

Adversarial Self-Supervised Contrastive Learning

Minseon Kim¹, Jihoon Tack¹, Sung Ju Hwang^{1,2}

KAIST¹, AITRICS²

{minseonkim, jihoontack, sjhwang82}@kaist.ac.kr

Abstract

Existing adversarial learning approaches mostly use class labels to generate adversarial samples that lead to incorrect predictions, which are then used to augment the training of the model for improved robustness. While some recent works propose semi-supervised adversarial learning methods that utilize unlabeled data, they still require class labels. However, do we really need class labels *at all*, for adversarially robust training of deep neural networks? In this paper, we propose a novel adversarial attack for unlabeled data, which makes the model confuse the instance-level identities of the perturbed data samples. Further, we present a self-supervised contrastive learning framework to adversarially train a robust neural network without labeled data, which aims to maximize the similarity between a random augmentation of a data sample and its *instance-wise* adversarial perturbation. We validate our method, *Robust Contrastive Learning (RoCL)*, on multiple benchmark datasets, on which it obtains comparable robust accuracy over state-of-the-art supervised adversarial learning methods, and significantly improved robustness against the *black box* and *unseen* types of attacks. Moreover, with further joint fine-tuning with supervised adversarial loss, RoCL obtains even higher robust accuracy over using self-supervised learning alone. Notably, RoCL also demonstrate impressive results in robust transfer learning.

1 Introduction

The vulnerability of neural networks to imperceptibly small perturbations [1] has been a crucial challenge in deploying them to safety-critical applications, such as autonomous driving. Various studies have been proposed to ensure the robustness of the trained networks against adversarial attacks [2–4], random noise [5], and corruptions [6, 7]. Perhaps the most popular approach to achieve adversarial robustness is adversarial learning, which trains the model with samples perturbed to maximize the loss on the target model. Starting from Fast Gradient Sign Method [8] which apply a perturbation in the gradient direction, to Projected Gradient Descent [9] that maximizes the loss over iterations, and TRADES [2] that trades-off clean accuracy and adversarial robustness, adversarial learning has evolved substantially over the past few years. However, conventional methods with adversarial learning all require *class labels* to generate adversarial attacks.

Recently, self-supervised learning [10–14], which trains the model on unlabeled data in a supervised manner by utilizing self-generated labels from the data itself, has become popular as means of learning representations for deep neural networks. For example, prediction of the rotation angles [10], and solving randomly generated Jigsaw puzzles [11] are examples of such self-supervised learning methods. Recently, instance-level identity preservation [12, 13] with contrastive learning has shown to be very effective in learning the rich representations for classification. Contrastive self-supervised learning frameworks such as [12–15] basically aim to maximize the similarity of a sample to its augmentation, while minimizing its similarity to other instances.

In this work, we propose a contrastive self-supervised learning framework to train an adversarially robust neural network *without* any class labels. Our intuition is that we can fool the model by generat-

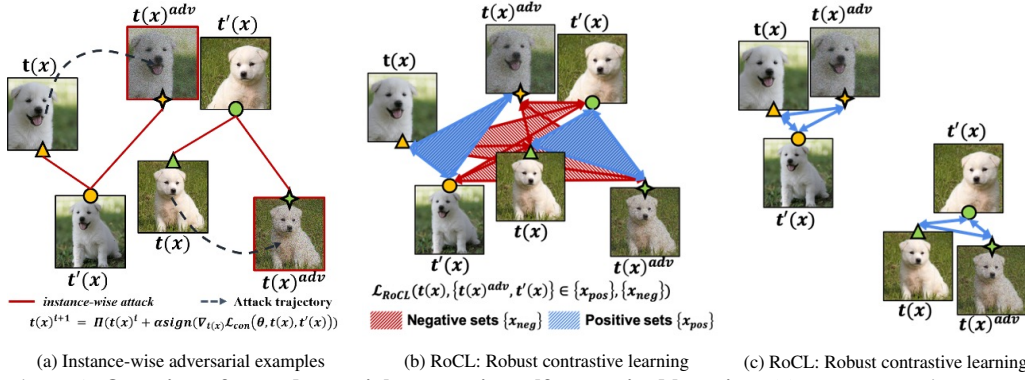


Figure 1: **Overview of our adversarial contrastive self-supervised learning.** (a) We generate instance-wise adversarial examples from an image transformed using a stochastic augmentation, which makes the model confuse the instance-level identity of the perturbed sample. (b) We then maximize the similarity between each transformed sample and their instance-wise adversaries using contrastive learning. (c) After training, each sample will have significantly reduced adversarial vulnerability in the latent representation space.

ing instance-wise adversarial examples (See Figure 1(a)). Specifically, we generate perturbations on augmentations of the samples to maximize their contrastive loss, such that the instance-level classifier becomes confused about the identities of the perturbed samples. Then, we maximize the similarity between clean samples and their adversarial counterparts using contrastive learning (Figure 1(b)), to obtain representations that suppress distortions caused by adversarial perturbations. This will result in learning representations that are robust against adversarial attacks (Figure 1(c)).

We refer to this novel adversarial self-supervised learning method as *Robust Contrastive Learning (RoCL)*. To the best of our knowledge, this is the first attempt to train robust neural networks *without any labels*, and to generate instance-wise adversarial examples. Recent works on semi-supervised adversarial learning [16, 17] or self-supervised adversarial learning [18] still require labeled instances to generate pseudo-labels on unlabeled instances or class-wise attacks for adversarial training, and thus cannot be considered as fully-unsupervised adversarial learning approaches.

To verify the efficacy of the proposed RoCL, we suggest a robust-linear evaluation for self-supervised adversarial learning and validate our method on benchmark datasets (CIFAR-10 and CIFAR-100) against supervised adversarial learning approaches. The results show that RoCL obtains comparable accuracy to strong supervised adversarial learning methods such as TRADES [2], although it does not use any labels during training. Further, when we extend the method to utilize class labels to fine-tune the network trained on RoCL with class-adversarial loss, we achieve even stronger robustness, *without* losing accuracy when clean samples. Moreover, we verify our rich robust representation with transfer learning which shows impressive performance. In sum, the contributions of this paper are as follows:

- We propose a novel **instance-wise** adversarial perturbation method which does not require any labels, by making the model confuse its instance-level identity.
- We propose a **adversarial self-supervised learning** method to explicitly suppress the vulnerability in the representation space by maximizing the similarity between clean examples and their instance-wise adversarial perturbations.
- Our method obtains **comparable** robustness to supervised adversarial learning approaches without using any class labels on the target attack type, while achieving **significantly better** clean accuracy and robustness on unseen type of attacks and transfer learning.

2 Related Work

Adversarial robustness Obtaining deep neural networks that are robust to adversarial attacks has been an active topic of research since Szegedy et al.[1] first showed their fragility to imperceptible distortions. Goodfellow et al.[8] proposed the fast gradient sign method (FGSM), which perturbs a target sample to its gradient direction, to increase its loss, and also use the generated samples to train the model for improved robustness. Follow-up works [9, 19–21] proposed iterative variants of the gradient attack with improved adversarial learning frameworks. After these gradient-based attacks have become standard in evaluating the robustness of deep neural networks, many more defenses followed, but Athalye et al. [22] showed that many of them appear robust only because they mask out

the gradients, and proposed new types of attacks that circumvent gradient obfuscation. Recent works focus on the vulnerability of the latent representations, hypothesizing them as the main cause of the adversarial vulnerability of deep neural networks. TRADES [2] uses Kullback-Leibler divergence loss between a clean example and its adversarial counterpart to push the decision boundary, to obtain a more robust latent space. Ilyas et al. [23] showed the existence of imperceptible features that help with the prediction of clean examples but are vulnerable to adversarial attacks. On the other hand, instead of defending the adversarial attacks, guarantee the robustness become one of the solutions to the safe model. Li et al. [24], "randomized smoothing" technique has been empirically proposed as certified robustness. Then, Cohen et al. [25], prove the robustness guarantee of randomized smoothing in ℓ_2 norm adversarial attack. Moreover, to improve the performance of randomized smoothing [26] directly attack the smoothed classifier. A common requirement of existing adversarial learning techniques is the availability of class labels, since they are essential in generating adversarial attacks. Recently, semi-supervised adversarial learning [16, 17] approaches have proposed to use unlabeled data and achieved large enhancement in adversarial robustness. Yet, they still require a portion of labeled data, and does not change the class-wise nature of the attack. Contrarily, in this work, we propose instance-wise adversarial attacks that do not require *any* class labels.

Self-supervised learning As acquiring manual annotations on data could be costly, self-supervised learning, which generates supervised learning problems out of unlabeled data and solves for them, is gaining increasingly more popularity. The convention is to train the network to solve a manually-defined (pretext) task for representation learning, which will be later used for a specific supervised learning task (e.g., image classification). Predicting the relative location of the patches of images [11, 27, 28] has shown to be a successful pretext task, which opened the possibility of self-supervised learning. Gidaris et al. [10] propose to learn image features by training deep networks to recognize the 2D rotation angles, which largely outperforms previous self-supervised learning approaches. Corrupting the given images with gray-scaling [29] and random cropping [30], then restoring them to their original condition, has also shown to work well. Recently, leveraging the instance-level identity is becoming a popular paradigm for self-supervised learning due to its generality. Using the contrastive loss between two different views of the same images [15] or two different transformed images from one identity [12, 13, 31] have shown to be highly effective in learning the rich representations, which achieve comparable performance to fully-supervised models. Moreover, even with the labels, the contrastive loss leverage the performance of the model than using the cross-entropy loss [32].

Self-supervised learning and adversarial robustness Recent works have shown that using unlabeled data could help the model to obtain more robust representations [16]. Moreover, [33] shows that a model trained with self-supervision improves the robustness. Using self-supervision signal in terms of perceptual loss also shows effective results in denoising the adversarial perturbation as purifier network [34]. Even finetuning the pretrained self-supervised learning helps the robustness [18], and self-supervised adversarial training coupled with the K-Nearest Neighbour classification improves the robustness of KNN [35]. However, to the best of our knowledge, none of these previous works explicitly target for adversarial robustness on unlabeled training. Contrarily, we propose a novel instance-wise attack, which leads the model to predict an incorrect instance for an instance-discrimination problem. This allows the trained model to obtain robustness that is on par or even better than supervised adversarial learning methods.

3 Adversarial Self-Supervised Learning with Instance-wise Attacks

We now describe how to obtain adversarial robustness in the representations *without* any class labels, using instance-wise attacks and adversarial self-supervised contrastive learning. Before describing ours, we first briefly describe supervised adversarial training and self-supervised contrastive learning.

Adversarial robustness We start with the definition of adversarial attacks under supervised settings. Let us denote the dataset $\mathbb{D} = \{X, Y\}$, where $x \in X$ is training sample and $y \in Y$ is a corresponding label, and a supervised learning model $f_\theta : X \rightarrow Y$ where θ is parameters of the model. Given such a dataset and a model, *adversarial attacks* aim towards finding the worst-case examples nearby by searching for the perturbation, which maximizes the loss within a certain radius from the sample (e.g., norm balls). We can define such adversarial attacks as follows:

$$x^{i+1} = \Pi_{B(x, \epsilon)}(x^i + \alpha \text{sign}(\nabla_{x^i} \mathcal{L}_{\text{CE}}(\theta, x^i, y)) \quad (1)$$

Algorithm 1 Robust Contrastive Learning (RoCL)

Input: Dataset \mathbb{D} , parameter of model θ , model f , parameter of projector π , projector g , constant λ
for all iter \in number of training iteration **do**
 for all $x \in$ minibatch $B = \{x_1, \dots, x_m\}$ **do**
 Generate adversarial examples from transformed inputs \triangleright instance-wise attacks
 $t(x)^{i+1} = \Pi_{B(t(x), \epsilon)}(t(x)^i + \alpha \text{sign}(\nabla_{t(x)^i} \mathcal{L}_{\text{con}, \theta, \pi}(t(x)^i, \{t'(x)\}, t(x)_{\text{neg}})))$
 end for
 $\mathcal{L}_{\text{total}} = \frac{1}{N} \sum_{k=1}^N [\mathcal{L}_{\text{RoCL}, \theta, \pi} + \lambda \mathcal{L}_{\text{con}, \theta, \pi}(t(x)_k^{\text{adv}}, \{t'(x)_k\}, \{t(x)_{\text{neg}}\})]$ \triangleright total loss
 Optimize the weight θ , π over $\mathcal{L}_{\text{total}}$
end for

where $B(x, \epsilon)$ is the ℓ_∞ norm-ball around x with radius ϵ , and Π is the projection function for norm-ball. The α is the step size of the attacks and $\text{sign}(\cdot)$ returns the sign of the vector. Further, \mathcal{L}_{CE} is the cross-entropy loss for supervised training, and i is the number of attack iterations. This formulation generalizes across different types of gradient attacks. For example, Projected Gradient Descent (PGD) [9] starts from a random point within the $x \pm \epsilon$ and perform i gradient steps, to obtain an attack x^{i+1} .

The simplest and most straightforward way to defend against such adversarial attacks is to minimize the loss of adversarial examples, which is often called *adversarial learning*. The adversarial learning framework proposed by Madry et al. [9] solve the following non-convex outer minimization problem and non-convex inner maximization problem where δ is the perturbation of the adversarial images, and $x + \delta$ is an adversarial example x^{adv} , as follow:

$$\underset{\theta}{\operatorname{argmin}} \mathbb{E}_{(x, y) \sim \mathbb{D}} \left[\max_{\delta \in B(x, \epsilon)} \mathcal{L}_{\text{CE}}(\theta, x + \delta, y) \right] \quad (2)$$

In standard adversarial learning framework, including PGD [9], TRADES [2], and many others, generating such adversarial attacks require to have a class label $y \in Y$. Thus, conventional adversarial attacks are inapplicable to unlabeled data.

Self-supervised contrastive learning The self-supervised contrastive learning framework [12, 13] aims to maximize the agreement between different augmentations of the same instance in the learned latent space while minimizing the agreement between different instances. Let us define some notions and briefly recap the SimCLR. To project the image into a latent space, SimCLR uses an encoder $f_\theta(\cdot)$ network followed by a projector, which is a two-layer multi-layer perceptron (MLP) $g_\pi(\cdot)$ that projects the features into latent vector z . SimCLR uses a stochastic data augmentation t , randomly selected from the family of augmentations \mathcal{T} , including random cropping, random flip, random color distortion, and random grey scale. Applying any two transformations, $t, t' \sim \mathcal{T}$, will yield two samples denoted $t(x)$ and $t'(x)$, that are different in appearance but retains the instance-level identity of the sample. We define $t(x)$'s positive set as $\{x_{\text{pos}}\} = t'(x)$ from the same original sample x , while the negative set $\{x_{\text{neg}}\}$ as the set of pairs containing the other instances x' . Then, the contrastive loss function \mathcal{L}_{con} can be defined as follows:

$$\begin{aligned} & \mathcal{L}_{\text{con}, \theta, \pi}(x, \{x_{\text{pos}}\}, \{x_{\text{neg}}\}) \\ & := -\log \frac{\sum_{\{z_{\text{pos}}\}} \exp(\text{sim}(z, \{z_{\text{pos}}\})/\tau)}{\sum_{\{z_{\text{pos}}\}} \exp(\text{sim}(z, \{z_{\text{pos}}\})/\tau) + \sum_{\{z_{\text{neg}}\}} \exp(\text{sim}(z, \{z_{\text{neg}}\})/\tau)}, \end{aligned} \quad (3)$$

where $z, \{z_{\text{pos}}\}$, and $\{z_{\text{neg}}\}$ are corresponding 128-dimensional latent vectors obtained by the encoder and projector $z = g_\pi(f_\theta(x))$, $\{x_{\text{pos}}\}$, and $\{x_{\text{neg}}\}$, respectively. The standard contrastive learning only contains a single sample in the positive set $\{x_{\text{pos}}\}$, which is $t(x)$. The $\text{sim}(u, v) = u^T v / \|u\| \|v\|$ denote cosine similarity between two vectors and τ is a temperature parameter.

We show that standard contrastive learning, such as SimCLR, is vulnerable to the adversarial attacks as shown in Table 1. To achieve robustness with such self-supervised contrastive learning frameworks, we need a way to adversarially train them, which we will describe in the next subsection.

3.1 Adversarial Self-supervised Contrastive Learning

We now introduce a simple yet novel and effective approach to adversarially train a self-supervised learning model, using unlabeled data, which we coin as *robust contrastive learning (RoCL)*. RoCL

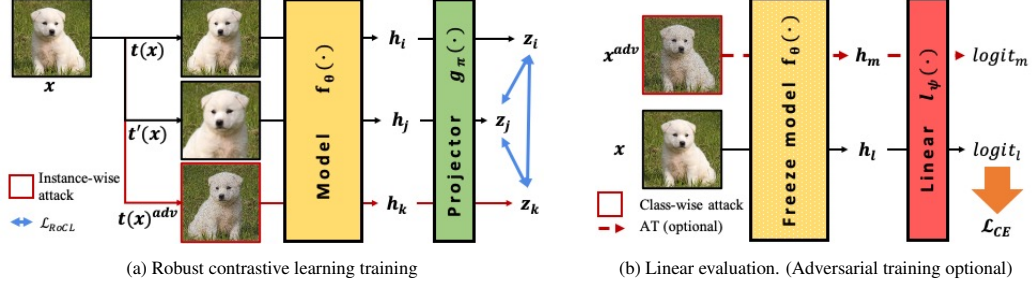


Figure 2: **Adversarial training and evaluation steps for RoCL.** During adversarial training, we maximize the similarity between two differently transformed examples $\{t(x), t'(x)\}$ and their adversarial perturbations $t(x)^{adv}$. After the model is fully trained to obtain robustness, then we evaluate the model on the target classification task by using linear model instead of projector. Here, we could either train the linear classifier only on clean examples, or adversarially train it with class-adversarial examples.

is trained without a class label by using *instance-wise* attacks, which makes the model confuse the instance-level identity of a given sample. Then, we use a contrastive learning framework to maximize the similarity between a transformed example and the instance-wise adversarial example of another transformed example. Algorithm 1 summarizes our robust contrastive learning framework.

Instance-wise adversarial attacks Since class-wise adversarial attacks for existing approaches are inapplicable to the unlabeled case we target, we propose a novel *instance-wise* attack. Specifically, given a sample of an input instance, we generate a perturbation to fool the model by confusing its instance-level identity; such that it mistakes it as an another sample. This is done by generating a perturbation that maximizes the self-supervised contrastive loss for discriminating between the instances, as follows:

$$t(x)^{i+1} = \Pi_{B(t(x), \epsilon)}(t(x)^i + \alpha \text{sign}(\nabla_{t(x)^i} \mathcal{L}_{\text{con}, \theta, \pi}(t(x)^i, \{t'(x)\}, \{t(x)_{\text{neg}}\}))) \quad (4)$$

where $t(x)$ and $t'(x)$ are transformed images with stochastic data augmentations $t, t' \sim \mathcal{T}$, and $\{t(x)_{\text{neg}}\}$ are the negative instances for $t(x)$, which are examples of other samples x' .

Robust Contrastive Learning (RoCL) We now present a framework to learn robust representation via self-supervised contrastive learning. The adversarial learning objective for an instance-wise attack, following the min-max formulation of [9] could be given as follows:

$$\underset{\theta, \pi}{\text{argmin}} \mathbb{E}_{(x) \sim \mathbb{D}} \left[\max_{\delta \in B(t(x), \epsilon)} \mathcal{L}_{\text{con}, \theta, \pi}(t(x) + \delta, \{t'(x)\}, \{t(x)_{\text{neg}}\}) \right] \quad (5)$$

where $t(x) + \delta$ is the adversarial image $t(x)^{adv}$ generated by *instance-wise* attacks (eq. 4). Note that we generate the adversarial example of x using a stochastically transformed image $t(x)$, rather than the original image x , which will allow us to generate diverse attack samples. This adversarial learning framework is essentially the same as the supervised adversarial learning framework, except that we train the model to be robust against m-way instance-wise adversarial attacks. Note that the proposed regularization can be interpreted as a denoiser. Since the contrastive objective maximize the similarity between clean samples: $t(x), t'(x)$, and generated adversarial example, $t(x)^{adv}$.

We generate label-free adversarial examples using instance-wise adversarial attacks in eq. 4. Then we use the contrastive learning objective to maximize the similarity between clean examples and their instance-wise perturbation. This is done using a simple modification of the contrastive learning objective in eq. 3, by using the *instance-wise* adversarial examples as additional elements in the positive set. Then we can formulate our *Robust Contrastive Learning* objective as follow:

$$\begin{aligned} \mathcal{L}_{\text{RoCL}, \theta, \pi} &:= \mathcal{L}_{\text{con}, \theta, \pi}(t(x), \{t'(x), t(x)^{adv}\}, \{t(x)_{\text{neg}}\}) \\ \mathcal{L}_{\text{total}} &:= \mathcal{L}_{\text{RoCL}, \theta, \pi} + \lambda \mathcal{L}_{\text{con}, \theta, \pi}(t(x)^{adv}, \{t'(x)\}, \{t(x)_{\text{neg}}\}) \end{aligned} \quad (6)$$

where $t(x)^{adv}$ are the adversarial perturbation of an augmented sample $t(x)$, $t'(x)$ is another stochastic augmentation, and λ is a regularization parameter. The $\{z_{\text{pos}}\}$, which is the set of positive samples in the latent feature space, is compose of z' and z^{adv} which are latent vectors of $t'(x)$ and $t(x)^{adv}$ respectively. The $\{z_{\text{neg}}\}$ is the set of latent vectors for negative samples in $\{t(x)_{\text{neg}}\}$.

Linear evaluation of RoCL With RoCL, we can adversarially train the model without any class labels (Figure 2(a)). Yet, since the model is trained for instance-wise classification, it cannot be directly used for class-level classification. Thus, existing self-supervised learning models leverage *linear evaluation* [12, 29, 36, 37], which learns a linear layer $l_\psi(\cdot)$ on top of the fixed $f_\theta(\cdot)$ embedding layer (Figure 2(b)) with clean examples. While RoCL achieves impressive robustness with this standard evaluation (Table 1), to properly evaluate the robustness against a specific type of attack, we propose a new evaluation protocol which we refer to as *robust-linear evaluation (r-LE)*. r-LE trains a *linear* classifier with class-level adversarial examples of specific attack (e.g. ℓ_∞) with the fixed encoder as follows:

$$\operatorname{argmin}_{\psi} \mathbb{E}_{(x,y) \sim \mathbb{D}} \left[\max_{\delta \in B(x,\epsilon)} \mathcal{L}_{\text{CE}}(\psi, x + \delta, y) \right] \quad (7)$$

where \mathcal{L}_{CE} is the cross-entropy that only optimize parameters of linear model ψ . While we propose r-LE as an evaluation measure, it could be also used as an efficient means of obtaining an adversarially robust network using network pretrained using self-supervised learning.

Transformation smoothed inference We further propose a simple inference method for robust representation. Previous works [26, 25] proposed *smoothed classifiers*, which obtain smooth decision boundaries for the final classifier by taking an expectation over classifiers with Gaussian-noise perturbed samples. This method aims to fix the problem with the sharp classifiers, which may result in misclassification of the points even with small perturbations. Similarly, we observe that our objective enforces to assemble all differently transformed images into the adjacent area, and propose a *transformation smoothed classifier* to obtain a smooth classifier for RoCL, which predicts the class c by calculating expectation \mathbb{E} over the transformation $t \sim \mathcal{T}$ for a given input x as follows:

$$S(x) = \operatorname{argmax}_{c \in Y} \mathbb{E}_{t \sim \mathcal{T}} (l_c(f(t(x))) = c) \quad (8)$$

where $l_c(\cdot)$ is the logit value of the class. We approximate the expectation over the transformation by multiple sampling the random transformation t and aggregating the penultimate feature $f(t(x))$.

4 Experimental Results

We now validate RoCL on benchmark datasets against existing adversarial learning methods. Specifically, we report the results of our model against white-box and black-box attacks and in the transfer learning scenario in Section 4.1, and conduct an ablation study to verify the efficacy of individual component of RoCL in Section 4.2.

Experimental setup For every experiments in the main text, we use ResNet18 or ResNet50 [38] trained on CIFAR-10 [39]. For all baselines and our method, we train with ℓ_∞ attacks with the same attack strength of $\epsilon = 8/255$. All ablation studies are conducted with ResNet18 trained on CIFAR-10, with the attack strength of $\epsilon = 8/255$. Regarding the additional results on CIFAR-100 and details of the optimization & evaluation, please see the Appendix A, and C. The code to reproduce the experimental results is available at <https://github.com/Kim-Minseon/RoCL>.

4.1 Main Results

We first report the results of baselines and our models against white-box attacks with linear evaluation, robust linear evaluation and finetuning in Table 1. We also report the results against black-box attacks in Table 2, where adversarial samples are generated by AT, TRADES, RoCL with the PGD attack, and RoCL model with the instance-wise attack. Then, we demonstrate the efficacy of the transformation smoothed classifier in Table 3. We further report the results of transfer learning, where we transfer the learned networks from from CIFAR-10 to CIFAR-100, and CIFAR-100 to CIFAR-10 in Table 4.

Results on white box attacks To our knowledge, our *RoCL* is the first attempt to achieve robustness in a fully self-supervised learning setting, since existing approaches used self-supervised learning as a pretraining step before supervised adversarial training. Therefore, we analyze the robustness of representation of RoCL which is acquire during the training only with linear evaluation including robust linear evaluation. Also, we discover that RoCL is also robust against unseen attacks. Lastly, we demonstrate the results of finetuning the RoCL.

Table 1: Experimental results with white box attacks on ResNet18 and ResNet50 trained on the CIFAR-10. r-LE denotes robust linear evaluation, and SCL is the supervised contrastive learning [32] which uses the labels in the contrastive loss. Baselines with * are the models with our data augmentation applied during training. AT denotes the supervised adversarial training[9], and SS denotes the self-supervised loss. Rot+pretrained is the model [18] which finetunes the network trained with rotation-prediction self-supervised learning. For a fair comparison, we report the single self-supervised model pretrained version with the ResNet50-v2 model. ⁺ is the reported performance of [18]. A_{nat} is the accuracy of clean images. All models are trained with ℓ_∞ ; thus the ℓ_∞ is the *seen* adversarial attack and ℓ_2 , and ℓ_1 attacks are *unseen*.

Train type	Method	ResNet18								ResNet50							
		<i>seen</i>				<i>unseen</i>				<i>seen</i>				<i>unseen</i>			
		A_{nat}	ℓ_∞		ℓ_2	ℓ_1		A_{nat}	ℓ_∞		ℓ_2	ℓ_1					
			ϵ	8/255		16/255	0.25		0.5	7.84		12	ϵ	8/255	16/255	0.25	0.5
Supervised	\mathcal{L}_{CE}	92.82	0.00	0.00	20.77	12.96	28.47	15.56	93.12	0.00	0.00	13.42	3.44	28.78	13.98		
	AT[9]	81.63	44.50	14.47	72.26	59.26	66.74	55.74	84.03	46.76	17.63	72.98	58.78	65.28	52.45		
	TRADES[2]	77.03	48.01	22.55	68.07	57.93	62.93	53.79	82.10	53.49	25.18	73.01	61.94	65.48	54.52		
	TRADES*[2]	73.26	42.71	17.71	65.25	56.13	62.89	55.95	75.65	46.20	20.96	67.02	57.12	62.46	55.09		
	SCL[32]	94.05	0.08	0.00	22.17	10.29	38.87	22.58	95.02	0.00	0.00	16.72	1.68	39.44	22.59		
Self-supervised	SimCLR[12]	91.25	0.63	0.08	15.3	2.08	41.49	25.76	92.69	0.07	0.00	25.13	3.85	50.17	31.63		
	RoCL	83.71	40.27	9.55	66.39	63.82	79.21	76.17	85.99	43.56	11.38	70.87	67.59	82.65	80.02		
	RoCL+rLE	80.43	47.69	15.53	68.30	66.19	77.31	75.05	80.79	45.33	16.85	67.14	64.61	77.54	75.76		
Self-supervised+finetune	Rot. Pretrained[18]	-	-	-	-	-	-	-	85.66 ⁺	50.40 ⁺	-	-	-	-	-		
	RoCL+AT	80.26	40.77	22.83	68.64	56.25	65.16	56.07	82.72	50.60	18.83	72.12	70.03	81.02	79.22		
	RoCL+TRADES	84.55	43.85	14.29	73.01	60.03	68.25	58.04	85.41	45.68	21.21	74.06	59.60	65.37	53.54		
	RoCL+AT+SS	91.34	49.66	14.44	70.75	61.55	83.08	81.18	84.67	52.44	19.53	76.61	66.38	72.76	64.56		

Table 2: Performance of RoCL against black box attacks on the CIFAR-10 dataset. Each column denotes the black box model used to generate the ℓ_∞ adversarial examples with $\epsilon = 8/255$ and $16/255$, respectively. We generate PGD attack examples (PGD) and instance-wise attack examples (*inst.*) from RoCL with linear layer and a projector, respectively. Each row shows the performance of the target model trained with ℓ_∞ .

		ResNet18							
Target	Source	8/255				16/255			
		AT	TRADES	RoCL(PGD)	RoCL(<i>inst.</i>)	AT	TRADES	RoCL(PGD)	RoCL(<i>inst.</i>)
AT [9]	-	-	77.48	69.83	47.25	-	63.87	48.99	47.42
TRADES [2]	60.73	-	64.81	46.22	41.87	-	48.07	45.73	
RoCL	66.76	77.33	-	-	41.97	62.98	-	-	

We first compare RoCL against SimCLR[12], which is a vanilla self-supervised contrastive learning model. The result shows that SimCLR is extremely vulnerable to adversarial attacks. However, RoCL achieves high robust accuracy (40.27) against the target ℓ_∞ attacks. This is an impressive result, which demonstrates that it is possible to train adversarially robust models without any labeled data. Moreover, RoCL+rLE outperform supervised adversarial training by Madry et al. [9] and obtains comparable performance to TRADES [2]. Note that while we used the same number of instances in this experiment, in practice, we can use any number of unlabeled data available to train the model, which may lead to larger performance gains. To show that this result is not due to the effect of using augmented samples for self-supervised learning, we applied the same set of augmentations for TRADES (TRADES*), but it obtains worse performance over the original TRADES.

Moreover, RoCL obtains significantly higher robustness over the supervised adversarial learning approaches against *unseen* types of attacks, except for ℓ_1 attack with small perturbation, and much higher clean accuracy (See the results on ℓ_2 , ℓ_1 attacks in Table 1). This makes RoCL more appealing over baselines in practice, and suggests that our approach to enforce a consistent identity over diverse perturbations of a single sample in the latent representation space is a more fundamental solution to ensure robustness against general types of attacks. This point is made more clear in the comparison of RoCL against RoCL with robust linear evaluation (RoCL+rLE), which trains the linear classifier with class-wise adversaries. RoCL+rLE improves the robustness against the target ℓ_∞ attacks, but degenerates robustness on unseen types of attacks (ℓ_1).

Existing works [40, 18] have shown that finetuning the supervised or self-supervised pretrained networks with adversarial training improves robustness. This is also confirmed with our results in Table 1, which show that the models fine-tuned with our method obtain even better robustness and higher clean accuracy over models trained from scratch. We observe that using self-supervised loss (SS loss eq. 3) during adversarial finetuning further improves robustness (RoCL + AT + SS). Moreover, our method outperforms Chen et al. [18], which uses self-supervised learning only for model pretraining, before supervised adversarial training.

Table 3: Experimental results of transformation smoothed classifier. We test the RoCL + transformation smoothed classifier against clean images, black box attack images which is generated from AT attack and expectation of transformation attack (EoT).

	Clean	Blackbox (AT)		EoT
	-	8/255	16/255	8/255
RoCL	84.11 (+0.40)	66.86 (+0.10)	42.50 (+0.43)	33.24 -

Table 5: Performance with different target images for generating instance-wise attacks.

	A_{nat}	8/255	16/255
original x	87.96	36.6	11.78
$t'(x)$	83.71	40.27	9.55

Table 6: Experimental results of RoCL against ℓ_∞ attack with different number of steps.

	20	40	100
RoCL	40.27	39.80	39.74

Table 4: Results of transfer learning across the CIFAR-10 and CIFAR-100 datasets with ResNet18. We compare against adversarial transfer learning results from [41], with a larger WRN 32-10 [42] architecture. ⁺ is the reported performance from [41].

source	target	Method	A_{nat}	ℓ_∞
CIFAR-100	CIFAR-10	Transfer ⁺ [41] RoCL	72.05 73.93	17.70 18.62
CIFAR-10	CIFAR-100	Transfer ⁺ [41] RoCL	41.59 45.84	11.63 15.33

Table 7: Performance of RoCLs with different attack loss types. The original RoCL maximizes the contrastive loss (Contrastive) to generate instance-wise attacks. We observe that Other types of losses, such as mean square error (MSE), cosine similarity, Manhattan distance (MD) are less effective.

$\mathcal{L}_{\theta, \pi}$	A_{nat}	8/255	16/255
Contrastive	83.71	40.27	9.55
MSE	88.35	40.12	7.88
Cosine similarity	73.49	9.30	0.06
MD	84.40	21.05	1.65

Results on black box attacks We also validate our models against black-box attacks. We generate adversarial examples using the AT, TRADES, and RoCL, perform black-box attacks across the methods. As shown in Table 2, our model is superior to TRADES [2] against AT black box attacks, and achieves comparable performance to AT [9] against TRADES black box attack samples. We also validate RoCL’s robustness by generating adversarial samples using our model and use them to attack AT and TRADES. We also generate black-box adversarial examples with RoCL by attacking the RoCL with a linear layer using the PGD attack (RoCL (PGD)), and the RoCL with a projector using the instance-wise attack (RoCL (*inst.*)). The low robustness of attacked models (AT, TRADES) shows that attacks with RoCL are strong. Specifically, RoCL with the PGD attack is stronger than TRADES attacks on AT, and RoCL with the instance-wise attacks is significantly stronger over both AT and TRADES black box attacks.

Transformation smoothed classifier Transformation smoothed classifier can enhance the model accuracy not only on the black-box adversarial examples, but also on clean examples (Table 3). Intuitively, since we enforce differently transformed samples of the same instance to have a consistent identity, they will be embedded in nearby places in the latent representation space. Therefore, we can calculate the transformation ball around the samples, that is similar to Gaussian ball in [25]. Accordingly, RoCL obtains a smoother classifier and acquires larger gains in both black-box robustness and clean accuracy (Table 3). As shown in Figure 3(d), as the number of samples ($t \sim \mathcal{T}$) increases, the model becomes increasingly more robust. We also test the transformation smoothed classifier with expectation of transformation (EoT) attack [22], which is a white box attack against models with test-time randomness. We found that although transformation smoothed classifier suffers from loss of robust accuracy with EoT attacks, it is still reasonably robust (Table 3). We provide the detailed settings of transformation smoothed classifier experiments in Section A of the Appendix.

Transfer learning Another advantage of our unsupervised adversarial learning, is that the learned representations can be easily transferred to diverse target tasks. We demonstrate the effectiveness of our works on transfer learning in Table 4, against the fully supervised adversarial transfer learning [41] with larger networks. Surprisingly, our model achieves even better accuracy and robustness in both cases (CIFAR-10→CIFAR-100 and CIFAR-100→CIFAR-10) without any other additional losses. The detailed settings for the transfer learning experiments are given in Section B of the Appendix .

4.2 Ablation studies

Effect of target images to generate attacks When generating instance-wise attacks, we can either attack the original x or the transformed instance $t'(x)$. The comparative study in Table 5 shows that our RoCL achieves high clean accuracy and robustness regardless of the target examples we select for instance-wise perturbation. This is because the our method aims at preserving the instance-level

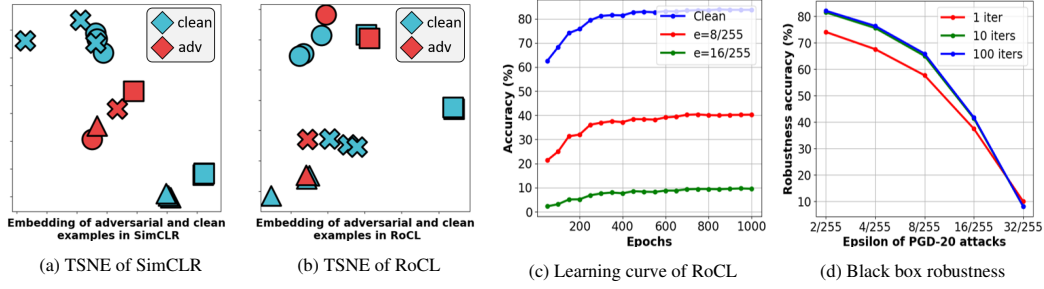


Figure 3: (a,b) Visualizations of the embedding of *instance-wise* adversarial examples and clean examples for SimCLR and RoCL after training. (c) The learning curve of ResNet18 RoCL. (d) The transformation smoothed classifier performance on AT’s black box attack over transformation iterations against different attack budgets.

identity regardless of the transformations applied to an instance. Therefore, our methods achieves consistent performance with any target instances that have the same identity.

Effect of attack loss type For instance-wise attacks, we can consider various losses to maximize the distance of adversarial samples from the target samples. We compare four different distance functions, namely mean square error (MSE), cosine similarity, Manhattan distance (MD), and contrastive loss. Table 7 shows that the contrastive loss is the most effective among all losses we considered.

Effect of the number of PGD attack iterations We further validate the robustness of RoCL under larger iteration steps of the PGD attack. Table 6 shows that RoCL remains robust with any number of PGD iterations (e.g., 39.74% under 100 iteration steps).

Visualizations of *instance-wise* attacks We further examine and visualize the samples generated with our instance-wise attacks on SimCLR in Figure 3(a)). The visualization of the samples in the latent embedding space shows that our attacks generate confusing samples (denoted with red markers) that are far apart from the original instances (denoted with blue markers) with the same identities. However, after we train the model with RoCL (Figure 3(b)), the instance-wise adversarial examples are pushed toward the samples with the same instance-level identity.

5 Conclusion

In this paper, we tackled a novel problem of learning robust representations without any class labels. We first proposed a *instance-wise attack* to make the model confuse the instance-level identity of a given sample. Then, we proposed a *robust contrastive learning* framework to suppress their adversarial vulnerability by maximizing the similarity between a transformed sample and its instance-wise adversary. Furthermore, we demonstrate an effective transformation smoothed classifier which boosts our performance during the test inference. We validated our method on multiple benchmarks with different neural architectures, on which it obtained comparable robustness to the supervised baselines on the targeted attack without any labels. Notably, RoCL obtained significantly better clean accuracy and better robustness against black box, unseen attacks, and transfer learning, which makes it more appealing as a general defense mechanism. We believe that our work opened a door to more interesting follow-up works on *unsupervised adversarial learning*, which we believe is a more fundamental solution to achieving adversarial robustness with deep neural networks.

Broader Impact

Achieving adversarial robustness against malicious attacks with deep neural networks, is a fundamental topic of deep learning research that has not yet been fully solved. Until now, supervised adversarial training, which perturbs the examples such that the target deep network makes incorrect predictions, has been a dominant paradigm in adversarial learning of deep neural networks. However, supervised adversarial learning suffers from lack of generalization to unseen types of attacks, or unseen datasets, as well as suffers from loss of accuracy on clean examples, and thus is not a fundamental, nor practical solution to the problem. Our adversarial self-supervised learning is a research direction that delved into the vulnerability of deep networks in the intrinsic representation space, which we believe is the root cause of fragility of existing deep neural networks, and we hope that more research is conducted in the similar directions.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-00153) and Samsung Research Funding Center of Samsung Electronics (No. SRFC-IT1502-51). We thank Sihyun Yu, Seanie Lee, and Hayeon Lee for providing helpful feedbacks and suggestions in preparing an earlier version of the manuscript. We also thank the anonymous reviewers for their insightful comments and suggestions.

References

- [1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.
- [2] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, “Theoretically principled trade-off between robustness and accuracy,” in *Proceedings of the 36th International Conference on Machine Learning*, 2019.
- [3] F. Tramèr and D. Boneh, “Adversarial training and robustness for multiple perturbations,” in *Advances in Neural Information Processing Systems*, pp. 5858–5868, 2019.
- [4] D. Madaan, J. Shin, and S. J. Hwang, “Adversarial neural pruning with latent vulnerability suppression,” in *Proceedings of the 37th International Conference on Machine Learning*, 2020.
- [5] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, “Improving the robustness of deep neural networks via stability training,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4480–4488, 2016.
- [6] D. Hendrycks and T. Dietterich, “Benchmarking neural network robustness to common corruptions and perturbations,” in *International Conference on Learning Representations*, 2019.
- [7] D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, “A fourier perspective on model robustness in computer vision,” in *Advances in Neural Information Processing Systems*, pp. 13255–13265, 2019.
- [8] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *International Conference on Learning Representations*, 2015.
- [9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *International Conference on Learning Representations*, 2018.
- [10] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” in *International Conference on Learning Representations*, 2018.
- [11] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in *European Conference on Computer Vision*, pp. 69–84, Springer, 2016.
- [12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning*, 2020.
- [13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [14] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.
- [15] Y. Tian, D. Krishnan, and P. Isola, “Contrastive multiview coding,” in *European Conference on Computer Vision*, 2020.
- [16] Y. Carmon, A. Raghuathan, L. Schmidt, J. C. Duchi, and P. S. Liang, “Unlabeled data improves adversarial robustness,” in *Advances in Neural Information Processing Systems*, pp. 11190–11201, 2019.
- [17] R. Stanforth, A. Fawzi, P. Kohli, *et al.*, “Are labels required for improving adversarial robustness?,” in *Advances in Neural Information Processing Systems*, 2019.
- [18] T. Chen, S. Liu, S. Chang, Y. Cheng, L. Amini, and Z. Wang, “Adversarial robustness: From self-supervised pre-training to fine-tuning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

- [19] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.
- [20] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” *arXiv preprint arXiv:1607.02533*, 2016.
- [21] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *2017 IEEE symposium on security and privacy (sp)*, pp. 39–57, IEEE, 2017.
- [22] A. Athalye, N. Carlini, and D. Wagner, “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples,” in *Proceedings of the 35th International Conference on Machine Learning*, 2018.
- [23] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, “Adversarial examples are not bugs, they are features,” in *Advances in Neural Information Processing Systems*, pp. 125–136, 2019.
- [24] B. Li, C. Chen, W. Wang, and L. Carin, “Certified adversarial robustness with additive noise,” in *Advances in Neural Information Processing Systems*, pp. 9459–9469, 2019.
- [25] J. Cohen, E. Rosenfeld, and Z. Kolter, “Certified adversarial robustness via randomized smoothing,” in *Proceedings of the 36th International Conference on Machine Learning*, pp. 1310–1320, 2019.
- [26] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang, “Provably robust deep learning via adversarially trained smoothed classifiers,” in *Advances in Neural Information Processing Systems*, pp. 11289–11300, 2019.
- [27] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, “Discriminative unsupervised feature learning with exemplar convolutional neural networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1734–1747, 2015.
- [28] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual representation learning by context prediction,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1422–1430, 2015.
- [29] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in *European Conference on Computer Vision*, pp. 649–666, Springer, 2016.
- [30] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.
- [31] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, “What makes for good views for contrastive learning,” in *Advances in Neural Information Processing Systems*, 2020.
- [32] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Supervised contrastive learning,” in *Advances in Neural Information Processing Systems*, 2020.
- [33] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, “Using self-supervised learning can improve model robustness and uncertainty,” in *Advances in Neural Information Processing Systems*, pp. 15637–15648, 2019.
- [34] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, “A self-supervised approach for adversarial robustness,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [35] K. Chen, H. Zhou, Y. Chen, X. Mao, Y. Li, Y. He, H. Xue, W. Zhang, and N. Yu, “Self-supervised adversarial training,” *arXiv preprint arXiv:1911.06470*, 2019.
- [36] P. Bachman, R. D. Hjelm, and W. Buchwalter, “Learning representations by maximizing mutual information across views,” in *Advances in Neural Information Processing Systems*, pp. 15509–15519, 2019.
- [37] A. Kolesnikov, X. Zhai, and L. Beyer, “Revisiting self-supervised visual representation learning,” *CoRR*, vol. abs/1901.09005, 2019.
- [38] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
- [39] A. Krizhevsky, G. Hinton, *et al.*, “Learning multiple layers of features from tiny images,” *Citeseer*, 2009.
- [40] D. Hendrycks, K. Lee, and M. Mazeika, “Using pre-training can improve model robustness and uncertainty,” in *Proceedings of the 36th International Conference on Machine Learning*, 2019.

- [41] A. Shafahi, P. Saadatpanah, C. Zhu, A. Ghiasi, C. Studer, D. Jacobs, and T. Goldstein, “Adversarially robust transfer learning,” in *International Conference on Learning Representations*, 2020.
- [42] S. Zagoruyko and N. Komodakis, “Wide residual networks,” *arXiv preprint arXiv:1605.07146*, 2016.
- [43] Y. You, I. Gitman, and B. Ginsburg, “Large batch training of convolutional networks,” *arXiv preprint arXiv:1708.03888*, 2017.
- [44] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, “Accurate, large minibatch sgd: Training imagenet in 1 hour,” *arXiv preprint arXiv:1706.02677*, 2017.
- [45] I. Loshchilov and F. Hutter, “Sgdr: Stochastic gradient descent with warm restarts,” *arXiv preprint arXiv:1608.03983*, 2016.
- [46] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *PMLR* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 448–456, 07–09 Jul 2015.
- [47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.
- [48] A. Krizhevsky, “Learning multiple layers of features from tiny images,” *University of Toronto*, 05 2012.

Appendix

Adversarial Self-Supervised Contrastive Learning

Organization Appendix is organized as follows. In section A, we describe the experimental details, including the descriptions of the datasets and the evaluation process. We then provide an algorithm which summarizes our RoCL in section B. Then, we further report the RoCL results on both CIFAR-10 and CIFAR-100 against PGD attacks and CW attacks in Section C. Finally, perform ablation studies of our RoCL in section D.

A Experimental Setup

A.1 Training detail and dataset

Training details We use ResNet18 and ResNet50 [38] as the base encoder network f_θ and 2-layer multi-layer perceptron with 128 embedding dimension as the projection head g_π . All models are trained by minimizing the final loss $\mathcal{L}_{\text{total}}$ with a temperature of $\tau = 0.5$. We set the regularization parameter to $\lambda = 1/256$. For the inner maximization step of RoCL i.e., instance-wise attack, we set the perturbation $\epsilon = 0.0314$ and step size $\alpha = 0.007$ under ℓ_∞ bound, with the number of inner maximize iteration as $K = 7$. For the rest, we follow the similar optimization step of SimCLR [12]. For optimization, we train RoCL with 1,000 epoch under LARS optimizer [43] with weight decay of $1e-6$ and momentum with 0.9. For the learning rate scheduling, we use linear warmup [44] for early 10 epochs until learning rate of 1.0 and decay with cosine decay schedule without a restart [45]. We use batch size of 512 for RoCL (we found out that batch size of 512 was sufficient for CIFAR-10 and CIFAR-100). Furthermore, we use global batch normalization (BN) [46], which shares the BN mean & variance in distributed training over the GPUs.

Data augmentation details We use SimCLR augmentations: Inception crop [47], horizontal flip, color jitter, and grayscale for random augmentations. The detailed description of the augmentations are as follows. *Inception crop*: Randomly crops the area of the original image with uniform distribution 0.08 to 1.0. After the crop, cropped image are resized to the original image size. *Horizontal flip*: Flips the image horizontally with 50% of probability. *Color jitter*: Change the hue, brightness, and saturation of the image. We transform the RGB (red, green, blue) channeled image into an HSV (hue, saturation, value) channeled image format and add noise to the HSV channels. We randomly apply color jitter transformation with 80% of probability. *Grayscale*: Convert into a gray scale image. We randomly apply the grayscale transformation with 20% of probability.

Dataset details For RoCL training, we use CIFAR-10 [48] and CIFAR-100 [48]. CIFAR-10 and CIFAR-100 consist of 50,000 training and 10,000 test images with 10 and 100 image classes, respectively.

A.2 Evaluation

Linear evaluation setup In the linear evaluation phase, we train the linear layer l_ψ on the top of the frozen encoder f_θ . We train the linear layer for 150 epochs with the learning rate of 0.1. The learning rate is dropped by a factor of 10 at 30, 50, 100 epoch of the training progress. We use stochastic gradient descent (SGD) optimizer with a momentum of 0.9, weight decay of $5e-4$, and train the linear layer with the cross-entropy (CE) loss.

Robust linear evaluation setup For robust linear evaluation, we train the linear layer l_ψ on the top of the frozen encoder f_θ , as done with linear evaluation. We train the linear layer for 150 epochs with an learning rate of 0.02. The learning rate scheduling and the optimizer setup is the same with the setup for linear evaluation. We use the project gradient descent (PGD) attack to generate class-wise adversarial examples. We perform ℓ_∞ attack with epsilon $\epsilon = 0.0314$ and the step size $\alpha = 0.007$ for 10 steps.

Robustness evaluation setup For evaluation of adversarial robustness, we use white-box project gradient descent (PGD) attack. We evaluate under PGD attacks with 20, 40, 100 steps. We set $\ell_\infty, \ell_2, \ell_1$ attacks with $\epsilon = 0.0314, 0.072$ for ℓ_∞ , $\epsilon = 0.25, 0.5$ for ℓ_2 , and $\epsilon = 7.84, 12$ for ℓ_1 for testing CIFAR 10 and CIFAR 100.

A.3 Transformation smoothed classifier setup

In the transformation smoothed classifier, we used same data augmentation that is used in training phase A.1. The probability is also same with training phase, yet we used fixed sized inception crop with 0.54 scale. For Table 3, we used 30 times iteration for all tests. For Figure 3(d) we differ the transformation iterations to 1, 10, and 100.

For the expectation of transformation (EoT), we evaluate under the perturbation $\epsilon = 0.0314$ and step size $\alpha = 0.00314$ under ℓ_∞ bound, with the number of inner maximize step iteration as $K = 20$.

A.4 Transfer learning setup

We first briefly describe robust transfer learning and our experiments in its experimental setting. Shafahi et al. [41] suggest that an adversarially trained model can be transferred to another model to improve upon its robustness. They used modified WRN 32-10 to train the fully supervised adversarial model. Moreover, they initialize the student network with an adversarially trained teacher network and utilize the distillation loss and cross-entropy loss to train the student network’s linear layer on the top of the encoder layer. We follow the experimental settings of Shafahi et al. [41], and train only the linear layer with cross-entropy loss. However, we did not use the distillation loss in order to evaluate the robustness of the encoder trained with our RoCL only (ResNet18). We train the linear model with CIFAR-100 on top of the frozen encoder, which is trained on CIFAR-10. We also train the linear layer with CIFAR-10 on top of the frozen encoder, which is trained on CIFAR-100. We train the linear layer for 100 epochs with a learning rate of 0.2. We use stochastic gradient descent (SGD) for optimization.

A.5 Training efficiency of RoCL

Training efficiency of RoCL RoCL takes about 41.7 hours to train 1000 epochs with two RTX 2080 GPUs. Moreover, ours acquires sufficiently high clean accuracy and robustness even after 500 epochs (Figure 3(c)).

Comparison to Semi-supervised learning in required dataset Recently, semi-supervised learning[16, 17] have been shown to largely enhance the adversarial robustness of deep networks, by exploiting unlabeled data. However, they eventually require labeled data, to generate pseudo-labels on the unlabeled samples, and to generate class-wise adversaries. Also, they assume the availability of a larger dataset to improve robustness on the target dataset and require extremely large computation resources.

B Algorithm of RoCL

We present the algorithm for RoCL in Algorithm 2. During training, we generate the instance-wise adversarial examples using contrastive loss and then train the model using two differently transformed images and their instance-wise adversarial perturbations. We also include a regularization term that is defined as a contrastive loss between the adversarial examples and clean transformed examples.

Algorithm 2 Robust Contrastive Learning (RoCL)

Input: Dataset \mathbb{D} , parameter of model θ , model f , parameter of projector π , projector g , constant λ
for all iter \in number of training iteration **do**
 for all $x \in$ minibatch $B = \{x_1, \dots, x_m\}$ **do**
 Generate adversarial examples from transformed inputs \triangleright instance-wise attacks
 $t(x)^{i+1} = \Pi_{B(t(x), c)}(t(x)^i + \alpha \text{sign}(\nabla_{t(x)^i} \mathcal{L}_{\text{con}, \theta, \pi}(t(x)^i, \{t'(x)\}, t(x)_{\text{neg}})))$
 end for
 $\mathcal{L}_{\text{total}} = \frac{1}{N} \sum_{k=1}^N [\mathcal{L}_{\text{RoCL}, \theta, \pi} + \lambda \mathcal{L}_{\text{con}, \theta, \pi}(t(x)_k^{\text{adv}}, \{t'(x)_k\}, \{t(x)_{\text{neg}}\})]$ \triangleright total loss
 Optimize the weight θ, π over $\mathcal{L}_{\text{total}}$
end for

C Results of CIFAR-10 and CIFAR-100

While we only report the performance of RoCL on CIFAR-10 in the main paper as the baselines we mainly compare against only experimented on this dataset, we further report the performance of RoCL on CIFAR-100 as well (Table 8) and performance against CW attacks [21] (Table 9). We observe that RoCL consistently achieves comparable performance to that of the supervised adversarial learning methods, even on the CIFAR-100 dataset. Moreover, when employing the robust linear evaluation, RoCL acquires better robustness over the standard linear evaluation. Finally, the transformation smoothed classifier further boosts the performance of RoCL on both datasets.

Table 8: Experimental results with white box attacks on ResNet18 trained on the CIFAR-10 and CIFAR-100 dataset. r-LE denotes robust linear evaluation. AT denotes the supervised adversarial training[9]. All models are trained with ℓ_∞ ; thus the ℓ_∞ is the *seen* adversarial attack and ℓ_2 , and ℓ_1 attacks are *unseen*.

Train type	Method	CIFAR10								CIFAR100							
		<i>seen</i>				<i>unseen</i>				<i>seen</i>				<i>unseen</i>			
		ℓ_∞			ℓ_2	ℓ_1			ℓ_∞			ℓ_2	ℓ_1				
		A_{nat}	ϵ	8/255	16/255	0.25	0.5	7.84	12	A_{nat}	ϵ	8/255	16/255	0.25	0.5	7.84	12
Supervised	\mathcal{L}_{CE}	92.82	0.00	0.00	20.77	12.96	28.47	15.56	71.35	0.00	0.00	6.54	2.31	11.14	5.86		
	AT[9]	81.63	44.50	14.47	72.26	59.26	66.74	55.74	53.97	20.09	6.19	43.08	32.29	40.43	33.18		
	TRADES[2]	77.03	48.01	22.55	68.07	57.93	62.93	53.79	56.63	17.94	4.29	44.82	33.76	43.70	37.00		
Self-supervised	SimCLR[12]	91.25	0.63	0.08	15.3	2.08	41.49	25.76	57.46	0.04	0.02	6.58	0.7	19.27	12.1		
	RoCL	83.71	40.27	9.55	66.39	63.82	79.21	76.17	56.13	19.31	4.30	38.65	35.94	50.21	46.67		
	RoCL + rLE	80.43	47.69	15.53	68.30	66.19	77.31	75.05	51.82	26.27	8.94	41.59	39.86	49.00	46.91		

Table 9: Experimental results with white box CW attacks [21] on ResNet18 trained on the CIFAR-10. r-LE denotes robust linear evaluation. All models are trained with ℓ_∞

Train type	Method	CIFAR-10		CIFAR-100	
		A_{nat}	CW	A_{nat}	CW
Self	RoCL	83.71	77.35	56.13	44.57
-supervised	RoCL+rLE	80.43	76.15	51.82	44.77

D Ablation

In this section, we report the results of several ablation studies of our RoCL model. For all experiments, we train the backbone network with 500 epochs and train the linear layer with 100 epochs, which yield models with sufficiently high clean accuracy and robustness. We first examine the effects of the target image when generating the instance-wise adversarial examples. Along with instance-wise attacks, the regularization term in algorithm 1 can also affect the final performance of the model. To examine lambda’s effect on the transformed images, we set lambda as $\lambda = 1/256$ for CIFAR-10 and CIFAR-100. We also examine the effects of lambda λ on the CIFAR-10 dataset.

D.1 Adversarial contrastive learning

We examine the effect of the transformation function on the instance-wise attack and the regularization. For each input instance x , we generated three transformed images $t(x)$, $t'(x)$, and $t(x)^{adv}$ and use them as the positive set. The results in Table 10 demonstrate that using any transformed images from the same identity for instance-wise attacks is equally effective. In contrast, for regularization, using images transformed with a different transformation function from the one used to generate attack helps obtain improved clean accuracy and robustness.

Instance-wise attack To generate instance-wise attacks, we can decide which identity we will use for instance-wise attack. Since the original transformed image $t(x)$ and image transformed with another transformation $t'(x)$ have the same identity, we can use both of them in instance-wise attacks. To find the optimal perturbation that maximizes the contrastive loss between adversarial examples and same identity images, we vary X in the following equation:

$$t(x)^{i+1} = \Pi_{B(t(x), \epsilon)}(t(x)^i + \alpha \text{sign}(\nabla_{t(x)^i} \mathcal{L}_{\text{con}, \theta, \pi}(t(x)^i, \{X\}, t(x)_{\text{neg}}))) \quad (9)$$

where X is either $t'(x)$ and $t(x)$.

Regularization To regularize the learning, we can calculate the contrastive loss between adversarial examples and clean samples with the same instance-level identity. We vary Y in the regularization term to examine which identity is the most effective, as follows:

$$\lambda \mathcal{L}_{\text{con}, \theta, \pi}(t(x)^{adv}_i, \{Y\}, \{t(x)_{\text{neg}}\}) \quad (10)$$

where Y can be $t'(x)$ and $t(x)$.

Table 10: Experimental results with white box attacks on ResNet18 trained on the CIFAR-10 and CIFAR-100 dataset. All models are trained with ℓ_∞ .

Method	instance-wise attack (X)		regularization (Y)		CIFAR-10		CIFAR-100	
	$t'(x)$	$t(x)$	$t'(x)$	$t(x)$	A_{nat}	ℓ_∞	A_{nat}	ℓ_∞
RoCL	✓	-	✓	-	82.79	36.71	55.64	17.56
	✓	-	-	✓	81.47	29.97	53.84	14.18
	-	✓	✓	-	82.43	34.93	55.61	17.42
	-	✓	-	✓	81.96	30.99	53.76	14.74

D.2 Lambda λ and batch size B

We observe that λ , which controls the amount of regularization in the robust contrastive loss, and the batch size for calculating the contrastive loss, are two important hyperparameters for our robust contrastive learning framework. We examine the effect of two hyperparameters in Table 11, and Table 12. We observe that the optimal lambda λ is different for each batch size B .

Table 11: lambda λ ablation experimental results with white box attacks on ResNet18 trained on the CIFAR-10 dataset. All models are trained with ℓ_∞ .

CIFAR-10	λ	A_{nat}	ℓ_∞	
			8/255	16/255
RoCL	1/16	82.05	35.12	8.05
	1/32	82.25	36.02	8.68
	1/64	83.00	36.26	8.19
	1/128	82.79	36.71	8.34
	1/256	82.12	38.05	8.52
	1/512	82.68	37.24	8.53

Table 12: Ablation study of the batch size B , for the white box attacks on ResNet18 trained on the CIFAR-10 dataset. All models are trained with ℓ_∞ attacks.

CIFAR-10	B	λ	A_{nat}	ℓ_∞	
				8/255	16/255
RoCL	256	1/128	82.70	37.13	8.98
	256	1/256	82.90	36.86	8.89
	512	1/256	82.12	38.05	8.52
	1024	1/256	81.48	34.98	7.42