Supplementary Material: Shape-biased CNNs are Not Always Superior in Out-of-Distribution Robustness

Xinkuan Qiu[1,3] Meina Kan[2,3] Yongbin Zhou[1,3,4] Yanchao Bi[5] Shiguang Shan[2,3,6]
[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China
[2] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[4] Nanjing University of Science and Technology, Nanjing 210094, China
[5] Beijing Normal University, Beijing 100875, China
[6] Peng Cheng Laboratory, Shenzhen 518055, China
qiuxinkuan@iie.ac.cn kanmeina@ict.ac.cn
zhouyongbin@njust.edu.cn ybi@bnu.edu.cn sgshan@ict.ac.cn

This supplementary material provides additional details on our work in three aspects. Section A elaborates on the design of Category Balanced ImageNet dataset. Section B presents further information on the training and evaluation process. Section C exhibits experiment details.

1. Details about Category Balanced ImageNet

To systematically investigate the effect of categorical factors (animal vs. artifact) on CNNs' shape and texture learning, we have meticulously designed this dataset to avoid disturbance from data imbalance.

1.1. Image Source

To avoid the image source becoming a confounding factor, all images in Category Balanced ImageNet were selected from a single dataset, namely ImageNet [3]. ImageNet is a suitable source for two main reasons. Firstly, ImageNet has a carefully designed hierarchy, based on which we can create similar hierarchies for animal and artifact to ensure fair comparisons. Secondly, ImageNet is a large dataset, and the subset we built is likely to be sufficiently large to conduct convincing experiments.

1.2. Dataset Hierarchy

Animal and artifact categories are designed to have similar hierarchical structures to ensure equal representations in the dataset for fair comparisons. The dataset hierarchy is shown in Figure 1. With this restriction, the dataset has a relatively small number of categories compared to ImageNet (128 vs. 1000). The number 128 is chosen according to the following rules.

The hierarchy consists of two levels only due to the complexity of ImageNet hierarchy [1], namely the fine-grained level and the coarse-grained level. Therefore, each image has two labels. **In this work, only fine-grained labels are used in the experiments**, while coarse-grained labels are only used to ensure that animal and artifact categories have similar hierarchical structures.

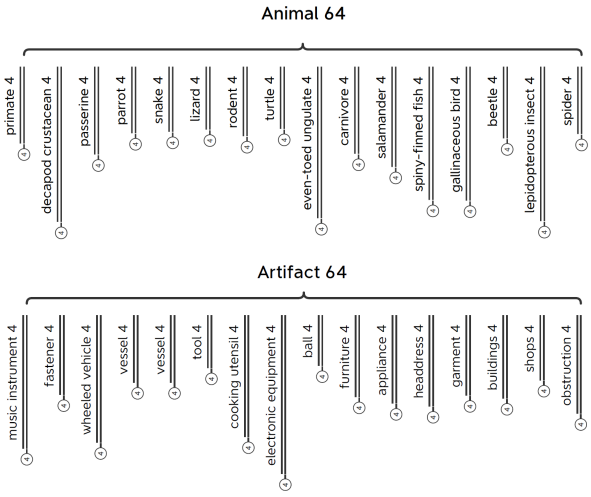

Figure 1. Hierarchy of Category Balanced ImageNet. The coarse level contains 16 animal categories and 16 artifact categories, each containing 4 fine-grained categories, resulting in a total of 128 fine-grained categories.

Fine-grained categories are selected from the 1000 categories of ImageNet2012. Since they are all leaf nodes in the hierarchy tree, it is reasonable to assume those categories have similar levels of granularity. Same as ImageNet, each fine category has about 1300 images. Choosing **coarse-grained categories** is a more elaborate work. Since all non-leaf nodes in the ImageNet hierarchy are valid

Coarse level	Coarse-to-fine ratio(M)	Max number of coarse categories(N)	Image number
phylum	-	2	-
class	-	7	-
order	3	18	70.2k
order	**4**	**16**	**83.2k**
order	5	14	91.0k
order	6	12	93.6k
order	7	9	81.8k
order	8	8	83.2k
order	9	6	70.2k
order	12	4	62.4k

Table 1. Illustration of candidates for the granularity setting of animal categories. Our decision is highlighted in bold.

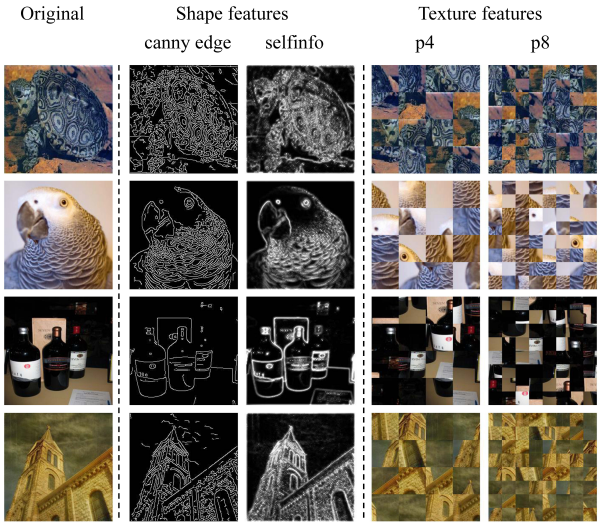

Figure 2. Illustrations of shape and texture features.

candidates, multiple factors are taken into consideration as shown in Table 1. Assuming we have N coarse categories, and each coarse category contains at least M fine categories, we want both N and M to be as large as possible. However, there is a trade-off between N and M as shown in Table 1. Taking the animal part as an example, we first select "order" as the suitable granularity level for coarse categories based on Linnaean taxonomy. "Phylum" and "class" are not appropriate, since the maximum number of coarse categories(N) would be 2 and 7 respectively, which is insufficient. While if the granularity level for coarse categories is finer than "order", such as "family", the difference in granularity between coarse and fine categories would be less apparent. Next, to decide the values of M and N, we gradually increase M and observe how N changes. We finally choose N as 16 and M as 4 according to the trade-off illustrated in Table 1. The artifact part is designed with similar considerations.

All categories in Category Balanced ImageNet are illustrated in Tables 2 for animals and Table 3 for artifacts, including the coarse-to-fine mapping.

2. Training and Evaluation Details

2.1. Shape and Texture Features

Shape and texture features play a crucial role in our main paper. For experiments presented in Table 1 of the main paper, these features are utilized for both model training and testing. For experiments presented in Table 2 of the main paper, they are employed to generate cue-conflicting images to assess texture bias. For experiments presented in Table 3 of the main paper, they are incorporated in joint training to obtain shape and texture-biased models. Visual examples of shape and texture features are provided in Figure 2.

2.2. Implementation Details

This work involves extensive experiments conducted under similar settings. To facilitate file management, we utilize MMpretrain [2], an open-source toolbox based on PyTorch [8], for training and testing our models. The training setup for models trained on different dataset is listed below.

Category-Balanced ImageNet. All models employ the default ResNet 50 architecture [5] with 100 training epochs. The following pre-processing steps are applied during training: resize (to 256x256), center crop (to 224x224), and channel-wise normalization. Models are trained using SGD with a momentum of 0.9 and weight decay of 0.0001. The learning rate is set as 0.025 per GPU, and decays by a factor of 10 at the 60th and 80th epochs. The batch size is 64.

Office-Home. Apart from the setting above, we employ the pretrained ResNet18 model to initialize training due to its limited dataset size.

CIFAR10. Models are trained using SGD with a momentum of 0.9 and weight decay of 0.0005 with 200 epochs. The learning rate is set as 0.025 per GPU, and decays by a factor of 5 at the 60th, 80th, 160th and 190th epochs. The batch size is set to 128.

Besides, the learning rate reduces correspondingly if comparable methods fail to converge. Also, the batch size and learning rate are reduced together correspondingly, for models trained with online augmentation, due to limited memory.

Index	Coarse categories	Fine categories
1	spider	tarantula, black widow, wolf spider, argiope aurantia
2	lepidopterous insect	monarch, admiral, ringlet, lycaenid
3	beetle	tiger beetle, dung beetle, ladybug, rhinoceros beetle
4	decapod crustacean	dungeness crab, spiny lobster, king crab, hermit crab
5	salamander	eft, axolotl, european fire salamander, common newt
6	turtle	terrapin, loggerhead, box turtle, mud turtle
7	lizard	green lizard, anole, banded gecko, gila monster
8	snake	ring snake, horned viper, thunder snake, boa constrictor
9	parrot	lorikeet, sulphur-crested cockatoo, macaw, african grey
10	passerine	goldfinch, jay, indigo bunting, water ouzel
11	gallinaceous bird	black grouse, quail, prairie chicken, ptarmigan
12	spiny-finned fish	puffer, anemone fish, lionfish, rock beauty
13	rodent	fox squirrel, marmot, porcupine, beaver
14	primate	madagascar cat, chimpanzee, howler monkey, guenon
15	even-toed ungulate	llama, hartebeest, warthog, ox
16	carnivore	samoyed, persian cat, black-footed ferret, brown bear

Table 2. Animal categories in Category-Balanced ImageNet

Index	Coarse categories	Fine categories
17	musical instrument	bassoon, steel drum, grand piano, sax
18	fastener	combination lock, nail, padlock, buckle
19	wheeled vehicle	racer, sports car, police van, jinrikisha
20	watercraft	liner, fireboat, submarine, catamaran
21	vessel	whiskey jug, beer bottle, beaker, mortar
22	tool	plunger, screwdriver, plane, hammer
23	cooking utensil	crock pot, teapot, coffeepot, spatula
24	electronic equipment	cd player, oscilloscope, monitor, computer keyboard
25	ball	soccer ball, rugby ball, basketball, punching bag
26	furniture	day bed, cradle, entertainment center, folding chair
27	home appliance	iron, espresso maker, washer, refrigerator
28	headdress	cowboy hat, mortarboard, shower cap, sombrero
29	garment	stole, sweatshirt, abaya, sarong
30	building	greenhouse, monastery, library, mosque
31	shop	tobacco shop, butcher shop, bakery, toyshop
32	barrier	worm fence, picket fence, grille, chainlink fence

Table 3. Artifact categories in Category-Balanced ImageNet

3. Experiment Details

3.1. Detailed results for models trained by individual shape or texture features

In our main paper, models are trained by the combined datasets of all shape features or all texture features. The results regarding individual shape (Selfinfo and Cannyedge) or texture (P4, P8 and P16) features are illustrated in Table 4 for i.i.d case and o.o.d case.

3.2. Details of texture bias evaluation for animals and artifacts

As animal and artifact categories demonstrate distinct characteristics, we guess that whether a model is shape or texture-biased heavily depends on the dataset categories. We further conduct experiments to investigate models' texture bias regarding animal and artifact categories. We train models on the original dataset and test on the cue-conflicting dataset, with images generated by adding up the shape and texture features of two randomly selected images. Texture bias is calculated as the ratio of the number of times

Train set	Test set	i.i.d case								o.o.d case							
		Shape			Texture				Ratio	Shape			Texture				Ratio
		SI	CE	avg	P4	P8	P16	avg		SI	CE	avg	P4	P8	P16	avg	
Animals	Animals	71	66	68	79	76	69	75	0.92	29	24	27	33	28	23	28	0.96
Artifacts	Artifacts	72	68	70	70	63	53	62	1.13	34	29	32	28	21	15	21	1.48
All	Animals	71	65	68	81	77	68	75	0.90	29	23	26	33	27	21	27	0.95
	Artifacts	72	69	71	72	66	55	64	1.10	32	28	30	28	21	15	21	1.41
	All	71	67	69	76	71	61	70	1.00	30	25	28	30	24	18	24	1.15

Table 4. Accuracy (%) of models trained and tested on individual shape/texture features in i.i.d and o.o.d cases.

Method	avg	Noise			Blur				Weather				Digital			
		Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
Vanilla ResNet	79.2	88.3	90.0	92.1	74.8	84.6	79.0	78.8	83.1	77.9	66.5	54.0	75.3	79.5	81.6	82.5
Cutout	79.7	93.2	94.1	96.6	73.7	86.6	79.9	77.7	82.3	77.2	65.0	53.7	73.9	79.8	81.0	80.5
Mixup	71.0	78.2	80.7	80.4	73.7	84.6	74.0	72.9	72.9	56.0	46.4	48.0	58.8	75.1	84.6	78.9
Cutmix	79.8	88.0	89.9	92.9	74.3	88.0	79.3	70.2	78.4	74.9	60.3	50.9	72.5	77.7	108.8	91.1
Patch Gaussian	74.8	67.0	71.6	71.0	75.5	85.5	80.1	76.8	84.9	78.3	65.0	54.8	74.6	79.9	77.7	78.8
Stylized IN	63.0	63.7	68.7	57.9	67.0	75.1	66.9	84.0	63.7	61.9	42.9	48.5	48.1	**68.2**	**59.4**	**69.1**
AutoAugment	68.9	64.3	63.8	63.9	74.4	86.6	77.9	86.8	73.6	64.2	43.8	41.5	**37.9**	89.8	85.0	79.9
Augmix	61.6	60.4	59.9	59.1	60.7	78.0	53.2	60.4	65.7	60.5	40.9	44.7	44.2	70.9	86.6	73.7
APR	58.9	**48.9**	**53.4**	**53.5**	61.4	83.3	57.1	72.6	**53.6**	**51.3**	**30.7**	40.3	42.3	81.5	72.6	81.4
STAR identi-heads	60.7	56.5	55.7	55.7	63.0	79.0	52.3	58.5	63.1	58.2	41.4	44.0	47.2	68.8	91.0	75.5
STAR(ours)	**57.4**	55.5	54.4	56.1	**55.3**	**73.7**	**47.4**	**53.8**	60.7	55.9	41.2	42.8	46.4	68.8	79.7	70.1

Table 5. The corruption errors (%) of comparable methods across 15 corruption scenarios. The best results are highlighted in bold and the second bests are underlined. Our proposed method consistently improves model robustness across diverse o.o.d scenarios.

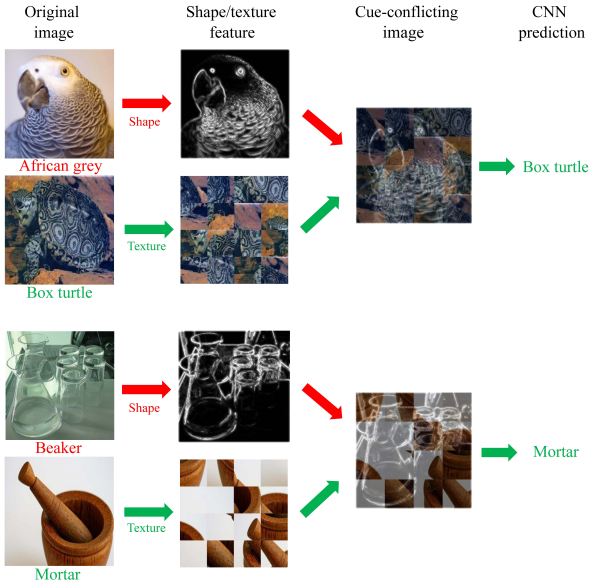

Figure 3. Illustrations of two cue-conflicting images and their classification results.

the model classifies images to their texture labels, to the number of times it classifies images to either shape or texture labels.

As proposed by Geirhos *et al*., researchers usually evaluate texture bias using a cue-conflicting test set [6, 10], which contains images generated by style transfer using shape information from one category and texture information from another.

However, this test set is not suitable for our experiments for several reasons. Firstly, it employs style transfer to generate images, which introduces potential biases as the pretrained generator may have learned shape and texture information from additional training datasets. Secondly, artifact texture is represented by repeated objects (*i.e.* a collection of bottles) while animal texture is represented by fur images. This discrepancy could act as a confounding factor for model biases. Thirdly, Geirhos's cue-conflicting dataset covers only approximately one-third of the categories in the training dataset, making it unclear whether the results can accurately represent the behavior of the entire model.

In this work, cue-conflicting images are generated by simply adding up the shape and texture features of two randomly selected images (i.e. cannyedge + p4, selfinfo + p16). In our approach, we avoid using additional datasets, employ unified generation methods for animals and artifacts, and ensure that all categories are covered during the testing stage. We provide visual examples of these cue-conflicting images in Figure 3.

As shown in Table 2 of the main paper, it is evident that models exhibit much higher texture bias on animal categories than on artifact categories. We believe that the differ-

ent levels of texture bias are related to the category's intrinsic discriminative feature. Extended experiments using architectures including AlexNet [7], Vgg16 [9] and VIT-B [4] are shown in Table 6, proving that our conclusions are architecture invariant (We only repeat the experiments on the setting where all 128 categories are used in training, with animal/artifact sub-datasets tested separately). These experiments reveal an essential fact for researchers investigating shape and texture bias: **categorical factors significantly impact the model's texture bias, and they should be considered seriously during texture bias analysis.** For instance, if a model has higher texture bias on dataset A than on dataset B, it is also likely that dataset A contains a higher proportion of animal images. **Besides, the recommendation of using category-balanced datasets is reinforced since shape-texture bias heavily depends on the training categories.**

Architecture	i.i.d case		o.o.d case	
	animal	artifact	animal	artifact
AlexNet	0.759	0.590	0.758	0.610
VGG16	0.785	0.669	0.740	0.620
VIT	0.925	0.780	0.846	0.665

Table 6. Texture bias ([0,1]) of various models.

3.3. Validation the role of H_t

As illustrated in Table 7, H_t shows higher accuracy for all blurs and elastic transformations over the other two heads.

Head	avg	defocus	glass	motion	zoom	elastic
H_s	62.0	58.1	49.8	68.0	65.5	68.7
H_t	**64.4**	**61.3**	**51.4**	**70.7**	**68.1**	**70.3**
H_d	63.5	60.0	50.5	69.9	67.3	69.7

Table 7. Accuracy (%) of H_s,H_t,H_d on test data with defocus, glass, motion, zoom blurs and elastic transformation.

3.4. Detailed results for individual corruption o.o.d scenarios

The complete results regarding individual o.o.d cases in ImageNet-C for all comparable methods are shown in Table 5. It can be seen that our proposed method consistently improves model robustness across diverse o.o.d scenarios. Especially for blurs, STAR outperforms other methods by a great margin, highlighting the advantages of using robust texture information extracted by the texture-biased head.

References

[1] Mike Bostock. Imagenet hierarchy. https://observablehq.com/@mbostock/imagenet-hierarchy, 2018. 1

[2] MMPreTrain Contributors. Openmmlab's pre-training toolbox and benchmark. https://github.com/open-mmlab/mmpretrain, 2023. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[6] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *In Advances in Neural Information Processing Systems*, 2020. 4

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 5

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *In Advances in neural information processing systems*, 32, 2019. 2

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[10] Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang Zhou, Zhongang Cai, Haiyu Zhao, Xianglong Liu, and Ziwei Liu. Delving deep into the generalization of vision transformers under distribution shifts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7277–7286, 2022. 4