Edges to Shapes to Concepts: Adversarial Augmentation for Robust Vision

Aditay Tripathi[1*] Rishubh Singh[2] Anirban Chakraborty[1] Pradeep Shenoy[2]
[1] CDS, Indian Institute of Science [2]Google Research India
shenoypradeep@google.com

Abstract

Recent work has shown that deep vision models tend to be overly dependent on low-level or "texture" features, leading to poor generalization. Various data augmentation strategies have been proposed to overcome this so-called texture bias in DNNs. We propose a simple, lightweight adversarial augmentation technique that explicitly incentivizes the network to learn holistic shapes for accurate prediction in an object classification setting. Our augmentations superpose edgemaps from one image onto another image with shuffled patches, using a randomly determined mixing proportion, with the image label of the edgemap image. To classify these augmented images, the model needs to not only detect and focus on edges but distinguish between relevant and spurious edges. We show that our augmentations significantly improve classification accuracy and robustness measures on a range of datasets and neural architectures. As an example, for ViT-S, We obtain absolute gains on classification accuracy gains up to 6%. We also obtain gains of up to 28% and 8.5% on natural adversarial and out-of-distribution datasets like ImageNet-A (for ViT-B) and ImageNet-R (for ViT-S), respectively. Analysis using a range of probe datasets shows substantially increased shape sensitivity in our trained models, explaining the observed improvement in robustness and classification accuracy.

1. Introduction

A growing body of research catalogues and analyzes apparent failure modes of deep vision models. For instance, work on texture bias [1,7,12] suggests that image classifiers are overdependent on textural cues and fail against simple (adversarial) texture substitutions. Relatedly, the idea of simplicity bias [25] captures the tendency of deep models to use weakly predictive "simple" features such as color or texture, even in the presence of strongly predictive complex features. In psychology & neuroscience, too, evidence suggests that deep networks focus more on "local" features rather than global features and differ from human behavior in related tasks [19]. More broadly speaking, there is a mismatch between the cognitive concepts and associated world knowledge implied by the category labels in image datasets such as Imagenet and the actual information content made available to a model via one-hot vectors encoding these labels. In the face of under-determined learning problems, we need to introduce inductive biases to guide

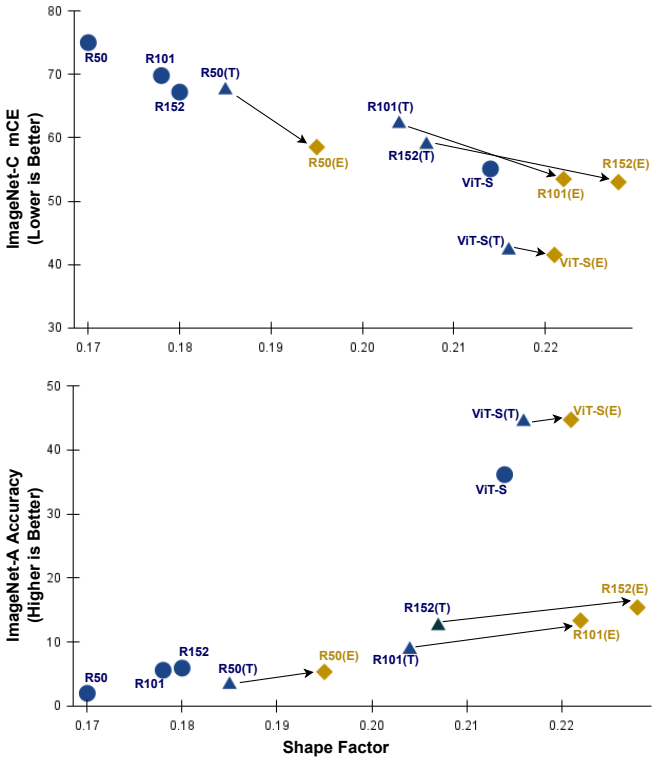

Figure 1. **Comparison of the models on robustness and shape-bias.** The shape factor gives the fraction of dimensions that encode shape cues [17]. Backbone(T) denotes texture shape debiased (TSD) models [21]. In comparison, ELEAS denoted by Backbone(E) is more shape biased and shows better performance on ImageNet-C and ImageNet-A datasets.

*Work done at Google Research India.

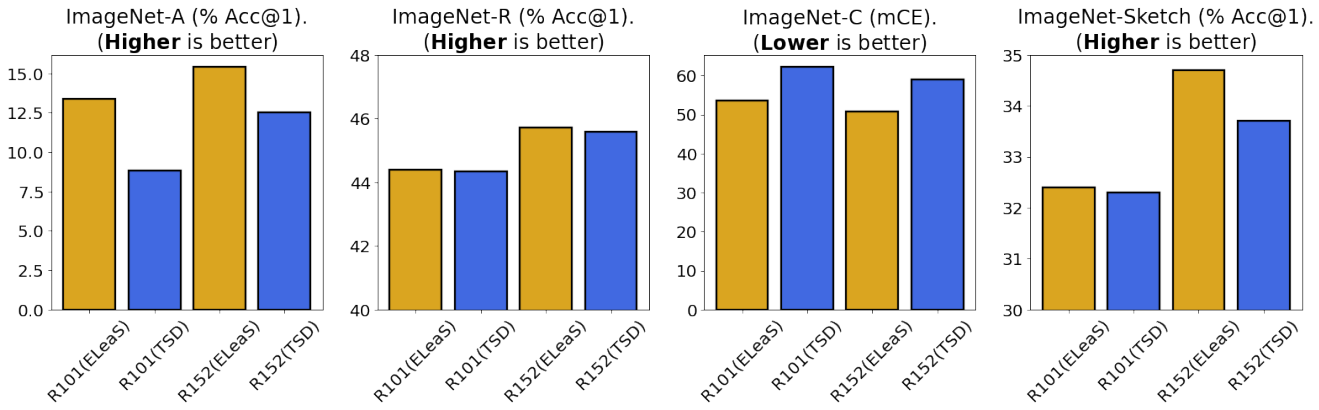

Figure 2. **Representative performance comparison** of our model, with the 'Debiased' models (TSD [21]) on ImageNet-A, R, C, and Sketch datasets. The models trained using ELEAS show improved performance on out-of-distribution robustness datasets. The large performance improvement on the ImageNet-A dataset indicates better robustness to natural adversarial examples.

the learning process. To this end, Geirhos et al. [7] proposed a data augmentation method wherein the texture of an image was replaced with that of a painting through stylization. Follow-on work improved upon this approach by replacing textures from other objects (instead of paintings) and teaching the model to separately label the outer shape and the substituted texture according to their source image categories [21]. Both these approaches discourage overdependence on textural features in the learned model; however, they do not explicitly incentivize shape recognition.

We propose a lightweight *adversarial augmentation* technique ELEAS (**E**dge **Lea**rning for **S**hape sensitivity) that is designed to increase shape sensitivity in vision models. Specifically, we augment a dataset with superpositions of random pairs of images from the dataset, where one image is processed to produce an edge map, and the other is modified by shuffling the location of image patches within the image. The two images are superposed using a randomly sampled relative mixing weight (similar to Mixup [30]), and the new superposed image is assigned the label of the edgemap image. ELEAS is designed to specifically incentivize not only edge detection but shape sensitivity: 1) classifying the edgemap image requires the model to extract and exploit edges – the only features available in the image, 2) distinguishing the edgemap object category from the superposed shuffled image requires the model to distinguish the overall edgemap object shape (relevant edges) from the shuffled image edges (irrelevant edges, less likely to be "shape like"). We perform extensive experiments over a range of model architectures, image classification datasets, and probe datasets, comparing ELEAS against recent baselines. Figure 1 provides a small visual sample of our findings and results; across various models, a measure of *shape sensitivity* [17] correlates very strongly with measures of *classifier robustness* [11, 12] (see Results for more details), validating the shape sensitivity inductive bias. In addition, for a number of model architectures, models trained with ELEAS significantly improve both measures compared to the previous SOTA data augmentation approach [21].

Summing up, we make the following contributions:

- We propose an adversarial augmentation technique, ELEAS, designed to incentivize shape sensitivity in vision models. Our augmentation technique is lightweight (needing only an off-the-shelf edge detection method) compared to previous proposals that require expensive GAN-based image synthesis [7, 21].
- In experiments, ELEAS shows increased shape sensitivity on a wide range of tasks designed to probe this property. Consequently, we obtain increased accuracy in object classification with a 6% improvement on ImageNet-1K classification accuracy for ViT-Small among others.
- ELEAS shows high generalizability and out of distribution robustness with 14.2% improvement in ImageNet-C classification performance and 5.89% increase in shape-bias for Resnet152.

2. Related work

Texture bias in vision models. In [1, 7], convolutional neural networks are shown to be more sensitive towards the texture present in the image to classify the object correctly. Further, in order to mitigate the texture bias, they have suggested a training strategy where they have used modified images with random texture along with the natural images during training. However, utilizing images with conflicting textures leads to an unnatural shift in data distribution, leading to drops in performance on natural images. Instead, sim-

ple and naturalistic data augmentations strategies can also lead to an increase in shape bias of the CNNs without much loss in performance on natural images [13]. An increase in shape-bias in the CNN models is associated with an increase in the robustness of the models [7, 26], however [22] suggested that a more balanced shape or texture biased models also leads to increase in the model robustness. The study of texture bias in vision models has been extended to the vision transformers (ViT) [3], where the transformer models are found to be less texture biased than the convolutional models [27].

A measure of shape/texture bias is proposed in [7], where they used images with conflicting textures obtained using style-transfer methods. However, this measure ignores a large portion of test images and is biased toward models trained using images modified by style-transfer methods. In [14], Hermann et al. proposed a linear classification layer in order to measure the shape decodability of the representations learned by the model. A more fine-grained shape evaluation measure was proposed in [18] where they compute the number of dimensions in the image representation that correlate with the shape features. Further, they proposed a shallow read-out module that takes feature representations from the model and predicts a segmentation map.

Data Augmentation for improving shape-bias In [2, 7], authors suggested a data augmentation strategy in order to increase the shape bias in the trained models. They proposed the Stylized-ImageNet dataset, where they strip each image of its original texture and replace it with the style of a randomly selected painting through AdaIN style transfer [16]. During training, the model is trained to predict the category corresponding to the shape in the model, ignoring the texture, thus leading to an increase in the shape bias of the models. However, instead of changing the distribution of training images through style transfer, naturalistic data augmentation techniques such as can also lead to improvement in the shape bias of the model [13].

3. Methodology

We now describe the training strategy used in ELEAS to increase the shape sensitivity of deep image classifiers.

Let $\mathcal{I}$, and $\mathcal{C}$ be the set of all images and their corresponding categories in the dataset. In standard classifier training, a classifier Θ is trained to predict the category $c \in \mathcal{C}$ of the image $i \in \mathcal{I}$. However, the image classifiers trained in this way are sensitive to the texture present in the natural images. In this work, we propose to use edge maps along with textures from natural images to increase the shape sensitivity of image classifiers.

Obtaining shapes and textures: We approximate the shapes of objects by extracting edge-maps from images and natural textures by shuffling patches within images. In particular, for each image $i \in \mathcal{I}$, an edge-map is constructed using an edge detection kernel (i.e., Laplacian kernel) to produce the set of all edge-maps or "shapes" $\mathcal{S}$. Similarly, a "texture" dataset $\mathcal{T}$ is generated by first dividing an image in $\mathcal{I}$ into 4×4 patches and then randomly shuffling the patches within the image to obtain a patch-shuffled image.

Generating augmentations: We superimpose randomly selected pairs of images $s \in \mathcal{S}$ and $t \in \mathcal{T}$ to create new images, or augmentations, to add to the training dataset. Each such augmented image is assigned the label of the shape image s. The superimposed image is obtained as follows:

$$i_s = \lambda * t + (1 - \lambda) * s \tag{1}$$

where λ is drawn randomly from a $\texttt{Beta}(\alpha, \beta)$ distribution. The parameters α and β are chosen such that a higher weight is given to the samples from set T. The random weighing parameter λ introduces variation in the augmented image set and helps to obtain better edge-map classification. Let's call the set of superimposed images $\mathcal{B}$.[1]

Training procedure: Training proceeds via minibatch gradient descent, as is standard in current machine learning literature. We construct each mini-batch to have half of its images from the set of natural images $I \subset \mathcal{I}$ and the rest from the augmented dataset $B \subset \mathcal{B}$. During training, we minimize the cross-entropy loss on natural image samples as well as our augmentations. In order to carefully control the degree of shape sensitivity induced by this process, we compute a weighted mixture of cross-entropy loss on these two image sets denoted here by L:

$$L(I, B, y_I, y_B) = \eta * CE(I, y_I) + 1 - \eta * CE(B, y_B) \tag{2}$$

where η is varied to control the shape sensitivity, CE is the cross-entropy loss, y_I, y_B are the labels corresponding to the natural and augmented image sets I, B respectively. To predict augmented images correctly, the model needs to interpret the edge-map present in the superimposed sample while at the same time ignoring distracting edges and textures from the other superimposed image. In this manner, we induce shape sensitivity in learned classifiers. The minibatch-mixing strategy encourages the model to generalize learned representations across natural and augmented images, thereby improving shape bias (and consequently overall accuracy) on natural images as well.

4. Experimental setup

4.1. Models and training setup

We train convolutional neural networks (CNNs) and Vision Transformers (ViTs) using our methodology. Among the CNN models, similar to [21], we show results on ResNet50, ResNet101, and ResNet152. For training the

[1] Refer to Fig. 1 in the supplementary material for examples of ELEAS images.

Model	Method	IN-A(↑)	IN-R(↑)	IN-C(↓)	IN-Sketch(↑)	IN-1K(↑)
Resnet50	Vanilla	2.0	36.2	75.0	23.5	76.4
	TSD [21]	3.3	40.8	67.5	28.3	76.9
	ELEAS	**5.4**	**41.7**	**58.5**	**29.7**	**77.1**
Resnet101	Vanilla	5.6	39.3	69.8	27.1	78.0
	TSD [21]	8.8	44.3	62.2	32.3	**78.8**
	ELEAS	**13.4**	**44.4**	**53.5**	**32.4**	78.6
Resnet152	Vanilla	5.9	41.3	67.2	28.4	78.6
	TSD [21]	12.5	45.5	58.9	33.3	79.7
	ELEAS	**15.4**	**45.7**	**53.0**	**34.7**	79.0
ViT-S	Vanilla	16.6	36.1	55.1	33.2	74.6
	TSD [21]	27.4	44.4	42.2	32.4	76.4
	ELEAS	**28.3**	**44.7**	**41.5**	**34.7**	**80.6**
ViT-B	Vanilla	34.6	40.7	50.8	45.8	79.5
	ELEAS	**62.9**	**56.4**	**35.9**	**46.4**	**85.5**
ViT-L	Vanilla	63.4	63.3	33.1	52.7	85.8
	ELEAS	**67.4**	**65.9**	**29.5**	**54.1**	**86.2**

Table 1. **Performance comparison on the ImageNet and the robustness datasets.** The models trained using ELEAS show an improvement in the ImageNet performance along with better robustness. Except for IN-C the performance is measured in Accuracy@1 (**higher is better**). For IN-C, the performance is measured in mean corruption error (mCE) (**lower is better**). (**Refer to Sections 5.1 and 5.2.1**)

ResNet models, we supplemented ImageNet data with an equal number of augmented images. We trained them for 100 epochs with a starting learning rate of 0.2 which is reduced by a factor of 10 at the 30th, 60th and 90th epoch. Similar to TSD, while training the ResNet models, we also use auxiliary batch norm [29]. As ViTs are compute-intensive, we finetune ImageNet pretrained ViT models for 20k steps with a cosine learning rate schedule with a starting learning rate of 0.01. The stochastic gradient descent (SGD) with a momentum of 0.9 is used to train the models. We train all our models on 8 A100 GPUs with a batch size of 512 for ResNets, 256 for ViT-Small and ViT-Base and 128 for ViT-Large. We find that values $\alpha = 4$, $\beta = 1$, and $\eta = 0.65$ produce the best results for all models.

4.2. Datasets and evaluation protocol

The models are trained on the ILSVRC 2012 [24] dataset with 1.28M training and 50k validation images. We evaluate and compare our trained models to answer the following questions: (i) Does ELEAS lead to an increase in the shape-sensitivity of a model? (ii) Does the increased shape sensitivity result in better classification performance? (iii) Does the robustness of the models to distribution shift improve with increased shape sensitivity? and (iv) Does an increase in shape sensitivity lead to improvement in downstream tasks such as object detection and instance segmentation?

The shape sensitivity of the models is first evaluated using the metric proposed in [7]. The authors first created an image dataset called *cue-conflict*, using a style-transfer method to transfer texture from one image to another. They then defined shape bias as the fraction of the shape decision when the model predicts either the shape or the texture category. However, this metric ignores many a lot of images and is biased toward the methods that used Stylized-ImageNet during the training of the models. Further, authors in [17] proposed a method to quantify the number of dimensions in the image representation that encode the object's shape. They take a pair of images with similar semantic concepts (i.e., shape) and then count the number of neurons that encode that specific concept. They have used Pascal VOC [4] and Stylized Pascal VOC dataset to create pair of images with particular semantic concepts. We have used these metrics to evaluate the shape bias in this work. Further, similar to [17], we also evaluate the quality of the shape encoding from these models by performing the binary and semantic segmentation on the frozen representations. PASCAL VOC is image segmentation dataset [4, 8], with 10,582 training images and 1,449 validation images spanning across 20 object categories.

The out-of-distribution generalization of the trained models is evaluated on four publicly available datasets, which are, ImageNet-A [12], ImageNet-R [10], ImageNet-C [11], and ImageNet-Sketch [28]. The ImageNet-A dataset consists of real-world natural adversarial images and is a challenging classification dataset. The ImageNet-R dataset

Model	Method	Edge	Silhouette	Cue-conflict	Sketch	Stylized-IN
Resnet50	Vanilla	13.75	54.38	18.20	59.62	37.13
	TSD [21]	22.5	**55.62**	21.40	67.0	**56.13**
	ELEAS	**35.62**	54.37	**21.41**	**67.88**	46.0
Resnet101	Vanilla	23.85	49.37	19.92	63.12	41.75
	TSD [21]	31.25	51.87	**24.92**	70.12	**59.37**
	ELEAS	**45.00**	**61.25**	24.06	**71.88**	48.25
Resnet152	Vanilla	20.63	56.25	20.70	66.75	41.63
	TSD [21]	22.5	**58.125**	**25.31**	69.13	**57.67**
	ELEAS	**41.88**	56.88	23.83	**73.38**	48.75
ViT-S	Vanilla	25.0	26.25	22.96	49.12	44.25
	TSD [21]	22.5	**46.87**	**31.71**	68.62	**77.5**
	ELEAS	**34.38**	43.13	27.66	**69.63**	53.00
ViT-B	Vanilla	13.75	12.05	28.05	51.63	50.08
	ELEAS	**41.25**	**66.25**	**34.92**	**82.00**	**59.38**
ViT-L	Vanilla	55.62	68.75	40.39	83.75	65.50
	ELEAS	**75.00**	**69.38**	**42.11**	**85.88**	**66.50**

Table 2. **Performance comparison on out-of-distribution benchmark proposed in [6].** The proposed method significantly improves performance over vanilla ResNet models. Debiased models, trained on stylized images, show high performance on datasets utilizing stylized images for evaluation, such as Cue-Conflict and Stylized-IN. (**Refer to Section 5.2.2**)

consists of images with various renditions, such as cartoon art, DeviantArt, graphics, paintings, origami, etc., of the objects present in the original ImageNet dataset. The ImageNet-Sketch dataset consists of sketches of 1000 ImageNet object categories. Both ImageNet-R and ImageNet-Sketch are used to evaluate the out-of-distribution generalization capability of the models. The ImageNet-C dataset consists of images with varying degrees of artificial distortions like 'Gaussian Noise', 'Motion Blur', 'Speckle Noise', etc. Hence, it evaluates the robustness of the model to added distortions.

Besides using classification accuracy for most datasets, ImageNet-C's performance is measured using mean corruption error (mCE). Additionally, we evaluated the models on the out-of-distribution benchmark proposed in [6].

4.2.1 Baselines

We compare the performance of ELEAS with Shape-Texture Debiased (TSD) trained models [21], which focus on both shape and texture cues. TSD uses stylized ImageNet images with conflicting shape and texture cues, and the models are trained using supervision from both semantic cues. Among ViTs, only ViT-S is trained with TSD due to the large computation overhead involved in training ViT-B and ViT-L. In contrast, ELEAS learns the overall object shape in the presence of conflicting, less 'shape-like' edges, leading to increased shape sensitivity in the trained models.

5. Results and discussion

Our experiments study the impact of ELEAS in the following stages: we show classification accuracy gains on the ImageNet dataset; we measure the robustness of trained models to a range of prediction challenges; and we examine ELEAS's influence on shape-sensitivity of trained models and its effect on performance.

5.1. Classification

Along with the increase in shape bias, ELEAS also improves classification performance on the ImageNet dataset, in many cases by very large margins (Refer Table 1). For instance, we see $\mathbf{5.96\%}$ **and** $\mathbf{5.93\%}$ **absolute improvement** in ImageNet classification accuracy for ViT-S and ViT-B respectively, compared to vanilla baseline, and a smaller but still significant 4.22% gain in accuracy compared to TSD on VIT-S. We also see 0.6%, 0.62%, and 0.44% increase in performance for ResNet50, ResNet101, and ResNet152 models respectively.

5.2. Out-of-distribution robustness

5.2.1 Evaluation on ImageNet-A, R, C, and Sketch datasets.

As shown in Table 1, ELEAS significantly improves accuracy over both vanilla and TSD-trained models at these robustness challenges. For ImageNet-A, the considerable improvement (**4.54%** vs. TSD on ResNet101) indicates bet-

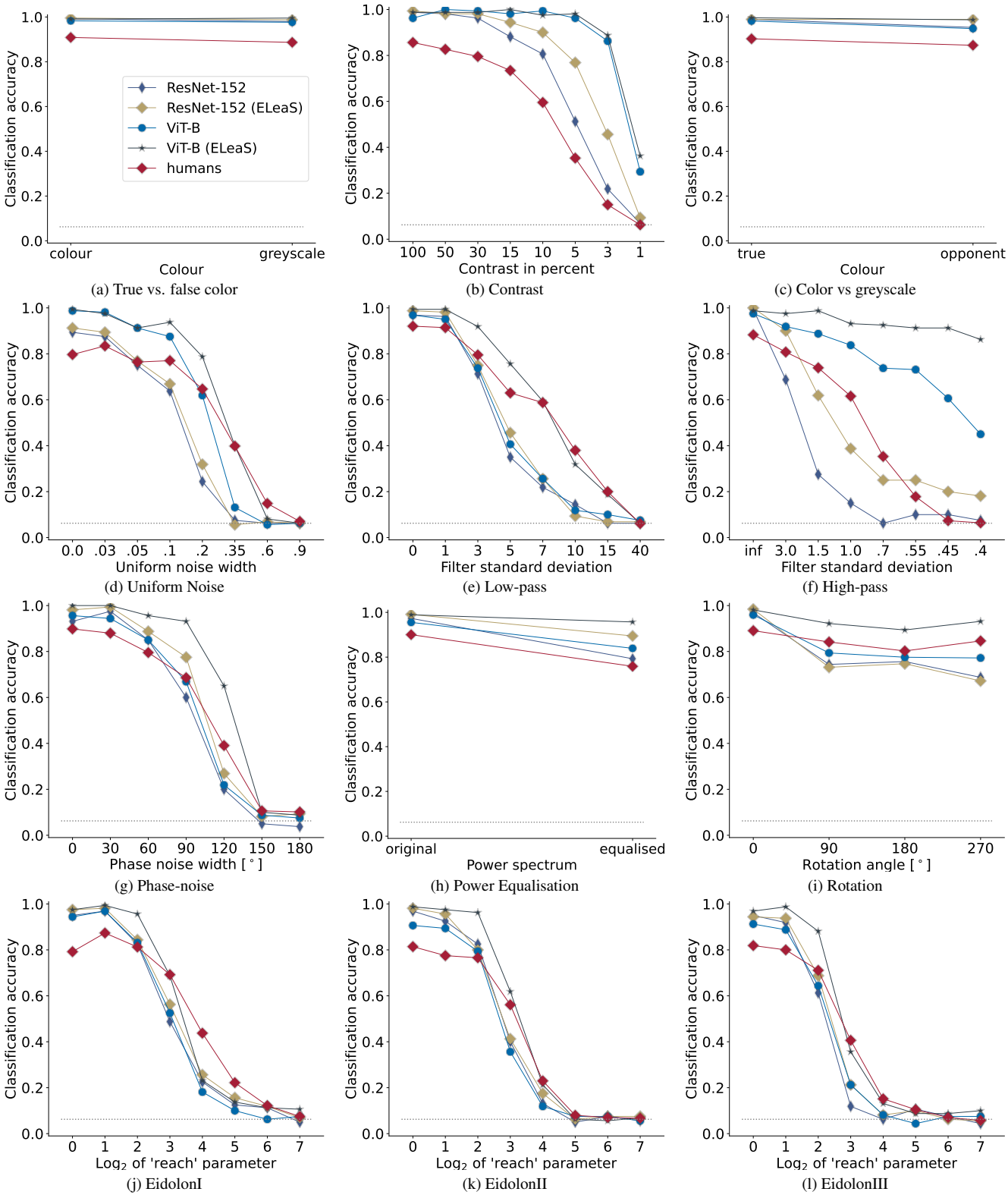

Figure 3. **Robustness comparison of** ELEAS **trained Resnet152 and ViT-B with the human subjects.** ELEAS leads to trained models which are more robust to added distortion and for many of the noise types it is more robust than the human subjects. (**Refer to Section 5.2.1**)

Model	Method	Shape-bias	Factor $\lvert z_k \rvert$	
			Shape dim	Texture dim
Res50	Vanilla	19.75	0.170	0.338
	TSD [21]	23.11	0.185	0.285
	ELEAS	23.53	0.195	0.299
Res101	Vanilla	22.44	0.178	0.323
	TSD [21]	27.69	0.204	0.256
	ELEAS	26.04	0.222	0.259
Res152	Vanilla	21.74	0.180	0.298
	TSD [21]	28.78	0.207	0.236
	ELEAS	27.86	0.228	0.249
ViT-S	Vanilla	33.33	0.214	0.210
	TSD [21]	36.56	0.216	0.213
	ELEAS	34.64	0.221	0.217
ViT-B	Vanilla	33.12	0.219	0.212
	ELEAS	37.90	0.239	0.205
ViT-L	Vanilla	47.04	0.216	0.221
	ELEAS	48.10	0.225	0.207

Table 3. **Comparison of the shape-bias of the models.** The vision models trained using our strategy shows an increase in the fraction of 'Shape' dimensions. However, the shape-bias metric proposed by Geirhos et al. [7] is biased towards the models trained using style-transfer datasets. (**Refer to Section 5.3**)

ter robustness to naturally occurring adversarial examples. Similarly, the improvement in mCE on ImageNet-C (**8.72%** vs. TSD on ResNet101) showcases robustness to added distortions of various degrees. Similarly, ELEAS shows robustness performance improvements on the ViT models. These large gains are driven primarily by ELEAS's ability to better encode object shape (Refer Figure 1 for evidence supporting this causal link.)

5.2.2 Evaluation on OOD benchmark [6]

We also evaluated trained models on the OOD benchmark proposed in [6]; results are in Table 2. The datasets in this benchmark consist of images from different domains such as 'Edge', 'Silhouette', 'Sketch', 'Stylized', and 'Cue-conflict.', which require stronger shape sensitivity for correct classification. The 'Cue-conflict' dataset is generated using iterative style transfer [5], where the texture from a source image is transferred to a target image. Since these images contain cues corresponding to different objects, it is challenging to classify them without good shape-sensitive representations. Similarly, the Stylized imagenet (SIN) is generated by replacing the texture present in an image with style from a randomly selected painting using the AdaIN style transfer method [16]. Models trained using ELEAS can better encode object shapes in images, leading to large performance improvements on these datasets over

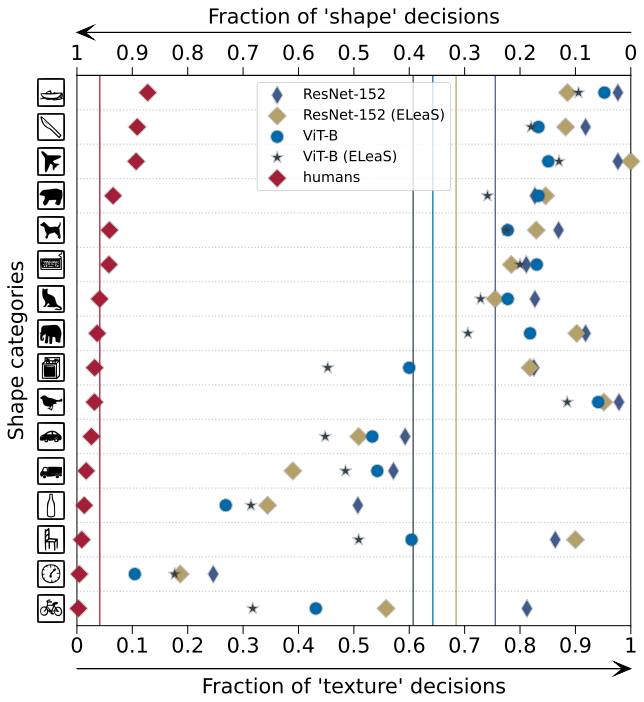

Figure 4. **Shape vs. texture bias for the vision models:** The images with conflicting shape and texture cues are used to calculate the shape bias of the trained models. The ELEAS trained models show improvement in shape-bias. We have shown results for ResNet152 and ViT-B in this figure.(**Refer to Section 5.3**)

the vanilla model. The better performance of TSD is expected as these models are trained using images from similar domains, i.e., the Stylized ImageNet dataset.

We further evaluated robustness to distortion using the proposed benchmarks (Refer to Figure 3), which shows improved robustness to added distortions of varying degrees for ELEAS.

5.3. Increasing shape sensitivity

We measure shape-sensitivity using the datasets and protocols described in Section 4.2; the results are in Tables 3 and 4. Models trained using ELEAS show a large 6.12% and 4.78% increase in shape-bias for ResNet152 and ViT-L models, respectively, compared to vanilla baselines and are comparable to the shape-bias of models trained using TSD. [2] We further added class-wise shape bias comparison with human subjects in Figure 4.

An alternative measure, the shape factor [17], counts the number of dimensions in the image representation that encodes shape. ELEAS substantially increases the number of

[2]Note that shape bias [7] has limitations: since it is defined in terms of a Stylized-ImageNet test set, the metric is biased towards methods that use SIN as a training augmentation. Further, it ignores images on which neither of the true shape/texture categories is predicted.

Model	Method	Bin	Sem
Resnet50	Vanilla	79.8	61.6
	TSD [21]	78.9	61.1
	ELEAS	79.2	61.9
Resnet101	Vanilla	80.4	63.4
	TSD [21]	80.0	64.7
	ELEAS	81.0	65.9
Resnet152	Vanilla	80.3	64.4
	TSD [21]	79.8	65.5
	ELEAS	81.0	66.6

Table 4. **Shape decodability of the learned representations:** To evaluate the amount of shape information, read-out modules are added to the frozen learned representation and used to predict binary (Bin) and semantic (Sem) segmentation masks [17], with performance measured by mIoU. The shape decodability of the models is evaluated using the Pascal VOC dataset. (**Section 5.3**)

shape dimensions in trained models (Table 3). We further evaluated shape decodability from learned representations by predicting the binary and semantic mask of the object in the image (Table 4). The segmentation is performed by adding a three-layer readout module on top of the learned representations. The read-out module is trained on PASCAL VOC 2012 dataset [4].

Resnet(TSD) [21] shows a significant worsening of binary segmentation performance for all models and semantic segmentation performance for ResNet50, suggesting that their learned representations cannot predict pixel-wise object categories. ELEAS-trained model shows an increase in Semantic segmentation performance, indicating better decodability of pixel-wise semantic categories of objects.

5.4. Instance Segmentation and Object detection performance

To evaluate the effectiveness of the shape-sensitive image representations in downstream tasks such as object detection and segmentation, we trained the MaskRCNN [9] model using ResNet101 as the backbone. Results are reported in Tables 5 for models trained on the MS-COCO dataset. The model trained using ELEAS outperformed other models by a significant margin in both object detection ($\approx$ 1.8 mAP) and instance segmentation tasks ($\approx$ 1.3 mAP), indicating that learned shape-sensitive representations can improve performance in these tasks. In contrast, TSD showed decreased performance even when compared to the Vanilla backbone.

5.5. Lightweightness of ELEAS

Compared to previous augmentation methods, ELEAS is lightweight. Edge maps and shuffled patch images can be precomputed and stored on disk, and obtaining adversarial augmented images requires only a few operations. In contrast, TSD [21] runs a GAN-based stylization to generate every augmented image, which is much slower.

Model	Object Detection			Instance Segmentation		
	mAP	AP@0.50	AP@0.75	mAP	AP@0.50	AP@0.75
Vanilla	39.87	60.21	43.33	36.35	57.39	38.79
TSD [21]	37.82	58.98	41.29	33.87	55.41	35.85
ELEAS	**41.65**	**61.83**	**45.43**	**37.63**	**58.99**	**40.33**

Table 5. **Effect of shape sensitivity on Object Detection and Segmentation performance.** The shapes-sensitive ResNet101 (i.e. ELEAS) backbone leads to improvement in object detection and instance segmentation performance. The models are evaluated on the *COCO-Val2017* dataset.

GAN-based stylized image generation takes almost twice the time (with all other factors constant – 8 A100 GPUs in our case) compared to computing all edge maps and shuffled patch images. Both ELEAS and TSD require new augmented images with different source and texture images every epoch, but GAN-based augmentation must be run every time, making it expensive. Therefore, the per epoch cost of ELEAS is negligible compared to TSD.

Further, once computed, the edge maps and shuffled patch images used by ELEAS can be reused to train other models, further reducing the average data generation cost.

6. Conclusion

We propose and evaluate ELEAS – a lightweight *adversarial augmentation* technique for image classification in order to induce a shape bias in learned models. Previous work in this area [7, 21] identified the need for such inductive biases due to the apparent dependence of deep models on textural features; however, their proposed augmentation techniques have drawbacks–first, they primarily sever the dependence on texture, while not explicitly enforcing processing and representation of shape, and second, the proposed augmentation process is very expensive. Although our work, too, proposes only data augmentation and does not come with guarantees on learned representations, the augmentations are simple to compute and are designed to encourage holistic shape representation. Extensive experimentation shows both the value of this inductive bias (shape factors correlate with model robustness and accuracy across models) and the substantial gains of our proposal, including over 5% absolute accuracy improvements on ImageNet using vision transformers. An interesting direction for future work is to explore richer ways of specifying "shape" to the models–for instance, using off-the-shelf segmentation or depth estimation models–and also to explore other inductive biases that can profitably be incorporated into vision models.

References

[1] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.*, 14(12), 2018.

[2] Sanghyuk Chun and Song Park. Styleaugment: Learning texture de-biased representations by style augmentation without pre-defined textures. *CoRR*, abs/2108.10549, 2021.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[5] Leon A. Gatys. *Texture synthesis and style transfer using perceptual image representations from convolutional neural networks*. PhD thesis, University of Tübingen, Germany, 2017.

[6] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems 34*, 2021.

[7] 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.

[8] Bharath Hariharan, Pablo Arbelaez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool, editors, *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 991–998. IEEE Computer Society, 2011.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society, 2017.

[10] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8320–8329. IEEE, 2021.

[11] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[12] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021.

[13] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19000–19015. Curran Associates, Inc., 2020.

[14] Katherine L. Hermann and Andrew K. Lampinen. What shapes feature representations? exploring datasets, architectures, and training. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[15] Minui Hong, Jinwoo Choi, and Gunhee Kim. Stylemix: Separating content and style for enhanced data augmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14857–14865, 2021.

[16] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1510–1519. IEEE Computer Society, 2017.

[17] Md Amirul Islam, Matthew Kowal, Patrick Esser, Sen Jia, Björn Ommer, Konstantinos G. Derpanis, and Neil Bruce. Shape or texture: Understanding discriminative features in {cnn}s. In *International Conference on Learning Representations*, 2021.

[18] Md. Amirul Islam, Matthew Kowal, Patrick Esser, Sen Jia, Björn Ommer, Konstantinos G. Derpanis, and Neil D. B. Bruce. Shape or texture: Understanding discriminative features in cnns. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[19] Georgin Jacob, R. T. Pramod, Harish Katti, and S. P. Arun. Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, 12, 2021.

[20] Sangjun Lee, Inwoo Hwang, Gi-Cheon Kang, and Byoung-Tak Zhang. Improving robustness to texture bias via shape-focused augmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4322–4330, 2022.

[21] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and cihang xie. Shape-texture debiased neural network training. In *International Conference on Learning Representations*, 2021.

[22] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan L. Yuille, and Cihang Xie. Shape-texture debiased neural network training. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[23] Chaithanya Kumar Mummadi, Ranjitha Subramaniam, Robin Hutmacher, Julien Vitay, Volker Fischer, and Jan Hen-

drik Metzen. Does enhanced shape bias improve neural network robustness to common corruptions? In *International Conference on Learning Representations*, 2021.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[25] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[26] Baifeng Shi, Dinghuai Zhang, Qi Dai, Zhanxing Zhu, Yadong Mu, and Jingdong Wang. Informative dropout for robust representation learning: A shape-bias perspective. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8828–8839. PMLR, 2020.

[27] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are convolutional neural networks or transformers more like human vision? *ArXiv*, abs/2105.07197, 2021.

[28] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

[29] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L. Yuille, and Quoc V. Le. Adversarial examples improve image recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 816–825. Computer Vision Foundation / IEEE, 2020.

[30] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.