information processing systems 31 (2018).

[3] S. Liang, Y. Li, R. Srikant, Enhancing the reliability of out-of-distribution image detection in neural networks, arXiv preprint arXiv:1706.02690 (2017).

[4] C. Geng, S.-j. Huang, S. Chen, Recent advances in open set recognition: A survey, IEEE transactions on pattern analysis and machine intelligence (2020).

[5] A. Meinke, M. Hein, Towards neural networks that provably know when they don't know, arXiv preprint arXiv:1909.12180 (2019).

[6] A. Meinke, J. Bitterwolf, M. Hein, Provably robust detection of out-of-distribution data (almost) for free, arXiv preprint arXiv:2106.04260 (2021).

[7] Z. Fang, J. Lu, A. Liu, F. Liu, G. Zhang, Learning bounds for open-set learning, in: International Conference on Machine Learning, PMLR, 2021, pp. 3122–3132.

[8] L. G. Valiant, A theory of the learnable, Communications of the ACM 27 (11) (1984) 1134–1142.

[9] S. Liu, R. Garrepalli, T. Dietterich, A. Fern, D. Hendrycks, Open category detection with pac guarantees, in: International Conference on Machine Learning, PMLR, 2018, pp. 3169–3178.

[10] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, et al., Analysis of representations for domain adaptation, Advances in neural information processing systems 19 (2007) 137.

[11] A. R. Dhamija, M. Günther, T. E. Boult, Reducing network agnostophobia, arXiv preprint arXiv:1811.04110 (2018).

[12] T. G. Dietterich, A. Guyer, The familiarity hypothesis: Explaining the behavior of deep open set methods, Pattern Recognition 132 (2022) 108931.

[13] J. Yang, K. Zhou, Y. Li, Z. Liu, Generalized out-of-distribution detection: A survey, arXiv preprint arXiv:2110.11334 (2021).

[14] L. Neal, M. Olson, X. Fern, W.-K. Wong, F. Li, Open set learning with counterfactual images, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 613–628.

[15] S. Vaze, K. Han, A. Vedaldi, A. Zisserman, Open-set recognition: A good closed-set classifier is all you need, CoRR abs/2110.06207 (2021). arXiv:2110.06207.
URL https://arxiv.org/abs/2110.06207

[16] A. Bendale, T. E. Boult, Towards open set deep networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1563–1572.

[17] L. Shu, H. Xu, B. Liu, Doc: Deep open classification of text documents, arXiv preprint arXiv:1709.08716 (2017).

[18] G. Chen, L. Qiao, Y. Shi, P. Peng, J. Li, T. Huang, S. Pu, Y. Tian, Learning open set network with discriminative reciprocal points, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, 2020, pp. 507–522.

[19] G. Chen, P. Peng, X. Wang, Y. Tian, Adversarial reciprocal points learning for open set recognition, arXiv preprint arXiv:2103.00953 (2021).

[20] H.-M. Yang, X.-Y. Zhang, F. Yin, Q. Yang, C.-L. Liu, Convolutional prototype network for open set recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[21] D.-W. Zhou, H.-J. Ye, D.-C. Zhan, Learning placeholders for open-set recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4401–4410.

[22] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, arXiv preprint arXiv:1710.09412 (2017).

[23] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, Y. Bengio, Manifold mixup: Better representations by interpolating hidden states, in: International Conference on Machine Learning, PMLR, 2019, pp. 6438–6447.

[24] J. Jang, C. O. Kim, Collective decision of one-vs-rest networks for open-set recognition, IEEE Transactions on Neural Networks and Learning Systems (2022).

[25] T. Kasarla, G. J. Burghouts, M. van Spengler, E. van der Pol, R. Cucchiara, P. Mettes, Maximum class separation as inductive bias in one matrix, arXiv preprint arXiv:2206.08704 (2022).

[26] J. Sokolić, R. Giryes, G. Sapiro, M. R. Rodrigues, Robust large margin deep neural networks, IEEE Transactions on Signal Processing 65 (16) (2017) 4265–4280.

[27] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, J. Sohl-Dickstein, Sensitivity and generalization in neural networks: an empirical study, arXiv preprint arXiv:1802.08760 (2018).

[28] D. Varga, A. Csiszárik, Z. Zombori, Gradient regularization improves accuracy of discriminative models, arXiv preprint arXiv:1712.09936 (2017).

[29] D. Jakubovitz, R. Giryes, Improving dnn robustness to adversarial attacks using jacobian regularization, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 514–529.

[30] J. Hoffman, D. A. Roberts, S. Yaida, Robust learning with jacobian regularization, arXiv preprint arXiv:1908.02729 (2019).

[31] G. Zhang, C. Wang, B. Xu, R. Grosse, Three mechanisms of weight decay regularization, arXiv preprint arXiv:1810.12281 (2018).

[32] A. P. Bradley, The use of the area under the roc curve in the evaluation of machine learning algorithms, Pattern recognition 30 (7) (1997) 1145–1159.

[33] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE transactions on pattern analysis and machine intelligence (2) (1979) 224–227.

[34] D. Zhang, Y. Li, Z. Zhang, Deep metric learning with spherical embedding, Advances in Neural Information Processing Systems 33 (2020).

[35] D. Hendrycks, M. Mazeika, S. Kadavath, D. Song, Using self-supervised learning can improve model robustness and uncertainty, arXiv preprint arXiv:1906.12340 (2019).

[36] I. Golan, R. El-Yaniv, Deep anomaly detection using geometric transformations, Advances in neural information processing systems 31 (2018).

[37] Z.-g. Liu, Y.-m. Fu, Q. Pan, Z.-w. Zhang, Orientational distribution learning with hierarchical spatial attention for open set recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[38] H. Huang, Y. Wang, Q. Hu, M.-M. Cheng, Class-specific semantic reconstruction for open set recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

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

[40] P. Perera, V. I. Morariu, R. Jain, V. Manjunatha, C. Wigington, V. Ordonez, V. M. Patel, Generative-discriminative feature representations for open-set recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11814–11823.

[41] S. Zagoruyko, N. Komodakis, Wide residual networks, arXiv preprint arXiv:1605.07146 (2016).

[42] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009).

[43] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, Reading digits in natural images with unsupervised feature learning (2011).

[44] Y. Le, X. Yang, Tiny imagenet visual recognition challenge, CS 231N 7 (7) (2015) 3.

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (3)

(2015) 211–252.
[46] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, J. Xiao, Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, arXiv preprint arXiv:1506.03365 (2015).
[47] Y. Sasaki, The truth of the f-measure, Teach Tutor Mater (01 2007).
[48] L. Van Der Maaten, Learning a parametric embedding by preserving local structure, in: Artificial Intelligence and Statistics, PMLR, 2009, pp. 384–391.

Appendix A. Assumptions to the Theory

For technical proofs of the theoretical parts, we assume the following: Firstly, the known classes C_i in the input space (e.g., image pixel space) follow the following regularity:

Assumption 1.

(a) Each known class C_k is a simple smooth, connected compact manifold with a nonzero volume $\mathrm{Vol}(C_k)$.

(b) For any linear path $\boldsymbol{\gamma} : [0,1] \to \mathcal{X}$ from C_j to C_k, there exist t_1 and t_2 such that $\boldsymbol{\gamma}([t_1,t_2]) \subseteq \mathcal{O}$ with $0 < t_1 < t_2 < 1$

Note that the volume, denoted by Vol, in Assumption 1a is the Lebesgue measure in the Euclidean space $\mathbb{R}^d$. Assumption 1b indicates that the in-between part of the linear interpolation between different known classes is a part of an open set. This assumption is reasonable since interpolating in a very high-dimensional space, such as image pixel space, always induces meaningless inputs in the middle part of the interpolation.

Now, as to the representation embedding function $\boldsymbol{f}$, we restrict our consideration to the neural network family with the following regularity.

Assumption 2.

(a) $\boldsymbol{f}$ is a bounded smooth parametrized neural network (i.e., $\boldsymbol{f} = \boldsymbol{f}_\theta$ with a parameter θ) with a sufficient complexity.

(b) For any linear simple smooth path $\boldsymbol{\gamma} : [0,1] \to \mathbb{R}^{d_z}$ from $\boldsymbol{x}_j \in C_j$ to $\boldsymbol{x}_k \in C_k$, the inter-class distance maximization ***strictly increases*** *the length of the path $\boldsymbol{f}(\boldsymbol{\gamma}([0,1]) \cap \mathcal{O})$ between $\boldsymbol{f}(\boldsymbol{x}_j)$ and $\boldsymbol{f}(\boldsymbol{x}_k)$.*

(c) For any simple smooth path $\boldsymbol{\gamma} : [0,1] \to \mathbb{R}^{d_z}$ from $\boldsymbol{x}_j \in C_j$ to $\boldsymbol{x}_k \in C_k$, the inter-class distance maximization does ***not decrease*** *the length of any sub-path $\boldsymbol{f}(\boldsymbol{\gamma}([t_1,t_2]) \cap \mathcal{O})$ between $\boldsymbol{f}(\boldsymbol{x}_j)$ and $\boldsymbol{f}(\boldsymbol{x}_k)$ for any $0 \le t_1 < t_2 \le 1$*

Assumption 2a is a standard regularity condition. Assumption 2b means that the inter-class separation is effective on the linear interpolating path γ. Assumption 2b is visualized in Fig. 1, which indicates that the length of projected (inter-class) path is strictly increased by metric learning. Assumption 2c means that the inter-class distance maximization involves no contradictory behavior when the inter-class path is observed locally. The assumption is reasonable and based on the empirical evidence given in Fig. 4.

In the below mathematical derivations, when we say that a quantity $Q(\boldsymbol{f})$ increases with respect to a function $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^{d_z}$, it formally means that there is a sequence $(\boldsymbol{f}^{(n)})_{n=0}^N$ of functions $\boldsymbol{f}^{(n)} : \mathbb{R}^d \to \mathbb{R}^{d_z}$ with $N \ge 1$ such that

$$Q(\boldsymbol{f}^{(n)}) \le Q(\boldsymbol{f}^{(n+1)}) \tag{A.1}$$

for all $0 \le n < N$. In the case of a strict increment, the inequality is replaced by the strict one. The decrement of $Q(\boldsymbol{f})$ is similarly defined.

Depending on the context, Q may include the vectors $\boldsymbol{w}_k \in \mathbb{R}^{d_z}$ (that serve as representation prototypes in our work): $Q = Q(\boldsymbol{f}, \{\boldsymbol{w}_k\}_{k=1}^K)$.

Appendix B. Proofs to the Theory

Proof of Proposition 1. Fix C_k. We prove a stronger result that

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \to \boldsymbol{0} \in \mathbb{R}^{d_z \times d} \tag{B.1}$$

for all $\boldsymbol{x} \in C_k$. In which case, $\frac{d\boldsymbol{f}(\boldsymbol{\gamma}(t))}{dt} \to \boldsymbol{0}$ for all $t \in \boldsymbol{\gamma}^{-1}(\boldsymbol{\gamma}([0,1]) \cap C_k)$ since $\frac{df_j(\boldsymbol{\gamma}(t))}{dt} = \sum_{i=1}^n \frac{\partial f_j}{\partial x_i}(\boldsymbol{\gamma}(t))\ \gamma_i'(t)$ for all j where $\boldsymbol{x} = (x_1, \ldots, x_d)$ and $\boldsymbol{f} = (f_1, \ldots, f_{d_z})$. Then, this implicates that the length of $\boldsymbol{f}(\boldsymbol{\gamma}([0,1]) \cap C_k)$ converges to 0 (as the pointwise convergence guarantees the L_p convergence when the functions are bounded).

Let $\boldsymbol{f}^{(n)} := \boldsymbol{f}_{\theta^{(n)}}$ be a sequence that minimizes $\mathcal{D}(\boldsymbol{f}^{(n)}(\boldsymbol{x}), \boldsymbol{w}_k)$ to 0 as $n \to N$. Let $\boldsymbol{x} \in C_k$. Since the quantity at hand is a partial derivative, without loss of generality, assume $\boldsymbol{f}(\boldsymbol{x}) = f(x)$ and $\boldsymbol{x} = x$ are scalar-valued (and also for $\boldsymbol{w}_k = w_k$). Fix $\epsilon > 0$. Then for some $\delta > 0$, we have

$$\left| \frac{d}{dx} f^{(n)}(x) - \left(\frac{f^{(n)}(x+h) - f^{(n)}(x)}{h} \right) \right| < \epsilon \tag{B.2}$$

for all $h \in [-\delta, \delta] \setminus \{0\}$. Taking $n \to N$, we obtain

$$\left| \lim_{n \to N} \frac{d}{dx} f^{(n)}(x) - 0 \right| < \epsilon \tag{B.3}$$

since $f^{(n)} \to w_k$. The arbitrariness of ϵ concludes the proof. □

Proof of Theorem 2. The intra-class minimization part is proved in the proof of Proposition 1.

For the inter-class maximization part, without loss of generality, redefine $\boldsymbol{f}$ such that $\boldsymbol{f}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{O}$, while to be the same as the original $\boldsymbol{f}$ over $\mathcal{O}$. Now, it suffices to prove the strict increment of $\int_{\mathcal{X}} \| \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}} \|_F \, d\boldsymbol{x}$ with respect to this $\boldsymbol{f}$.

Note that Assumption 2c implies that $\| \frac{\partial \boldsymbol{f} \circ \boldsymbol{\gamma}}{\partial t} \|_2$ as a function of t is non-decreasing with respect to the changing f for any simple smooth path γ. Hence, $\| \frac{\partial \boldsymbol{f}}{\partial x_l}(\boldsymbol{x}) \|_2^2$ is non-decreasing for any $l = 1, \ldots, d$. We use this property freely in the following.

Since $\mathrm{Vol}(C_i) > 0$ for all known classes C_i, for any pair of different known classes C_i and C_j, we have a $(d-1)$-dimensional hyperplane $P \subseteq \mathcal{X}$ such that

$$\mathrm{Vol}_{d-1}\left(\rho(C_i) \cap \rho(C_j) \right) > 0 \tag{B.4}$$

where Vol_{d-1} is the $(d-1)$-dimensional volume, and $\rho(C_i)$ is the projection of C_i to the hyperplane P. Since a coordinate change under rotation and translation does not change the volume integral of a function, we assume without loss of generality that

$$\boldsymbol{e}_k \perp P; \tag{B.5}$$

that is, $\boldsymbol{e}_k$ is perpendicular to P where $\boldsymbol{e}_k$ is the k-th standard basis element of $\mathbb{R}^d$.

Now, observe

$$\int_{\mathcal{X}} \left\| \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}} \right\|_F^2 d\boldsymbol{x} = \sum_{l=1}^d \int_{\mathcal{X}} \left\| \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_l} \right\|_2^2 d\boldsymbol{x}. \tag{B.6}$$

Since $\int_{\mathcal{X}} \|\frac{\partial f(\boldsymbol{x})}{\partial x_l}\|_2^2 \, d\boldsymbol{x}$ is non-decreasing for $l \neq k$, it suffices to show that $\int_{\mathcal{X}} \|\frac{\partial f(\boldsymbol{x})}{\partial x_k}\|_2^2 \, d\boldsymbol{x}$ is strictly increasing. Let

$$R = \{\widehat{\boldsymbol{x}}_k \in [-1,1]^{d-1} : \boldsymbol{x} \in \rho(C_i) \cap \rho(C_j)\}. \tag{B.7}$$

where $\widehat{\boldsymbol{x}}_k$ denotes $\widehat{\boldsymbol{x}}_k := (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d)$ that removes the k-th element of $\boldsymbol{x}$. Note that

$$\int_{\mathcal{X}} \left\| \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_k} \right\|_2^2 d\boldsymbol{x} = \int_R \int_{-1}^1 \left\| \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_k} \right\|_2^2 dx_k d\widehat{\boldsymbol{x}}_k + \int_{R^c} \int_{-1}^1 \left\| \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_k} \right\|_2^2 dx_k d\widehat{\boldsymbol{x}}_k \tag{B.8}$$

where $d\widehat{\boldsymbol{x}}_k := dx_1 \cdots dx_{k-1} dx_{k+1} \cdots dx_d$. Since the second term on the RHS of the above equation is non-decreasing, we consider the first only, whose inner term

$$\int_{-1}^1 \|\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_k}\|_2^2 \, dx_k. \tag{B.9}$$

is decomposed into

$$\int_{-1}^a + \int_a^b + \int_b^1 \left\| \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_k} \right\|_2^2 dx_k \tag{B.10}$$

where the scalars $a = a(\widehat{\boldsymbol{x}}_k)$ and $b = b(\widehat{\boldsymbol{x}}_k)$ are the infimum and supremum of $\{x_k : \boldsymbol{x} \in C_i \cup C_j\}$, respectively, with $\widehat{\boldsymbol{x}}_k \in R$. The first and third terms non-decrease, thus ignored. To compute the mid term $\int_a^b \|\frac{\partial f(\boldsymbol{x})}{\partial x_k}\|_2^2 dx_k$, consider a path

$$\boldsymbol{\gamma}(t) = (x_1, \ldots, x_{k-1}, a + (b-a)t, x_{k+1}, \ldots, x_d) \tag{B.11}$$

that depends on $\widehat{\boldsymbol{x}}_k = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d) \in R$. Then, $\boldsymbol{\gamma}$ is a path from C_i to C_j or the other way, and

$$\int_a^b \left\| \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_k} \right\|_2^2 dx_k = (b-a) \int_0^1 \left\| \frac{d\boldsymbol{f} \circ \boldsymbol{\gamma}(t)}{dt} \right\|_2^2 dt \tag{B.12}$$

$$= (b-a)\ell(\boldsymbol{f} \circ \boldsymbol{\gamma}) \tag{B.13}$$

where $\ell(\boldsymbol{f} \circ \boldsymbol{\gamma}) = \int_0^1 \|\frac{df \circ \gamma(t)}{dt}\|_2^2 dt$. In summary,

$$\int_O \left\| \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}(\boldsymbol{x}) \right\|_2^2 d\boldsymbol{x} = A(\boldsymbol{f}) + \int_R [b(\widehat{\boldsymbol{x}}_k) - a(\widehat{\boldsymbol{x}}_k)] \ell(\boldsymbol{f} \circ \boldsymbol{\gamma}) d\widehat{\boldsymbol{x}}_k. \tag{B.14}$$

Here, the inter-class maximization does not change R, $\widehat{\boldsymbol{x}}_k$, γ, $a(\widehat{\boldsymbol{x}}_k)$, and $b(\widehat{\boldsymbol{x}}_k)$. On the other hand, the inter-class maximization does not decrease the term $A(\boldsymbol{f})$. Moreover, note that $\gamma \cap O$ is not empty and contains an interval by Assumption 1b. Thus, by Assumption 2b, the inter-class maximization strictly increases the term $\ell(\boldsymbol{f} \circ \boldsymbol{\gamma})$, thereby strictly increasing the global integral of the Jacobian norm over the open set. This finishes the proof. □

Proof of Corollary 4. By the above theorem, the inter-class distance maximization increases the integral $\int_O \|\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}\|_F \, d\boldsymbol{x}$. Now, the integral $\int_O \|\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}\|_F \, d\boldsymbol{x}$ can be decomposed into

$$\int_O \|\frac{\partial f}{\partial \boldsymbol{x}}(\boldsymbol{x})\|_F \, d\boldsymbol{x} = \mathrm{Vol}(S) \cdot \mathop{\mathbb{E}}_{\boldsymbol{x} \sim S} [\|\frac{\partial f}{\partial \boldsymbol{x}}(\boldsymbol{x})\|_F] \tag{B.15}$$

where $\boldsymbol{x} \sim S$ is uniformly sampled is the support set of the Jacobian norm over the open set O. Based on the decomposition, the inter-class distance maximization increases the volume $\mathrm{Vol}(S)$ of the set S and/or the expected Jacobian norm over S. □

Appendix C. Proofs to the Method

Propositions 6 and 7. *With the optimal prototypes $\{\boldsymbol{w}_k\}_{k=1}^K$ for a single sample reprsentation $\boldsymbol{f}(\boldsymbol{x})$ paired with label y, we have the collapse $\boldsymbol{w}_k = -\boldsymbol{w}_y$ for all $k \neq y$ if $m_n = 0$, while $\measuredangle(\boldsymbol{w}_k, -\boldsymbol{w}_y) \geq m_n$ with $k \neq y$ if $0 < m_n < \pi/2$.*

Proof of Propositions 6 and 7. For the optimal prototypes $\{\boldsymbol{w}_k\}_{k=1}^K$, we have $s_y = \max_{\boldsymbol{w}_y} s_y$ and $s_k = -1$. Regardless of whether $m_n > 0$ or not, we have $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{w}_y$. For the negative pair, if $m_n = 0$, then $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{w}_k$, and hence $-\boldsymbol{w}_y = \boldsymbol{w}_k$ for all $k \neq y$. If $m_n > 0$, then $s_k = -1$ when the angle between $\boldsymbol{w}_k$ and $\boldsymbol{f}(\boldsymbol{x})$ is $\pi - m_n$, implying that $\measuredangle(\boldsymbol{w}_k, -\boldsymbol{w}_y) = \measuredangle(\boldsymbol{w}_k, -\boldsymbol{f}(\boldsymbol{x})) = m_n$, finishing the proof. □

Proof of Proposition 8. Observe that

$$\frac{\partial \mathcal{L}_{\mathrm{SCE}}}{\partial \theta} = -c\left(\sum_{j \neq y} e^{s_j - s_y}\right) \frac{\partial s_y}{\partial \theta} + c \sum_{k \neq y} e^{s_k - s_y} \frac{\partial s_k}{\partial \theta} \tag{C.1}$$

where $\mathcal{L}_{\mathrm{SCE}} = \log(1 + \sum_{k \neq y} e^{s_k - s_y})$ and $c = (1 + \sum_{j \neq y} e^{s_j - s_y})^{-1}$. On the other hand,

$$\frac{L_{\text{m-OvR}}}{\partial \theta} = -[1 + e^{s_y}]^{-1} \frac{\partial s_y}{\partial \theta} + \sum_{k \neq y} [1 + e^{s_k}]^{-1} \frac{\partial s_k}{\partial \theta}. \tag{C.2}$$

Hence, the inter-class gradient $\frac{\partial s_k^{\mathrm{SCE}}}{\partial \theta}$ for SCE is

$$\frac{\partial s_k^{\mathrm{SCE}}}{\partial \theta} = c e^{s_k - s_y} \frac{\partial s_k}{\partial \theta}, \tag{C.3}$$

while the inter-class gradient for $\frac{\partial s_k^{\text{m-OvR}}}{\partial \theta}$ is

$$\frac{\partial s_k^{\text{m-OvR}}}{\partial \theta} = [1 + e^{s_k}]^{-1} \frac{\partial s_k}{\partial \theta}. \tag{C.4}$$

To prove our claim, it suffices to show that $\frac{\partial s_k^{\text{m-OvR}}}{\partial \theta} > \frac{\partial s_k^{\mathrm{SCE}}}{\partial \theta}$. To this end, observe that

$$\frac{e^{s_k - s_y}}{1 + \sum_{j \neq y} e^{s_j - s_y}} < \frac{1}{1 + e^{-s_k}}, \tag{C.5}$$

which is equivalent to

$$e^{s_k} + 1 < \sum_j e^{s_j} = e^{s_k} + \sum_{j \neq k,y} e^{s_j} + e^{s_y}, \tag{C.6}$$

which holds if $s^{s_y} > 0$ due to $\sum_{j \neq k,y} e^{s_j} \geq 0$. This completes our proof. □

Appendix D. Additional Empirical Results

The results given in D.16 show that the Jacobian norm trend that we observed in the main sections holds the same way for the softmax cross entropy models.

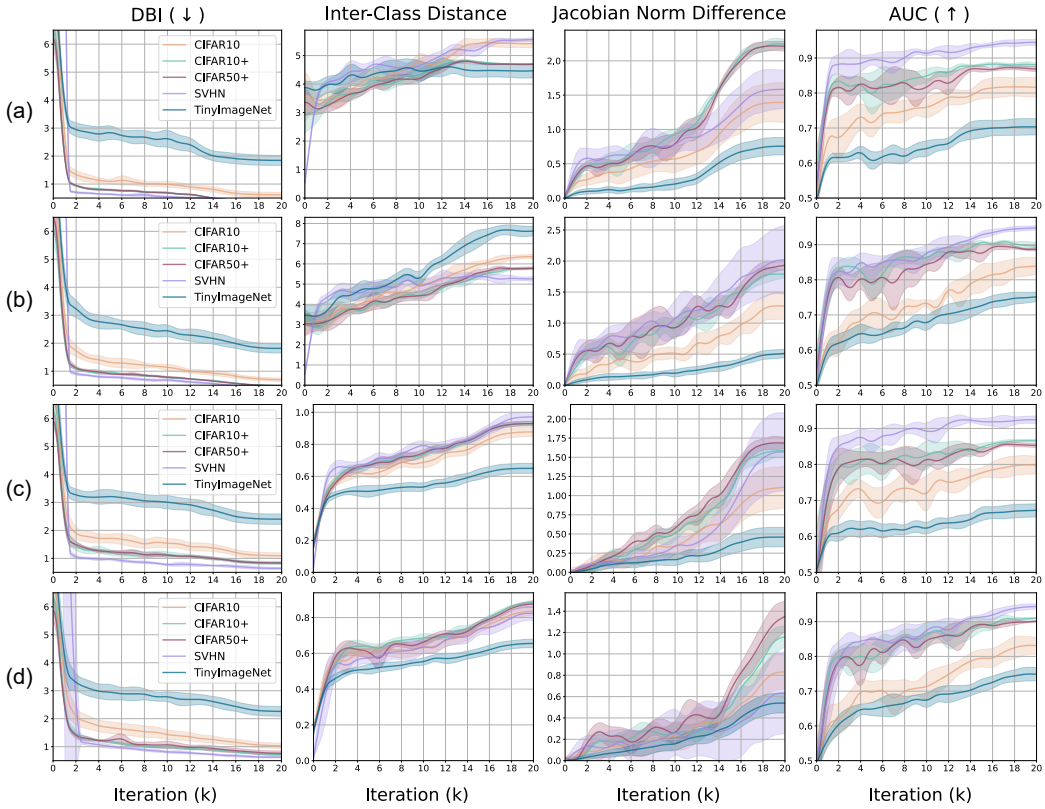

Figure D.16: Several metrics measured while different discriminative models are begin trained. (a) SCE without data augmentation, (b) SCE with data augmentation, (c) SCE with normalized embedding but without data augmentation, (d) SCE with normalized embedding and data augmentation.