

Adversarial Contrastive Learning via Asymmetric InfoNCE

Qiyong Yu^{1,2*}, Jieming Lou², Xianyuan Zhan¹, Qizhang Li², Wangmeng Zuo²,
Yang Liu^{1,3}, and Jingjing Liu^{1*}

¹ Institute for AI Industry Research, Tsinghua University, China

² School of Computer Science and Technology, Harbin Institute of Technology, China

³ Department of Computer Science and Technology, Tsinghua University, China
yuqy22@mails.tsinghua.edu.cn, jjliu@air.tsinghua.edu.cn

Abstract. Contrastive learning (CL) has recently been applied to adversarial learning tasks. Such practice considers adversarial samples as additional positive views of an instance, and by maximizing their agreements with each other, yields better adversarial robustness. However, this mechanism can be potentially flawed, since adversarial perturbations may cause instance-level *identity confusion*, which can impede CL performance by pulling together different instances with separate identities. To address this issue, we propose to treat adversarial samples unequally when contrasted, with an asymmetric InfoNCE objective (*A-InfoNCE*) that allows discriminating considerations of adversarial samples. Specifically, adversaries are viewed as *inferior positives* that induce weaker learning signals, or as *hard negatives* exhibiting higher contrast to other negative samples. In the asymmetric fashion, the adverse impacts of conflicting objectives between CL and adversarial learning can be effectively mitigated. Experiments show that our approach consistently outperforms existing Adversarial CL methods across different finetuning schemes. The proposed A-InfoNCE is also a generic form that can be readily extended to other CL methods. Code is available at <https://github.com/yqy2001/A-InfoNCE>.

Keywords: Adversarial Contrastive Learning, Robustness, Self-supervised Learning

1 Introduction

Well-performed models trained on clean data can suffer miserably when exposed to simply-crafted adversarial samples [42,20,4,15]. There has been many adversarial defense mechanisms designed to boost model robustness using labeled data [29,40,50,47,51,52,2]. In practice, however, obtaining large-scale annotated data can be far more difficult and costly than acquiring unlabeled data. Leveraging easily-acquired unlabeled data for adversarial learning, thus becomes particularly attractive.

* Corresponding authors

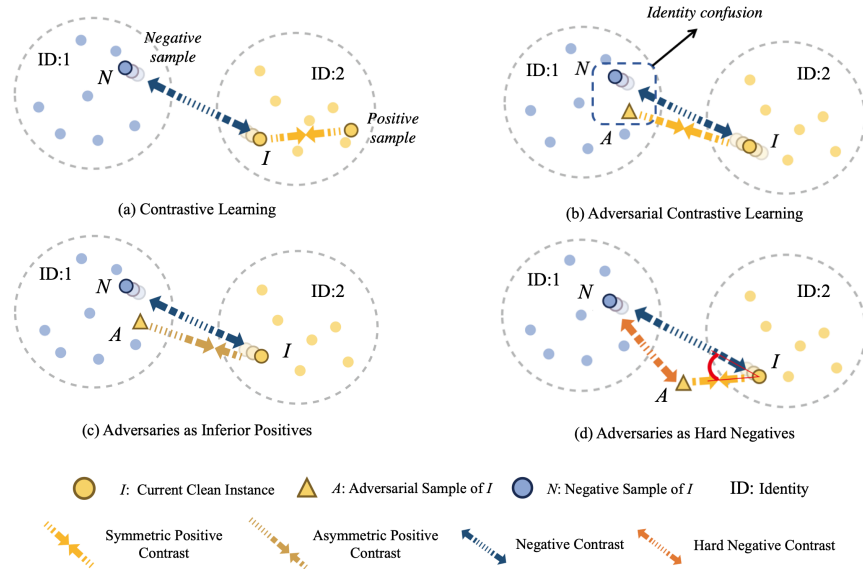


Fig. 1. Illustrations of (a) Contrastive Learning; (b) Adversarial Contrastive Learning; and our proposed methods for viewing adversarial samples asymmetrically as: (c) Inferior Positives (asymmetric contrast), and (d) Hard Negatives. In each circle, data points are augmentations of the same instance, sharing the same *Identity*. In (b), the Adversarial sample (A) shares the same *Identity* ($ID:2$) as the current Instance (I), but resides close to a different *Identity* ($ID:1$), thus *Identity Confusion* problem occurs. Specifically, the Adversarial sample (A) of Instance (I) exhibits similar representations to the Negative sample (N) of (I), which makes the positive contrast ($A \leftrightarrow I$) and negative contrast ($N \leftrightarrow I$) undermine each other in the training process (colored figure).

Contrastive Learning (CL) [23], which performs instance discrimination [48] (Figure 1 (a)) by maximizing agreement between augmentations of the same instance in the learned latent features while minimizing the agreement between different instances, has made encouraging progress in self-supervised learning [9,24,11,22]. Due to its effectiveness in learning rich representations and competitive performance over fully-supervised methods, CL has seen a surge of research in recent years, such as positive sampling [9,44,3,45], negative sampling [24,28,13,48], pair reweighting [13,39], and different contrast methods [22,6,33].

Recently, contrastive learning has been extended to adversarial learning tasks in a self-supervised manner, leading to a new area of *adversarial contrastive learning* (Adversarial CL) [32,18,27,21]. The main idea is to generate adversarial samples as additional positives of the same instance [32,18,27] for instance-wise attack, and maximize the similarity between clean views of the instance and their adversarial counterparts as in CL, while also solving the min-max optimization problem following canonical adversarial learning objective [35,40,50,47,51,52]. For example, RoCL[32] first proposed an attack mechanism against contrastive

loss to confuse the model on instance-level identity, in a self-supervised adversarial training framework. AdvCL[18] proposed to minimize the gap between unlabeled contrast and labeled finetuning by introducing pseudo-supervision in the pre-training stage.

Although these Adversarial CL methods showed improvement on model robustness, we observe that a direct extension from CL to adversarial learning (AL) can introduce ineffective CL updates during training. The core problem lies in that they add worst-case perturbations δ that no longer guarantee the preservation of instance-level identity [32] (*i.e.*, different from other data augmentation methods, adversarial samples can reside faraway from the current instance in the feature space after several attack iterations, because the attack objective is to make adversaries away from the current instance while approximating other instances, against the CL objective). As illustrated in Figure 1(b), when the adversarial sample (A) of the current instance (I) are in close proximity to negative samples (N), CL objective minimizes the agreement between negative samples and current instance (I and N are pushed away from each other), while AL objective maximizes the agreement between adversarial samples and current instance (A and I are pulled together as A is considered as an augmented view of I). Meanwhile, A and N share similar representations, which renders the two objectives contradicting to each other. We term this conflict as “*identity confusion*”, it means A attracts and ‘confuses’ I with a false identity induced by N , which impedes both CL and AL from achieving their respective best performance.

To address this issue of *identity confusion*, we propose to treat adversarial samples unequally and discriminatingly, and design a generic asymmetric InfoNCE objective (A -InfoNCE), in order to model the asymmetric contrast strengths between positive/negative samples. Firstly, to mitigate the direct pull between adversarial sample (A) and current instance (I) (Figure 1 (c)) that might dampen the effectiveness of CL, we propose to treat adversarial samples as *inferior positives* that induce weaker learning signals to attract their counterparts in a lower degree when performing positive contrasts. This asymmetric consideration in AL promises a trade-off and reduces conflicting impact on the CL loss.

Secondly, to encourage adversarial samples (A) to escape from false identities induced by negative samples (N) that share similar representations to (A) (pushing A away from N) (Figure 1(d)), we consider adversarial samples (A) as *hard negatives* [39] of other negative samples (N), by strengthening the negative contrast between A and N in CL computation. To effectively sample true adversarial negatives and re-weight each sample, we follow positive-unlabeled learning [16,17] and contrastive negatives reweighting [39,13] practice.

Our contributions are summarized as follows: 1) We propose an generic asymmetric InfoNCE loss, A -InfoNCE, to address the *identity confusion* problem in Adversarial CL, by viewing adversarial samples as *inferior positives* or *hard negatives*. 2) Our approach is compatible to existing Adversarial CL methods, by simply replacing standard CL loss with A -InfoNCE. 3) Experiments on CIFAR-

10, CIFAR-100 and STL-10 show that our approach consistently outperforms existing Adversarial CL methods.

2 Asymmetric InfoNCE

2.1 Notations

Contrastive Learning (CL) CL aims to learn generalizable features by maximizing agreement between self-created positive samples while contrasting to negative samples. In typical contrastive learning, each instance x will be randomly transformed into two views (x_1, x_2) , then fed into a feature encoder f with parameters θ to acquire normalized projected features, *i.e.*, $z_i = f(x_i; \theta)$. Let $\mathcal{P}(i)$ denote the set of positive views of x_i , containing the views transformed from x with the same instance-level *identity* (*e.g.*, augmentations of the original image x_i); $\mathcal{N}(i)$ denotes the set of negative views of x_i , containing all the views from other instances. The conventional InfoNCE loss function [36] used in CL for a positive pair (x_i, x_j) is defined as:

$$\mathcal{L}_{\text{CL}}(x_i, x_j) = -\log \frac{\exp(\text{sim}(z_i, z_j)/t)}{\exp(\text{sim}(z_i, z_j)/t) + \sum_{k \in \mathcal{N}(i)} \exp(\text{sim}(z_i, z_k)/t)} \quad (1)$$

where x_i serves as the anchor, $\text{sim}(z_i, z_j)$ denotes a similarity metric (*e.g.*, cosine similarity) between z_i and z_j , and t is a temperature parameter. The final loss of the CL problem is averaged over all positive pairs of instances.

Adversarial CL Adversarial CL can be regarded as an extension of CL by adding adversarial samples into the positive sets $\mathcal{P}(\cdot)$ to contrast. Adversarial CL is typically modeled as the following min-max optimization formulation to incorporate instance-wise attack [35,18]:

$$\min_{\theta} \mathbb{E}_{x \in \mathcal{X}} \max_{\|\delta\|_{\infty} \leq \epsilon} \sum_i \sum_{j \in \mathcal{P}(i)} \mathcal{L}_{\text{CL}}(x_i, x_j), \quad \mathcal{P}(i) \leftarrow \mathcal{P}(i) \cup \{\hat{x}_i + \delta\} \quad (2)$$

where \hat{x}_i is the view of x_i used to generate adversarial samples, δ is the adversarial perturbation whose infinity norm is constrained as less than ϵ . In the above formulation, the inner maximization problem constructs adversarial samples by maximizing the contrastive loss, and the outer minimization problem optimizes the expected worst-case loss w.r.t. the feature encoder f .

2.2 Asymmetric InfoNCE: A Generic Learning Objective

Current Adversarial CL frameworks directly inherit CL’s conventional contrastive loss (*e.g.*, InfoNCE) to evaluate the similarity between adversarial and clean views in a symmetric fashion. This can result in ineffective or even conflicting updates during CL training as aforementioned. To address this challenge, we

propose a generic Asymmetric InfoNCE loss (*A-InfoNCE*) to incorporate the asymmetric influences between different contrast instances, given by:

$$\mathcal{L}_{\text{CL}}^{\text{asym}}(x_i, x_j; \alpha, \lambda^p, \lambda^n) = -\log \frac{\lambda_j^p \cdot \exp(\text{sim}^\alpha(z_i, z_j)/t)}{\lambda_j^p \cdot \exp(\text{sim}^\alpha(z_i, z_j)/t) + \sum_{k \in \mathcal{N}(i)} \lambda_k^n \cdot \exp(\text{sim}^\alpha(z_i, z_k)/t)} \quad (3)$$

where $\text{sim}^\alpha(\cdot)$ is a generalized similarity metric that enables the incorporation of asymmetric relationships (a concrete instantiation is described in the next section); λ^p and λ^n are asymmetric weighting factors for positive and negative pairs, respectively. It is worth noting that although A-InfoNCE is proposed to address the *identity confusion* issue in Adversarial CL, it can be easily extended to other CL settings when the asymmetric characteristics between different views need to be captured. A-InfoNCE can also be generalized to many existing CL methods, for example, $\mathcal{P}(i)$ and $\mathcal{N}(i)$ can be altered to different choices of positive and negative views; $\text{sim}^\alpha(z_i, z_j)$ is also changeable to a symmetric similarity metric for z_i and z_j . λ^p and λ^n control the weights of different positive/negative pairs. Generalization strategies are itemized below:

- If $\text{sim}^\alpha(z_i, z_j)$ is a symmetric similarity metric and $\lambda^p, \lambda^n = 1$, it degrades to the conventional InfoNCE loss used in CL [9].
- If $\mathcal{P}(i)$ is altered, it corresponds to positives sampling [44,3,45]. When we add adversaries into $\mathcal{P}(i)$, it degenerates to the conventional Adversarial CL objectives, where $\lambda^p, \lambda^n = 1$ with symmetric $\text{sim}^\alpha(z_i, z_j)$ [32,27,18].
- If we seek better $\mathcal{N}(i)$, it echoes negative sampling methods [39,28] such as Moco [24], which maintains a queue of consistent negatives; or mimics DCL [13] that debiases $\mathcal{N}(i)$ into true negatives.
- If we change λ^p and λ^n , it mirrors the pair reweighting works [13,39] that assign different weights to each pair according to a heuristic measure of importance such as similarity.

While most existing methods adopt a symmetric similarity metric, we claim that in some scenarios the asymmetric similarity perspective needs to be taken into account, especially when the quality and property of different views vary significantly. In this paper, we focus on the study of Adversarial CL, and demonstrate the benefits of capturing the asymmetric relationships between adversaries and clean views. Specifically, we design two instantiations to model the asymmetric relationships between adversarial and clean samples, as detailed in next section. Both instantiations can be integrated into the proposed *A-InfoNCE* framework.

3 Adversarial Asymmetric Contrastive Learning

This section explains the instantiations of the *A-InfoNCE* loss for Adversarial CL. From the *inferior-positive* perspective, to reduce the impact of identity confusion, we first design a new asymmetric similarity metric $\text{sim}^\alpha(z_i, z_j^{\text{adv}})$ for modeling the asymmetric relationships and weakening the learning signals from adversarial examples. From the *hard-negative* perspective, we view adversaries as hard negatives for other negative samples, and reweight each negative pairs by assigning similarity-dependent weights to ease the identity confusion.

3.1 Adversarial Samples as Inferior Positives

Adversarial samples with different identities may attract their anchors (clean samples) in a contradicting manner to the exertion of CL. By weakening the learning signal from these adversarial examples in positive contrast (as *inferior positives* that attract the anchors less), we can effectively mitigate the undesired pull from clean samples via an adaptive gradient stopping strategy.

Asymmetric Similarity Function. As the symmetric nature of InfoNCE can bring conflicts in Adversarial CL, we design a new asymmetric similarity function $\text{sim}^\alpha(z_i, z_j)$ for *A-InfoNCE*, by manipulating the scale of gradient for each contrasted branch. We decompose it into two parts for each branch:

$$\text{sim}^\alpha(z_i, z_j) = \alpha \cdot \overline{\text{sim}}(z_i, z_j) + (1 - \alpha) \cdot \overline{\text{sim}}(z_j, z_i) \quad (4)$$

where $\overline{\text{sim}}(a, b)$ means the one-sided similarity of a to b , *i.e.*, when maximizing $\overline{\text{sim}}(a, b)$, we freeze b and only move a towards b . This can be implemented by stopping the gradient back-propagation for b and only optimizing a .

We use a hyperparameter α to control how much z_i and z_j head towards each other. For a clean sample and an adversarial sample, we let α denote the coefficient of the clean branch’s movement. If α is 0, it performs total gradient freezing on the clean branch and only adversarial representations are optimized through training. Our empirical analysis finds that α is relatively easy to tune for boosted performance. We show that any value lower than 0.5 brings reasonable performance boost (see Figure 2), when clean samples move less towards adversaries, following the intrinsic asymmetric property of Adversarial CL.

Adaptive α -annealing. When the *identity confusion* is at play, it is necessary to treat adversarial samples inferior to ensure model robustness. But as training progresses, when model learns robust representations and the negative identity-changing impact of adversarial perturbation wanes, we consider adversarial perturbation as strong augmentations, equal to other typical transformations [9].

The question is how to measure the reduction of instance confusion effect. Here we take a geometry perspective and propose to adaptively tune the proportional coefficient α on-the-fly based on Euclidean distance. Let $d_{i,j} = \|z_i - z_j\|_2$ denote the distance between an original image and its adversarial view in the representation space. Given α_{min} , d_{max} , α_{max} , d_{min} , the goal is for α to be α_{max} when the distance approximates d_{min} , and α_{min} to be close to d_{max} . During training, we first compute the current representation distance d , then use a simple linear annealing strategy to compute α :

$$\alpha = \alpha_{min} + (d_{max} - d) \frac{\alpha_{max} - \alpha_{min}}{d_{max} - d_{min}} \quad (5)$$

d_{min} and α_{min} can be treated as hyperparameters. α_{max} is 0.5, indicating adversarial perturbation is equal to other transformations and $\text{sim}^\alpha(z_i, z_j)$ degrades to the symmetric similarity. Moreover, we use the first N epochs as a warm-up to compute the average distance as d_{max} , in which period α is fixed.

Adversarial CL Loss with Inferior Positives. With the above asymmetric similarity function $\text{sim}^\alpha(\cdot)$ and the *A-InfoNCE* loss function $\mathcal{L}_{\text{CL}}^{\text{asym}}(x_i, x_j; \alpha, \lambda^p, \lambda^n)$, the complete Adversarial CL loss with *inferior positives* (IP) can be written as:

$$\mathcal{L}^{\text{IP}} = \sum_i \sum_{j \in \mathcal{P}(i)} \mathcal{L}_{\text{CL}}^{\text{asym}}(x_i, x_j; 0.5, 1, 1) + \gamma \cdot \sum_i \sum_{j \in \mathcal{P}(i)} \mathcal{L}_{\text{CL}}^{\text{asym}}(x_i, x_j^{\text{adv}}; \alpha, 1, 1) \quad (6)$$

where the first part stands for standard CL loss that maximizes the similarity between two clean views, which is symmetric ($\alpha = 0.5$) with $\lambda^p = \lambda^n = 1$, degrading to the conventional InfoNCE loss. The second part is a robust CL loss that maximizes the agreement between clean and adversarial views, but uses the asymmetric similarity function (4) with a hyperparameter α that gives weaker learning signals to the counterparts of inferior adversarial samples. The hyperparameter γ balances the robustness and accuracy objectives.

3.2 Adversarial Samples as Hard Negatives

Besides inferior positives, we also propose an alternative view of adversaries as *hard negatives* [39] that be pushed away from surrounding data points with higher weights. This can potentially assuage the confusion brought by adversarial samples of the current instance residing too close to the negative samples of the same instance (as illustrated in Figure 1 (d)). Furthermore, this strategy encourages the model towards more robustness-aware, by giving adversarial samples possessing indiscriminating features higher weights in the pretraining stage, further enhancing Adversarial CL.

In practice, we assign a weight of similarity to each pair. To set a basis for weight assigning, we adopt a simple and adaptive weighting strategy used in [39], *i.e.*, taking each pair’s similarity as its weight, with $w_{i,j} = \exp(\text{sim}(z_i, z_j)/t)$. By doing so, the adversaries with bad instance-level identity (greater similarity to negative samples) can be automatically assigned with higher weights. The weights can adaptively decay as the instance identity recovers during training.

However, as the commonly-used $\mathcal{N}(i)$ is uniformly sampled from the entire data distribution $p(x)$ [13] (*e.g.*, SimCLR [9] uses other instances in the current batch as negative samples), simply taking similarities as weights may heavily repel semantically-similar instances whose embeddings should be close. To estimate the true negatives distribution $p^-(x)$, we take advantage of PU-learning [16,17] and resort to DCL,HCL [13,39] to debias negative sampling.

PU-learning [16] decomposes the data distribution as: $p(x) = \tau p^+(x) + (1 - \tau)p^-(x)$, where $p^+(x), p^-(x)$ denote the distribution of data from the same or different class of x , and τ is the class prior. Thus $p^-(x)$ can be rearranged as $p^-(x) = (p(x) - \tau p^+(x))/(1 - \tau)$. We can use all instances and positive augmentations containing adversarial samples of x to estimate $p(x)$ and $p^+(x)$, respectively. Following [13], we debias the negative contrast part in (3) as:

$$\frac{1}{1 - \tau} \left(\sum_{k \in \mathcal{N}(i)} w_{i,k}^n \cdot \exp(\text{sim}^\alpha(z_i, z_k)/t) - \frac{N}{M} \cdot \tau \sum_{j \in \mathcal{P}(i)} w_{i,j}^p \cdot \exp(\text{sim}^\alpha(z_i, z_j)/t) \right) \quad (7)$$

where M, N are the numbers of positives and negatives, $w_{i,k}^n$ is the aforementioned weights for negatives, $w_{i,j}^p$ is an expandable weight for positives (set as 1 in our implementation, other choices can be further explored in the future work).

Adversarial CL Loss with Hard Negatives. We substitute (7) into the *A-InfoNCE* loss function (3) and rearrange it, acquiring the instantiation of *A-InfoNCE* loss with *hard negatives* (HN), with concrete forms of λ^p and λ^n as:

$$\mathcal{L}^{HN} = \sum_i \sum_{j \in \mathcal{P}(i)} \mathcal{L}_{\text{CL}}^{\text{asym}}(x_i, x_j; \alpha, \frac{M - (M + N)\tau}{M - M\tau} w_{i,j}^p, \frac{1}{1 - \tau} w_{i,k}^n), \quad k \in \mathcal{N}(i) \quad (8)$$

Due to the lack of class information, we treat τ as a hyperparameter and set as [13] suggested.

Combined Adversarial CL Loss. Finally, we can view adversaries both as inferior positives and hard negatives for other negative samples. This leads to following combined Adversarial CL loss:

$$\begin{aligned} \mathcal{L}^{IP+HN} = & \sum_i \sum_{j \in \mathcal{P}(i)} \mathcal{L}_{\text{CL}}^{\text{asym}}(x_i, x_j; 0.5, \frac{M - (M + N)\tau}{M - M\tau} w_{i,j}^p, \frac{1}{1 - \tau} w_{i,k}^n) + \\ & \gamma \cdot \sum_i \sum_{j \in \mathcal{P}(i)} \mathcal{L}_{\text{CL}}^{\text{asym}}(x_i, x_j^{\text{adv}}; \alpha, \frac{M - (M + N)\tau}{M - M\tau} w_{i,j}^p, \frac{1}{1 - \tau} w_{i,k}^n), \quad k \in \mathcal{N}(i) \end{aligned} \quad (9)$$

4 Experiments

To demonstrate the effectiveness and generalizability of the proposed approach, we present experimental results across different datasets and model training strategies. Our methods are compatible with existing Adversarial CL frameworks, and can be easily incorporated by replacing their CL loss. We choose two baselines and replace their loss with \mathcal{L}^{IP} (in Equation 6), \mathcal{L}^{HN} (8) and \mathcal{L}^{IP+HN} (9) for evaluation.

Datasets. We mainly use CIFAR-10 and CIFAR-100 for our experiments. Each dataset has 50,000 images for training and 10,000 for test. STL-10 is also used for transferability experiments. Following previous work [18], we use ResNet-18 [25] as the encoder architecture in all experiments.

Baselines. We compare with two baselines: RoCL [32], the first method to combine CL and AL; and AdvCL [18], the current state-of-the-art framework. During experiments, we observe severe overfitting of AdvCL when training 1000 epochs (experiment setting in the original paper), with performance inferior to training for 400 epochs. Thus, we pre-train 400 epochs on AdvCL at its best-performance setting. All other settings are the same as original papers except for some hyperparameter tuning. Our methods are also compatible with some recent

Table 1. Results for replacing the objectives of the two baselines with \mathcal{L}^{IP} , \mathcal{L}^{HN} and \mathcal{L}^{IP+HN} , in Standard Accuracy (SA) and Robust Accuracy (RA). The pre-trained methods are evaluated under the Linear Probing (LP), Adversarial Linear Finetuning (ALF) and Adversarial Full Finetuning (AFF) strategies. Supervised methods are trained under conventional adversarial training scheme

Dataset	Pre-training Methods	Finetuning Strategies								
		Linear Probing		Adversarial Linear Finetuning		Adversarial Full Finetuning				
		SA	RA	SA	RA	SA	RA			
CIFAR 10	Supervised	AT [35]	-	-	-	-	78.99	47.41	1	
		TRADES [51]	-	-	-	-	81.00	53.27	2	
	Self-	RoCL [32]	83.84	38.98	79.23	47.82	77.83	50.54	3	
		w/ \mathcal{L}^{IP}	87.63	41.46	84.15	50.08	78.97	50.29	4	
		w/ \mathcal{L}^{HN}	84.14	40.00	79.40	48.31	78.84	51.73	5	
		w/ \mathcal{L}^{IP+HN}	85.69	42.96	81.91	50.90	80.06	52.95	6	
		Supervised	AdvCL [18]	81.35	51.00	79.24	52.38	83.67	53.35	7
			w/ \mathcal{L}^{IP}	82.37	52.33	80.05	53.22	84.12	53.56	8
	w/ \mathcal{L}^{HN}		81.34	52.61	78.69	53.20	83.44	54.07	9	
	w/ \mathcal{L}^{IP+HN}		83.15	52.65	80.41	53.19	83.93	53.74	10	
CIFAR 100	Supervised	AT [35]	-	-	-	-	49.49	23.00	11	
		TRADES [51]	-	-	-	-	54.59	28.43	12	
	Self-	RoCL [32]	55.71	18.49	49.30	25.84	51.19	26.69	13	
		w/ \mathcal{L}^{IP}	59.30	21.34	54.49	30.33	52.39	27.84	14	
		w/ \mathcal{L}^{HN}	58.77	21.17	56.38	28.03	55.85	29.57	15	
		w/ \mathcal{L}^{IP+HN}	59.74	22.54	57.57	29.22	55.79	29.92	16	
		Supervised	AdvCL [18]	47.98	27.99	47.45	28.29	57.87	29.48	17
			w/ \mathcal{L}^{IP}	49.48	28.84	45.39	28.40	59.44	30.49	18
	w/ \mathcal{L}^{HN}		49.44	29.01	47.32	28.69	58.41	29.93	19	
	w/ \mathcal{L}^{IP+HN}		50.59	29.12	45.72	28.45	58.70	30.66	20	

work like SwARo [46] and CLAF [38], by modeling the asymmetry between clean and adversarial views as aforementioned.

Evaluation. Following [27] and [18], we adopt three finetuning strategies to evaluate the effectiveness of contrastive pre-training: 1) Linear Probing (LP): fix the encoder and train the linear classifier; 2) Adversarial Linear Finetuning (ALF): adversarially train the linear classifier; 3) Adversarial Full Finetuning (AFF): adversarially train the full model. We consider two evaluation metrics: 1) Standard Accuracy (SA): classification accuracy over clean images; 2) Robust Accuracy (RA): classification accuracy over adversaries via PGD-20 attacks [35]. Robustness evaluation under more diverse attacks is provided in the appendix.

Table 2. Transferring results from CIFAR-10/100 to STL-10, compared with AdvCL [18], evaluated in Standard accuracy (SA) and Robust accuracy (RA) across different finetuning methods with ResNet-18

Dataset	Pre-training Methods	Finetuning Strategies					
		Linear Probing		Adversarial Linear Finetuning		Adversarial Full Finetuning	
		SA	RA	SA	RA	SA	RA
CIFAR10 ↓ STL10	AdvCL [18]	64.45	37.25	60.86	38.84	67.89	38.78
	w/ \mathcal{L}^{IP}	64.83	37.30	61.95	38.90	68.25	39.03
	w/ \mathcal{L}^{HN}	65.24	38.18	62.83	39.70	67.88	39.75
	w/ \mathcal{L}^{IP+HN}	67.19	37.00	61.34	39.35	67.95	39.12
CIFAR100 ↓ STL10	AdvCL [18]	52.28	30.01	49.84	32.14	63.13	35.24
	w/ \mathcal{L}^{IP}	52.65	31.33	50.18	33.15	63.26	35.34
	w/ \mathcal{L}^{HN}	51.88	31.29	50.73	33.62	62.91	34.88
	w/ \mathcal{L}^{IP+HN}	53.41	31.30	51.10	33.23	63.69	35.09

4.1 Main Results

In Table 1, we report standard accuracy and robust accuracy of each model, learned by different pre-training methods over CIFAR-10 and CIFAR-100. Following previous works [32,27,18] and common practice in contrastive learning [9,24], we first use unlabeled images in CIFAR-10/-100 to pre-train, then introduce labels to finetune the model. As shown in Table 1, our methods achieve noticeable performance improvement over baselines in almost all scenarios, when replacing the original loss with our proposed adversarial CL loss.

In comparison with RoCL, \mathcal{L}^{IP} brings significant performance boost on both standard and robust accuracy consistently across different training methods (row 4 vs. 3, row 14 vs. 13) (except for RA of AFF on CIFAR10). Comparing to AdvCL, \mathcal{L}^{IP} also brings noticeable margin (row 8 vs. 7, row 18 vs. 17). This can be attributed to that \mathcal{L}^{IP} aims to lower the priority of adversaries and prevent clean samples moving towards other instances, which results in better instance discrimination and improves clean [48] and robust accuracy. \mathcal{L}^{HN} also yields substantial boost on robust and standard accuracy (*e.g.*, row 15 vs. 13). We hypothesize this is due to that \mathcal{L}^{HN} helps alert the model to adversarial samples by assigning higher weights for adversaries in negative contrast. When combined together, in most settings both standard and robust accuracy are further boosted, especially for Linear Probing. This is because directly mitigating the negative impact of *identity confusion* by \mathcal{L}^{IP} and helping adversarial get rid of false identities by \mathcal{L}^{HN} can complement each other, bringing further performance boost.

4.2 Transferring Robust Features

Learning robust features that are transferable is a main goal in self-supervised adversarial learning. It is of great significance if models pre-trained with a huge amount of unlabeled data possess good transferability by merely light-weight

finetuning. For example, Linear Probing is often $10\times$ quicker than conventional adversarial training, with only a linear classifier trained.

Here we evaluate the robust transferability of the proposed approach, by transferring CIFAR-10 and CIFAR-100 to STL-10, *i.e.*, use unlabeled images in CIFAR-10/-100 to pretrain, then use STL-10 to finetune and evaluate the learned models. As shown in Table 2, our methods yield both clean and robust accuracy gains in most settings, up to 1.48% (33.62% vs. 32.14%) in robust accuracy and 2.74% (67.19% vs. 64.45%) in clean accuracy.

4.3 Ablation studies

We design a basic adversarial contrastive model, named CoreACL, to study the effect of each component in our proposed methods. CoreACL only contains the contrastive component with three positive views: two clean augmented views and one adversarial view of the original image.

Fixed α for Asymmetric Similarity Function. We first use fixed α without adaptive annealing to explore the effectiveness of *inferior positives*. Figure 2 presents the results with different α values when training models for 200 epochs. Recall that α represents the tendency of the clean sample heading towards the adversarial sample. $\alpha < 0.5$ means clean samples move less toward the adversaries (vice versa for $\alpha > 0.5$), and $\alpha = 0.5$ degenerates to the original symmetric similarity function form.

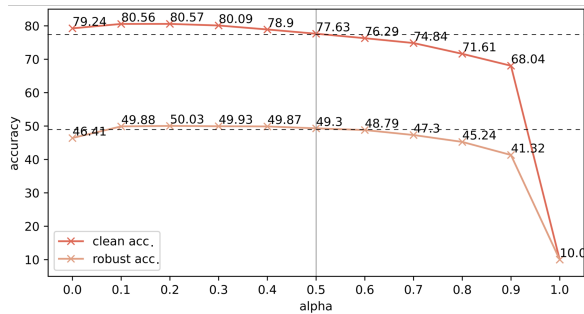


Fig. 2. Deep probing for asymmetric similarity function with different α .

Compared with symmetric CoreACL ($\alpha = 0.5$), our approach achieves better robustness and accuracy when $\alpha < 0.5$ (adversarial examples are treated as *inferior positives*). Intriguingly, when $\alpha = 1.0$, the extreme case when only clean samples are attracted by adversaries, we observe the presence of a trivial solution [12], that is all images collapse into one point. This validates our observation that adversaries with false identities are indeed pulling their positives towards other instances in the positive contrasts, with the risk of drawing all samples together. It is also worth noting that when $\alpha < 0.2$, performance begins to drop, showing that a small but non-zero α is the optimal setting empirically.

Fixed α vs. α -Annealing. As shown in Table 3, compared to CoreACL, fixed α obtains higher clean accuracy (81.29% vs. 78.90%) but with no gain on robust

accuracy. Adaptive annealing α achieves both higher robust accuracy (50.24% vs. 51.27%) and better clean accuracy (79.46% vs. 78.90%).

Comparison with AdvCL. Table 3 reports the performance and computation cost comparisons with AdvCL. CoreACL with \mathcal{L}^{IP+HN} achieves similar performance to AdvCL, which is equivalent to integrate additional components (high frequency view and pseudo-supervision) into CoreACL. The computation time of AdvCL is almost twice than that of w/\mathcal{L}^{IP+HN} , which could due to extra computation on contrasting high frequency views and the pseudo-labeled adversarial training. Our methods only need to compute pair-wise Euclidean distance for α -annealing in \mathcal{L}^{IP} , and no extra cost introduced in \mathcal{L}^{HN} .

Table 3. Ablation studies, evaluated in SA, RA and time cost. Trained for 400 epochs on 2 Tesla V100 GPUs.

Methods	SA	RA	Time Cost (s/epoch)
CoreACL	78.90	50.27	96
w/fixed α	81.29	50.24	96
w/annealing α	79.46	51.37	101
w/ \mathcal{L}^{IP+HN}	81.19	51.31	101
AdvCL	81.35	51.00	182

Effect of Hard Negatives. To investigate the effect of hard negatives, we evaluate each component (negatives debiasing [13], reweighting [39]) as shown in Table 4. With negatives-debiasing removed, we observe decrease in robust accuracy, with slightly increased standard accuracy. We hypothesize that without debiasing, semantically similar adversarial representations that should be mapped closely are pushed away instead. In addition, the removal of negatives reweighting results in a sharp performance drop, showing that viewing adversarial views as *hard negatives* with higher weights plays a key role in discriminating adversarial samples.

Table 4. Ablation studies for AdvCL with hard negatives (AdvCL-HN), evaluated under Linear Probing (LP), Adversarial Linear Finetuning (ALF) and Adversarial Full Finetuning (AFF)

Methods	LP		ALF		AFF	
	SA	RA	SA	RA	SA	RA
AdvCL-HN	81.34	52.96	78.69	53.20	83.44	54.07
w/o debias	81.52	51.61	78.89	52.34	83.73	54.01
w/o reweight	76.93	50.01	73.49	49.86	81.74	52.60

4.4 Qualitative Analysis

Figure 3 shows the distribution of normalized Euclidean distance over all negative pairs. We take AdvCL [18] as the baseline and compare it with its enhanced versions with our methods. Generally, our methods can shift the original distribution curve right (larger distance), meaning that treating adversaries as inferior positives or hard negatives encourages the model to separate negative pairs further apart and induce better instance discrimination. This suggests that our proposed methods effectively mitigate the negative impacts of *identity confusion*.

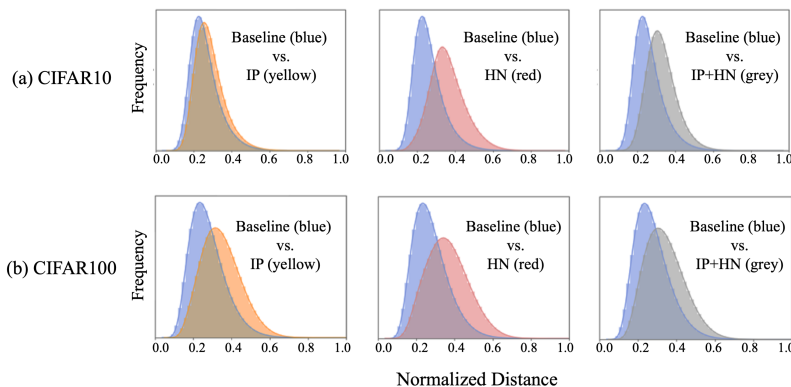


Fig. 3. Histograms of Euclidean distance (normalized) distribution of all negative pairs learned by different objectives in (a) CIFAR10 (first row) and (b) CIFAR100 (second row). Baseline is AdvCL [18]; IP: baseline with Inferior Positives; HN: baseline with Hard Negatives. On each dataset, our methods are better at differentiating different instances (with larger distance between negative pairs)

Figure 4 provides 2-D visualization (t-SNE [34] on CIFAR-10) for the embeddings learnt by SimCLR [9], RoCL [32] and RoCL enhanced by \mathcal{L}^{IP} (RoCL-IP). Each class is represented in one color. Compared to SimCLR, RoCL representations are corrupted by adversaries and exhibit poor class discrimination. RoCL-IP yields better class separation compared with RoCL. This shows that asymmetric similarity consideration eases instance-level identity confusion.

5 Related Work

Contrastive Learning CL has been widely applied to learn generalizable features from unlabeled data [9,24,44,22,11,6,3,36,10,7,31]. The basic idea is in-

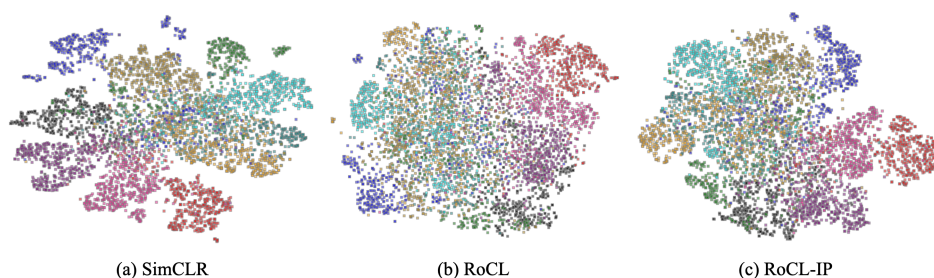


Fig. 4. t-SNE visualizations in a global view on CIFAR-10 validation set. The embeddings are learned by different self-supervised pre-training methods (SimCLR(a), RoCL(b) and RoCL-IP(c)) (colored figure)

stance discrimination [48]. Representative works include CMC [44], SimCLR[9], MoCo[24], SwAV[6], BYOL[22]. There is also a stream of work focusing on refined sampling on different views for improved performance [44,28,13,39,43]. For example, DCL[13] proposed to *debias* the assumption that all negative pairs are true negatives. HCL[39] extended DCL and proposed to mine hard negatives for contrastive learning, whose embeddings are uneasy to discriminate.

Adversarial Training Adversarial training (AT) stems from [20] and adopts a min-max training regime that optimizes the objective over adversaries generated by maximizing the loss [35,51,40,50,47,52,19,37]. Some recent work introduced unlabeled data into AT [26,8,5,1,32]. By leveraging a large amount of unlabeled data, [5,1] performed semi-supervised self-training to first generate pseudo-supervisions, then conducted conventional supervised AT. Our work explores how to learn robust models without any class labels.

Adversarial Contrastive Learning Some recent studies applied CL on adversarial training [32,27,18,21], by considering adversaries as positive views for contrasting, such that the learned encoder renders robust data representations. RoCL [32] was the first to successfully show robust models can be learned in an unsupervised manner. AdvCL [18] proposed to empower CL with pseudo-supervision stimulus. Same as CL, these Adversarial CL methods perform symmetric contrast for all pairs, which could potentially induces conflicts in CL and AT training objectives. We are the first to investigate the asymmetric properties of Adversarial CL, by treating adversaries discriminatively.

6 Conclusions

In this work, we study enhancing model robustness using unlabeled data and investigate the *identity confusion* issue in Adversarial CL, *i.e.*, adversaries with different identities attract their anchors together, contradicting to the objective of CL. We present a generic asymmetric objective *A-InfoNCE*, and treat adversaries discriminatively as *inferior positives* or *hard negatives*, which can overcome the identify confusion challenge. Comprehensive experiments with quantitative and qualitative analysis show that our methods can enhance existing Adversarial CL methods effectively. Further, it lies in our future work to extend the proposed asymmetric form to other CL settings to take into consideration the asymmetric characteristics between different views.

Acknowledgement

This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0112100, partly by Baidu Inc. through Apollo-AIR Joint Research Center. We would also like to thank the anonymous reviewers for their insightful comments.

References

1. Alayrac, J.B., Uesato, J., Huang, P.S., Fawzi, A., Stanforth, R., Kohli, P.: Are labels required for improving adversarial robustness? *Advances in Neural Information Processing Systems* **32** (2019)
2. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: *International conference on machine learning*. pp. 274–283. PMLR (2018)
3. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. *Advances in neural information processing systems* **32** (2019)
4. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: *2017 IEEE Symposium on Security and Privacy (SP)*. pp. 39–57. IEEE (2017)
5. Carmon, Y., Raghuathan, A., Schmidt, L., Duchi, J.C., Liang, P.S.: Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems* **32** (2019)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems* **33**, 9912–9924 (2020)
7. Chen, S., Niu, G., Gong, C., Li, J., Yang, J., Sugiyama, M.: Large-margin contrastive learning with distance polarization regularizer. In: *International Conference on Machine Learning*. pp. 1673–1683. PMLR (2021)
8. Chen, T., Liu, S., Chang, S., Cheng, Y., Amini, L., Wang, Z.: Adversarial robustness: From self-supervised pre-training to fine-tuning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 699–708 (2020)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: III, H.D., Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 119, pp. 1597–1607. PMLR (13–18 Jul 2020), <https://proceedings.mlr.press/v119/chen20j.html>
10. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems* **33**, 22243–22255 (2020)
11. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020)
12. Chen, X., He, K.: Exploring simple siamese representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 15750–15758 (2021)
13. Chuang, C.Y., Robinson, J., Lin, Y.C., Torralba, A., Jegelka, S.: Debiased contrastive learning. *Advances in neural information processing systems* **33**, 8765–8775 (2020)
14. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: *International conference on machine learning*. pp. 2206–2216. PMLR (2020)
15. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 9185–9193 (2018)
16. Du Plessis, M.C., Niu, G., Sugiyama, M.: Analysis of learning from positive and unlabeled data. *Advances in neural information processing systems* **27** (2014)

17. Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 213–220 (2008)
18. Fan, L., Liu, S., Chen, P.Y., Zhang, G., Gan, C.: When does contrastive learning preserve adversarial robustness from pretraining to finetuning? *Advances in Neural Information Processing Systems* **34** (2021)
19. Gan, Z., Chen, Y.C., Li, L., Zhu, C., Cheng, Y., Liu, J.: Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems* **33**, 6616–6628 (2020)
20. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014)
21. Gowal, S., Huang, P.S., van den Oord, A., Mann, T., Kohli, P.: Self-supervised adversarial robustness for the low-label, high-data regime. In: *International Conference on Learning Representations* (2020)
22. Grill, J.B., Strub, F., Althé, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent—a new approach to self-supervised learning. *Advances in Neural Information Processing Systems* **33**, 21271–21284 (2020)
23. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*. vol. 2, pp. 1735–1742. IEEE (2006)
24. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9729–9738 (2020)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
26. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems* **32**, 15663–15674 (2019)
27. Jiang, Z., Chen, T., Chen, T., Wang, Z.: Robust pre-training by adversarial contrastive learning. In: *NeurIPS* (2020)
28. Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D.: Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems* **33**, 21798–21809 (2020)
29. Kannan, H., Kurakin, A., Goodfellow, I.: Adversarial logit pairing. *arXiv preprint arXiv:1803.06373* (2018)
30. Kantipudi, J., Dubey, S.R., Chakraborty, S.: Color channel perturbation attacks for fooling convolutional neural networks and a defense against such attacks. *IEEE Transactions on Artificial Intelligence* **1**(2), 181–191 (2020)
31. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. *Advances in Neural Information Processing Systems* **33**, 18661–18673 (2020)
32. Kim, M., Tack, J., Hwang, S.J.: Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems* **33** (2020)
33. Li, J., Zhou, P., Xiong, C., Hoi, S.C.: Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966* (2020)
34. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of machine learning research* **9**(11) (2008)

35. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018)
36. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
37. Pang, T., Xu, K., Dong, Y., Du, C., Chen, N., Zhu, J.: Rethinking softmax cross-entropy loss for adversarial robustness. arXiv preprint arXiv:1905.10626 (2019)
38. Rahamim, A., Naeh, I.: Robustness through cognitive dissociation mitigation in contrastive adversarial training. arXiv preprint arXiv:2203.08959 (2022)
39. Robinson, J.D., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. In: International Conference on Learning Representations (2020)
40. Shafahi, A., Najibi, M., Ghiasi, M.A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G., Goldstein, T.: Adversarial training for free! Advances in Neural Information Processing Systems **32** (2019)
41. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation **23**(5), 828–841 (2019)
42. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations, ICLR 2014 (2014)
43. Tao, Y., Takagi, K., Nakata, K.: Clustering-friendly representation learning via instance discrimination and feature decorrelation. arXiv preprint arXiv:2106.00131 (2021)
44. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: European conference on computer vision. pp. 776–794. Springer (2020)
45. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? Advances in Neural Information Processing Systems **33**, 6827–6839 (2020)
46. Wahed, M., Tabassum, A., Lourentzou, I.: Adversarial contrastive learning by permuting cluster assignments. arXiv preprint arXiv:2204.10314 (2022)
47. Wong, E., Rice, L., Kolter, J.Z.: Fast is better than free: Revisiting adversarial training. arXiv preprint arXiv:2001.03994 (2020)
48. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
49. Xiao, C., Zhu, J.Y., Li, B., He, W., Liu, M., Song, D.: Spatially transformed adversarial examples. arXiv preprint arXiv:1801.02612 (2018)
50. Zhang, D., Zhang, T., Lu, Y., Zhu, Z., Dong, B.: You only propagate once: Accelerating adversarial training via maximal principle. Advances in Neural Information Processing Systems **32** (2019)
51. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International conference on machine learning. pp. 7472–7482. PMLR (2019)
52. Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., Liu, J.: Freeb: Enhanced adversarial training for natural language understanding. arXiv preprint arXiv:1909.11764 (2019)

Appendix

A Results under more attacks

In order to verify the effectiveness of the proposed method, in this section, we further evaluate the robustness of our method under a broader range of powerful attacks: 1) AutoAttack [14] (an ensemble of four strong diverse attacks, which is widely considered as the strongest attack for robustness evaluation), 2) CW attack [4] (CW-200), 3) PGD attack with restart [35] (PGD-200), 4) One-pixel attack [41], 5) Spatial Transformation attack [49], as well as 6) Color Channel attack [30]. PGD-200 and CW-200 both restart 5 times with 40 optimization steps each restart.

In Table 5, we report the robust accuracy under these attacks with AdvCL serving as baseline on CIFAR100. The results show that our methods can improve robustness under all different attacks across almost all settings, *e.g.*, 21.43% vs. 19.57% under AutoAttack and 29.56% vs. 27.13% under PGD-200 attack, with loss function \mathcal{L}^{IP+HN} (Equation 9), under Linear Probing.

Table 5. Robustness evaluation under diverse attacks on CIFAR100 with AdvCL as baseline.

Training Methods		PGD-200	CW-200	AA	One-pix.	Spatial-Tr.	Color-Ch.
Linear Probing	AdvCL	27.13	21.85	19.57	72.10	47.94	25.62
	w/ \mathcal{L}^{IP}	27.87	22.10	19.80	69.60	49.31	25.88
	w/ \mathcal{L}^{HN}	29.43	23.10	21.23	73.20	51.57	28.01
	w/ \mathcal{L}^{IP+HN}	29.56	23.60	21.43	73.00	52.62	28.94
Adversarial Linear Finetuning	AdvCL	27.29	22.01	20.09	72.80	47.31	24.98
	w/ \mathcal{L}^{IP}	27.84	22.37	20.06	71.60	46.22	24.23
	w/ \mathcal{L}^{HN}	29.79	23.79	21.52	70.80	51.04	27.84
	w/ \mathcal{L}^{IP+HN}	29.58	23.64	21.66	71.70	49.87	27.14
Adversarial Full Finetuning	AdvCL	29.48	25.73	24.46	72.20	57.86	25.12
	w/ \mathcal{L}^{IP}	30.10	26.05	24.73	71.00	58.95	25.55
	w/ \mathcal{L}^{HN}	30.46	26.60	25.22	69.30	59.04	26.02
	w/ \mathcal{L}^{IP+HN}	30.46	26.54	25.06	69.00	59.33	25.70

Table 6 provides results on CIFAR10 under canonical optimization-based attack methods: PGD-200, CW-200 and AutoAttack. Our methods also yield robustness gain in almost all settings.

Besides, we also report results compared with RoCL under PGD-200, CW-200 and AutoAttack in Table 7, which further validate the effectiveness of the proposed methods. For instance, 25.09% vs. 23.51% under CW-200 attack, Adversarial Full Finetuning scheme, on CIFAR100.

Table 6. Robustness evaluation under optimization-based attacks on CIFAR10, with AdvCL as baseline.

Training Methods		PGD-200	CW-200	AutoAttack
Linear Probing	AdvCL	51.05	45.65	43.48
	w/ \mathcal{L}^{IP}	51.99	46.02	43.57
	w/ \mathcal{L}^{HN}	52.36	46.09	43.68
	w/ \mathcal{L}^{IP+HN}	52.01	45.35	42.92
Adversarial Linear Finetuning	AdvCL	52.30	46.04	43.93
	w/ \mathcal{L}^{IP}	52.77	46.60	44.22
	w/ \mathcal{L}^{HN}	53.22	46.44	44.15
	w/ \mathcal{L}^{IP+HN}	52.77	45.55	43.01
Adversarial Full Finetuning	AdvCL	52.90	50.92	49.58
	w/ \mathcal{L}^{IP}	53.61	51.25	49.90
	w/ \mathcal{L}^{HN}	53.25	51.11	49.93
	w/ \mathcal{L}^{IP+HN}	53.51	51.46	50.28

Table 7. Robustness evaluation under optimization-based attacks, with RoCL as baseline, on CIFAR-10 and CIFAR-100.

Dataset	Training Methods		PGD-200	CW-200	AutoAttack
CIFAR10	Linear Probing	RoCL	32.47	33.33	24.11
		w/ \mathcal{L}^{IP+HN}	34.13	34.59	24.58
	Adversarial Linear Finetuning	RoCL	42.58	40.21	31.81
		w/ \mathcal{L}^{IP+HN}	43.54	41.26	30.37
	Adversarial Full Finetuning	RoCL	50.33	47.57	46.69
		w/ \mathcal{L}^{IP+HN}	51.47	48.26	47.05
CIFAR100	Linear Probing	RoCL	14.93	14.75	7.58
		w/ \mathcal{L}^{IP+HN}	17.95	16.57	8.58
	Adversarial Linear Finetuning	RoCL	22.59	18.99	11.93
		w/ \mathcal{L}^{IP+HN}	24.46	20.69	11.69
	Adversarial Full Finetuning	RoCL	27.95	23.51	22.70
		w/ \mathcal{L}^{IP+HN}	29.37	25.09	24.01