Machine Learning (2023) 112:2105–2130
https://doi.org/10.1007/s10994-022-06269-7

Adversarial supervised contrastive learning

Zhuorong Li[1] · Daiwei Yu[1] · Minghui Wu[1] · Canghong Jin[1] · Hongchuan Yu[2]

Received: 2 March 2022 / Revised: 16 August 2022 / Accepted: 7 October 2022 /
Published online: 1 November 2022

Abstract
Contrastive learning is prevalently used in pre-training deep models, followed with fine-tuning in downstream tasks for better performance or faster training. However, pre-trained models from contrastive learning are barely robust against adversarial examples in downstream tasks since the representations learned by self-supervision may lack the robustness and also the class-wise discrimination. To tackle the above problems, we adapt the contrastive learning scheme to adversarial examples for robustness enhancement, and also extend the self-supervised contrastive approach to the supervised setting for the ability to discriminate on classes. Equipped with our new designs, we proposed adversarial supervised contrastive learning (ASCL), a novel framework for robust pre-training. Despite its simplicity, extensive experiments show that ASCL achieves significant margins in adversarial robustness over the prior arts, proceeding towards either the lightweight standard fine-tuning or adversarial fine-tuning. Moreover, ASCL also shows benefits for robustness to diverse natural corruptions, suggesting the wide applicability to all sorts of practical scenarios. Notably, ASCL demonstrate impressive results in robust transfer learning.

Keywords Adversarial robustness · Adversarial attack · Self-supervised learning · Contrastive learning · Consistency regularization

Editors: Dana Drachsler Cohen, Javier Garcia, Mohammad Ghavamzadeh, Marek Petrik, Philip S. Thomas.

✉ Zhuorong Li
lizr@zucc.edu.cn

Daiwei Yu
ydw.ccm@gmail.com

Minghui Wu
mhwu@zucc.edu.cn

Canghong Jin
jinch@zucc.edu.cn

Hongchuan Yu
hyu@bournemouth.ac.uk

[1] School of Computer and Computing Science, Zhejiang University City College, Hangzhou 310015, China

[2] National Centre for Computer Animation, Bournemouth University, Poole BH12 5BB, UK

1 Introduction

Whereas neural networks are being developed in a broad spectrum of applications with great success, they are not intrinsically robust. In particular, by imposing imperceptible but carefully chosen deviations on the inputs, which are also known as adversarial perturbations, the resulting images can drastically change the prediction of the neural network (Biggio et al., 2013; Pei et al., 2019; Goodfellow et al., 2014). Worse still, these crafted inputs can be transferred across different models, enabling the black-box attacks where the target model is even hidden from the attackers (Szegedy et al., 2014; Papernot et al., 2016, 2017). It is thus of great importance to ensure that the deployed models are robust and generalize to diverse input perturbations, especially in safety-critical and security-sensitive applications, e.g., autonomous driving and identity authentication system (Biggio & Roli, 2018; Mirjalili & Ross, 2017).

Since the first observation on the high vulnerability to adversarial examples of the neural networks, there has been a flurry of activity on crafting sophisticate adversarial perturbations (Gowal et al., 2020; Athalye et al., 2018; Carlini & Wagner, 2017), which has thus encouraged a great deal of investigation on building defenses against such perturbations (Goodfellow et al., 2015; Tramèr et al., 2018; Yan et al., 2018; Cissé et al., 2017). Among them, adversarial training is widely regarded as one of the most effective methods to achieve robustness, which trains robust models by generating adversarial examples at each training steps and injecting them into the training set (Goodfellow et al., 2015; Madry et al., 2018).

Despite great success of adversarial training and its variants, the accuracy of deep models on adversarial inputs is still far below that on normal inputs. This gap might be traced back to an increased sample complexity that induced by the non-convex nature of the min–max formulation for adversarial training and most of its variants (Schmidt et al., 2018). An intuitive scheme is to leverage more training data for greater robustness (Schmidt et al., 2018; Alayrac et al., 2019). However, this is impractical in many real world applications as the annotations and data efficiency challenges are further exacerbated in the context of adversarial training.

Self-supervised and unsupervised training techniques attempt to address this challenge by eliminating the need for manually labeled data. Among them, contrastive learning (CL) has advanced self-supervised representation learning and achieved state-of-the-art performance (Chen et al., 2020; Wu et al., 2018; Hénaff, 2020; van den Oord et al., 2018; Hjelm et al., 2019; He et al., 2020). Despite a recent surge in activity, CL-based self-supervised learning remains underestimated and has only recently been leveraged for gaining robustness. The prior work (Chen et al., 2020) is the first to incorporates adversarial training with self-supervised learning, nevertheless, it heavily relies on multiple ad-hoc pretext tasks. To address this issue, a new family of methods have been proposed (Jiang et al., 2020; Kim et al., 2020; Fan et al., 2021). Intuition behind these methods is to encourage the pre-trained representation to be robust through contrastive learning. However, it still leaves many unanswered questions, especially with respect to the inadequacy of class-discrimination, which is a characteristic issue of conventional self-supervised learning.

Our work attempts to make a rigorous and comprehensive study on addressing the above issues. Inspired by a successful discrimination enhancement solution (Khosla et al., 2020), we propose a framework for the robust pre-training, which advances the self-supervised contrastive learning by incorporating with a supervision for discrimination enhancement, and take aim at adversarial robustness.

Our method is motivated by two observations. First, contrastive learning approaches alone have already been able to acquire expressive representations of data, yet not adversarially robust. Second, adversarial self-supervised learning is somewhat effective for robust generalization while is lack of class-discrimination. Therefore, we adapt the contrastive learning scheme to adversarial examples for robustness enhancement, and also extend the self-supervised contrastive approach to the supervised setting for class-wise discrimination. In the empirical part, we verify the effectiveness of our novel framework on adversarial training benchmarks following the common protocols in Zhang et al. (2019), Chen et al. (2020). Concretely, we make the following contributions:

1. We propose ASCL, an **A**dversarial **S**upervised **C**ontrastive **L**earning framework, which adapts the contrastive learning scheme to adversarial examples and further extends the self-supervised approach to supervised setting, leading to both adversarial robustness and class-discrimination.
2. We propose to simultaneously inject label-independent and label-based attacks into the pipeline, which are generated by self-supervised contrastive loss and supervised cross-entropy respectively, to further boost the generalization and calibration.
3. We verify the proposed ASCL through extensive evaluation under attacks of different setups and also in highly challenging scenarios, e.g., transferring across datasets, defending against diverse unforeseen corruptions. The proposed method shows consistent superiority, proceeding towards either the lightweight standard fine-tuning or adversarial fine-tuning, and on both quantitative and qualitative evaluations.

2 Related work

Our method is deeply rooted in the recent surge of studies on self-supervised representation learning, contrastive learning and adversarial training. Here we focus on the most relevant work.

2.1 Contrastive learning

Numerous approaches for self-supervised representation learning have been developed in recent years. Most adopt objective functions similar to those for the supervised learning, while train the models on handcrafted pretext tasks where the labels are derived from unlabeled data. Pretext tasks including rotation prediction (Gidaris et al., 2018), Jigsaw puzzle (Noroozi & Favaro, 2016), Selfie (Trinh et al., 2019) and region filling (Criminisi et al., 2004), heavily rely on heuristics and thus suffering from limited generality of representations.

A recently proposed family of self-supervised methods, contrastive learning (Chen et al., 2020; Tian et al., 2020b; He et al., 2020; Chen et al., 2020), have demonstrated impressive ability in learning generalizable representations. The general idea of contrastive learning is to acquire invariant representations by maximizing the agreement between positive samples while contrasting with the negatives, where the positives are different augmented views of the same sample. SimCLR (Chen et al., 2020) has been demonstrated a simply yet powerful contrastive learning framework for representation learning, on which we elaborate our formulation. Another work (Khosla et al., 2020) extended the self-supervised contrastive approaches to the supervised setting, for a fully utilization of label information. Despite existing literature

on contrastive learning (Hénaff, 2020; Wu et al., 2018; Hjelm et al., 2019; Tian et al., 2020a; Bachman et al., 2019; Misra & van der Maaten, 2020) show improvement on either the generalization or discrimination, most of them perform standard training and therefore do not tackle adversarial attacks.

2.2 Adversarial training

There has been a flurry of activity on building defense against the adversarial attacks in recent years. Methods range from input denoising (Cissé et al., 2017; Guo et al., 2018; Liao et al., 2018), adversarial detection (Lee et al., 2018; Ma et al., 2018), and adversarial training (Madry et al., 2018; Wang et al., 2019; Kannan et al., 2018; Qin et al., 2019; Zhang et al., 2019; Wang et al., 2020). Among them, adversarial training is widely regarded as one of the most effective methods to achieve robustness. The classical version of adversarial training (abbreviated as AT for simplicity) was proposed by Madry et al. (2018), which has withstood intensive scrutiny and is so effective that it is the de facto standard for training models robust against adversarial examples (Gowal et al., 2020; Rebuffi et al., 2021). The main idea behind adversarial training is to trains robust models by generating adversarial examples at each training steps and injecting them into the training set (Goodfellow et al., 2015; Madry et al., 2018).

Another line of work Kannan et al. (2018), Zhang et al. (2019) have proposed consistency regularization to boost the adversarial robustness over (Madry et al., 2018) and its variants. One notable work, TRADES (Zhang et al., 2019), measures the Kullback–Leibler divergence on the softmax outputs for pairs of natural-adversarial images, providing sufficient theoretical guarantees for a better trade-off between standard and robust accuracy. Their success inspires our nature-adversarial contrastive views in robust pre-training.

2.3 Self-supervised adversarial training

Self-supervised learning has only recently been connected to the study of adversarial robustness. Several recent studies have attempted to use contrastive representation learning in adversarial robustness field (Chen et al., 2020), in the most straightforward way to inject adversarial samples yet yield unsatisfactory results. Other concurrent work (Jiang et al., 2020; Fan et al., 2021) improve the robustness by less aggressive adversarial contrastive views.

Despite better robustness, it still leaves unanswered questions on the class-discriminative ability, which is unacquirable in the self-supervised pre-training but is required by a robust prediction on downstream tasks. In the literature (Jiang et al., 2020; Kim et al., 2020), the final adversarial robustness on downstream tasks usually heavily relies on advanced techniques in the phase of fine-tuning, and thus makes the advantages of contrastive-based pre-training less significant. For example, Jiang et al. (2020) suggested an adversarial full fine-tuning, where the pre-trained model is only used as an initialization and all the weights require updating by adversarial training.

3 Our proposal

In this section, we present a novel framework for robust pre-training, which advances supervised contrastive learning in the adversarial scenario. We incorporate the supervision as a complementary objective, which is co-optimized with the self-supervised contrastive loss through adversarial training.

3.1 Preliminaries

3.1.1 Problem statement

For a L-class classification problem with a given dataset $\mathcal{D} = \left\{(x_i, y_i)\right\}_{i=1}^{n} \subseteq \mathcal{X} \times \mathcal{Y}$, the goal is to learn a classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$ that maps the image $x \in \mathcal{X} \subseteq \mathbb{R}^d$ to one-hot label $y \in \mathcal{Y} \subseteq \mathbb{R}^L$ according to certain quality metrics, such as the error probability $\text{err}(f_\theta) := \mathbb{P}_{(x,y)\sim\mathcal{P}_{x,y}}\big(f_\theta(x) \neq y\big)$ in standard training, where $\mathcal{P}_{x,y}$ denotes the underlying joint distribution over (x, y) pairs. While in the scenario of adversarial training, the aim is to train f_θ to correctly classify not only x but also the adversarial perturbed data $x + \delta$. Here δ denotes the perturbation that subjects to ℓ_p-norm budget as $\|\delta\|_p \leq \epsilon$.

3.1.2 Adversarial training

The high-altitude idea of adversarial training(AT) is to directly augment the training set with perturbed samples that generated on-the-fly, and thus the model becomes robust to such attacks (Goodfellow et al., 2015; Madry et al., 2018; Athalye et al., 2018). Essentially, the adversarial training can be formulated as an alternative min–max optimization, whose goal is to minimize the adversarial risk:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{P}_{x,y}}\left[\max_{\delta\in\mathbb{S}} \ell(f_\theta(x+\delta), y)\right], \tag{1}$$

where $f_\theta(\cdot)$ is the output vector of the θ-parameterized learning model, x is the input image and y is the corresponding label-indicator vector. Let $\mathcal{P}_{x,y}$ denotes the underlying joint distribution over (x, y) pairs, and $\mathbb{S}$ denotes the set of allowed perturbations. For example, for adversarial perturbations within the ϵ-ball and bounded by l_p-norm, the adversarial set can be denoted as $\mathbb{S}_p = \{\delta \mid \|\delta\|_p \leq \epsilon\}$ for $\epsilon > 0$. The symbol ℓ is a suitable classification loss (e.g., the 0–1 loss in the context of classification task).

More concretely, to find the optimum of the inner maximization problem above, which is NP-hard, Madry et al. (2018) proposed to approximately optimize the inner maximization by project gradient descent (PGD) method. They compute the perturbation in K gradient ascent steps of size α as:

$$\begin{aligned} \hat{\delta} &= \delta^{(k)} + \alpha \cdot \text{sign}\nabla_x \ell\big((f(x), y\big) \\ \delta^{(k+1)} &= \max\big(\min(\hat{\delta}, \epsilon), -\epsilon\big) \end{aligned} \tag{2}$$

where $\delta^{(0)}$ is chosen at random within $\mathbb{S}$, and the symbol $\nabla_x \ell$ denotes gradient of the loss ℓ with respect to the input image x. We will refer to this inner optimization procedure with K steps as PGD-K.

For each sample x with ground-truth label y, one of the most basic form of AT (Madry et al., 2018) replaces the non-differentiable 0–1 loss with the softmax cross-entropy loss ℓ_{xent} and minimizes the loss given in (1) by the following implementation:

$$\mathcal{L}_{\text{AT}} := \max_{\|\delta\|_p \leq \epsilon} \ell_{\text{xent}}\big(f_\theta(x+\delta), y\big). \tag{3}$$

3.1.3 Contrastive-style adversarial training

Contrastive learning is an important class of the self-supervised learning algorithms, which is a powerful approach to learning effective representations for better performance or faster training on downstream tasks, without requiring labeled data (Wu et al., 2018; Chen et al., 2020; Hénaff, 2020; Hjelm et al., 2019; Tian et al., 2020a). The general idea is to learn effective representations by maximizing the agreement between different augmentations of the same sample, via a contrastive loss:

$$\ell_{\mathrm{CL}}(\widetilde{x}_i, \widetilde{x}_j) = -\sum_{i \in I} \log \frac{\exp(z_i \cdot z_j/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)}, \tag{4}$$

where $\widetilde{x}_i$ and $\widetilde{x}_j$ are augmented views of sample x, and $(\widetilde{x}_i, \widetilde{x}_j)$ denotes a contrastive pair. The notation $i \in I \equiv \{1 \cdots 2N\}$ is the index of an arbitrary augmented view, j is the index of a different augmented view of the same sample, and $A(i) \equiv I \backslash \{i\}$ is the set that with i excluded. The index i and j is called the anchor and the positive, respectively, with the other $2(N-1)$ indices called the negatives. The symbol $\cdot$ denotes the inner product, $\tau > 0$ denotes a temperature parameter, z_k denotes the projected representation under the k-th augmented view, and $\exp(\cdot)$ denotes the exponential function.

When applying the contrastive learning to the field of adversarial defense, namely, robust pre-training, the optimization objective (1) is given by:

$$\min_{\theta} \mathbb{E}_{x \in \mathcal{X}} \max_{\|\delta\|_p \leq \epsilon} \ell_{\mathrm{CL}}(\widetilde{x}_i + \delta_i, \widetilde{x}_j + \delta_j), \tag{5}$$

where the adversarial perturbation δ_i, δ_j w.r.t. the contrastive view $\widetilde{x}_i$, $\widetilde{x}_j$ can be computed by the PGD-K optimization procedure (2) with the loss ℓ designed as the contrastive loss ℓ_{CL}.

Generally, in the contrastive-style adversarial training, a supervised fine-tuning will usually immediately follow the self-supervised pre-training (5). Specifically, the supervised fine-tuning can be formulated as:

$$\min_{\theta_c} \mathbb{E}_{(x,y) \in D} \ell_{\mathrm{sup}}\left(\phi_{\theta_c} \circ f_{\theta}(x), y\right), \tag{6}$$

where ℓ_{sup} denotes a supervised loss (e.g., the cross-entropy loss), the notation $\phi_{\theta_c} \circ f_{\theta}$ denotes the classifier that with a linear prediction head ϕ_{θ_c} (to be learned in the fine-tuning) on top of a feature encoder f_{θ}, which is learned in the pre-training phrase. Note that one can also apply the worse-case cross-entropy loss for the supervised fine-tuning, that is, using $\phi_{\theta_c} \circ f_{\theta}(x + \delta)$ in (6) instead.

New challenge arises is that in the aforementioned self-supervised contrastive approaches, a class-discriminative ability specific to the target categories is unacquirable but is required by a robust prediction on downstream tasks. This might be attributed to a task mismatch from the self-supervised pre-training to the fully-supervised fine-tuning.

3.2 Proposed method: ASCL

3.2.1 Equips contrastive pre-training with class-wise discrimination

As we target for class-wise discrimination, we adopt a supervised recipe in the adversarial contrastive learning, as oppose to the self-supervised contrastive learning. While this is a simple extension to the self-supervised setup, it is non-obvious how to properly setup the training objective. On the one hand, the sample x is commonly restricted to be unlabeled data in the previous robust pre-training methods, as (4) is incapable of handling the labeled data. On the other hand, a supervised fine-tuning will usually immediately follow the robust pre-training, with the use of a supervised loss, e.g., cross-entropy loss or adversarial classification loss, and train over target dataset that contains labels. In what follows, we will elaborate a solution.

Specifically, we borrow ideas from the recently proposed supervised contrastive learning (SCL) techniques (Khosla et al., 2020), which has achieved substantial gains in discrimination over the conventional contrastive learning by extending the self-supervised contrastive learning to the supervised setting. The supervised contrastive loss with respect to the augmented views of the sample data x can be given by the following formulation:

$$\ell_{\mathrm{SCL}}(\widetilde{x}_i, \widetilde{x}_p) = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a\in A(i)} \exp(z_i \cdot z_a/\tau)}. \tag{7}$$

Being similar to the notation of $A(i)$ as aforementioned in (4), $P(i)$ denotes the set of indices of all positives distinct from i, i.e., $P(i) \equiv \{p \in A(i) : y_p = y_i\}$, and the symbol $|\cdot|$ denotes the cardinality. In practical, $\widetilde{x}_i$ and $\widetilde{x}_p$ here are common augmentations of the original sample x, such as random cropping, random color distortion and their composition.

Note that SCL is originally devised for the classification on natural images, where the source sample x as well as the corresponding augmented views $\widetilde{x}_i$ and $\widetilde{x}_p$ are benign. Nevertheless, our goal is to develop robustness enhancement solutions. To this end, the most direct way is to replace the common augmentations with their adversarial versions and then use for contrastive learning, in a similar way with self-supervised contrastive pre-training. Thereby, an SCL loss that adapts to adversarial samples can be given by a SCL loss over an adversarial contrastive pair $(\widetilde{x}_i + \delta_i, \widetilde{x}_p + \delta_p)$, which has the following form:

$$\sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp(z_i^{\mathrm{adv}} \cdot z_p^{\mathrm{adv}}/\tau)}{\sum_{a\in A(i)} \exp(z_i^{\mathrm{adv}} \cdot z_a/\tau)}, \tag{8}$$

where z_i^{adv}, z_p^{adv} denotes the projected representation of the adversarial sample $\widetilde{x}_i + \delta_i$, $\widetilde{x}_p + \delta_p$, respectively. More specifically, a PGD adversary can iteratively adjust perturbation by gradient descent in a same manner as (2) to maximize the modified SCL loss (8).

Intuitive as it might look, there is a pitfall for this implementation. By minimizing (8), the representation we learn is invariant to different cascades of augmentation and adversarial perturbation, i.e., $\widetilde{x} + \delta$, while what we hope to defend against are adversarial samples $x + \delta$. As observed in Xie and Yuille (2020), in the context of deep learning, the representation statistics of natural sample x and the corresponding adversarial version $x + \delta$ can be very different. This observation is also hold true when referring to $\widetilde{x} + \delta$. Thus, the invariance we yield by training on $\widetilde{x} + \delta$ can not guarantee smooth boundary for us to make correct prediction on $x + \delta$.

Several recent studies Mao et al. (2019), Kim et al. (2020) have attempted to leverage metric learning in adversarial training, specifically, bringing the natural and adversarial samples of the same class (Mao et al., 2019) or same instance (Kim et al., 2020) closer while enlarging the margins between distinct class or instance. Inspired by this, we propose to enforce the supervised contrastive loss over a pair of common augmentation and the adversarial attack, which is essentially aligned with the intuition of the classical adversarial training (Zhang et al., 2019). This is expected to mitigate the aggressive consistency between different worst-case based perturbations, thus to prevent the degradation of the representation quality. Formally, the proposed adversarial supervised contrastive loss can be written as:

$$\ell_{\text{ASCL}}(\widetilde{x}_i, x+\delta) = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z^{\text{adv}}/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)}, \tag{9}$$

where z^{adv} denotes the projected representation under adversarial sample $x+\delta$.

Using adversarial contrastive views and the proposed supervision scheme in conjunction, adversarial perturbation δ_{ASCL} can be generated according to procedure (2) with a maximized ℓ_{ASCL} over the proposed contrastive pair $(\widetilde{x}_i, x+\delta)$.

Equipped with such supervision, we expect to train a robust model and empower the class-awareness in the contrastive learning framework.

3.2.2 Ensemble of adversarial samples

Recent study empirically shows that different pre-training tasks tend to induce models with different adversarial vulnerabilities (Chen et al., 2020). They thereby proposed leveraged multiple pretext tasks, including selfie (Trinh et al., 2019), jigsaw (Gidaris et al., 2018) and rotation (Carlucci et al., 2019; Noroozi & Favaro, 2016) to achieve complementary traits. Despite better defense performance can be achieved by the ensemble of self-supervised learning tasks, these handcrafted pretext tasks require heuristic scheme. Moreover, it is extremely computation consuming for models ensemble.

To tackle this problem, we proposed to use the ensemble of adversarial samples, rather than the ensemble of multiple pre-training tasks, which corresponds to training several deep models simultaneously. The rationale behind our proposal is that, adversarial samples can be used as a kind of data augmentation and lead to gains in robustness (Tack et al., 2021; Xie et al., 2020).

To this end, we use different attack generation strategies to produce ensemble of adversarial samples. In particular, we leverage the proposed supervised adversarial contrastive loss ℓ_{ASCL} and the self-supervised contrastive loss ℓ_{CL} for the label-wise and label-independent attack generation, respectively. As will be verified later, such different adversarial samples can offer complementary benefits, i.e., calibration and generalization to model performance as they are statistically distinct.

3.2.3 Overall training objective

Now we derive an overall training objective for our method, i.e., the proposed adversarial supervised contrastive loss in addition to the self-supervised robust training, with hyper-parameter λ_1 and λ_2 to control the strength of self-supervision and supervision regularization, respectively. Formally, the total loss can be given by:

$$\min_{\theta}\left[\mathbb{E}_{x\in\hat{\mathcal{X}}}\max \ell_{\mathrm{AT}}+\lambda_1\mathbb{E}_{x\in\mathcal{X}}\max_{\|\delta_{\mathrm{CL}}\|_\infty\leq\epsilon}\ell_{\mathrm{CL}}+\lambda_2\mathbb{E}_{x\in\hat{\mathcal{X}}}\max_{\|\delta_{\mathrm{ASCL}}\|_\infty\leq\epsilon}\ell_{\mathrm{ASCL}}\right] \tag{10}$$

where $\hat{\mathcal{X}}$ and $\mathcal{X}$ are labeled and unlabeled datasets, respectively.

Note that our method is agnostic to the implementation of adversarial training. Therefore, we can use either the basic version or any other variants of the adversarial training objective for ℓ_{AT} [e.g., AT (Madry et al., 2018), TRADES (Zhang et al., 2019) and MART (Wang et al., 2020), and then combined with the proposed loss term].

4 Experiments and results

In this part, we conduct comprehensive evaluations to verify the effectiveness of our proposed ASCL, including its benchmarking robustness, ablation studies and analysis, so as to provide some additional insights.

4.1 Experiment setups

To be comparable to the baselines, we follow the common protocols and experiment setups suggested by previous literature (Chen et al., 2020; Tack et al., 2021; Jiang et al., 2020; Kim et al., 2020; Fan et al., 2021)

4.1.1 Training details

We empirically benchmark our models on CIFAR-10 (Krizhevsky et al., 2009), which is widely used in adversarial training. We adopt ResNet-18 (He et al., 2016) as the backbone in all experiments, as was also used by Zhang et al. (2019). We apply stochastic gradient descent optimizer with a momentum of 0.9 and a weight decay of 2×10^{-4} during training. We employ a piece-wise learning rate schedule, which is initially set to 0.1 and decayed by a factor of 10 at 60% and 80% of the training progress, and a constant learning rate of o.1 to train the linear layer for 10 epochs unless otherwise specified. Batch size is set 512 in all experiments. As for the attacks during training, we generate the adversarial samples by standard Projected Gradient Descent (PGD) optimization the budget $\epsilon=8/255$ and step size $\alpha=2/255$ with l_∞ constraint.

By default, we first use ResNet-18 (He et al., 2016) (with the last fully connected layer excluded) as the the encoder architecture in the adversarial pre-training, and then the pre-trained encoder is frozen and added with a zero-initialized fully connected layer fine-tuning. We apply the widely used Cross-Entropy for standard fine-tuning, and the TRADES loss (Zhang et al., 2019) for adversarial fine-tuning. The hyper-parameters λ_1 and λ_2 are respectively set 0.5 and 0.2 for a well balance, basing on our experimental observations.

4.1.2 Evaluation setups

We evaluate the models by attacking them with untargeted adversarial samples in white-box setting, which is more challenging than the targeted ones to defense against. In particular, we apply 20-step PGD optimization with l_∞ constraint to generate the adversarial perturbations for all the competing methods to ensure a fair comparison.

Evaluation metrics we employ are: (1) Robust Accuracy (RA), i.e., the classification accuracy on the perturbed testing set; and (2) Standard Accuracy (SA), i.e., the accuracy on natural images without adversarial perturbations. Both RA and SA of models under attacks are reported to avoid sacrificing the nominal performance for adversarial robustness.

Unless otherwise specified, we evaluate the models by attacking them with both PGD attacks (Fan et al., 2021; Madry et al., 2018) and the more challenging Auto-Attacks (Croce & Hein, 2020b). In particular, we apply 20-step PGD optimization with l_∞ constraint to generate the adversarial perturbations for all the competing methods to ensure a fair comparison. Accordingly, we use RA-PGD and RA-AA to denote RA of models under PGD attacks and Auto-Attacks respectively, for simplicity.

Baselines we compare to can be roughly divided into supervised adversarial training and self-supervised pretraining with finetuning, including: AT (Madry et al., 2018), a solid baseline of supervised adversarial training; SCL (Khosla et al., 2020), a contrastive learning method with the use of labeled data; SimCLR (Chen et al., 2020), which is a vanilla self-supervised CL method; ACL (Jiang et al., 2020), AdvCL (Fan et al., 2021), RoCL and its enhanced version RoCL + rLE (Kim et al., 2020) are robust CL-based pretraining methods that followed with finetuning, where rLE denotes the additional use of robust Linear Evaluation; on the contrary, Selfie and Selfie + DPE (Chen et al., 2020) are self-supervised adversarial training methods that without using contrastive loss.

4.2 Comparing ASCL with the state-of-the-arts

To close in on the true robustness, we evaluate trained models under a wide range of attacks. Tables 1, 2 and Fig. 1 presents the comparison results of model robustness against white-box, black-box and unforeseen attacks, respectively, all showing that our ASCL offers consistently better performance than others. In what follows, we analyze these results and provide additional insights.

4.2.1 White-box robustness

4.2.1.1 Setup To extensively evaluate the robustness without gradient obfuscation (Athalye et al., 2018), we first consider a wide range of adversarial attacks in white-box setting, i.e., PGD attacks with 20 iterations (Madry et al., 2018) and the Auto-Attacks (Croce & Hein, 2020b), which further consists of untargeted/targeted FAB (Croce & Hein, 2020a) and square attack (Andriushchenko et al., 2020) with 5000 quires, making it a well-recognized strong attack to defend.

4.2.1.2 Results and analysis Table 1 reports the performance of the proposed ASCL and the competitive methods. The most direct baseline that any new adversarial training method should be compared with, is the plain adversarial training, AT (Madry et al., 2018). We can observe that ASCL yields a significant improvement over AT on robustness, leading a clear margin of 9.04% and 5.63% on RA-PGD and RA-AA, respectively. This is an impressive result which provides strong evidence that self-supervised representation learning is effective for robustness enhancement. Moreover, ASCL achieves comparable performance to another solid baseline, TRADES (Zhang et al., 2019), which is a supervised method with strong regularization. Note that both AT and TRADES require much more epochs to reach a plateaus as they train from scratch. This makes our method more appealing in practice.

Table 1 White-box robustness of ASCL compared with supervised and self-supervised baselines, in terms of RA-PGD, RA-AA and SA on the benchmark CIFAR-10 dataset

Training scheme	Method	Finetuning type			RA-PGD (%)	RA-AA (%)	SA (%)
		SPF	APT	AFF			
Supervised	AT				44.05	40.07	84.48
	TRADES				51.41	45.41	82.2
	SCL				0.11	0	**92.01**
Self-supervised CL + finetuning	SimCLR	✓			0.27	0	90.6
	RoCL	✓			40.27	28.38	83.71
	RoCL + rLE		✓		47.69	–	80.43
	Selfie		✓		37.65	–	74.3
	Selfie + DPE			✓	52.22	40.24	83
	ACL			✓	52.82	45.61	82.19
	AdvCL		✓		52.01	43.52	79.39
Supervised CL + finetuning	ASCL (ours)		✓		**53.09**	**45.7**	81.67

Models trained by self-supervised CL are followed with specific finetuning as suggested in original paper, e.g., standard partial finetuning (SPF), adversarial partial training (APT) and adversarial full finetuning (AFF), whereas the supervised baselines use end-to-end training without finetuning. The bold denotes the best performance. We use the published results of the baseline methods where possible

Table 2 Black-box robustness of ASCL compared with baseline methods on the benchmark CIFAR-10 dataset

Source	Target			
	AT	ACL	AdvCL	ASCL
AT	–	67.03	61.89	*59.39*
ACL	52.98	–	51.61	*50.78*
AdvCL	58.61	62.87	–	*56.98*
ASCL	**62.16**	**67.32**	**62.86**	–

Results in bold indicate top performance achieved by the proposed method. Moreover, ASCL can also be used to generate stronger black-box attacks compared to baseline methods, which can be implied by the lowest values in each rows, as denoted in italic type

We then compare against SimCLR (Chen et al., 2020) and SCL (Khosla et al., 2020), where the former is a powerful contrastive learning framework that our method builds on, and the latter one is an extension of traditional contrastive loss to the supervised setting. Result shows that SCL and SimCLR can obtain high accuracy on natural samples, nevertheless, both are extremely vulnerable to adversaries. This is not surprising as SCL and SimCLR are vanilla CL-based learning methods that without training with adversarial samples. We have also noticed that both RoCL and Selfie lead considerable improvement in robustness. Nevertheless, RoCL shows a frustrating gap between RA-PGD and RA-AA, and Selfie heavily relies on heuristics for pretext tasks (Trinh et al., 2019). Although Selfie + DPE (Trinh et al., 2019) improves over its origin and yields comparable performance to our ASCL, it requires prohibitively expensive ensemble training. Methods stand out among the rest are ACL (Jiang et al., 2020) and AdvCL (Fan et al., 2021), which indeed bring improvement over the prior arts but with much overhead, for a thorough finetuning and pseudo labeling respectively. On the contrary, ASCL suggests a lightweight finetuning

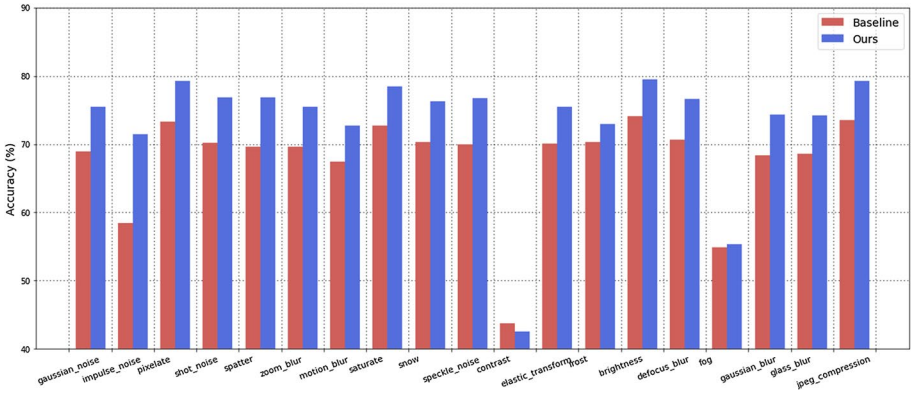


Fig. 1 Robustness against unforeseen attacks on CIFAR-10-C. The proposed ASCL outperforms the baseline on most types, demonstrating more general robustness against various perturbations

without requiring any plug-in technique or extra offline training. As a whole, the proposed ASCL offers substantial improvement over almost all baseline methods under white-box attacks.

4.2.2 Black-box robustness

4.2.2.1 Setup Previous works (Carlini & Wagner, 2017; Uesato et al., 2018) show that defenses relied on obfuscated gradients may give a false sense of robustness. In this part, we verify our models in a black-box setting, where the target models are hidden, so as the gradients for generating perturbations. To this end, we first train the ResNet-18 with baseline methods and our ASCL as target models for evaluation. Then, we use models that trained with different methods as the source models, and compute the perturbations according to gradients that provided by the source models on the inputs. Specifically, we craft adversarial samples on AT, ACL and AdvCL model to test our ASCL trained model, and the rest can be done in the same manner.

4.2.2.2 Results and analysis As shown in Table 2, our ASCL consistently achieves the best robustness under black-box attacks, as denoted in the bold. Moreover, the proposed ASCL can also be used to generate stronger black-box attacks compared to baseline methods, which can be implied by the lowest values in each rows. For instance, the target model AT yields the worst accuracy when it defenses against attacks that generated by the gradients of ASCL, with a significant drop of 7.64% and 2.50% on robust accuracy compared to that sourced by ACL and AdvCL, respectively.

4.2.3 Robustness against unforeseen attacks

4.2.3.1 Setup In addition to evaluating at adversarial attacks, which includes white-box and black-box attacks, we also verify if our robustness enhancement can consistently hold against unforeseen attacks. As suggested by previous works (Kang et al., 2020; Hendrycks & Dietterich, 2019), we employ 19 common corruptions on Cifar-10-C as unforeseen attacks for testing, e.g., impulse noise, JPEG compression and elastic transform. Figure 1

presents the performance comparison between our proposed ASCL and the baseline method in Hendrycks et al. (2019), which is not specially trained for adversarial defense but rather a more generalized robustness against various corruptions.

4.2.3.2 Results and analysis As we can see, our method achieves consistent robustness gains in defending most of the unforeseen perturbations (18 out of 19 types), where the gain ranges from 0.46 to 12.99% while a slight drop of 1.05% on the contrast corruption. Remarkably, our method brings about an overall gain of 5.55% when averaging on 19 unforeseen attacks, indicating improvement in robustness are spread across perturbations. We note that as inputs might be attacked in various ways that not have been encountered in training, such robustness should be an indispensable property for models deployed in real-world.

4.3 Ablation study and analysis

4.3.1 The transfer of learned representation across datasets

4.3.1.1 Setup In what follows, we verify our method on transfer learning in the scenario of adversarial defense, in order to gain some insight into the transferability of robust representations from adversarial pre-training to fine-tuning. We closely follow the protocols as previous literature (Kim et al., 2020; Fan et al., 2021). Specifically speaking, models are first pre-trained on dataset A with the same setup as described in Sect. 4.1. We froze the trained network except the logits layer, where we use a zero-initialized fully connected layer instead and then fine-tune it on another dataset B. We perform such an evaluation on two transfer learning tasks, namely, Cifar-10 → STL-10 and Cifar-10 → Cifar-100. The former one transfers the learned representation from a larger dataset to a smaller one while the latter is the opposite.

4.3.1.2 Results and analysis Table 3 shows how our models compare to one-shot adversarial training as well as baselines that similar to us in separating the adversarial training into pre-training and fine-tuning. We observe that ASCL leads a comfortable margin over one-shot training while RoCL (Kim et al., 2020) could barely yield comparable results. This implies that our improvement over AT (Madry et al., 2018) essentially benefits from the richer representations learned by our robust pre-training, rather than the two-step learning. The advantage of integrating the supervised contrastive loss into adversarial training can be confirmed with the substantial improvement over the prior arts ACL (Jiang et al., 2020) in [RA, SA] by [6.05%, 9.94%] on Cifar-10 → STL-10, and [4.97%, 3.34%] on Cifar-10 → Cifar-100, respectively. More intriguingly, our proposal improves RA as well as SA, which is non-trivial as previous studies have reported a common trade-off between the robust and clean accuracy in adversarial training (Tsipras et al., 2019; Zhang et al., 2019).

4.3.2 Visual interpretability of the learned representation

4.3.2.1 Setup To further demonstrate the effectiveness of our proposal, we visualize the feature inversion map (Mahendran & Vedaldi, 2016) of neuron response. It has been shown in Boopathy et al. (2020), Engstrom et al. (2019), Kaur et al. (2019) that representations of robust models are more aligned with human-recognizable features than those of the standard networks. Therefore, the feature inversion map can be use as a good indicator for robustness. Concretely, we acquire such maps by maximizing the activation of a specific component of

Table 3 Evaluation results on transfer learning in the scenario of adversarial defense

Source	Target	Training type	Method	RA	SA
Cifar-10	STL-10	One-shot	AT	30.45	54.7
		Pre-training	RoCL	28.18	54.56
		+	ACL	31.8	55.81
		Fine-tuning	ASCL	**37.85**	**65.75**
Cifar-10	Cifar-100	One-shot	AT	17.63	47.59
		Pre-training	RoCL	15.33	45.84
		+	ACL	18.69	47.13
		Fine-tuning	ASCL	**23.66**	**50.47**

The bold indicates the best results. Our models significantly improve state-of-the-art

the representation vector with respect to the input images. Formally, we solve the optimization problem:

$$x' = \arg\max_{\delta} g(x_0 + \delta)_i, \tag{11}$$

where x_0 denotes random seed inputs (images/noise), and i denotes the ith component of the vector.

4.3.2.2 Results and analysis Fig. 2 shows that for model trained with our method, different neurons consistently represent similar pattern across highly distinct input images. That is in starFk contrast to the case of standard models, where the representations tend to be perceptually meaningless and change drastically with the inputs, implying poor model explanation and classifier smoothness. Remarkably, we can still observe consistency of feature representations even when we change the seed input from natural image to random noise, not just across different natural images. This justifies that ASCL equips the model with more perceptually-aligned representation thus to better robustness.

4.3.3 Class-wise discrimination

Furthermore, we demonstrate the advantages of the proposed method from the perspective of class-discriminate ability, which can be illustrated by the visualization of the latent representations learned by models. Figure 3 shows the t-SNE visualization (Arora et al., 2018; Rauber et al., 2016) on sampled CIFAR-10 images, with different colors representing different ground-truth label. We visualize the penultimate fully connected layer of model trained with Rocl (Kim et al., 2020), ACL (Jiang et al., 2020) and the proposed ASCL, respectively.

As we can see in Fig. 3, RoCL (Kim et al., 2020) and ACL (Jiang et al., 2020) show indistinct boundary, with representations getting closer together. Note that a good discriminative model will push apart clusters of samples from different classes, so that it will be difficult to misclassify the images into wrong categories. In striking contrast to (a) and (b), we can observe clearly separated clusters in (c), where samples belonging to the same class are pulled together while points from different classes are pushed apart, indicating that ASCL is able to better discriminate images among different classes than the baselines.

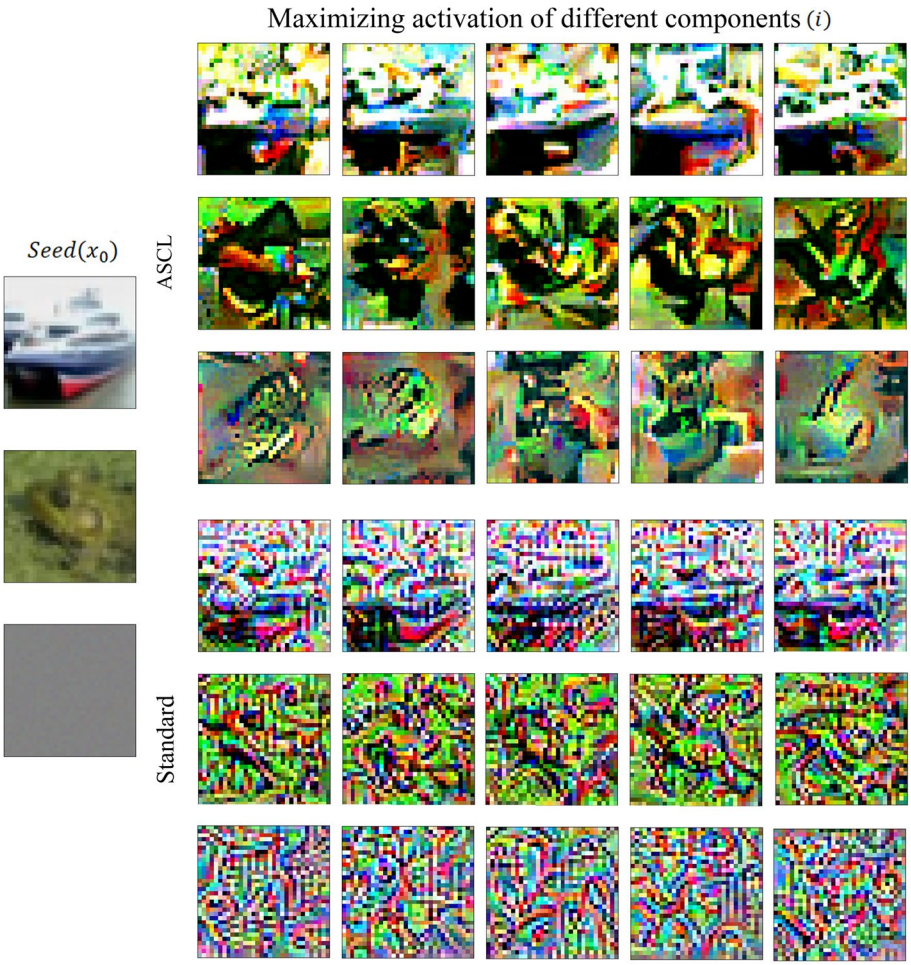
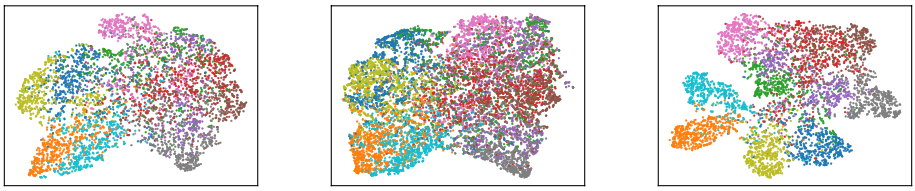


Fig. 2 The comparison on feature inversion map of ResNet-18 neurons when the proposed ASCL and standard training is applied. In the left column we present the seed inputs that randomly selected from CIFAR-10, and in the subsequent columns we visualize the activation of a few components of the representation vector with respect to seed inputs. The striking contrast between the visual interpretability of our model (top) and that of the standard model (bottom) demonstrates significant robustness enhancement achieved by our proposal



(a) RoCL **(b)** ACL **(c)** ASCL(ours)

Fig. 3 t-SNE visualization of latent representations learned with different methods. The proposed ASCL shows a much clearer separation among different classes

4.3.4 Linear separability of the learned representations

4.3.4.1 Setup To further evaluate the learned representations, we carry out the linear evaluation. We follow the evaluation protocol in Chen et al. (2020). Specifically speaking, we attach a linear prediction head on top of the pre-trained encoder, with a stop gradient on the input to the linear head, so as to prevent the label from affecting the pre-trained encoder during evaluation. We separately perform standard training and adversarial training on the linear classification head, and the evaluate the representation learned by ASCL and baseline methods in terms of both SA and RA, which are use as an indicator for the representation quality. Higher SA and RA implies better *standard linear separability* and *adversarial linear separability* respectively.

4.3.4.2 Results and analysis As shown in Table 4, our method yields the best results with clear performance margins over the others under different evaluation settings, as denoted in bold with respect to each column. Furthermore, we analyze the results by row, to acquire insights from the pre-trained representations. It is notable that the gap induced by Selfie (Trinh et al., 2019) between RA of standard training and that of adversarial training is extremely large, which indicates that the eventual robustness of model is heavily relied on adversarial fine-tuning rather than the self-supervised pretraining. Though significant improvement on RA (from 6.3 to 37.65%) can be obtained when we switch the tuning to the adversarial mode, its robustness is still lags behind other methods. On the contrary, we can observe the tightest gap (0.71%) achieved by ASCL, demonstrating that representations learned by our proposed supervised CL-based pre-training is superior to baseline methods, all of which are in self-supervision. This makes our ASCL more appealing as it is much less dependent on the retaining, and thus enabling a merely lightweight finetuning in downstream tasks.

4.3.5 Robust generalization analysis through loss surface

There exists a large body of work investigating the correlation between robust generalization and salient properties of the function surface of DNNs (Keskar et al., 2017; Li et al., 2018; Wu et al., 2020). It has been empirically veried and commonly accepted that atter loss surface tends to yield better generalization, and this understanding is further utilized to design regularization (e.g., Garipov et al. 2018; Wei et al. 2020; Ishida et al. 2020). Here, we leverage this agreement to analyze the robust generalization. In particular, we visualize two types of loss landscape to illustrate the function surface, i.e., *input* loss landscape and *weight* loss landscape. The former indicates the change of loss in the vicinity of input data,

Table 4 Comparison of linear separability of ASCL with baseline methods

Method	Standard training		Adversarial training	
	SA (%)	RA (%)	SA (%)	RA (%)
Selfie	78.93	6.3	74.30	37.65
ACL	79.76	30.31	74.22	44.22
RoCL	83.71	40.27	80.43	47.69
AdvCL	80.85	50.45	79.39	52.01
ASCL(ours)	**81.35**	**52.38**	**81.67**	**53.09**

The bold indicates the best results

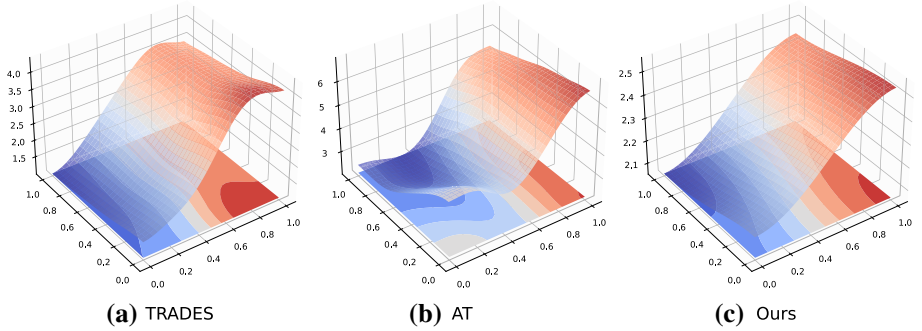


Fig. 4 Visualization of loss landscape on CIFAR-10 test set for various models

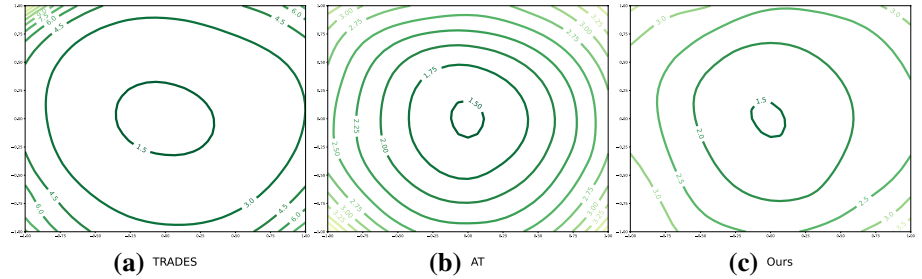


Fig. 5 Visualization of loss landscape w.r.t. model weights using different methods on the CIFAR-10 test set

and the latter depicts the geometry of the loss landscape around model weights instead of the randomly sampled input (Wu et al., 2020).

Figure 4 shows that the resultant model of AT (Madry et al., 2018) exhibits a convolved surface, implying the occurrence of robust overfitting that caused by supervision. While TRADES (Zhang et al., 2019) significantly improves AT on robust generalization, it is not so strong as our method when defending against perturbation of larger radius. Among these subplots, Fig. 4c presents the smoothest and flattest input loss landscape, as well as the lowest level of loss.

We further present a comparison on robust generalization through the lens of weight loss landscape. Similar observation can be found in Fig. 5, where the proposed ASCL produces a flatter surface w.r.t. weight perturbation than other competing methods, i.e., AT and TRADES. These results confirm our conjectures that the proposed ASCL, which is rooted in contrastive learning, shows great advantages in learning generalized representations over the supervised baselines.

4.3.6 Towards larger steps and larger radius

To further examine whether our trained models can be well generalized, we perform evaluation under various PGD attacks, with the amount of PGD steps varied from 0 to 100 and the perturbation budget varied from 0 to 32/255, respectively. Figure 6 shows that the

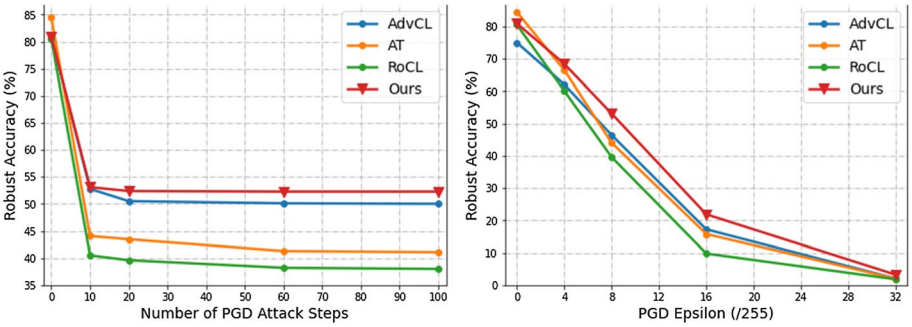


Fig. 6 Robustness accuracy of models evaluated under attacks with different strength, in terms of varied amount of PGD steps (left) and perturbation budgets (right)

Table 5 Ablation study on each component of the proposed training objective

Component	SA	RA-PGD	RA-AA
AT objective	79.74	40.51	36.74
AT + Self-sup CL	76.97	48.1	42.52
AT + Sup CL	81.5	48.09	42.46
Full model	**81.67**	**53.09**	**45.7**

We report the clean accuracy as well as the robust accuracy on PGD-20 attacks and Auto-Attack. The bold indicates that the full mode offers the best performance

proposed ASCL consistently outperforms the baselines with a non-trivial margin, ensuring security under a wide range of attacks and thus a good generalization.

4.3.7 Ablation on components

In this part, we compare against the ablations of our full model to explore if each component is essential. Quantitative results reported in Table 5 demonstrate that each component of the proposed training object is indispensable for the final performance, as the robust accuracy under both PGD attacks and Auto-Attack gets improved step-by-step with the integration of the component. It is interesting to find that the standard accuracy first gets reduced when added with the self-supervised contrastive objective, yet it goes up as the proposed supervision is incorporated. This suggests that our method greatly benefits from the class-discrimination that empowered by the supervised contrastive learning. This is in consistent with Table 1, showing that our method sacrifices less standard accuracy than the baseline for robustness.

To delve into the proposed ASCL, we go beyond adversarial robustness and investigate how the ablations of our full model affect the robustness against a broad range of corruptions (e.g., Gaussian blur, impulse noise). As shown in Fig. 7, simply applying self-supervised contrastive loss to the plain AT objective can substantially improve the performance in defending all types of perturbations. Moreover, adapting the supervised contrastive loss can boost the accuracy still higher. This observation demonstrates that the class-discrimination brought by the supervision indeed enhance the robustness

against not only the adversarial samples but also those common perturbations, making our full objective more appealing in real-world applications than its ablations.

4.3.8 Ablation on attack generation in pre-training

We now investigate the performance difference when different attack generation strategies are applied. The label-based strategy trains the model only with attacks that generated as the convention, i.e., by solving the maximization of the classification loss with PGD optimizer, while the label-free counterpart crafts the perturbation by maximizing the self-supervised contrastive loss without using any label.

As shown in Table 6, whereas the label-free strategy improves robustness accuracy over the conventional label-based generation by 2.84% and 4.06%, it degrades standard accuracy. That is not surprising, as contrastive learning considers no label information and provides no advantage in discrimination on standard classification. It is interesting that such label-free attacks boosts the robustness on $\epsilon = 16/255$ by a larger margin than the case where $\epsilon = 8/255$, which thanks to the more generalized representation that obtained by contrastive learning. Remarkably, when inject both the class-wise attacks and also the label-free generated attacks into our pretraining, we are able to achieve the best performance in terms of SA and TA simultaneously. We consider this to consist with another recent study that connecting the adversarial perturbation with data augmentation for performance gains (Tack et al., 2021), and we will leave it to our future work.

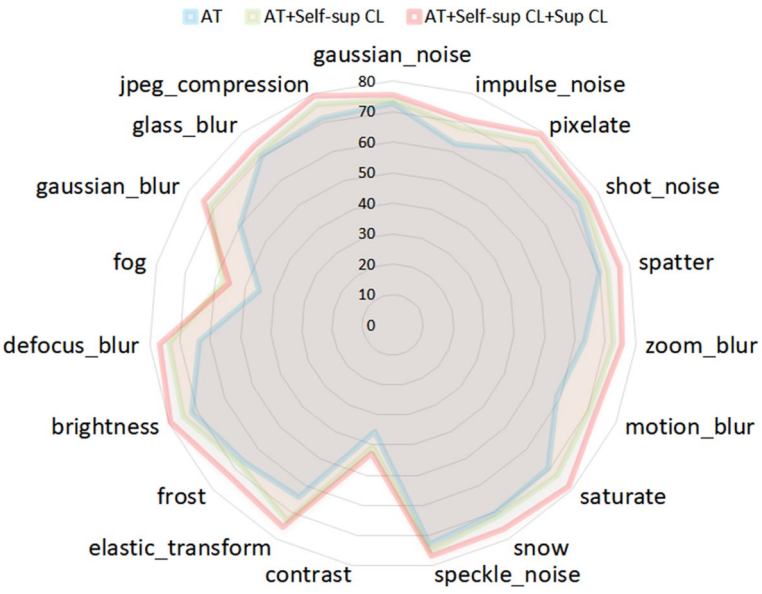


Fig. 7 Component analysis of the proposed method. Reported performance on each unforeseen corruption is measured on ResNet-18 trained with standard AT, standard AT with self-supervised contrastive loss, standard AT with both self-supervised and supervised contrastive loss, respectively

Table 6 Comparison of pretrained model with different attack generation schemes for ASCL

Attack generation choice	SA (%)	RA (%)	
		$\epsilon = 8/255$	$\epsilon = 16/255$
Label-based	81.5	48.09	17.75
Label-independent	80.12	50.93	21.81
Ensemble	**81.67**	**53.09**	**23.01**

Models are trained with l_∞ of $\epsilon = 8/255$ constraint. Results are measured by SA as well as RA under different sized l_∞ balls. The best results under each evaluation metrics are highlighted in bold

4.3.9 Ablation of defense approaches in fine-tuning strategy

Next, we analyze how the finetuning strategy affects the eventual robustness. Table 7 reports SA and RA that under attacks with the perturbation budget of 8/255 and 16/255, respectively. We begin by verifying the effectiveness of adversarial finetuning. Comparison between LSF and LAF, between FSF and FAF show that the use of adversarial finetuning is advantageous, as it yields consistent improvement in RA and/or SA either on the scenario of lightweight linear setting or full model finetuning.

In the next step, it is natural to compare the impacts of linear finetuning with full finetuning. As shown in Table 7, the utilization of full finetuning harms the robustness, leading to a significant drop by 6.07% in robust accuracy (FAF vs. LAF). The robustness even hits the ground (close to 0%) when full standard finetuning is applied. We conjecture that one serious problem when we using full fine-tuning is that most learned embedding has been wiped away, which will further present scalability challenges on downstream tasks. Interestingly, our experimental results somewhat show a different conclusion from the previous work (Chen et al., 2020; Fan et al., 2021), which conclude that the full adversarial finetuning is able to gain substantial robustness boost while only the standard finetuning in full mode weakens the robust representation learned in pretraining. On the whole, LAF offers better robustness (than standard finetuning) yet it consumes much less computation (than full model finetuning) in our observation, and thus it is adopted as our default finetuning scheme.

4.3.10 Batch normalization strategy

We further study the performance difference when different batch normalization (BN) strategies are applied, i.e., (1) using the same BN for both branches of the contrastive learning;

Table 7 Ablation results on finetuning strategies

Finetuning strategy		SA (%)	RA(%)	
			$\epsilon = 8/255$	$\epsilon = 16/255$
Linear	Standard fine-tuning (LSF)	80.85	50.45	32.2
	Adversarial fine-tuning (LAF)	**81.67**	**53.09**	**23.01**
Full	Standard fine-tuning (FSF)	92.06	0.21	0
	Adversarial fine-tuning (FAF)	85.4	47.02	15.8

The bold indicates the best results

Table 8 Performance comparison with different batch normalization strategies for ACL (Jiang et al., 2020) and ASCL

BN option	ACL		ASCL (ours)	
	RA (%)	SA (%)	RA (%)	SA (%)
Same BN	35.6	75.15	50.15	80.06
Separated BN	52.11	80.82	53.09	81.67

and (2) using separated BN for different branches. Table 8 shows the performance comparison of models with different BN options. Remarkably, the performance gap between two settings induced by the proposed method is much smaller than the baseline. We conjecture that since these BN strategies are applied in the contrastive pre-training while the performance is evaluated after the fine-tuning completed, a smaller gap might indicate that the proposed pre-training inherits strong discriminative capability from supervised contrastive learning. This also echos the observation that impressing performance can be achieved by the proposed robust pre-training with just a lightweight fine-tuning scheme followed.

5 Conclusion

In this work, we propose ASCL, which is an extension of contrastive learning that targets for adversarial robustness. We show that injecting supervision component into the contrastive-style robust pre-training is benefit for class-discrimination. We further show that the incorporation of more diverse adversarial attacks could offer complementary benefits. Comprehensive evaluations demonstrate that ASCL is able to bring consistent gains over state-of-the-art methods in diverse scenarios, e.g., white-box attacks, black-box attacks, and also natural corruptions. Moreover, ASCL shows impressive results in robust transfer learning. Hence, it may help building a more robust and secure learning system in practice.

Adversarial samples we consider in this paper are based on the extensively studied threat models, i.e., l_∞, and thus the resistance is also put up on them. Meanwhile, the deployed systems in real world are confronted with adversarial attacks from all sides, where we still far from complete robustness. Nevertheless, we hope that this work inspires others to consider extending advanced contrastive learning methods into adversarial defense scenarios. Potential future work includes the scalability of our proposed method to larger datasets and models, and a broader fusion of other self-supervised learning techniques.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by ZL, DY, MW, CJ and HY. The first draft of the manuscript was written by ZL and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LQ21F020006 and No. LY21F020003, Zhejiang Provincial Key Research and Development Program of China under Grant No. 2021C01164, and Innovative Project for High-level Talents and Overseas Students of Hangzhou City.

Availability of data and material We conduct experiments on public datasets which are available on the official website.

Code availability Our source code is not yet public available.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethics approval This paper is approved in ethics.

Consent to participate This paper is consented to participate.

Consent for publication This paper is consented for publication.

References

Alayrac, J., Uesato, J., Huang, P., Fawzi, A., Stanforth, R., & Kohli, P. (2019). Are labels required for improving adversarial robustness? In *Advances in neural information processing systems 32: Annual Conference on neural information processing systems 2019, NeurIPS 2019*, December 8–14, 2019, Vancouver, BC, Canada, pp. 12192–12202.

Andriushchenko, M., Croce, F., Flammarion, N., & Hein, M. (2020). Square attack: A query-efficient black-box adversarial attack via random search. In *Computer vision—ECCV 2020—16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XXIII. Lecture Notes in Computer Science* (Vol. 12368, pp. 484–501). Springer.

Arora, S., Hu, W., & Kothari, P.K. (2018). An analysis of the t-sne algorithm for data visualization. In *Conference on learning theory, COLT 2018, Stockholm, Sweden, 6–9 July 2018. Proceedings of machine learning research* (Vol. 75, pp. 1455–1462). PMLR.

Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th international conference on machine learning. Proceedings of machine learning research* (Vol. 80, pp. 274–283). PMLR.

Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. In *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, NeurIPS 2019*, December 8–14, 2019, Vancouver, BC, Canada, pp. 15509–15519.

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84,* 317–331.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic, N., Laskov, P., et al. (2013). Evasion attacks against machine learning at test time. In *Machine learning and knowledge discovery in databases—European conference, ECML PKDD 2013, Prague, Czech Republic, September 23–27, 2013, proceedings, part III* (Vol. 8190, pp. 387–402). Springer.

Boopathy, A., Liu, S., Zhang, G., Liu, C., Chen, P.-Y., Chang, S., et al. (2020). Proper network interpretability helps adversarial robustness in classification. In *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 1014–1023). PMLR.

Carlini, N., & Wagner, D. A. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security, AISec@ CCS 2017*, Dallas, TX, USA, November 3, 2017 (pp. 3–14). ACM.

Carlini, N. & Wagner, D.A. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy, SP 2017*, San Jose, CA, USA, May 22–26, 2017, pp. 39–57. IEEE Computer Society.

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B., & Tommasi, T. (2019). Domain generalization by solving jigsaw puzzles. In *IEEE conference on computer vision and pattern recognition, CVPR 2019*, Long Beach, CA, USA, June 16–20, 2019, pp. 2229–2238. Computer Vision Foundation/IEEE.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th international conference on machine learning. proceedings of machine learning research* (Vol. 119, pp. 1597–1607). PMLR.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. E. (2020). Big self-supervised models are strong semi-supervised learners. In *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, NeurIPS 2020*, December 6–12, 2020, Virtual.

Chen, T., Liu, S., Chang, S., Cheng, Y., Amini, L. & Wang, Z. (2020). Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Cissé, M., Bojanowski, P., Grave, E., Dauphin, Y. N., & Usunier, N. (2017). Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th international conference on machine learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017. Proceedings of machine learning research* (Vol. 70, pp. 854–863). PMLR.

Criminisi, A., Pérez, P., & Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing, 13*(9), 1200–1212.

Croce, F., & Hein, M. (2020a). Minimally distorted adversarial examples with a fast adaptive boundary attack. In *Proceedings of the 37th international conference on machine learning, ICML 2020, 13–18 July 2020, virtual event. Proceedings of machine learning research* (Vol. 119, pp. 2196–2205). PMLR.

Croce, F., & Hein, M. (2020b). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th international conference on machine learning. Proceedings of machine learning research* (Vol. 119, pp. 2206–2216). PMLR.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B., & Madry, A. (2019). Adversarial robustness as a prior for learned representations.

Fan, L., Liu, S., Chen, P.-Y., Zhang, G., & Gan, C. (2021). When does contrastive learning preserve adversarial robustness from pretraining to finetuning? In A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems*.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., & Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of DNNS. In *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, NeurIPS 2018*, December 3-8, 2018, Montréal, Canada, pp. 8803–8812.

Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. In *6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, conference track proceedings*. OpenReview.net.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems 27: Annual conference on neural information processing systems 2014*, December 8–13 2014, Montreal, Quebec, Canada, pp. 2672–2680.

Goodfellow, I.J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, conference track proceedings*.

Gowal, S., Qin, C., Uesato, J., Mann, T. A., & Kohli, P. (2020). Uncovering the limits of adversarial training against norm-bounded adversarial examples. CoRR abs/2010.03593 arXiv:2010.03593.

Guo, C., Rana, M., Cissé, M., & van der Maaten, L. (2018). Countering adversarial images using input transformations. In *6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, conference track proceedings*. OpenReview.net.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R.B. (2020). Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020*, Seattle, WA, USA, June 13–19, 2020 (pp. 9726–9735). Computer Vision Foundation/IEEE.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *International conference on learning representations*.

Hendrycks, D., Mazeika, M., Kadavath, S. & Song, D. (2019). Using self-supervised learning can improve model robustness and uncertainty. In *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.

Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., & Bengio, Y. (2019). Learning deep representations by mutual information estimation and maximization. In *7th international conference on learning representations, ICLR 2019*, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., et al. (2019). Learning deep representations by mutual information estimation and maximization. In *7th International Conference on learning representations, ICLR 2019*, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net.

Hénaff, O. J. (2020). Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th international conference on machine learning, ICML 2020, 13–18 July 2020, virtual event. Proceedings of machine learning research* (Vol. 119, pp. 4182–4192). PMLR.

Ishida, T., Yamane, I., Sakai, T., Niu, G., & Sugiyama, M. (2020). Do we need zero training loss after achieving zero training error? In *Proceedings of the 37th international conference on machine learning, ICML 2020, 13–18 July 2020, virtual event. Proceedings of machine learning research* (Vol. 119, pp. 4604–4614). PMLR.

Jiang, Z., Chen, T., Chen, T. & Wang, Z. (2020). Robust pre-training by adversarial contrastive learning. In *Advances in neural information processing systems* (Vol. 33, pp. 16199–16210).

Kang, D., Sun, Y., Hendrycks, D., Brown, T. & Steinhardt, J. (2020). Testing robustness against unforeseen adversaries.

Kannan, H., Kurakin, A., & Goodfellow, I. J. (2018). Adversarial logit pairing. CoRR abs/1803.06373 arXiv:1803.06373.

Kaur, S., Cohen, J., & Lipton, Z.C. (2019). Are perceptually-aligned gradients a general property of robust classifiers.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. In *5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, conference track proceedings*. OpenReview.net.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., et al. (2020). Supervised contrastive learning. In *Advances in neural information processing systems* (Vol. 33, pp. 18661–18673). Curran Associates, Inc.

Kim, M., Tack, J., & Hwang, S. J. (2020). Adversarial self-supervised contrastive learning. In *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, NeurIPS 2020*, December 6–12, 2020.

Krizhevsky, A., et al. (2009). Learning multiple layers of features from tiny images.

Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, NeurIPS 2018*, December 3–8, 2018, Montréal, Canada, pp. 7167–7177.

Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. (2018) Visualizing the loss landscape of neural nets. In *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018*, NeurIPS 2018, December 3–8, 2018, Montréal, Canada, pp. 6391–6401.

Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., & Zhu, J. (2018). Defense against adversarial attacks using high-level representation guided denoiser. In *2018 IEEE conference on computer vision and pattern recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18–22, 2018, pp. 1778–1787. Computer Vision Foundation/IEEE Computer Society.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S. N. R., Schoenebeck, G., et al. (2018) Characterizing adversarial subspaces using local intrinsic dimensionality. In *6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, conference track proceedings*. OpenReview.net.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International conference on learning representations*.

Mahendran, A., & Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural preimages. *International Journal of Computer Vision, 120*(3), 233–255.

Mao, C., Zhong, Z., Yang, J., Vondrick, C., & Ray, B. (2019). Metric learning for adversarial robustness. In *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, NeurIPS 2019*, December 8–14, 2019, Vancouver, BC, Canada, pp. 478–489.

Mirjalili, V., & Ross, A. (2017). Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In *2017 IEEE international joint conference on biometrics, IJCB 2017*, Denver, CO, USA, October 1–4, 2017 (pp. 564–573). IEEE.

Misra, I., & van der Maaten, L. (2020). Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020*, Seattle, WA, USA, June 13–19, 2020, pp. 6706–6716. Computer Vision Foundation/IEEE.

Noroozi, M., & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer vision—ECCV 2016—14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, proceedings, part VI. Lecture Notes in Computer Science* (Vol. 9910, pp. 69–84). Springer.

Papernot, N., McDaniel, P. D., & Goodfellow, I. J. (2016) Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. CoRR abs/1605.07277, arXiv:1605.07277.

Papernot, N., McDaniel, P. D., Goodfellow, I. J., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security, AsiaCCS 2017*, Abu Dhabi, United Arab Emirates, April 2–6, 2017 (pp. 506–519). ACM.

Pei, K., Cao, Y., Yang, J., & Jana, S. (2019). Deepxplore: Automated whitebox testing of deep learning systems. *Communications of the ACM, 62*(11), 137–145.

Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., et al. (2019). Adversarial robustness through local linearization. In *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, NeurIPS 2019*, December 8–14, 2019, Vancouver, BC, Canada, pp. 13824–13833.

Rauber, P. E., Falcão, A. X., & Telea, A. C. (2016) Visualizing time-dependent data using dynamic t-sne. In *18th Eurographics conference on visualization, EuroVis 2016—Short papers*, Groningen, The Netherlands, June 6–10, 2016, pp. 73–77.

Rebuffi, S., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., & Mann, T. A. (2021). Fixing data augmentation to improve adversarial robustness. CoRR abs/2103.01946, arXiv:2103.01946.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., & Madry, A. (2018). Adversarially robust generalization requires more data. In *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, NeurIPS 2018*, December 3–8, 2018, Montréal, Canada, pp. 5019–5031.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., et al. (2014) Intriguing properties of neural networks. In *2nd international conference on learning representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, conference track proceedings*.

Tack, J., Yu, S., Jeong, J., Kim, M., Hwang, S. J., & Shin, J. (2021). Consistency regularization for adversarial robustness. In *ICML 2021 workshop on adversarial machine learning*.

Tian, Y., Krishnan, D., & Isola, P. (2020a). Contrastive multiview coding. In *Computer vision—ECCV 2020—16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XI. Lecture Notes in Computer Science* (Vol. 12356, pp. 776–794). Springer.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., & Isola, P. (2020b). What makes for good views for contrastive learning? In *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, NeurIPS 2020*, December 6–12, 2020, Virtual.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I. J., Boneh, D., & McDaniel, P. D. (2018) Ensemble adversarial training: Attacks and defenses. In *6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, conference track proceedings*. OpenReview.net.

Trinh, T. H., Luong, M.-T., & Le, Q. V. (2019). Selfie: Self-supervised pretraining for image embedding.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A. & Madry, A. (2019). Robustness may be at odds with accuracy. In *International conference on learning representations*.

Uesato, J., O'Donoghue, B., Kohli, P., & van den Oord, A. (2018) Adversarial risk and the dangers of evaluating against weak attacks. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning. Proceedings of machine learning research* (Vol. 80, pp. 5025–5034). PMLR.

Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., & Gu, Q. (2019). On the convergence and robustness of adversarial training. In *Proceedings of the 36th international conference on machine learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA. Proceedings of machine learning research* (Vol. 97, pp. 6586–6595). PMLR.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., & Gu, Q. (2020). Improving adversarial robustness requires revisiting misclassified examples. In *8th international conference on learning representations, ICLR 2020*, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net.

Wei, C., Kakade, S. M. & Ma, T. (2020). The implicit and explicit regularization effects of dropout. In *Proceedings of the 37th international conference on machine learning, ICML 2020, 13–18 July 2020, virtual event. Proceedings of machine learning research* (Vol. 119, pp. 10181–10192). PMLR.

Wu, D., Xia, S., & Wang, Y. (2020). Adversarial weight perturbation helps robust generalization. In *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, NeurIPS 2020*, December 6–12, 2020, Virtual.

Wu, Z., Xiong, Y., Yu, S. X. & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE conference on computer vision and pattern recognition,*

CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, pp. 3733–3742. Computer Vision Foundation/IEEE Computer Society.

Xie, C., & Yuille, A. L. (2020). Intriguing properties of adversarial training at scale. In *8th international conference on learning representations, ICLR 2020*, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net.

Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A. L., & Le, Q. V. (2020). Adversarial examples improve image recognition. In *2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020*, Seattle, WA, USA, June 13–19, 2020, pp. 816–825. Computer Vision Foundation/IEEE.

Yan, Z., Guo, Y., & Zhang, C. (2018). Deep defense: Training DNNS with improved adversarial robustness. In *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, NeurIPS 2018*, December 3–8, 2018, Montréal, Canada, pp. 417–426.

Zhang, H., Yu, Y., Jiao, J., Xing, E., Ghaoui, L. E., & Jordan, M. I. (2019). Theoretically principled trade-off between robustness and accuracy. In *ICML* (pp. 7472–7482).

van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. CoRR abs/1807.03748, arXiv:1807.03748.