We thank all the reviewers for their constructive comments. We are encouraged that reviewers found our method novel (R1, R3, R4), simple yet effective (R3, R4), the experimental results to be encouraging and thorough (R1, R3), and the paper well written (R2, R3). We address the individual comments from each reviewer below.

[Reviewer #1] Comments regarding transformation smoothed inference. We first want to emphasize that RoCLs work well without smoothed classifiers (RoCL, RoCL +rLE). Nonetheless, we tested transformation smoothed classifier (sampled 30 times, with random-fixed size crop, random color distortion) against EoT as suggested, and observed degradation of performance on CIFAR-10 with ResNet-18 **(33.24%)**. This is still **robust**, compared with other randomized defenses (0.03%) [21]. Also, please note that RoCL + smoothed can still defend against black box attack (+9.8%, Fig.3(d)) and gain clean accuracy (+0.5%) under the trade-off with white-box EoT accuracy. Moreover, RoCL w/o smoothed classifier is robust against EoT attack **(37.28%)**. We thank you for your insightful suggestion and will clarify this. **What comparison is in "Comparison to semi-supervised learning"?** We intended to discuss the time-efficiency of RoCL in that section, and will revise the title to "Training efficiency∼" for improved clarity. We further report the training time for ResNet18 trained on CIFAR10 to reach convergence for the two methods, with two RTX 2080 Ti GPUs (Table B).

Table A: Training cost

	Ours	Semi[15]
Dataset	**50K**	150K
Time	**41.7h**	66.7h
Epoch	1000	200

Table B: Blackbox attack

	RoCL Source	
Target	our attack	PGD
AT	**42.87**	69.13
Trades	**41.59**	64.81

What if you use RoCL as the source model for the blackbox attacks? Adversarial examples generated from RoCL with instance-wise attack and PGD attack both succeed in attacking AT and TRADES. (Table B; epsilon=8/255).

Why train models with 16/255? The number of iterations? This seems like a misunderstanding since as described in L436, we trained all models including ours with 8/255 with 7 iterations. We used 20 iterations for test (L424,L442).

Conventional adversarial perturbations are instance-wise. We will rename our attack as "instance-identity attack".

[Reviewer #2] Novelty of rLE and transformation smoothed classifier. Our main contributions are 1) the proposal of a **novel adversarial perturbation** which makes the model to confuse a sample to another, 2) the **contrastive learning** framework for **unsupervised adversarial learning**. We proposed the rLE as an evaluation measure for robustness of unsupervised adversarial learning, and t-smoothed inference an alternative for existing smoothed classifier, but they are rather technical details and we do not claim them as our contributions (Please see L60-67 for our claims).

Inconsistency between claimed benefits in certain lines and Table 1. We will fix the inaccurate descriptions as follows: (a) We will rename RoCL to RoCL+rLE, which does outperform AT and obtain comparable performance to TRADES. (b) We will clarify that RoCL achieves higher robustness against "AT black-box attacks", as described in L262-265. (c) We will revise this as RoCL + finetuning "sometimes" outperform models trained from scratch.

Colors in Fig.3(a,b) The markers in Fig.3 with different shapes denote instances belonging to different instances, and the green and red color denote clean and adversarial (L286) instances, respectively. We will clarity this in the revision.

Purpose of Table 5. Table 5 shows that RoCL remains robust even with an increased number of attack iterations.

Missing formula for RoCL+AT+SS, typos in Eq.1 and Eq.4, a missing reference Thank you for pointing them out. We used both AT loss and Eq.3 for RoCL+AT+SS. We will fix the typos ("-") in Eq.1 and Eq.4 to "+", and cite and discuss [Naseer et al. 20] which is different from ours that proposes purifier network trained with Euclidean adversary.

[Reviewer #3] The methodology in section 3.1 is confusing. Here we generate adversarial examples using Eq.5 which fools the instance identity, and explicitly train the model using Eq.6 (L169-172) using the generated adversarial examples. The regularization term in the Algorithm 1 yields small gain (+1.07%) on robustness, and the most important loss is $\mathcal{L}_{con}$ (L187-193). We will include the descriptions of the regularizer and the Algorithm in the main paper.

"Unsupervised" could be misleading since label is used for downstream tasks. We will revise the texts to clarify that we do not require any lables to learn adversarially robust representations, but need labels for downstream tasks.

The description of the experimental results should be specific. We apologize and will revise the inaccurate descriptions. Please see the comments to Reviewer #2 (Inconsistency between claimed benefits and Table 1).

What is the backbone of [38]? [38] used WRN 32-10 (L446), which has much larger number of parameters and depths compared to ResNet18 used by our model that outperforms [38]. This demonstrates the effectiveness of RoCL.

[Reviewer #4] Do you need new linear layers on top of supervised [9,2] to compare against RoCL? Since supervised methods already have a linear layer at training time, it is unnecessary to add an additional linear layer.

Why use linear evaluation? Finetuning makes more sense. Since linear evaluation "freezes" the representation, it is the most direct way to measure the robustness of the learned representations (L196-198). Finetuning (L247-253) will change the lower layer representations and will make it difficult to separate the effect of supervised adversarial finetuning from effect of RoCL. **Which data augmentation were used?** As described in L147-148,173, we used random crop, and random color distortion. **Batchsize 256 too small.** SimCLR [12] experimented on ImageNet, and we empirically found that bathsize of 256 is sufficient for CIFAR-10 and CIFAR-100. Note that we use adversarial examples as additional positive examples as well. **Why consider both $t'(x)$ and $t'(x)^{adv}$ as positive samples?** Using both of them is essential since we are targeting for both clean and adversarial accuracies.

Transfer learning setup might not be the best example of TL. We agree. However, since the baseline [38] is using the described experimental setup, we had to follow it for a fair and direct comparison (L449).