We thank the reviewers for the positive feedback: new state-of-the-art results (R1,2&3), first to explore cross-domain transferability (R1), high significance to the community (R3), very well written and clear presentation (R1,2&3).

**Code**: Code will be made public. Fig.(1)[E][E] best viewed in zoom.

**R1.2&3**: Significance of Relativistic Cross-Entropy (RCE): Adversarial perturbations are crafted via loss function gradients. An effective loss helps in adversary generation by back-propagating stronger gradients. Below, we show that RCE ensures this requisite and thus leads to better performance than CE.

**Notation**: classifier $\mathcal{F}$, clean sample $x$, adversarial example $x'$, output scores $a = \mathcal{F}(x)$, $a' = \mathcal{F}(x')$.

**Gradient Perspective**: Let $CE(a', y) = -\log(e^{a'}/\sum_k e^{a_k})$ be the CE loss for input $x'$. For clarity, we define $p'_y = e^{a'_y}/\sum_k e^{a'_k}$. The derivative of $p'_y$ w.r.t $a'_i$ is $\partial p'_y/\partial a'_i = p'_y [i=y] - p'_i$. From chain rule, $\partial CE/\partial a'_i = p'_i - [i=y]$ (Eq. 1). For relativistic loss, $RCE(a', y) = -\log(e^{a'_y} - \sum_k e^{a'_k})/\sum_k e^{a'_k}$, we define $r_y = (e^{a'_y} - \sum_k e^{a'_k})/\sum_k e^{a'_k}$. The derivative of $r_y$ w.r.t $a'_i$ is $\partial r_y/\partial a'_i = r_i ([i=y] - r_y)$. From chain rule, $\partial RCE/\partial a'_i = r_i - [i=y]$ (Eq. 2).

In light of above relations, RCE has three important properties: (a) Comparing (Eq. 2) with (Eq. 1) shows that RCE gradient is a function of ‘difference’ $(a'_y - a_y)$ as opposed to only scores $a'_y$ in CE loss. Thus it measures the relative change in prediction as an explicit objective during optimization. (b) RCE loss back-propagates larger gradients compared to CE, resulting in efficient training and stronger adversaries (see Fig.1 for empirical evidence). **Sketch Proof**: We can factorize the denominator in (Eq. 2) as follows: $\partial RCE/\partial a'_i = (e^{a'_y} - \sum_k e^{a'_k})/\sum_k e^{a'_k}$. For clarity, we define $\sum_k e^{a'_k} = y$. Consider the fact that maximization of RCE is only possible when $e^{a'_y} - \sum_k e^{a'_k}$ decreases and $\sum_k e^{a'_k} = y$ increases. Generally, $a'_y > a_k \neq y$ for the score generated by a pre-trained model and $a'_y < a_k \neq y$. Thus, $\partial RCE/\partial a'_i > CE/\partial a'_i$, since $e^{a'_y} - \sum_k e^{a'_k} > \sum_k e^{a'_k}$. In simple words, the gradient strength of RCE is higher than CE.

(c) In case $x$ is misclassified by $\mathcal{F}(\cdot)$, the gradient strength of RCE is still higher than CE (here noise update with the CE loss will be weaker since adversary’s goal is already achieved i.e., $x$ is misclassified). We will add it in final version.

**Evaluation**: We further validate (see Fig. 2) the significance of RCE compared to CE in terms of three criterion (accuracy, logits difference and transfer to unseen classes). For the test on unseen classes, we divide ImageNet into two mutually exclusive sets (500 classes each), namely IN1 and IN2. VGG16 is trained on IN1 & IN2 from scratch.

**R1**: 1) RCE Justification: See R1.2&3 above. 2) Relation with Style-Transfer: We visualize the intermediate feature space of cross-domain perturbed images and compare it with original and stylized images (Fig. 3). We note that the feature space of perturbed images is fairly shifted from the original and stylized images. This shows that although some of the generated patterns resemble “style” of a specific domain (e.g., in Fig. 3 main paper), the overall behaviour of our proposed approach is distinct from style transfer. This is potentially due to the existence of “non-robust features” defined as ‘features that are highly predictive but brittle and incomprehensible to humans’ [A1]. Since, our generated perturbations are bounded (as opposed to unbounded style transfer), the attacker is likely to focus on the non-robust features. We will add further qualitative examples on other domains in final version (Fig. 6 in supp. material).

**R2**: Theoretical Result: See R1.2&3 earlier. **Typo**: We thank R2 & fix it.

**R3**: 1) Use of Instance-Agnostic: We used this term to differentiate the one-time training feature of our attack as opposed to instance-specific attacks. However, we acknowledge R3’s point and will replace this term with domain-agnostic for clarity. 2) Comparison with [1,19]: [1] trains conditional generators to learn original data manifold and searches the latent space conditioned on the human recognizable target class that is mis-classified by a target classifier. Different to [1], our approach learns to add adversarial noise to the original samples. [19] produces adversarial images by employing a separate discriminator alongside classifier. Different to [19], we train a generator to first produce unbounded adversaries and then project them to nearby original images. We thank R3 and will add further discussion in final version.

---