A Limitations of Theorem 1

While Theorem 1 provides an understanding of the tradeoff between adversarial and natural distributional robustness, there are some limitations. Firstly, the results consider a setting where the core and spurious features are completely disentangled, i.e, they each represent different parts of the input. In practice, spurious features may be entangled with the core features (e.g., the color of an image may represent a spurious feature.) In addition, our results mainly consider the *goal* of adversarial training as we focus on the expected loss $L_{p,\epsilon}(\theta)$, rather than its finite-sample variant. This is because even for an ℓ_2 adversary, characterizing the finite-sample behaviour of adversarial training is difficult and requires careful assumptions on the asymptotic behaviour of the parameters (e.g., see Theorem 3.3 in [26]). We leave exploring these directions to future work. Even so, we believe our theoretical results are of interest to the community since disjoint features already capture a wide variety of spurious correlations, e.g., background correlations, as well as examples where a spurious object is present in the image. The main goal of our theoretical analysis is to show the existence of explicit tradeoffs between adversarial and distributional robustness and build practical insights using those results.

B Societal Impact

Our work touches on two important notions of robustness for the safe and fair deployment of deep models in the wild. We hope our results lead to careful analysis of all modes of robustness, and the interplay between them, before deep models are used in sensitive applications. While our results create tension with some previous works [71, 75, 25], we stress that we do not wish to diminish their work; instead, we hope our work reveals the vast nuance associated with spurious correlations, which can help and hurt models in various ways. Lastly, we release all code to encourage future work.

C Additional results for the ℓ_1 norm

In this section, we further analyze the plateauing behaviour of the performance of the linear model observed in Figure 2 when using ℓ_1 adversarial training. To this end, we consider different values for the number of core features c and total features m and measure NFS for different values of adversarial budget ϵ as in Figure 2. The matrix Σ is constructed using Equation (4) as before, with modified number of rows and columns based on the values of c, p. Similarly, θ^{opt} is constructed as before, with the core coordinates set to 1 and the spurious coordinates set to 0. The value of η is fixed at 0.5. The results are shown in Figure 10.

As shown in the Figure, when using m total features and c core features, NFS plateaus at $\frac{m-c}{m}$ for large values of ϵ. Intuitively, this is because of the structure of the optimization problem (2). Recall that when using the ℓ_1 norm, the value of q in (2) equals ∞. As such, adversarial training tries to find a parameter θ that has a low ℓ_∞ norm and is "close" (as measured by σ_θ) to θ^{opt}. The ℓ_∞ penalty encourages values of θ that are uniform across the coordinates. Since there are $m - c$ spurious features and m total features, this leads to models that have an NFS value of $\frac{m-c}{m}$.

D Proof of Theorem 1

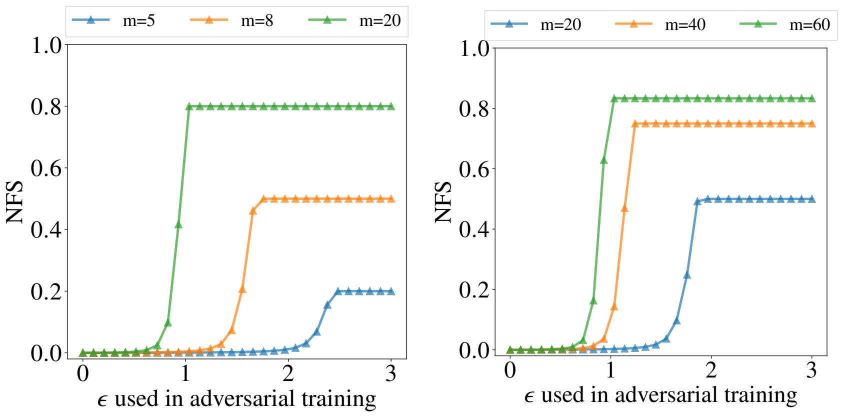
Proof. We first claim that

$$\max_{\|\delta\|\leq\epsilon} (Y - \langle X + \delta, \theta\rangle)^2 = \Big(|\langle y, x\rangle| + \epsilon \cdot \|\theta\|_q\Big)^2$$

To see why this holds, note that for all δ satisfying $\|\delta\|_p \leq \epsilon$,

$$\begin{aligned} |Y - \langle X + \delta, \theta\rangle| &\overset{(a)}{\leq} |Y - \langle X, \theta\rangle| + |\langle \delta, \theta\rangle| \\ &\overset{(b)}{\leq} |Y - \langle X, \theta\rangle| + \epsilon \cdot \|\theta\|_q, \end{aligned}$$

where (a) follows from the triangle inequality and (b) follows from Hölder's inequality. With a suitable choice of θ, we can achieve equality for (b). As $\|\theta\|_q = \|-\theta\|q$, at least one of $\{\theta, -\theta\}$



(a) 4 core feautres

(b) 10 core feautres

Figure 10: Analysis of NFS for the linear model when using the ℓ_1 norm in adversarial training. Each figure measures the reliance of the model on spurious features (measured by NFS) while varying the adversarial training budget ϵ, using different number of *total* features m. The number of core features is kept constant and set to 4 in Figure **(a)** and to 10 in Figure **(b)**.

would further achieve equality for (a). As maximizing $|.|$ is equivalent to maximizing $(.)^2$, (3) is proved.

Given (3), we can rewrite (1) as

$$
\begin{aligned}
L_{p,\epsilon} &= \mathbb{E}\left[\left(|Y - \langle X,\theta\rangle| + \epsilon \cdot \|\theta\|_q\right)^2\right] \\
&= \mathbb{E}\left[(Y - \langle X,\theta\rangle)^2\right] + \epsilon^2 \cdot \|\theta\|_q^2 + 2 \cdot \epsilon \cdot \mathbb{E}\left[|Y - \langle X,\theta\rangle|\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\left(\left\langle X, \theta - \theta^{\text{opt}}\right\rangle + W\right)^2\right] + \epsilon^2 \cdot \|\theta\|_q^2 + 2 \cdot \epsilon \cdot \mathbb{E}\left[\left|\left\langle X, \theta - \theta^{\text{opt}}\right\rangle + W\right|\right],
\end{aligned}
$$

Where for (a) we have used the fact that $Y = \langle X,\theta\rangle + W$.

Define v_θ as $\langle X, \theta - \theta^{\text{opt}}\rangle + W$. As X was assumed to be sampled from $N(0,\Sigma)$, v_θ is distributed as $N(0,\sigma_\theta^2)$. It follows that

$$
\begin{aligned}
L_{p,\epsilon} &= \mathbb{E}\left[\left(\left\langle X, \theta - \theta^{\text{opt}}\right\rangle + W\right)^2\right] + \epsilon^2 \cdot \|\theta\|_q^2 + 2 \cdot \epsilon \cdot \mathbb{E}\left[\left|\left\langle X, \theta - \theta^{\text{opt}}\right\rangle + W\right|\right], \\
&= \mathbb{E}\left[v_\theta^2\right] + \epsilon^2 \cdot \|\theta\|_q^2 + 2 \cdot \epsilon \cdot \mathbb{E}\left[|v_\theta|\right] \\
&\overset{(a)}{=} \sigma_\theta^2 + \epsilon^2 \cdot \|\theta\|_q^2 + 2 \cdot \epsilon \cdot \sigma_\theta \\
&= (c_1^2 + c_2) \cdot \sigma_\theta^2 + \epsilon^2 \cdot \|\theta\|_q^2 + 2 \cdot \epsilon \cdot \sigma_\theta \\
&= c_2 \cdot \sigma_\theta^2 + (c_1 \sigma_\theta + \epsilon \cdot \|\theta\|_q)^2
\end{aligned}
$$

where for (a) we have used the fact that $\mathbb{E}\left[\left|N(0,\sigma^2)\right|\right] = c_1 \cdot \sigma$. This proves (2) as claimed.

As for convexity, σ_θ is convex in θ since it can be written as $\left\|[\Sigma^{\frac{1}{2}}(\theta - \theta^{\text{opt}}), \sigma_w]\right\|_2$ where $[.,.]$ denotes the vector stacking operation. As $c_1\sigma_\theta + \epsilon \cdot \|\theta\|_q$ is always positive and $x \to x^2$ is convex and increasing for $x \geq 0$, this implies that $(c_1\sigma_\theta + \epsilon \cdot \|\theta\|_q)^2$ is convex as well. Finally $c_2\sigma_\theta^2$ is convex as $c_2 > 0$ and therefore (1) is convex in θ. $\square$

E Additional Details on Reverse Effect (Section 4.3)

Our final empirical observation is that the presence of a spurious feature (in both training and test distributions) can lead to increased adversarial robustness. This more directly creates tension with claims that adversarial vulnerability is born out of spurious feature reliance. We refer to this as the

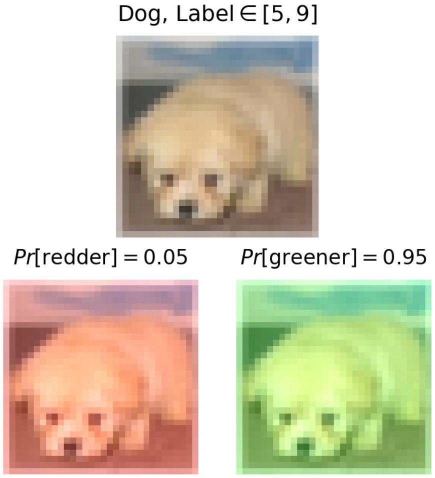


Figure 11: Color Shift, $\rho = 19 : 1$

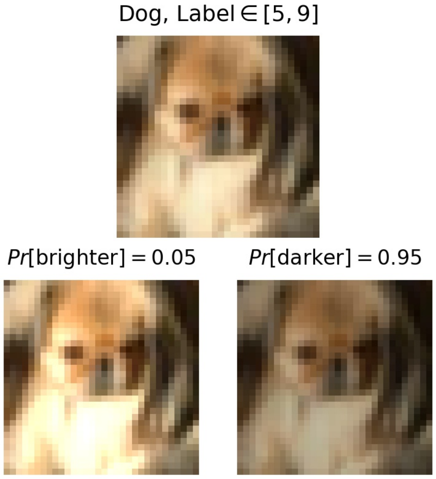


Figure 12: Lighting Shift, $\rho = 19 : 1$

'reverse effect', in relation to our primary empirical and theoretical finding that adversarial training increases spurious feature reliance. We now elaborate on the experimental setup discussed in Section 4.3, reproduce the results with a different spurious feature, and finally appeal to ImageNet-9 to demonstrate this effect using a more realistic spurious feature (i.e. backgrounds).

E.1 Experimental Setup

Overview. We inject spurious correlations to the CIFAR10 dataset. Based on the class label, we adjust half the images (i.e. with class label from 5 to 9) to shift in one direction with high probability. For example, a dog image is made greener with probability 0.95, corresponding to a majority-to-minority group ratio of $\rho = 19 : 1$. With probability 0.05, we shift in the other direction (e.g. make redder). We then standardly train a ResNet18 from scratch on the dataset with the spurious feature injected for the 10-way CIFAR classification task. Importantly, we evaluate robust accuracy with the spurious feature retained, and then compare adversarial robustness of models trained under data with different strengths of the injected spurious correlation. Figure 9 and 13 show that for two distinct spurious features (color and lighting), robust accuracy is higher when the spurious correlation is stronger. Notably, the gain is larger than the gain in standard accuracy. Intuitively, we see that relying on the predictive power of the spurious feature is helpful for standard accuracy, and especially for acccuracy under adversarial attack. Despite being irrelevant to the true labeling function, the spurious feature can improve model performance, and indeed even lead to better adversarial robustness.

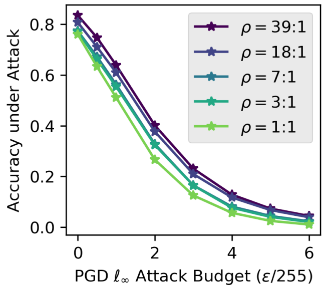


Figure 13: Reverse effect using spurious feature of *lighting*. Main text figure uses color as spurious feature.

Details. Color shift is achieved by increasing all pixel intensities along one channel by 0.25. Lighting shift is achieved by simply scaling an input by 1.25 to make brighter or 0.75 to make darker. All images are clamped to remain in the $[0, 1]$ pixel range after spurious feature injection. Models are trained for 20 epochs using an Adam optimizer with a learning rate of 0.001 and weight decay of $1e - 4$.

E.2 Leveraging ImageNet-9

We now demonstrate the observed reverse effect on the higher resolution ImageNet-9 dataset, leveraging the natural and ubiquitous spurious feature of backgrounds. We finetune pretrained models on MIXED-SAME and MIXED-RAND separately, and evaluate each model's accuracy under attack on the same split that they were trained over. Further, we leverage the adversarially trained models from test suite in this experiment. This way, accuracy under attack is more informative, as the models

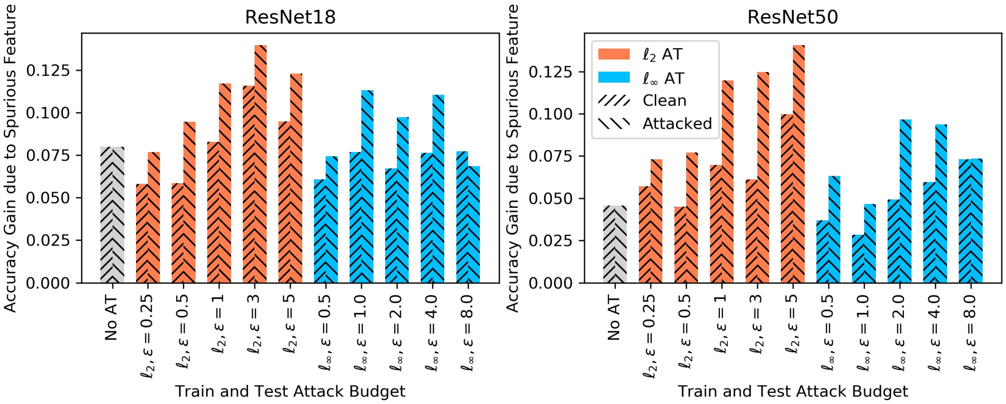


Figure 14: Background Gap (difference in accuracy on MIXED-SAME and MIXED-RAND) for clean and adversarially attacked images. Across models, background gap is larger when considering accuracy under attack, suggesting that the presence of a spurious correlation in training data makes the model more adversarially robust over the same distribution.

are trained to expect attacks (i.e. we are not imposing any distribution shifts that would lead to unexpected model behavior). Along this vain, we attack each backbone with the same norm and ϵ that it was pretrained over.

Figure 14 shows the gain in accuracy for the models trained and evaluated on MIXED-SAME compared to those using MIXED-RAND. We see that the presence of background correlations increases both standard and robust accuracy for all models (i.e. gains are positive). Further, gains in accuracy under attack are larger than gains in standard accuracy in nearly all cases. Thus, it seems like the added predictive power of the spurious background feature has a significantly nontrivial impact on improving adversarial robustness, contradicting many existing arguments on the link between spurious correlations and adversarial vulnerability.

F Adversarially Robust Model Test Suite (Section 3)

F.1 Model Details

We utilize the treasure trove of open-source adversarially trained models, contributed by [49], accessible at `https://github.com/Microsoft/robust-models-transfer`. For completeness, we now provide details on the models we use, though we refer readers to Appendix A.1 of the original text, where the information we share now is sourced.

Training All models were trained on ImageNet in batches of 512 samples, using SGD optimizer with momentum of 0.9 and weight decay of $1e-4$, for a total of 90 epochs, with learning rate dropping by a factor of 10 every 30 epochs. The standard procedure of [38] was performed to adversarially train models, using 3 projected gradient descent steps with a step size $\frac{2}{3}\epsilon$ for the attack budget ϵ.

Selected Models We focus our empirical study on the ResNet architecture [20] because of its wide spread popularity. Specifically, we study ResNet18s and ResNet50s that are adversarially trained under the ℓ_2 norm, for $\epsilon \in \{0.25, 0.5, 1, 3, 5\}$, and ℓ_∞ norm, for $\epsilon \in \{0.5/255, 1/255, 2/255, 4/255, 8/255\}$, as well as standardly trained baselines.

Table 1 shows the standard accuracies for these models. Note that we at times compare between the ℓ_2 and ℓ_∞ adversarially trained models (e.g. figure 6). We acknowledge that direct comparisons are challenging because the threat model under which adversarial robustness is optimized for are different. However, we note that standard accuracies of the i^{th} ℓ_2 AT model is roughly the same as that of the i^{th} ℓ_∞ AT model, suggesting that those models lie in similar points of the accuracy-robustness tradeoff.

Additional Models. We extend our analysis to other architectures. We replicate all pretrained-model experiments on the Wide ResNet50 (2x) backbones, for which we have checkpoints for each of the five ϵ values for both ℓ_2 and ℓ_∞ norms. We also inspect MobileNetv2 [51], DenseNet161 [24],

AT Norm	ϵ	ResNet18	ResNet50	Wide ResNet50 (2x)
No Adv Training		69.79	75.80	76.97
ℓ_2	0.25	67.43	74.14	76.21
ℓ_∞	$0.5/255$	66.13	73.73	75.82
ℓ_2	0.5	65.49	73.16	75.11
ℓ_∞	$1/255$	63.46	72.05	74.65
ℓ_2	1	62.32	70.43	73.41
ℓ_∞	$2/255$	59.63	69.10	72.35
ℓ_2	3	53.12	62.83	66.90
ℓ_∞	$4/255$	52.49	63.86	68.41
ℓ_2	5	45.59	56.13	60.94
ℓ_∞	$8/255$	42.11	54.53	60.82

Table 1: Clean ImageNet accuracy for test suite of ℓ_2 and ℓ_∞ adversarially trained ResNets over varying ϵ. Observe that the i^{th} ℓ_2 AT model has similar clean accuracy to the i^{th} ℓ_∞ AT model.

	ShuffleNet	MobileNet	VGG	DenseNet	ResNeXt
No AT	64.25	65.26	73.66	77.37	77.38
ℓ_2 AT, $\epsilon = 3$	43.32	50.40	57.19	66.98	66.25

Table 2: Clean ImageNet accuracy for five additional architectures considered.

ResNeXt5050_32x4d [68], ShuffleNet [74], and VGG16_bn [56]. For each of these five architectures, we compare an ℓ_2 adversarially trained model with $\epsilon = 3$ to a standardly trained baseline.

F.2 Experimental Details

ObjectNet and ImageNet-C [5, 22]. We report raw accuracies under noise, blur, and digital corruption types for ImageNet-C, as opposed to relative corruption error. For ObjectNet, we map ImageNet predictions to the set of 113 overlapping classes in ObjectNet. **RIVAL10 (*RFS*) and Salient ImageNet-1M (*RCS*)** [40, 59]. RFS computation requires finetuning a final linear layer over fixed features for the coarse-grained ten way RIVAL10 classification. RCS operates on models off the shelf, directly inspecting accuracies over ImageNet classes (and samples, with region-based noise corruption). **ImageNet-9 and Waterbirds** [67, 48]. ImageNet-9 accuracies are obtained by mapping off-the-shelf model predictions to the nine coarse labels deterministically. Waterbirds requires finetuning, which we do over fixed features. For RIVAL10 and Waterbirds finetuning, we use Adam with learning rate of $1e-4$ and weight decay of $1e-5$ for 20 and 15 epochs respectively.

F.3 Results on Extended Model Test Suite

We now corroborate all our empirical findings on new backbones, expanding our analysis to 21 new models (including 10 AT WideResNet50s over both ℓ_2 and ℓ_∞ norms) over six architectures.

WideResNets. We corroborate all our empirical findings on ResNet18s and ResNet50s on the WideResNet50 (2x) architecture. Figure 15 shows that accuracy drop in AT models is more severe on distribution shifts that break spurious correlations (ObjectNet), unlike the accuracy drop due to corruption of both core and spurious features (ImageNet-C), which can likely be explained by the reduced standard accuracy of AT models.

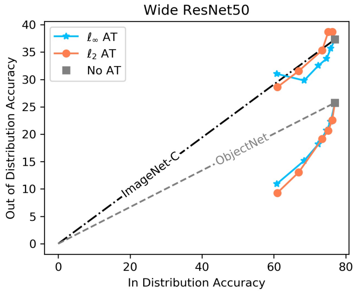


Figure 15: ObjectNet, ImageNet-C, and ImageNet accuracies for WideResNet50s.

Figure 16 shows reduced sensitivity to core and foreground regions for AT models. Again, the effect is more pronounced for ℓ_2 adversarially training and for larger ϵ. Also, we again see that decrease in RCS is less consistent than the drop in RFS. We conjecture that

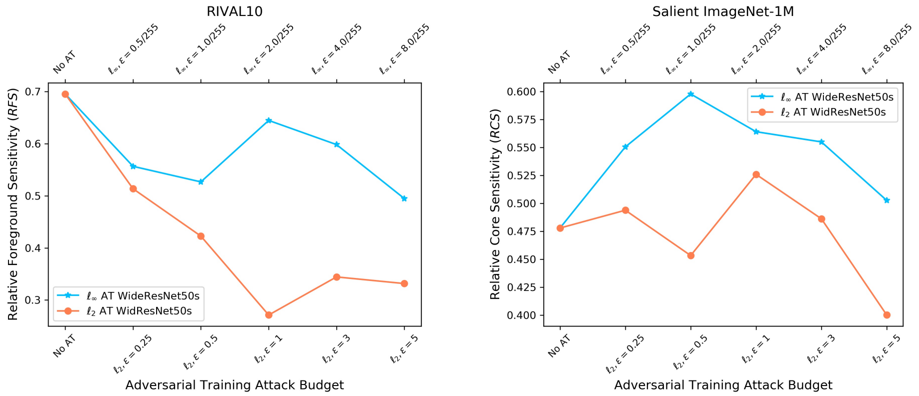


Figure 16: RFS and RCS for WideResNet50s. Sensitivity to core and foreground regions are reduced for higher ϵ, especially for ℓ_2 AT models and for RFS, computed over the RIVAL10 dataset, where background correlations are stronger.

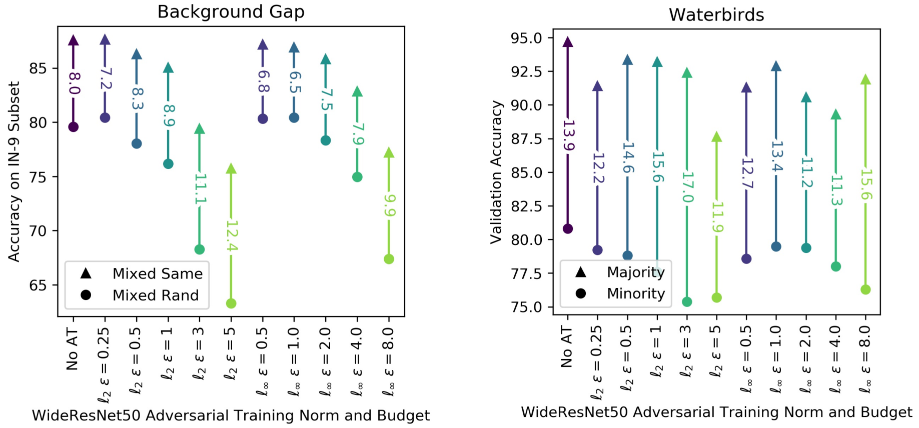
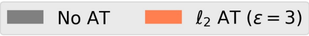


Figure 17: Background Gap (IN-9) and Waterbirds gap for WideResNet50s. AT models, especially under ℓ_2 norm, see larger accuracy drops when spurious correlations are broken.

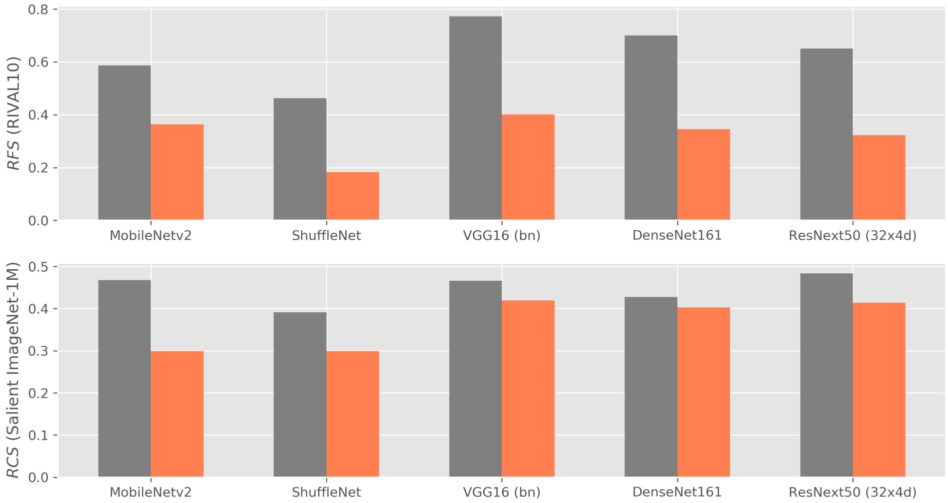
the diversity and fine-grain Salient ImageNet classification task reduces the strength of spurious correlations present in the data, thus diluting our observed effects of adversarial training on spurious feature reliance.

Lastly, figure 17 shows the drop in accuracy due to breaking spurious background correlations is larger for AT models. Indeed, the absolute background gap (IN-9) for the WideResNet50 under ℓ_2 AT with $\epsilon = 5$ is 50% larger than the gap for the standardly trained baseline. We note that the absolute gaps are smaller in some cases. We believe the lower standard accuracy of AT models may contribute to this, as there is less accuracy to drop from. Nonetheless, it is intriguing that in some cases, ℓ_∞ adversarial training seems to reduce spurious feature reliance; while our theory explains how a spurious feature can be completely ignored under ℓ_∞ training, it does not explain cases where spurious feature reliance is reduced compared to standard training. We believe this is an interesting direction for future work.

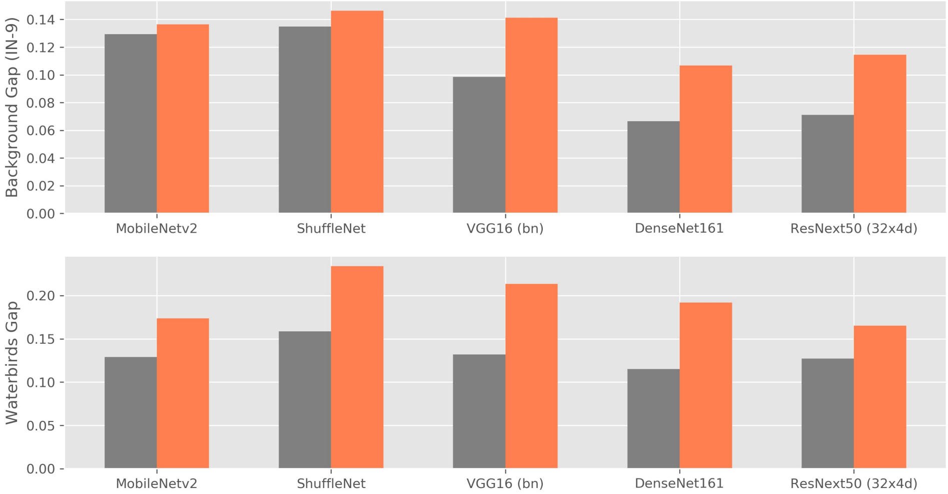
Other backbones. We now show results for ten other models, half of which are ℓ_2 adversarially trained with $\epsilon = 3$, while the others are standardly trained. Figure 18 summarizes our results, corroborating each of our empirical findings.



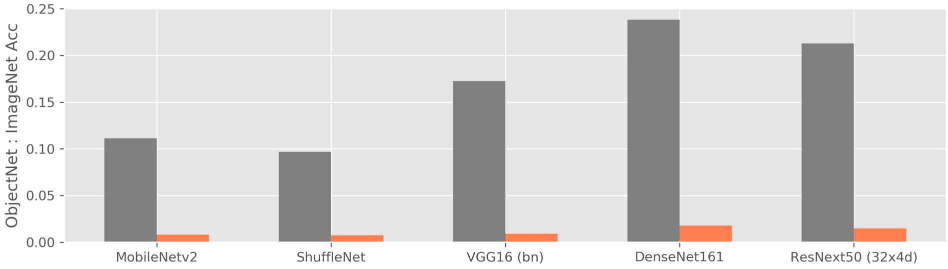
(a) Legend. We compare ℓ_2 adversarially trained models to standardly trained baselines for five new backbones.



(b) Lower RFS (RCS) entails Lower Foreground (Core Feature) Sensitivity



(c) Higher Gap entails Greater Background/Spurious Sensitivity



(d) Lower Ratio entails Lower Natural Distributional Robustness

Figure 18: Corroborating findings on additional backbones.