

1 We thank the reviewers for their insightful feedback.

2 **Strength of attack** Some of the reviewers raised concerns about the strength of the ReColorAdv attack relative to
 3 existing attacks. We perform two new experiments to demonstrate the strength of ReColorAdv. First, we found an
 4 implementation issue with the projection step of PGD for non-RGB color spaces in the original paper; we fixed it and
 5 found that the strength of the attack increased. In addition, we evaluated all attacks with 300 iterations of PGD. The
 6 results are shown in the table below. Note that the combined ReColorAdv + StAdv + delta attack reduces the accuracy
 7 of an adversarially trained classifier to just 3.6%, less than half of the previous best combined attack. Finally, the
 8 success rate of the ReColorAdv attack is not its only advantage; the perturbations it produces are less noticeable as well.

Defense	PGD iters	Attack						
		C	D	S	C+S	C+D	S+D	C+S+D
None	100	4.4	0.0	2.2	1.6	0.0	0.0	0.0
None	300	3.3	0.0	1.2	0.9	0.0	0.0	0.0
Adv. training	100	50.2	32.8	29.9	15.4	14.9	9.6	8.8
Adv. training	300	45.8	30.1	26.2	8.7	5.2	7.6	3.6
TRADES	100	64.8	53.7	31.2	23.2	29.0	11.1	8.1
TRADES	300	59.2	53.6	26.6	17.5	22.0	8.7	5.7













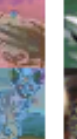

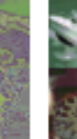









10 **Related work (R1)** Like our work, both Song et al.¹ and Zhang et al.² aim to construct adversarial examples outside
 11 of the usual ℓ_p ball perturbation. However, Song et al. uses a generative model to craft adversarial examples directly
 12 without perturbing other images; in contrast, we aim to augment the space of adversarial *perturbations* of images in
 13 the train or test sets. Zhang et al. is more similar to our work; they apply a single affine function to all pixels in an
 14 input before performing PGD. However, unlike our work, they do not consider more complex functions and they do not
 15 present their "scale and shift" technique as an attack in itself, only as a way of discovering images far from the training
 16 manifold. We will incorporate a discussion of similarities and differences to these works in the revised draft.

17 **Other functional attacks (R1)** The functional adversarial threat model is applicable to many domains but we choose
 18 to focus on image classification. This problem is widely viewed as a benchmark in machine learning and adversarial
 19 robustness research and focusing on it allows us to report in-depth results. While we hope to extend the threat model to
 20 other domains in the future, we believe it is beyond the scope of this paper.

21 **Perturbations of triples vs. separate channels (R3)** Indeed ReColorAdv does operate on triples rather than
 22 perturbing each channel independently, i.e. each pixel is mapped $(R, G, B) \rightarrow f(R, G, B)$.

23 **Other color spaces (R3)** Upon your suggestion, we experimented with the HSV (hue, saturation, value) and YPbPr
 24 color spaces in addition to RGB and CIELUV. HSV presents difficulties when performing PGD because the derivative
 25 of the transformation from RGB is highly discontinuous; thus we use an approximation HSV' which maps colors into
 26 a hexagonal pyramid instead of the standard HSV cone. A disadvantage of both HSV and YPbPr is that they were
 27 originally designed for transmitting video signals rather than as an accurate representation of how humans view colors.

28 Below are adversarial examples based on suggestions from R3, which we will also include in the paper. In these
 29 experiments, we present ReColorAdv using four color spaces; see C-{LUV, RGB, YPbPr, HSV'}. We also apply
 30 ReColorAdv separately to each channel. i.e. $(R, G, B) \rightarrow (f_1(R), f_2(G), f_3(B))$; see C-{LUV, RGB}-Sep-Channels.
 31 Finally, we use separate bounds on each RGB channel based on the sensitivity of the human eye ($\epsilon_R = 0.1, \epsilon_G =$
 32 $0.05, \epsilon_B = 0.15$); see C-RGB-Sep-Bounds. Note that we already applied separate bounds in CIELUV color space, as
 33 detailed in appendix B.1. For each of these variations, the accuracy of an undefended model under the attack is shown.

Original	C-LUV	C-LUV-Sep-Channels	C-RGB	C-RGB-Sep-Channels	C-RGB-Sep-Bounds	C-YPbPr	C-HSV'
92.3%	4.4%	8.4%	8.2%	9.3%	8.4%	2.6%	2.1%
							
							
							

36 **Minor comments (R3)** We appreciate these suggestions and will fix them in the revised draft.

¹Yang Song et al. Constructing Unrestricted Adversarial Examples with Generative Models. *NeurIPS*, 2018.

²Huan Zhang et al. The Limitations of Adversarial Training and the Blind-Spot Attack. *ICLR*, 2019.