NOISE OR SIGNAL: THE ROLE OF IMAGE BACKGROUNDS IN OBJECT RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

We assess the tendency of state-of-the-art object recognition models to depend on signals from image backgrounds. We create a toolkit for disentangling foreground and background signal on ImageNet images, and find that (a) models can achieve non-trivial accuracy by relying on the background alone, (b) models often misclassify images even in the presence of correctly classified foregrounds—up to 88% of the time with adversarially chosen backgrounds, and (c) more accurate models tend to depend on backgrounds less. Our analysis of backgrounds brings us closer to understanding which correlations machine learning models use, and how they determine models' out of distribution performance.

1 INTRODUCTION

Object recognition models are typically trained to minimize loss on a given dataset, and evaluated by the accuracy they attain on the corresponding test set. In this paradigm, model performance can be improved by incorporating any generalizing correlation between images and their labels into decision-making. However, the actual model reliability and robustness depend on the specific set of correlations that is used, and on how those correlations are combined. Indeed, outside of the training distribution, model predictions can deviate wildly from human expectations either due to relying on correlations that humans do not perceive (Jetley et al., 2018; Ilyas et al., 2019; Jacobsen et al., 2019), or due to overusing correlations, such as texture (Geirhos et al., 2019; Baker et al., 2018) and color (Yip & Sinha, 2002), that humans do use (but to a lesser degree). Characterizing the correlations that models depend on thus has important implications for understanding model behavior, in general.

Image backgrounds are a natural source of correlation between images and their labels in object recognition. Indeed, prior work has shown that models may use backgrounds in classification (Zhang et al., 2007; Ribeiro et al., 2016; Zhu et al., 2017; Rosenfeld et al., 2018; Barbu et al., 2019; Shetty et al., 2019; Sagawa et al., 2020; Geirhos et al., 2020), and suggests that even human vision makes use of image context for scene and object recognition (Torralba, 2003). In this work, we aim to obtain a deeper understanding of how current state-of-the-art image classifiers utilize image backgrounds. Specifically, we investigate the extent to which models rely on them, the implications of this reliance, and how models' use of backgrounds has evolved over time. Concretely:

- We create a variety of datasets that help disentangle the impacts of foreground and background signals on classification. The datasets and code for evaluating models on these datasets are available in this anonymized repository: `git.io/bg-role-iclr`.
- Using the aforementioned toolkit, we characterize models' reliance on image backgrounds. We find that image backgrounds alone suffice for fairly successful classification and that changing background signals decreases average-case performance. In fact, we further show that by choosing backgrounds in an adversarial manner, we can make standard models misclassify 88% of images as the background class.
- We demonstrate that standard models not only use but *require* backgrounds for correctly classifying large portions of test sets (35% on our benchmark).
- We study the impact of backgrounds on classification for a variety of classifiers, and find that more accurate models tend to simultaneously exploit background correlations more and have greater robustness to changes in image background.

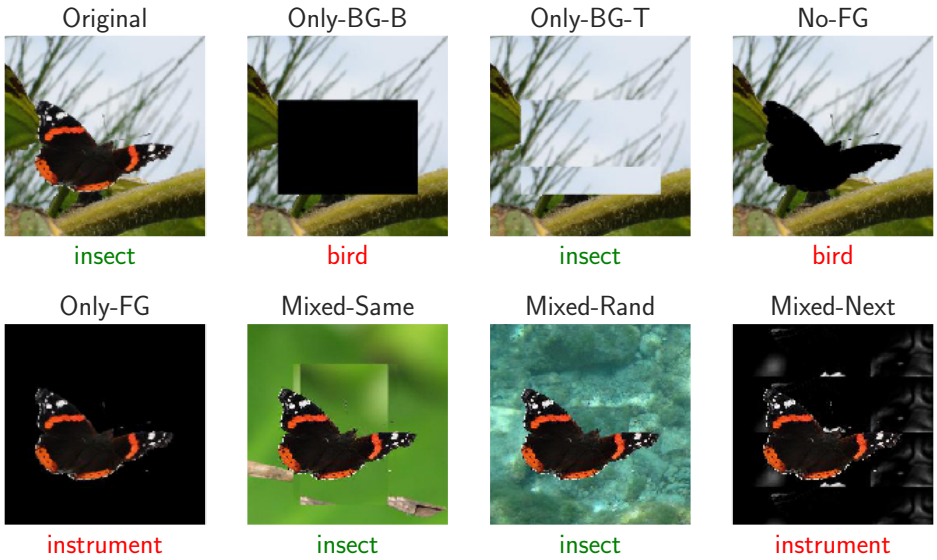

Figure 1: Variations of the synthetic dataset ImageNet-9, as described in Table 1. We label each image with its pre-trained ResNet-50 classification—green, if corresponding with the original label; red, if not. The model correctly classifies the image as "insect" when given: the original image, only the background, and two cases where the original foreground is present but the background changes. Note that, in particular, the model fails in two cases when the original foreground is present but the background changes (as in MIXED-NEXT or ONLY-FG).

2 METHODOLOGY

To properly gauge image backgrounds' role in image classification, we construct a synthetic dataset for disentangling background from foreground signal: ImageNet-9.

Base dataset: ImageNet-9. We organize a subset of ImageNet into a new dataset with nine coarse-grained classes and call it ImageNet-9 (IN-9) [1]. To create it, we group together ImageNet classes sharing an ancestor in the WordNet (Miller, 1995) hierarchy. We use coarse-grained classes because there are not enough images with annotated bounding boxes (which we need to disentangle backgrounds and foregrounds) to use the standard labels. The resulting IN-9 dataset is class-balanced and has 45405 training images and 4050 testing images. While we can (and do) apply our methods on the full ImageNet dataset as well, we choose to focus on this coarse-grained version of ImageNet because of its higher-fidelity images. We describe the dataset creation process in detail and discuss the advantages of focusing on IN-9 in Appendix A.

Variations of ImageNet-9 From this base set of images, which we call the ORIGINAL version of IN-9, we create seven other synthetic variations designed to understand the impact of backgrounds. We use both rectangular bounding boxes and the foreground segmentation algorithm GrabCut (Rother et al., 2004), as implemented in OpenCV, to disentangle backgrounds and foregrounds. We visualize these variations in Figure 1, and provide a detailed reference in Table 1. These subdatasets of IN-9 differ only in how they process the foregrounds and backgrounds of each constituent image.

Larger dataset: IN-9L We finally create a dataset called IN-9L that consists of all the images in ImageNet corresponding to the classes in ORIGINAL (rather than just the images that have associated bounding boxes). This dataset has about 180k training images in total. We leverage this larger dataset to train better generalizing models, and prefer to analyze models trained on IN-9L whenever possible.

[1] These classes are `dog`, `bird`, `vehicle`, `reptile`, `carnivore`, `insect`, `instrument`, `primate`, and `fish`.

Table 1: The 8 modified subdatasets created from ImageNet-9, which are visualized in Figure 1. The foreground detection method refers to how the pixels corresponding to the foreground are found. GrabCut refers to the foreground segmentation algorithm implemented in OpenCV. Random backgrounds in the last three datasets are taken from ONLY-BG-T. For more details see Appendix A.

Name	Foreground	Background	Foreground Detection Method
ORIGINAL	Unmodified	Unmodified	—
ONLY-BG-B	Black	Unmodified	Bounding Box
ONLY-BG-T	Tiled background	Unmodified	Bounding Box
NO-FG	Black	Unmodified	GrabCut
ONLY-FG	Unmodified	Black	GrabCut
MIXED-SAME	Unmodified	Random BG of the same class	GrabCut
MIXED-RAND	Unmodified	Random BG of a random class	GrabCut
MIXED-NEXT	Unmodified	Random BG of the next class	GrabCut

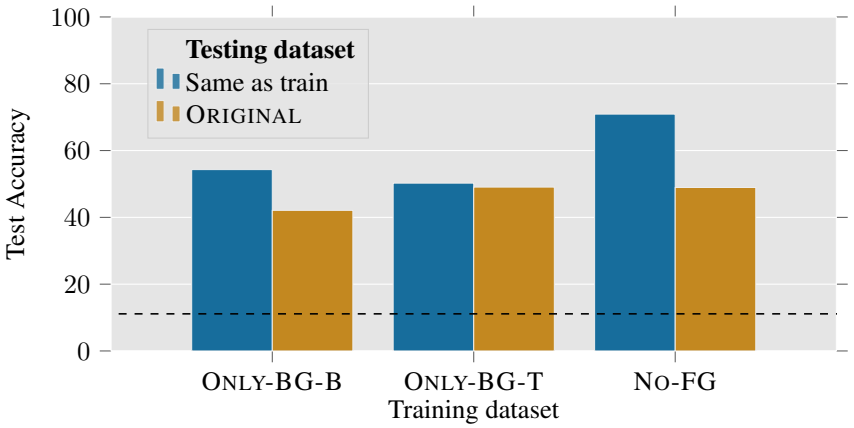

Figure 2: We train models on each of the "background-only" datasets, then evaluate each on its corresponding test set as well as the ORIGINAL test set. Even though the model only learns from background signal, it achieves (much) better than random performance on *both* the corresponding test set and ORIGINAL. Here, random guessing would give 11.11% (the dotted line).

3 QUANTIFYING RELIANCE ON BACKGROUND SIGNALS

With ImageNet-9 in hand, we now assess the role of image backgrounds in classification.

Backgrounds suffice for classification. Prior work has found that models are able to make accurate predictions based on backgrounds alone; we begin by directly quantifying this ability. Looking at the ONLY-BG-T, ONLY-BG-B, and NO-FG datasets, we find (cf. Figure 2) that models trained on these background-only training sets generalize reasonably well to both their corresponding test sets and to unmodified images from the ORIGINAL test set (around 40-50% for every model, far above the baseline of 11% representing random classification). Our results confirm that image backgrounds contain signal that models can accurately classify standard images with.

Models exploit background signal for classification. We discover that models can misclassify due to background signal, especially when the background class does not match that of the foreground. As a demonstration, we study model accuracies on the MIXED-RAND dataset, where image backgrounds are randomized and thus provide no information about the correct label. By comparing test accuracies on MIXED-RAND and MIXED-SAME [2], where images have class-consistent backgrounds, we can measure classifiers' dependence on the correct background. We denote the resulting accuracy gap between MIXED-SAME and MIXED-RAND as the BG-GAP; this difference represents the drop in

[2] MIXED-SAME controls for artifacts from image processing presented in MIXED-RAND. For further discussion, see Appendix D.

Table 2: Performance of state-of-the-art computer vision models on select test sets of ImageNet-9. We include both pre-trained ImageNet models and models of different architectures that we train on IN-9L. The BG-GAP is defined as the difference in test accuracy between MIXED-SAME and MIXED-RAND and helps assess the tendency of such models to rely on background signal. Architectures are sorted by their test accuracies on ImageNet and ORIGINAL for pre-trained and IN-9L-trained models, respectively. Shaded in grey are the two architectures that can be directly compared across datasets (ResNet-50 and Wide-ResNet-50x2).

	Pre-trained on ImageNet					**Trained on IN-9L**				
Test dataset	MobileNet-v3	EfficientNet-b0	ResNet-50	WRN-50x2	DPN-92	AlexNet	ShuffleNet	ResNet-50	WRN-50x2	VGG16-BN
ImageNet	67.9%	77.2%	77.6%	78.5%	80.0%	——				
ORIGINAL	95.5%	96.1%	96.9%	96.6%	97.2%	86.7%	95.7%	96.3%	97.2%	97.6%
ONLY-BG-T	16.3%	16.5%	17.4%	18.8%	17.6%	41.5%	43.6%	43.6%	45.1%	45.7%
MIXED-SAME	84.0%	86.2%	91.0%	88.3%	90.5%	76.2%	86.7%	89.9%	90.6%	91.0%
MIXED-RAND	73.2%	76.3%	84.3%	81.4%	86.1%	54.2%	69.4%	75.6%	78.0%	78.0%
BG-gap	10.8%	9.9%	6.7%	6.9%	4.4%	22.0%	17.3%	14.3%	12.6%	13.0%

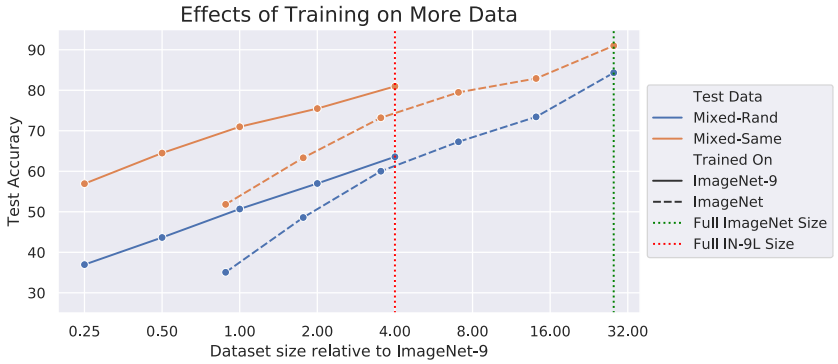

Figure 3: We compare test accuracies on MIXED-SAME and MIXED-RAND and observe that training with more data reduces the BG-GAP (BG-GAP measures the effect of backgrounds on model predictions). While this trend is true for models trained on both IN-9 and ImageNet, the trend is more noticeable on ImageNet, especially when the training dataset size is largest. In particular, note that the gap decreases much more on the right side of the graph.

model accuracy due to changing the class signal from the background. In Table 2, we observe a BG-GAP of 13-22% and 4-11% for models trained on IN-9L and ImageNet, respectively, suggesting that backgrounds often mislead state-of-the-art models *even when the correct foreground is present*.

More Training Data can reduce the BG-GAP. Our results indicate that ImageNet-trained models are less dependent on backgrounds than their IN-9L-trained counterparts—they have a smaller (but still significant) BG-GAP, and perform worse when predicting solely based on backgrounds (i.e., on the ONLY-BG-T dataset). We explore two ways that ImageNet differs from IN-9L to understand this phenomena—ImageNet has (a) more datapoints than IN-9L, and (b) a more fine-grained class structure. Figure 3 shows that more training data reduces the BG-GAP, particularly when the training dataset size approaches the size of ImageNet. This indicates that training on much more data (and thus, more backgrounds) can reduce (but not eliminate) the effect of backgrounds on model predictions. An ablation study of ImageNet's more fine-grained class structure does not find strong evidence supporting its helpfulness (cf. Appendix B).

Models are vulnerable to adversarial backgrounds. To understand how worst-case backgrounds impact models' performance, we evaluate model robustness to adversarially chosen backgrounds. We

find that 88% of foregrounds are susceptible to such backgrounds; that is, for these foregrounds, there is a background that causes the classifier to classify the resulting foreground-background combination as the background class. For a finer grained look, we also evaluate image backgrounds based on their attack success rate (ASR), i.e., how frequently they cause models to predict the (background) class in the presence of a conflicting foreground class. As an example, Figure 4 shows the five backgrounds with the highest ASR for the insect class—these backgrounds (extracted from insect images in ORIGINAL) fool a IN-9L-trained ResNet-50 model into predicting insect on up to 52% of non-insect foregrounds. We plot a histogram of ASR over all insect backgrounds in Figure 24 of the Appendix—it has a long tail. Similar results are observed for other classes as well (cf. Appendix G).

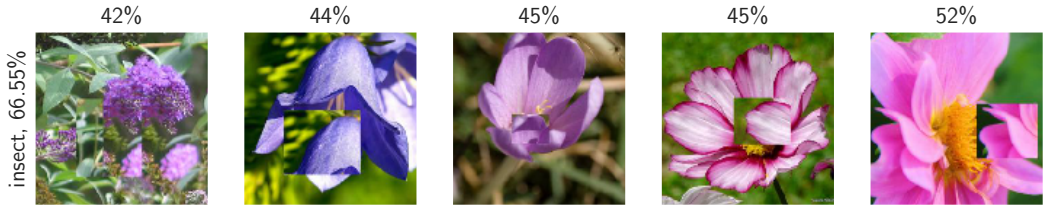

Figure 4: The adversarial backgrounds that most frequently fool IN-9L-trained models into classifying a given foreground as insect, ordered by the percentage of foregrounds fooled. The total portion of images that can be fooled (by any background from this class) is 66.55%.

Training on MIXED-RAND reduces background dependence. Next, we explore how to reduce models' dependence on background. To this end, we train models on MIXED-RAND, a synthetic dataset where background signals are decorrelated from class labels. As we would expect, MIXED-RAND-trained models extract less signal from backgrounds: evaluation results show that MIXED-RAND models perform poorly (15% accuracy—barely higher than random) on datasets with only backgrounds and no foregrounds, (ONLY-BG-T or ONLY-BG-B).

Indeed, such models are also more accurate on datasets where backgrounds do not match foregrounds. In Figure 5, we observe that a MIXED-RAND-trained model has 17.3% higher accuracy than its ORIGINAL-trained counterpart on MIXED-RAND, and 22.3% higher accuracy on MIXED-NEXT, a dataset where background signals class-consistently mismatch foregrounds. (Recall that MIXED-NEXT images have foregrounds from class y mixed with backgrounds from class $y + 1$, labeled as class y.) The MIXED-RAND-trained model also has little variation (at most 3.8%) in accuracy across all five test sets that contain the correct foreground.

Qualitatively, the MIXED-RAND-trained model also appears to place more relative importance on foreground pixels than the ORIGINAL-trained model; the saliency maps of the two models in Figure 6 show that the MIXED-RAND-trained model's saliency maps highlight more foreground pixels than those of ORIGINAL-trained models.

A fine grained look at dependence on backgrounds. We now analyze models' reliance on backgrounds at an image-by-image level and ask: for which images does introducing backgrounds help or hurt classifiers' performance? To this end, for each image in ORIGINAL, we decompose how models use foreground and background signals by examining classifiers' predictions on the corresponding image in MIXED-RAND and ONLY-BG-T. Here, we use the MIXED-RAND and ONLY-BG-T predictions as a proxy for which class the foreground and background signals (alone) point towards, respectively. We categorize each image based on how its background and foreground signals impact classification; we list the categories in Table 3 and show the counts for each category as a histogram per classifier in Figure 7. Our results show that while few backgrounds induce misclassification (see Appendix H for examples), a large fraction of images require backgrounds for correct classification—approximately 35% on the ORIGINAL trained classifiers, as calculated by combining the "BG Required" and "BG+FG Required" categories.

Further insights derived from IN-9 are discussed in the Appendix D. We focus on key findings in this section, but also include more comprehensive results and examples of other questions that can be explored by using the toolkit of IN-9 in the Appendix.

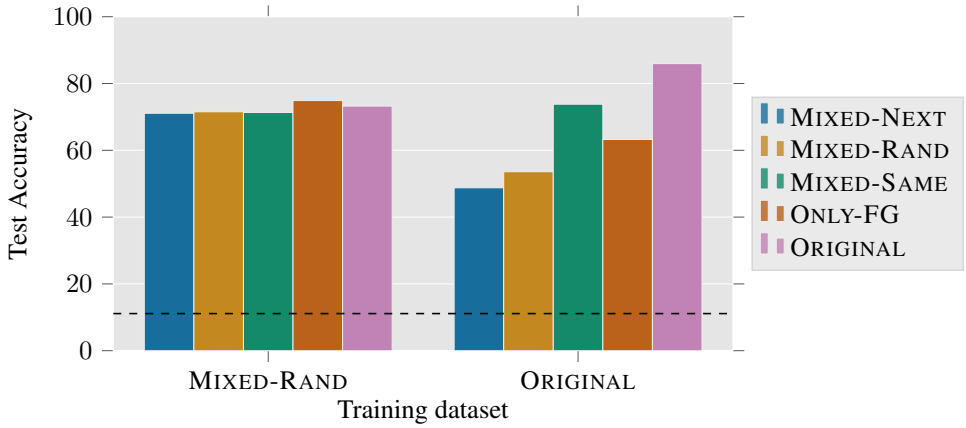

Figure 5: We compare the test performance of a model trained on the synthetic MIXED-RAND dataset with a model trained on ORIGINAL. We evaluate these models on variants of IN-9 that contain identical foregrounds. For the ORIGINAL-trained model, test performance decreases significantly when the background signal is modified during testing. However, the MIXED-RAND-trained model is robust to background changes, albeit at the cost of lower accuracy on images from ORIGINAL.

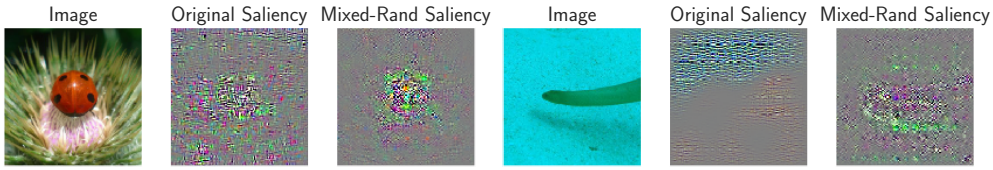

Figure 6: Saliency maps for the the ORIGINAL and MIXED-RAND models on two images. As expected, the MIXED-RAND model appears to place more importance on foreground pixels.

Table 3: Prediction categories we study for a given image-model pair. For a given image, a model can make differing predictions based on the presence or absence of its foreground/background. We label each possible case based on how the background classification relates to the original image classification and the foreground classification. To proxy classifying full images, foregrounds, and backgrounds separately, we classify on ORIGINAL, MIXED-RAND, and ONLY-BG-T (respectively). "BG Irrelevant" demarcates images where the foreground classification result is identical to that of the full image (in terms of correctness).

Label	Correct Prediction on Full Image	Correct Prediction on Foreground	Correct Prediction on Background
BG Required	✓	✗	✓
BG Fools	✗	✓	✗
BG+FG Required	✓	✗	✗
BG+FG Fools	✗	✓	✓
BG Irrelevant	✓/✗	✓/✗	—

4 BENCHMARK PROGRESS AND BACKGROUND DEPENDENCE

In the previous sections, we demonstrated that standard image classification models exploit signals from backgrounds. Considering that these models result from progress on standard computer vision benchmarks, a natural question is: *to what extent have improvements on image classification benchmarks resulted from exploiting background correlations?* And relatedly, *how has model robustness to misleading background signals evolved over time?*

As a first step towards answering these questions, we study the progress made by ImageNet models on our synthetic IN-9 dataset variations. In Figure 8 we plot accuracy on our synthetic datasets

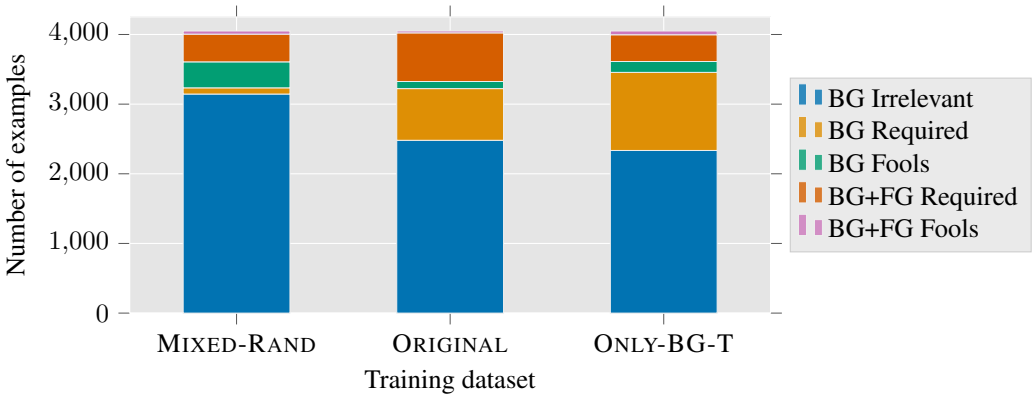

Figure 7: We categorize each test set image based on how a model classifies the full image, the background alone, and the foreground alone (cf. Table 3). The model trained on ORIGINAL needs the background for correct classification on 35% of images (measured by adding "BG Required" and "BG+FG Required), while a model trained on MIXED-RAND is much less reliant on background. The model trained on ONLY-BG-T requires the background most, as expected; however, the model often misclassifies both the full image and the background, so the "BG Irrelevant" subset is still sizable.

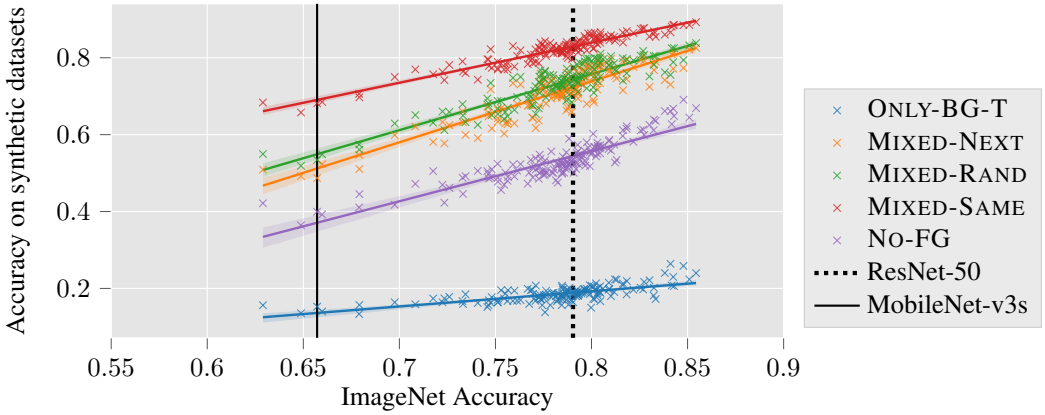

Figure 8: Measuring progress on each of the synthetic ImageNet-9 datasets with respect to progress on the standard ImageNet test set. Higher accuracy on ImageNet generally corresponds to higher accuracy on each of the constructed datasets, but the rate at which accuracy grows varies based on the types of features present in each dataset. Each pre-trained model corresponds to a vertical line on the plot—we mark ResNet-50 and MobileNet-v3s models for reference.

against ImageNet accuracy for each of the architectures considered. As evidenced by the lines of best fit in Figure 8, accuracy increases on the original ImageNet benchmark generally correspond to accuracy increases on all of the synthetic datasets. This includes the ONLY-BG datasets—indicating that models *do* improve at extracting correlations from image backgrounds.

Indeed, the ONLY-BG trend observed in Figure 8 suggests that either (a) image classification models can only attain their reported accuracies in the presence of background signals; or (b) these models carry an implicit bias towards features in the background, as a result of optimization technique, model class, etc.—in this case, we may need explicit regularization (e.g., through distributionally robust optimization (Sagawa et al., 2020) or related techniques) to obtain models invariant to these background features.

Still, models' *relative* improvement in accuracy across dataset variants is promising—models improve on classifying ONLY-BG-T at a slower (absolute) rate than MIXED-RAND, MIXED-SAME and MIXED-NEXT. Furthermore, the performance gap between the MIXED datasets and the others (most notably, between MIXED-RAND and MIXED-SAME; between MIXED-NEXT and MIXED-RAND;

and consequently between MIXED-NEXT and MIXED-SAME) trends towards closing, indicating that models not only are becoming better at using foreground features, but also are becoming more robust to misleading background features (MIXED-RAND and MIXED-NEXT).

Overall, the accuracy trends observed from testing ImageNet models on our synthetic datasets reveal that better models (a) are capable of exploiting background correlations, but (b) are increasingly robust to changes in background, suggesting that invariance to background features may not necessarily come at the cost of benchmark accuracy.

5 RELATED WORK

We focus our discussion here on the works most closely related to ours, specifically those investigating contextual bias or background dependence in computer vision (for a more extensive survey of and explicit comparison to prior work, see Appendix E). Prior work has explored the more general phenomenon of contextual bias (Torralba & Efros, 2011; Khosla et al., 2012; Myung Jin et al., 2012; Shetty et al., 2019), including studying its prevalence and developing methods to mitigate it. For image backgrounds specifically, prior works show that background correlations can be predictive (Torralba, 2003), and that backgrounds can influence model decisions to varying degrees (Zhang et al., 2007; Ribeiro et al., 2016; Beery et al., 2018; Rosenfeld et al., 2018; Barbu et al., 2019; Sagawa et al., 2020). These works all focus on different datasets that are much smaller than ImageNet.

The work most similar to ours is that of Zhu et al. (2017), who also analyze ImageNet classification (focusing on the older, AlexNet model). They find that the AlexNet model achieves small but non-trivial test accuracy on a dataset similar to our ONLY-BG-B dataset. While sufficient for establishing that backgrounds can be used for classification, this dataset also introduces biases by adding large black rectangular patches to all of the images (which our ONLY-BG-T dataset fixes). In comparison to Zhu et al. (2017) and the prior works mentioned earlier, we: (a) create a more extensive toolkit that allows us to measure not just model performance without foregrounds but also the relative influence of foregrounds and backgrounds on model predictions; (b) control for the effect of image artifacts via the MIXED-SAME dataset; (c) study model robustness to adversarial backgrounds; (d) study a larger and more recent set of classifiers (He et al., 2016; Zagoruyko & Komodakis, 2016; Tan & Le, 2019); and (e) show how improvements they give on ImageNet relate to background dependence.

6 DISCUSSION AND CONCLUSION

In this work, we study the extent to which classifiers rely on image backgrounds. To this end, we create a toolkit for measuring the precise role of background and foreground signal that involves constructing new test datasets that contain different amounts of each. Through these datasets we establish both the usefulness of background signal and the tendency of our models to depend on backgrounds, even when relevant foreground features are present. Our results show that our models are not robust to changes in the background, either in the adversarial case, or in the average case.

As most ImageNet images have human-recognizable foreground objects, our models appear to rely on background more than humans on that dataset. The fact that models can be fooled by adversarial background changes on 88% of all images highlights how poorly computer vision models may perform in an out-of-distribution setting. However, contextual information like the background can still be useful in certain settings. After all, humans do use backgrounds as context in visual processing, and the background may be necessary if the foreground is blurry or distorted (Torralba, 2003). Therefore, reliance on background is a nuanced question that merits further study.

On one hand, our findings provide evidence that models succeed by using background correlations, which may be undesirable in some applications. On the other hand, we find that advances in classifiers have given rise to models that use foregrounds more effectively and are more robust to changes in the background. To obtain even more robust models, we may want to draw inspiration from successes in training on the MIXED-RAND dataset (a dataset designed to neutralize background signal—cf. Table 1), related data-augmentation techniques (Shetty et al., 2019), and training algorithms like distributionally robust optimization (Sagawa et al., 2020) and model-based robust learning (Robey et al., 2020). Overall, the toolkit and findings in this work help us to better understand models and to monitor our progress toward the goal of reliable machine learning.

REFERENCES

Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman. Deep convolutional networks do not classify based on global object shape. In *PLOS Computational Biology*, 2018.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, C Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Neural Information Processing Systems*, 2019.

Sara Beery, Grant van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, 2018.

Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. In *IEEE Transactions on Image Processing*, 2004.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. In *arXiv preprint arXiv:2004.07780*, 2020.

Michael Harville, Gaile Gordon, and John Woodfill. Foreground segmentation using adaptive mixture models in color and depth. In *IEEE Workshop on Detection and Recognition of Events in Video*, 2001.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Andrew Ilyas, Shibani Santurkar, Logan Engstrom, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Jorn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. In *International Contemporary on Learning Representations (ICLR)*, 2019.

Saumya Jetley, Nicholas Lord, and Philip Torr. With friends like these, who needs adversaries? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A. Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision (ECCV)*, 2012.

George Miller. Wordnet: a lexical database for english. In *Communications of the ACM*, 1995.

Choi Myung Jin, Antonio Torralba, and Alan S. Willsky. Context models and out-of-context objects. In *Pattern Recognition Letters*, 2012.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining*, 2016.

Alexander Robey, Hamed Hassani, and George J. Pappas. Model-based robust deep learning. In *arXiv preprint arXiv:2005.10247*, 2020.

Amir Rosenfeld, Richard S. Zemel, and John K. Tsotsos. The elephant in the room. In *arXiv preprint arXiv:1808.03305*, 2018.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics*, 2004.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Rakshith Shetty, Mario Fritz, and Bernt Schiele. Adversarial scene editing: Automatic object removal from weak supervision. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Antonio Torralba. Contextual priming for object detection. In *International Journal of Computer Vision (IJCV)*, 2003.

Antonio Torralba and Alexei Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528, 2011.

Andrew W. Yip and Pawan Sinha. Contribution of color to face recognition. In *Perception*, 2002.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016.

Jianguo Zhang, Marcin Marszalek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. In *International Journal of Computer Vision (IJCV)*, 2007.

Zhuotun Zhu, Lingxi Xie, and Alan Yuille. Object recognition without and without objects. In *International Joint Conference on Artificial Intelligence*, 2017.