We truly appreciate all reviewers (R1, R2, R3)’ valuable comments, which will greatly help improve our final paper. We have addressed all raised questions below. The final version will be carefully polished up.

---

1. Concerns about baselines (R1): For clarifying, the general idea of adversarially robust pre-training was first explored in [1]. This work’s main goal is to improve this idea by leveraging contrastive learning, for the first time. We believe the contrastive learning to fit the goal of robustness better than the ad-hoc "pre-text" tasks employed in [1], for which our intuitions were elaborated in Sections 1 and 4 (directly enforcing feature smoothness). Hence naturally, our main comparison subject in the supervised fine-tuning scheme is also [1], which is the most recent SOTA. We also included SimCLR because it is a straightforward baseline , but it is not any main competitor on robustness results

Also, we already followed [1] to employ TRADE [23] in our supervised fine-tuning. We will make this clear in paper.

2. Unforeseen attacks robustness (R1): As you can see from Figure. 2, we already tested the robustness against the set of 19 unforeseen attacks (the most standard benchmark for unforeseen robustness) for models pre-trained with ACL (DS). Comparing to [1], our method achieved performance gains on 18 out of 19 attacks. Per your suggestion, we further test the robustness of ACL (DS)-pretrained models, to ℓ_1 ($\epsilon = \frac{2000}{255}$) and ℓ_2($\epsilon = 0.5$) attacks, respectively. We find ACL (DS) still leads to higher robustness of [46.65%, 61.70%], compared to [43.14%, 58.56%] when using [1].

---

1. Why the performance of 1% low label rate is so good? (R2): An important reason for our outstanding semi-supervised results is that our pre-training inherits the strong label-efficient learning property from contrastive learning. Thus, in the *stage ii* of semi-supervised training, the generated pseudo label demonstrates higher accuracy of 86.73% even with only 1% labels available. This greatly contributes to the final robustness. In contrast, other pre-training methods like Selfi in Table 2 can only generate pseudo labels with 46.75% accuracy.

2. More backbones besides ResNet18 (R2): We further include results on Wide-Resnet-32-10, which is also a standard model employed for testing adversarial robustness. Using this model, on the supervised fine-tuning experiment with CIFAR10, ACL(DS) leads to [TA, RA] of [85.12%, 56.72%], while random initialization yields [84.33%, 55.46%][1]. Our pre-training still contributes to a nontrivial [0.79%, 1.26%] margin. More results will be added to final version.

3. ImageNet experiments (R2): We agree ImageNet experiments would be a nice addition to our paper. As you also kindly acknowledged, due to the short rebuttal time , we will try our best to obtain those results for the final version.

General response to R3: We thank R3 for suggesting a few useful points in improving our writeup quality. However, **we respectively disagree** that a clear rejection recommendation could be solely grounded on them.

We find that most presentation issues you kindly pointed out can be easily fixed through revision - please see our point-to-point reply below. We are also humbled to note that the other two expert reviewers think quite positively of our writing quality (**“excellent”, “rather clear”, “well sorted”**, etc.)

- **The "PGD" in Figure 1:** we will replace PGD with another more general word, e.g., “Attack".
- **Skipped details on supervised adversarial training and only refers to [1]**: we will happily add 2-3 lines to explain the loss function, which is essentially training with the same TRADE loss [23].
- **Loss of semi-supervised steps i and ii:** *Step i* is identical to the ACL pre-training in the supervised fine-tuning case (as indicated in lines 133). *Step ii* is the standard training with the vanilla cross-entropy loss. We can make a table to summarize the three steps’ loss functions, if that is considered to improve clarity.
- **The distilling model of semi-supervised step iii :** We employ the standard distilling model with temperature, which was introduced in (Hinton et. al., arxiv May 2015). We will make this clear in paper.
- **Definition of “TA, RA" hidden in Table 1:** No, this is a wrong accusation. Please read lines 149-150 where they appear for the first time and are clearly defined, before they are later used (line 176, table 1, etc.)
- **lambda definition:** Lambda was inherited from [23], and is defined by us in line 146.
- **“Supervised contrastive learning"**: Just want to clarify here: “supervised contrast learning" is another different method’s name (Khosla et. al., arxiv April 2020), and has nothing to do with our paper. Section 3.2 is on our standard procedure: self-supervised contrastive pre-training, followed by supervised fine-tuning.
- **Revealing identity in line 71-72?** No. What we meant was simply “among the few, now we discuss [9] in more details as it is most relevant to our semi-supervised setting". The word “introduce” might be misleading, and we will revise to “‘discuss". We welcome R3 to check back on the identity issue if this paper is accepted.
- **Hiding details for the presented approach?** We are really confused by this critique, and we do not find more details provided by your review on this regard (besides the minor issues that have been addressed above). We are confident that this paper is fully reproducible based on its own content and reference pointers, which has been agreed by the other two reviewers. We also promise to release codes upon acceptance.

[1] We run the official code of [23] for baseline. The RA drops by 1% compared to [23] reported, since we test with the PGD attack started from random initializations, which leads to stronger adversarial attack. Details will be fully specified in final paper.