Self-Supervised Adversarial Training via Diverse Augmented Queries and Self-Supervised Double Perturbation

Ruize Zhang
Institute of Computing Technology,
Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
zhangruize21b@ict.ac.cn

Sheng Tang*
Institute of Computing Technology,
Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
ts@ict.ac.cn

Juan Cao
Institute of Computing Technology,
Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
caojuan@ict.ac.cn

Abstract

Recently, there have been some works studying self-supervised adversarial training, a learning paradigm that learns robust features without labels. While those works have narrowed the performance gap between self-supervised adversarial training (SAT) and supervised adversarial training (supervised AT), a well-established formulation of SAT and its connections with supervised AT are under-explored. Based on a simple SAT benchmark, we find that SAT still faces the problem of large robust generalization gap and degradation on natural samples. We hypothesize this is due to the lack of data complexity and model regularization and propose a method named as DAQ-SDP (Diverse Augmented Queries Self-supervised Double Perturbation). We first challenge the previous conclusion that complex data augmentations degrade robustness in SAT by using diversely augmented samples as queries to guide adversarial training. Inspired by previous works in supervised AT, we then incorporate a self-supervised double perturbation scheme to self-supervised learning (SSL), which promotes robustness transferable to downstream classification. Our work can be seamlessly combined with models pretrained by different SSL frameworks without revising the learning objectives and helps to bridge the gap between SAT and AT. Our method also improves both robust and natural accuracies across different SSL frameworks. Our code is available at https://github.com/rzzhang222/DAQ-SDP.

1 Introduction

Deep neural network has shown its power in various machine learning tasks. In spite of its beneficial properties in optimization and generalization, deep neural network is vulnerable to adversarial attack as samples with carefully designed tiny perturbations may cause significant deviations of model

*Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1389

predictions [16, 25, 12, 36]. One of the most successful defenses tackling this problem is adversarial training, which generates perturbations that makes the largest output deviation and trains the model with perturbed samples [40, 25]. Many following methods [21, 10, 39, 32, 40] have further developed more advanced adversarial training techniques based on this adversarial training framework.

The adversarial defenses mentioned above require full supervision. However, in real situations full labels may not be available. For the semi-supervised setting, some works [26, 3, 2] suggested that unlabeled data can be useful in improving model robustness by designing auxiliary pseudo label losses. However, the performance is largely affected by the amount of available labels. Later, some works studied adversarial robustness in self-supervised learning (SAT). In this scenario, adversarial training (AT) is integrated with self-supervised learning (SSL) to get robust features that can be efficiently finetuned [22, 20, 14, 38, 18, 24, 37, 23]. Some works [22, 20, 14, 24, 37] combined contrastive learning (CL) with AT to tackle this task. In this paper we term those works as ACL methods.

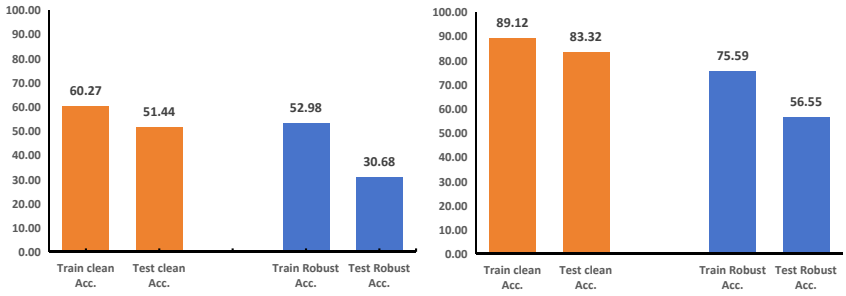

Figure 1: **Motivation:** Large robust generalization gap and reduced clean accuracy for SAT (SimCLR) on CIFAR-100 and CIFAR-10. The left part of the figure is the results on CIFAR-100 and the right part is for CIFAR-10. The robust generalization gap is over 20% on both datasets.

The ACL methods above were restricted to the CL framework and could not generalize to other SSL methods. Previously, Zhang et al. [38] proposed to disentangle the task of SSL and AT into two stages of learning, which first trains SSL models with natural data and then enables AT under the pseudo-supervision of the naturally trained SSL features. This learning framework brings consistent performance boost for both contrastive and non-contrastive methods. Moreover, the feature pseudo-supervision in this framework takes a form similar to the supervision in supervised and semi-supervised AT, thus providing a unified perspective to analyze AT paradigms under different supervision. In this paper, we regard this learning framework as a simple benchmark for SAT.

While previous works on ACL and SAT have achieved remarkable performance that is comparable to supervised AT, there is neither a well-established formulation nor an analysis of the general learning process as in the supervised counterpart. To better understand the limitations of this learning paradigm, we start from the SAT benchmark proposed by Zhang et al. [38], which advocates an SAT process that is much more efficient than previous ACL works (100 vs. 1000 epoches of adversarial training). As shown in Figure 1, we record finetuned train-set and test-set accuracies after SAT pretraining on CIFAR-10 with ResNet-34 as backbone and find that there is a robust generalization gap of around 19% and clean accuracy gap of around 6%. The results are similar for CIFAR-100, which gives a robust generalization gap of around 22% and clean accuracy gap of around 9%. The results show that there exists a large robust generalization gap and natural performance degradation, similar to the AT counterpart.

Given that both SSL and AT are hard tasks [38], we hypothesize these problems to be caused by insufficient data complexity and model regularization. In order to solve the problems, we propose a method termed as DAQ-SDP (Diverse Augmented Queries Self-supervised Double Perturbation) and handle the problems from two aspects. First, while previous papers [24, 38] concluded that complex augmentation techniques are crucial for SSL but harmful for SAT robustness, we argue that strong and diverse augmentations could help SAT if used properly. Specifically, we find that every training distribution is worth one set of BatchNorm layers in SAT and propose an Augmentation-Adversary (Aug-Adv) Pairwise-BatchNorm adversarial training method. It turns out that naturally trained SSL models with a single set of batch norm layers can provide effective guidance for multi-branch adversarial training. Second, while previous works on SAT focused on designing sample adversarial

perturbation for SSL, we find that self-supervised model perturbation also contributes to downstream robustness with a proper training scheme. Despite the task mismatch from pretraining to finetuning, there exists a cross-task transferability of robust generalization. Our work finds a general method that can be directly applied to different SSL frameworks without revising learning objectives, providing insights that contribute to the understanding of SAT.

As far as we know, we are the first to analyze SAT as a general learning paradigm in an SSL-framework-agnostic way and solves its problems by revealing its traits related to clean performance degradation and robust generalization. The main contributions are summarized as follows:

- In contrast to the conclusion of previous works [38, 24], we suggest that strong and diverse augmentations can boost self-supervised robustness and propose an Aug-Adv Pairwise-BatchNorm technique for better robust generalization and less natural degradation.
- Different from previous works [22, 20, 14, 38, 18, 24, 37, 23] that focused on introducing sample adversarial perturbation from supervised AT to SSL, we use model perturbation in SSL pre-training to boost downstream robustness. We then propose a self-supervised double perturbation scheme in the later stage of SAT to improve robust generalization without affecting the learning of natural features.
- We conduct experiments on CIFAR-10 and CIFAR-100, the commonly used datasets in previous works. On CIFAR-100, our proposed method improves over 2% on AutoAttack [7] and clean data results with ResNet-34. On CIFAR-10, our method improves over 1% on AutoAttack [7] and over 2% on clean data results with ResNet-34. The experimental results demonstrate the effectiveness of our method across SSL frameworks, models and datasets.

2 Related Works

The task of SAT integrates AT into SSL and aims at learning robust feature representations that enable efficient finetuning. Here we will give a brief summary of previous works related to this field.

Supervised Adversarial Training (supervised AT) Given deep neural network's vulnerability to adversarial perturbations, many defense methods have been proposed. Among them adversarial training, originally proposed by Goodfellow et al. [16], has become a prevalent method. Adversarial training simulates a min-max process that finds the perturbation with largest distortion and then minimize the training loss over the adversarial data. Madry et al. [25] proposed a representative adversarial training method and uses random initialized multi-step projected gradient descent to generate adversarial samples. Many following works [21, 10, 39, 32, 40] further revised the adversarial learning framework and applied more advanced techniques such as logits pairing, boundary guidance and consistency regularization for improving robustness while keeping the accuracy on clean data.

Self-Supervised Learning (SSL) Self-supervised learning is the task of learning feature representations with no label available. In this case, a discriminative self-supervised method generally relies on a pretext task for pretraining to learn useful information. Previously, many pretext methods have been proposed [15, 11, 28]. One of the most successful pretext task is instance discrimination [6, 4]. Contrastive learning defines a instance discrimination task and learns the features using similarities between positive and negative pairs. The recent development from contrastive learning methods MoCo [6] and SimCLR [4] to non-contrastive methods BYOL [17] and SimSiam [5] has revised the contrastive loss to a positive-pair loss and further simplified the framework. Moreover, some works [29, 13, 19] combined the contrastive framework with some more advanced techniques including positive and negative pair mining, prototypes and customized contrastive view crafting to enrich the information learned by the contrastive framework for effective and efficient learning.

Self-Supervised Adversarial Training (SAT) The development of self-supervised learning provides a new direction for acquiring robust features. In self-supervised learning, the emergence of instance discrimination as the new state of art self-supervised pretext task provides a natural setting for adversarial robustness to fit in. While constructing reliable decision boundary using ground truth labels is not feasible, previous works [22, 20, 14, 24, 37] proposed to exploit contrastive loss adversarial training for promoting robustness. The assumption is that if the feature space near the data sample is smooth enough, the feature prediction of adversarially perturbed data will be consistent with its clean counterpart. Gupta et al. [18] suggested that contrastive learning has intrinsic sensitivity to adversarial perturbations and proposed a simple method to remove false negative pairs. This

https://doi.org/10.52202/079017-1389

method enhances robustness in SSL without revising AT and is orthogonal to ours. Xu et al. [37] proposed to use causal reasoning in ACL and used adversarial invariance regularization to enhance ACL. While achieving an impressive performance, the methods above were still restricted to the contrastive learning framework.

Later, Some works have started to explore robustness in the broader SSL picture. Zhang et al. [38] formulated SAT as a two-stage framework that first trains an SSL model and then uses the learned features as guidance for AT. This work has set a strong baseline for this task and can be directly generalized to different SSL frameworks. Thus we take this two-stage SAT framework as the baseline to explore robustness in the broader picture of SSL. Kim et al. [23] proposed an interesting idea that carefully crafted targeted adversarial perturbations can help enhancing robustness for non-contrastive SSL methods. However, the improvements on contrastive frameworks are not as consistent as in the non-contrastive case. Moreover, some works [38, 24] suggested that complex augmentations are crucial for SSL but destructive for SAT. In this work, we approach the task of SAT by studying its learning process and drawing an analogy to supervised AT, with the goal of acquiring a better understanding of the difference and similarity between these two learning paradigms. We then propose a method to tackle the potential problems in SAT. In the following sections, we will introduce the problem statement and then describe our motivation and method.

3 Preliminary

Self-Supervised Learning As the most representative contrastive learning framework, SimCLR [4] uses data in the same batch as negative pairs and optimizes the following objective:

$$\ell_{CL}(\tau_1(x), \tau_2(x)) = -\log\left(\frac{\exp(\text{sim}(z_i, z_j)/t)}{\exp(\text{sim}(z_i, z_j)/t) + \sum_{k\neq i}^{\text{N}} \exp(\text{sim}(z_i, z_k)/t)}\right). \tag{1}$$

In the equation above,$\tau_1(x), \tau_2(x)$ are two augmented views of the same image. $z_i = g \cdot f(\tau_i(x))$ is the projected feature of the corresponding view. z_i and z_j are a positive pair. N is the number of negative samples.

Contrastive SSL relies on large batch size or extra maintained queue for negative pairs and can be computationally expensive. In contrast, positive-only frameworks only include positive pairs. The learning objective of SimSiam [5], a representative positive-only method, can be formulated as:

$$\ell_{ss}(\tau_1(x), \tau_2(x)) = -\frac{1}{2}\frac{p_1 \cdot stopgrad(z_2)}{\|p_1\|_2\|z_2\|_2} - \frac{1}{2}\frac{p_2 \cdot stopgrad(z_1)}{\|p_2\|_2\|z_1\|_2}. \tag{2}$$

In the equation above, $\tau_1(x), \tau_2(x)$ are two augmented views of the same image. $z_i = g \cdot f(\tau_i(x))$ and $p_i = h \cdot z_i$ are the projected and predicted feature of the corresponding view, in which g is a projector helps to preserve instance dicriminative features and h is a predictor helps to prevent model callapse.

Adversarial Contrastive Learning Based on the framework of contrastive learning (CL), multiple previous works have proposed adversarial contrastive learning (ACL) methods, which aims at improving the robustness of the learned features. Pevious works have adopted such a learning framework but each had some revision of the loss term [22, 20, 14]. In general, those methods can be formulated as:

$$\ell_{CL}^{\text{adv}} = \ell_{CL}(\tau_1(x), \tau_2(x), x^{\text{adv}}), \tag{3}$$

where

$$x^{\text{adv}} = x + \arg\max_{\delta} \ell_{CL}(\tau_1(x), \tau_2(x), x + \delta). \tag{4}$$

In the equations above, ℓ_{CL}^{adv} is the adversarial contrastive loss with an extra variable for the adversarial view. It is often calculated as the average of the pairwise contrastive loss [14]. x^{adv} is the adversarial sample generated with this loss.

Self-Supervised Adversarial Training To find a general method that can improve robustness for different SSL frameworks, we start from a basic SAT framework [38]:

$$\ell_{\text{stage1}} = \ell_{SSL}, \tag{5}$$

and

$$\ell_{\text{stage2}} = \text{Sim}(f_2(x), f_1(x)) + \lambda \cdot \text{Sim}(f_2(x^{\text{adv}}), f_2(x)), \tag{6}$$

where

$$x^{\text{adv}} = x + \arg\max_{\delta}(-\text{Sim}(f_2(x+\delta), f_1(x))). \tag{7}$$

This SAT framework separates the adversarial training process into two stages. In the first stage, an SSL model f_1 is trained with clean data. Then the features predicted by the clean model are used as pseudo-supervision for adversarial training of f_2 in the second stage. The clean samples providing guiding features can be regarded as queries that help to distill useful information from the clean model. Compared to ACL methods mentioned above, this type of method is more general and also more computationally efficient. In our work, we adopt this framework as the baseline.

4 Method

In this paper, we aim at finding a method that can solve the problem of robust generalization and clean accuracy degradation in SAT. Note that our work proposes a general method that can be directly combined with different pretrained SSL models for improvements instead of requiring adversarially re-training the models from scratch. Figure 2 demonstrates the overall framework of our method. In the following sections, we will introduce each part of our method.

4.1 Diverse Augmented Query

Previously in supervised AT, some works have discussed about the effects of complex augmentation strategies on robustness [31, 30, 1]. The idea is to fit the model to the labels on the generalized sample distributions to reduce overfitting. However, in SAT, there is no ground truth label to provide supervision on training or generalized data distributions. We hypothesize that SSL models trained with natural data, especially instance discrimination based ones, already contain certain level of generalization capability as the feature space is learned with large amount of data under strong and complex augmentations. However, this capability of generalization can be lost during AT. Thus we propose to use the diversely augmented clean features as supervision for AT.

In the field of ACL and SAT, previous works [38, 24] have concluded that strong and diverse augmentations are essential for SSL but harmful for robustness. Thus Luo et al. [24] proposed to gradually reduce the strength of augmentation during the training and Zhang et al. [38] proposed to only use random resized crop and horizontal flip in adversarial training. In this paper, we argue that diversely augmented samples help SAT given that the model has sufficient capacity, as it's actually essential for the robust model to distill rich information to keep the capability of generalization.

Note that data augmentation strategies including AutoAugment [9] and RandAugment [8] require labels to calculate validation accuracy in the process of searching for optimal augmentation policies, thus can not be directly applied to SAT. In this paper, we propose to use TrivialAugment [27], which is a dataset-independent and search-free method that randomly samples the augmentation policy and strength. Since we need to improve the generalization while also specialize on the testing distribution, we need to fit the adversarial model with clean model features both on the diversely augmented and basic augmented distributions. Inspired by previous work in supervised AT [1], we adopt a multi-stream structure that takes samples from different distributions as inputs. In the previous ACL and SAT works, there were no clear conclusion on the usage of BatchNorm layers. While some works [22, 14] suggested to use different BatchNorm layers for adversarial sample, Zhang et al. [38] suggested that separate BatchNorm layers are not necessary. In this work, we find that each training distribution is worth one set of BatchNorm parameters in SAT and even if the clean model only contains one set of BatchNorm layers, it can provide effective guidance for the multi-BatchNorm adversarial model. Specifically, we define four different input streams based on the combination of their adversarial type (adversarial vs. natural) and augmentation type (strong vs. weak), where the basic augmentation contains random resized crop and horizontal flip while diverse augmentation contains Trivial Augmentation [27] and basic augmentation. In the first stage

 https://doi.org/10.52202/079017-1389

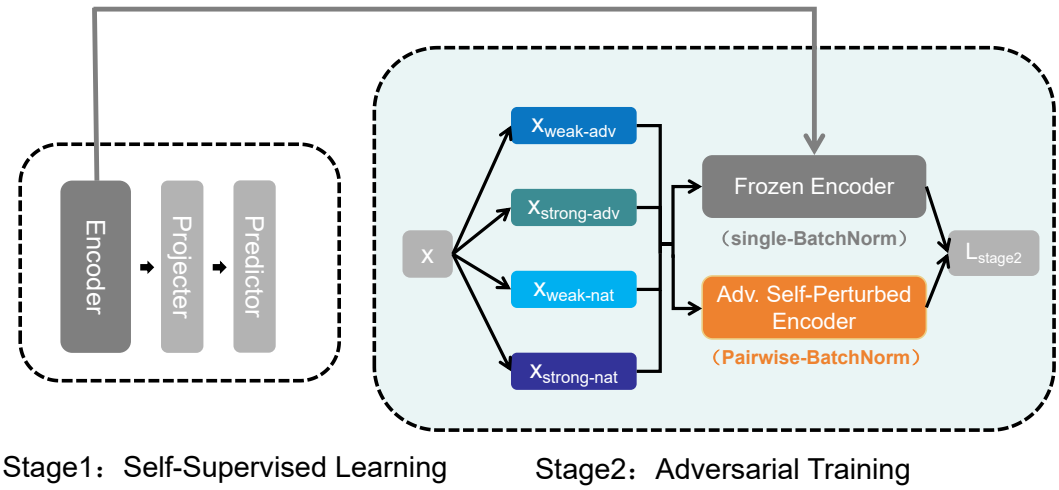

Figure 2: A demonstration of our proposed DAQ-SDP. The single BatchNorm encoder of pretrained model is extracted as teacher for our Pairwise-BatchNorm robust encoder for adversarial training.

of our method, we can either use a single set of BatchNorm to train a clean model or directly use a pretrained SSL model with no consideration for robustness. Then in the second stage, we propose an Aug-Adv pairwise-BatchNorm strategy for adversarial training, where each stream of features is pseudo-supervised by the features predicted by the clean model. After training, we only keep the basic-adv BatchNorm layers.

The formulation of our method is:

$$\ell_{diverse-aug} = \ell_{clean} + \lambda \cdot \ell_{adv}, \tag{8}$$

where

$$\ell_{clean} = \sum_{aug_i} \text{Sim}(f_{2-aug_i-clean}(x_{\text{aug}_i}), f_1(x_{\text{aug}_i})), \tag{9}$$

and

$$\ell_{adv} = \text{Sim}(f_{2-aug_i-adv}(x^{\text{adv}}_{\text{aug}_i}), f_{2-aug_i-clean}(x_{\text{aug}_i})), \tag{10}$$

and

$$x^{\text{adv}}_{\text{aug}_i} = x + \arg\max_{\delta}(-\text{Sim}(f_2(x^{\text{adv}}_{\text{aug}_i}), f_1(x_{\text{aug}_i}))). \tag{11}$$

In the equations above, $\text{Sim}(.,.)$ is the cosine similarity between two features and aug_i corresponds to the augmentation type mentioned above. f_{2-aug_i-adv} and $f_{2-aug_i-clean}$ are the student model with the corresponding pairwise-BatchNorm layers. Our method forces the pseudo-supervised adversarial model to inherit the generalization capability from clean model with diversely augmented queries. The AugAdv pairwise-BN helps features of each training distribution to fit the clean feature counterparts without interfering with each other. The experimental results demonstrate that diverse and complex augmentations can improve SAT robustness. This finding helps to narrow the gap between improving SAT and supervised AT.

4.2 Adversarial Self-Perturbed Weight

Previous works in ACL and SAT have borrowed the idea of adversarial sample perturbation from supervised AT to SAT and made revisions either on the specific training loss term [20, 22] or the effective way of generating sample perturbations [14, 23]. However, whether more advanced ideas in supervised AT can bring improvements in SAT is under-explored. In this section, we suggest that adversarial weight perturbation can be introduced into SSL pretext task for downstream robustness. Note that in the method proposed by Wu et al. [35], adversarial weight perturbation is applied to different supervised AT frameworks including vanilla PGD [25], TRADES [39], RST [3] and MART [34]. However, all those methods require full or partial labels and are based on classification loss.

The adversarial weight perturbation can be expressed as:

$$L_{\text{AWP}} = \max_{\hat{\theta} \in \mu(\theta)} L_{\text{CE}}(f_{\hat{\theta}}(x, y)) + \beta L_{\text{adv}}(f_{\hat{\theta}}(x^{\text{adv}}, y)), \tag{12}$$

where μ is the perturbation size of the model weight.

In contrast, in our work the weight perturbation is introduced into the SSL pretext task. Thus the perturbation doesn't regularize the weight classification-loss landscape, but works on the feature similarity loss in a label-free paradigm instead. This transition of learning paradigm makes it interesting to see whether such perturbations respect to the SSL objective can benefit downstream robust generalization. Specifically, the self-supervised weight adversarial perturbation perturbs the model weight in the direction of enlarging the self-supervised cosine similarity loss and increases the smoothness of this similarity loss landscape. The weight perturbation can be formulated as:

$$\hat{\theta}_2 = \arg\min_{\theta_2 \in \mu(\theta)} \text{Sim}(f_2(x_{\text{aug-weak}}), f_1(x_{\text{aug-weak}})) + \lambda \cdot \text{Sim}(f_2(x^{\text{adv}}_{\text{aug-weak}}), f_2(x_{\text{aug-weak}})). \tag{13}$$

The weight perturbation finds the "worst" adversarial scenario which is beneficial for model robustness. However, this extra adversarial component also further increases the difficulty of the task. In supervised AT, the existence of ground truth labels helps the model to converge despite the enlarged difficulty. However, both AT and SSL are difficult tasks. In the early stage of SAT, regulating the model to learn this rather difficult objective with respect to insufficiently learned adversarial features could impede the learning of natural features and we propose to apply this weight adversarial perturbation only in the later stage of learning when the pseudo-supervised learning of clean and adversarial features is stabilized. Without our weight self-perturbation scheme, there is a clean accuracy drop of 0.7% and PGD robust accuracy drop of 0.8% on CIFAR-10 with ResNet-34.

The adversarial weight perturbation calculated on weakly augmented data is combined with sample adversarial perturbations over the four sample distributions in the previous section. The overall learning objective is:

$$\ell_{swap-diverse-aug} = \sum_{aug_i} \ell_{clean} + \lambda \cdot \ell_{adv}, \tag{14}$$

where

$$\ell_{clean} = \text{Sim}(f_{\hat{\theta}_2 - aug_i - clean}(x_{aug_i}), f_{\theta_1}(x_{aug_i})), \tag{15}$$

and

$$\ell_{adv} = \text{Sim}(f_{\hat{\theta}_2 - aug_i - adv}(x^{\text{adv}}_{\text{aug}_i}), f_{\hat{\theta}_2 - aug_i - clean}(x_{aug_i})). \tag{16}$$

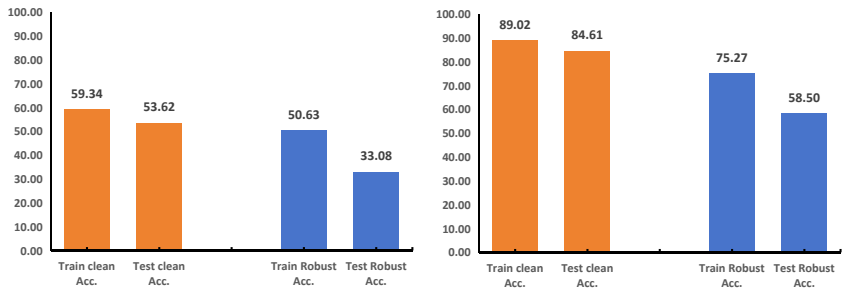

Figure 3: The generalization gap of SAT (SimCLR) with our proposed DAQ-SDP. The left part is the result on CIFAR-100. The right part is the result on CIFAR-10.

As shown in Figures 1 and 3, the robust generalization gap is reduced by around 3% and the test clean accuracy improves by more than 1.5% on average with our proposed method, which means properly regulating the smoothness of weight SSL-loss landscape in pre-training can improve the robust generalization of downstream classification despite the lack of labels.

4.3 Towards Unified Understanding of SAT and supervised AT

Despite the task difference between the learning paradigms, our method steps forward to an unified understanding of SAT and supervised AT by revealing their similar characteristics with respect to

https://doi.org/10.52202/079017-1389

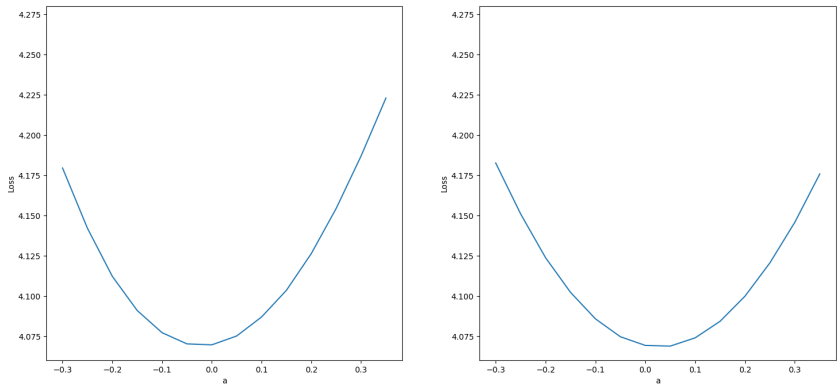

Figure 4: 1D visualization of the downstream weight loss landscape. The plot on the left is for the baseline method and the plot on the right shows that for our method with self-perturbed weight. The x-axis represents the magnitude to move the model weight.

Table 1: Results on ResNet-34 trained on CIFAR-10 with SimCLR framework.

Evaluation	Method	Clean	PGD	AutoAttack
Simple Linear Finetuning	DynACL[24]+AIR[37]	79.79	51.07	47.61
	TARO[23]	84.23	53.36	45.68
	DecoupledACL[38]	82.46	56.86	47.99
	DAQ-SDP (ours)	**84.57**	**58.57**	**49.22**

generalization. In this work, We also provide a “higher" level of perspective than previous narrower methods that focus on SAT with single SSL framework. We look forward to seeing future works with general approaches that improve adversarial training across different supervision settings.

5 Experiment

In this section, we demonstrate the effectiveness of our method. First, we evaluate the effectiveness when DAQ-SDP is plugged into contrastive and positive-pair only SSL frameworks. Then we conduct an ablation study for each part of our method. We also visualize the representation learned by our method through t-SNE [33] in Appendix.

Table 2: Results on ResNet-34 trained on CIFAR-100 with SimCLR framework.

Evaluation	Method	Clean	PGD	AutoAttack
Simple Linear Finetuning	DynACL[24]+AIR[37]	47.02	23.91	20.66
	TARO[23]	51.28	29.46	21.14
	DecoupledACL[38]	51.44	30.68	21.31
	DAQ-SDP (ours)	**53.54**	**33.09**	**23.42**

Experimental Setup: We apply our method on top of SimCLR [4], SimSiam [5] and BYOL [17] to evaluate the performance improvements across SSL frameworks. We also conduct extensive experiments on ResNet-18, ResNet-34 and ResNet-50. We find that the improvements are less significant on the smaller model ResNet-18. The rationale may be that simply increasing data complexity for ResNet-18 by using one extra strongly-augmented view could decrease the training set clean accuracy from 54.45% to 52.78% on CIFAR-100, suggesting insufficient model capacity to distill richer information while fitting well on clean data. This is understandable as model capacity takes a significant role in adversarial robustness and many techniques in supervised AT [30, 1, 35] require models with sufficient capacity to be effective. Given the complexity of AT and SSL, we expect a model larger than ResNet-18 is needed to learn rich information from the training data.

All SSL models in our method are first trained with clean data for 1000 epochs, then adversarially trained with 5-step PGD attack with the epsilon size of 8/255. Methods in previous works are adversarially trained for 1000 epoches as mentioned in their papers. The robustness is evaluated with

Table 3: Results on other SSL frameworks with ResNet-34 backbone. Note that most previous works are based on SimCLR and can not be used in Positive-Pair only SSL frameworks. The dataset is CIFAR-10.

SSL Framework	Method	Clean	PGD	AutoAttack
SimSiam	TARO[23]	**81.71**	52.61	44.46
SimSiam	DecoupledACL[38]	78.40	57.17	47.20
SimSiam	DAQ-SDP (ours)	80.42	**58.53**	**47.69**
BYOL	TARO [23]	**86.84**	52.01	44.76
BYOL	DecoupledACL[38]	83.15	55.22	47.67
BYOL	DAQ-SDP (ours)	85.91	**56.53**	**49.07**

Table 4: Results on ResNet-18 trained on CIFAR-10 with SimCLR framework.

Method	Clean	PGD	AA
DynACL[24]+AIR[37]	78.08	49.12	45.17
TARO[23]	**82.86**	52.44	43.99
DecoupledACL[38]	80.17	53.95	**45.31**
DAQ-SDP(ours)	81.76	**55.15**	45.12

AutoAttack [7] and PGD attack with 20 iterations and epsilon size of 8/255. λ is set to 2. We use double adversarial perturbation after 60 epochs of training and weight perturbation size constraint of 0.002. The SLF and AFF finetuning details are the same as previous works [14, 38] with 25 steps of training and initial learning rate of 0.1. The experimental results in our method is the average of 5 runs, with a maximal variation range of ± 0.5 for clean accuracy and ± 0.35 for robust accuracy. All experiments are conducted on 2 RTX 3090 GPUs.

5.1 Effectiveness of DAQ-SDP across SSL Frameworks, Models and Datasets

a) We first conduct experiments with different SSL frameworks on ResNet-34. As shown in Table 1 and Table 2, our method contributes to significant improvements on both clean and robust accuracy on both CIFAR-100 and CIFAR-10 with SimCLR [4]. Note that TARO [23] is an adversarial sample generation method that needs to be combined with specific SAT baselines. In this work we combine it with the same baseline framework we use for better performance and fair comparison. Although TARO [23] gives slightly better clean accuracy on CIFAR-10, our method outperforms TARO [23] on robust accuracy by a large margin. On CIFAR-100, both our clean and robust accuracy outperforms TARO [23]. Also note that methods except TARO [23], DecoupledACL [38] and ours are contrastive based methods and don't generalize to positive-pair only SSL frameworks. In Table 3, we compare our method with previous works on SimSiam [5] and BYOL [17]. As shown in the results, our work provides a consistent improvements for different SSL frameworks. This is because we treat differently trained models as the teacher that provides supervision for the clean data feature space. Once we obtain the supervision, we take an SSL-framework agnostic process for robustness improvements.

b) We then conduct experiments on ResNet-18 and ResNet-50. From Table 4 and Table 5, our method shows improvement across model sizes.

c) We also provide cross-dataset transfer learning results. Table 6 shows transfer learning from CIFAR-100 to CIFAR-10, in which our work outperforms other SAT methods. In Table 7, we provide the transfer learning SLF results from CIFAR-10 to STL-10, which shows that our method can transfer well across datasets from more different domains.

Table 5: Results on ResNet-50 trained on CIFAR-10 with SimCLR framework.

Method	Clean	PGD	AA
DynACL[24]+AIR[37]	80.67	/	47.56
TARO[23]	84.57	53.60	46.86
DecoupledACL[38]	83.32	55.70	48.24
DAQ-SDP(ours)	**85.22**	**58.05**	**49.49**

https://doi.org/10.52202/079017-1389

Table 6: Cross-dataset transfer learning from CIFAR-100 to CIFAR-10. Note that the methods compared here are not restricted to specific SSL framework. We use both simple linear finetuning (SLF) and adversarial full finetuning (AFF) in this experiment. We use ResNet-34 as the backbone model.

	SLF		AFF	
Method	Clean	PGD	Clean	PGD
DecoupledACL[38]	55.03	25.78	86.16	52.97
TARO[23]	57.13	23.99	86.00	52.71
DAQ-SDP (ours)	**57.66**	**26.83**	**86.83**	**53.08**

Table 7: Cross-dataset transfer learning from CIFAR-10 to STL-10. We use ResNet-34 as the backbone model.

Method	Clean	PGD
Baseline	63.84	40.66
DAQ-SDP(ours)	**66.79**	**40.75**

Table 8: Ablation study for our method with SimCLR on CIFAR-100. We use ResNet-34 as the backbone model.

Method	Clean	PGD
Baseline	51.44	30.68
DAQ(single-BN)	52.27	31.56
Diverse Augmented Query	52.67	32.11
Weight Self-Perturbed Scheme	51.56	32.37
DAQ-SDP(ours)	**53.54**	**33.09**

5.2 Ablation Study

In this section we evaluate each part of our method. Table 8 shows the results on CIFAR-100. With our Diverse Augmented Query, the clean accuracy increases by 1.23% and the PGD robust accuracy increases by 1.43%. With our Diverse Augmented Query and Weight Self-Perturbed Scheme, the clean accuracy increases by 2.10% and the robust accuracy increases by 2.41% . As shown in Figures 1 and 3, both robust generalization and clean accuracy are improved compared to the baseline. These improvements brought by our proposed method demonstrate the effectiveness of the enhanced sample complexity and model regularization with the SSL loss term.

The smoothness of the weight loss landscape has been shown to be important for robust generalization in supervised AT [35]. In Figure 4 we also analyze the effect of our method to downstream weight loss landscape through 1D visualization as previous work in supervised AT [35] did. From Figure 4 we can see that the regularizing effect transferred to downstream loss landscape is actually much attenuated compared with directly regularizing classification loss in supervised AT [35], showing difficulty of such a transferred improvement from SSL pretext tasks.

6 Conclusion

In this paper, we observe that SAT has the similar problem of large robust generalization gap and clean accuracy degradation as in supervised AT. We then propose a general method to solve this problem, which can be directly combined with different pretrained SSL models without further changing learning objectives. We first challenge the previous conclusion that diverse and strong augmentations harms SAT and propose a diversely augmented query based method with Aug-Adv Pairwise-BatchNorm to distill generalizable and diverse information from clean model. Second, different from previous works that focused on introducing sample perturbation to the SSL pretext task, we suggest that regulating the smoothness of the SSL loss landscape by adversarial weight self-perturbation boosts robust generalization transferable to downstream classification. Our method not only improves the performance across different SSL frameworks, but also provides insights for narrowing the gap between the study of these two adversarial learning paradigms.

Acknowledgements

This research work is supported by the Project of Chinese Academy of Sciences (E141020).

References

[1] Sravanti Addepalli, Samyak Jain, and Venkatesh Babu R. Efficient and effective augmentation strategy for adversarial training. In *NeurIPS*, pages 1488–1501, 2022.

[2] Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? In *NeurIPS*, pages 1–24, 2019.

[3] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *NeurIPS*, pages 1–12, 2019.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. pages 1597–1607, 2020.

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning, 2020. arXiv:2011.10566v1.

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020. arXiv:2003.04297.

[7] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, pages 2206–2216, 2020.

[8] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. arxiv:1909.13719.

[9] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] Jiequan Cui, Shu Liu, Liwei Wang, and Jiaya Jia. Learnable boundary guided adversarial training. In *ICCV*, pages 15721–15730, 2021.

[11] Carl Doersch, Abhinav Gupta, and A Alexei Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015.

[12] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, pages 1–9, 2018.

[13] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: nearest-neighbor contrastive learning of visual representations. In *ICCV*, pages 9588–9597, 2021.

[14] Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, and Chuang Gan. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? In *NeurIPS*, pages 21480–21492, 2021.

[15] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018. arXiv:1803.07728.

[16] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, pages 1–11, 2015.

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *NeurIPS*, pages 21271–21284, 2020.

 https://doi.org/10.52202/079017-1389

[18] Rohit Gupta, Naveed Akhtar, Ajmal Mian, and Mubarak Shah. Contrastive self-supervised learning leads to higher adversarial susceptibility, 2023. arXiv:2207.10862.

[19] Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *CVPR*, pages 1074–1083, 2021.

[20] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. In *NeurIPS*, pages 16199–16210, 2020.

[21] Harini Kannan, Alexey Kurakin, and J. Ian Goodfellow. Adversarial logit pairing, 2018. CoRR:abs/1803.06373.

[22] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. In *NeurIPS*, pages 2983–2994, 2020.

[23] Minseon Kim, Hyeonjeong Ha, Sooel Son, and Sung Ju Hwang. Effective targeted attacks for adversarial self-supervised learning. In *NeurIPS*, pages 56885–56902, 2023.

[24] Rundong Luo, Yifei Wang, and Yisen Wang. Rethinking the effect of data augmentation in adversarial contrastive learning, 2023. arXiv:2303.01289.

[25] Aleksander Madry, Aleksander Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, pages 1–28, 2018.

[26] Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE TPAMI*, 41(8): 1979–1993, 2019.

[27] Samuel G. Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 774–782, October 2021.

[28] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84, 2016.

[29] Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, and Yang You. Crafting better contrastive views for siamese representation learning. In *CVPR*, pages 16031–16040, 2022.

[30] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in NeurIPS*, volume 34, pages 29935–29948, 2021.

[31] Jihoon Tack, Sihyun Yu, Jongheon Jeong, Minseon Kim, Sung Ju Hwang, and Jinwoo Shin. Consistency regularization for adversarial robustness, 2022. arXiv:2103.04623.

[32] Florian Tramér, Alexey Kurakin, Nicolas Papernot, J. Ian Goodfellow, Dan Boneh, and D. Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *ICLR*, 2018.

[33] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[34] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan. Gu. Improving adversarial robustness requires revisiting misclassifed examples. In *ICLR*, 2020.

[35] Dongxian Wu, Shu-tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization, 2020. arXiv:2004.05884.

[36] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, pages 2730–2739, 2019.

[37] Xilie Xu, Jingfeng ZHANG, Feng Liu, Masashi Sugiyama, and Mohan S Kankanhalli. Enhancing adversarial contrastive learning via adversarial invariant regularization. In *NeurIPS*, pages 16783–16803, 2023.

[38] Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Axi Niu, Jiu Feng, Chang D. Yoo, and In So Kweon. Decoupled adversarial contrastive learning for self-supervised adversarial robustness. In *ECCV*, 2022.

[39] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, El Laurent Ghaoui, and I Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019.

[40] Haizhong Zheng, Ziqi Zhang, Juncheng Gu, Honglak Lee, and Atul Prakash. Efficient adversarial training with transferable adversarial examples. In *CVPR*, pages 1–10, 2020.

 https://doi.org/10.52202/079017-1389

A Appendix / supplemental material

A.1 Feature Visualization

In this section, we visualize the the learned features of the test set of CIFAR-10 with t-SNE[33]. Each data point is colored with its label. As shown in 5, the features of different classes learned by our DAQ-SDP has a clearer boundary than the baseline method.

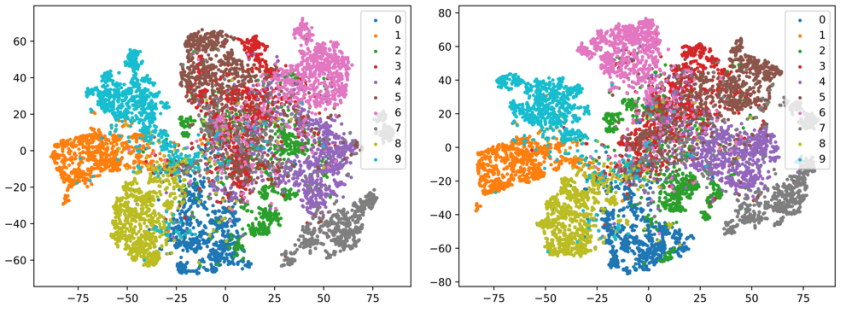

Figure 5: T-SNE results of the test set features. The left part of the figure is the features predicted by our adversarial training baseline and the right part of the figure is the features predicted by our DAQ-SDP.

A.2 Societal Impacts

Our work is useful for pretraining robust models with no labels. However, it generated one more adversarial data and perturbed weights, which takes more computation. So it can cause more consumption of energy and pollution. Despite this limitation, we believe our method is still beneficial for the society and promotes model robustness in real life.

https://doi.org/10.52202/079017-1389

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim that we find a general method to solve the robust generalization and clean accuracy reduction problem for different SSL frameworks. This claim is accurately reflected and supported in the analysis and experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mentioned in the introduction that to distill diverse and generalizable information from clean model, we need to have a backbone model that has sufficient capacity.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

https://doi.org/10.52202/079017-1389

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: In this work we analyzed the problem of current SAT methods and provided experimental results for our method.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provided all the hyperparameters and formulas for our method.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The URL of the code will be released if got accepted.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the training and testing details are provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: For the computation complexity of adversarial training and self-supervised learning, we didn't include confidence interval or sigma error bars. However, we run each experiments for 5 times and took the average results and also included the range of variation of our results in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

https://doi.org/10.52202/079017-1389

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We included the compute resources in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conformed to the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work helps to build robust models under no label conditions, which can be useful in real life.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The contents used in this paper are cited

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

https://doi.org/10.52202/079017-1389

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The model and code will be released if got accepted.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1389