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

Abstract

In recent years, Out-of-Distribution (o.o.d) Robustness has garnered increasing attention in Deep Learning, and shape-biased Convolutional Neural Networks (CNNs) are believed to exhibit higher robustness, attributed to the inherent shape-based decision rule of human cognition. In this work, we delve deeper into the intricate relationship between shape/texture information and o.o.d robustness by leveraging a carefully curated "Category-Balanced ImageNet" dataset. We find that shape information is not always superior in distinguishing distinct categories and shape-biased model is not always superior across various o.o.d scenarios. Motivated by these insightful findings, we design a novel method named Shape-Texture Adaptive Recombination (STAR) to achieve higher o.o.d robustness. A category-balanced dataset is firstly used to pretrain a debiased backbone and three specialized heads, each adept at robustly extracting shape, texture, and debiased features. Subsequently, an instance-adaptive recombination head is trained to adaptively adjust the contributions of these distinctive features for each given instance. Through comprehensive experiments, our proposed method achieves state-of-the-art o.o.d robustness across various scenarios such as image corruptions, adversarial attacks, style shifts, and dataset shifts, demonstrating its effectiveness.

1. Introduction

Convolutional Neural Networks perform quite well in the seen scenarios, while degrade significantly in o.o.d scenarios. One plausible explanation is that CNNs learn shortcut decision rules based on training data statistics rather than capturing intrinsic true decision rules [19], and therefore fail to deal with distribution shifts in the open world. To improve CNNs' performance in unseen scenarios, researchers have focused on enhancing their o.o.d robustness.

Mimicking the human visual system, which exhibits high robustness, is a promising direction to overcome this limitation [5, 6]. Previous studies have discovered an unintuitive phenomenon known as texture bias: in ImageNet classification task, CNNs use local texture as the preliminary cue, while humans extract global shape information [20]. Motivated by this observation, researchers further analyzed the texture bias phenomenon [27, 40] and reduced the texture bias to improve model robustness [2, 19, 41, 54]. A consensus has formed that shape-biased models exhibit high robustness [1, 11, 20, 37, 41, 49, 54], which also means that shape information remains the most discriminative cue when test distribution shifts.

These works merely demonstrate the superiority of shape-biased models on public o.o.d benchmarks in general cases. To deeply investigate the intriguing relationship between shape/texture information and o.o.d robustness, in this work, we carefully establish Category-Balanced ImageNet dataset to conduct comprehensive control experiments. Unexpectedly, we find that the shape information is not always most discriminative in distinguishing different categories and the shape-biased model is not always superior across different o.o.d scenarios, for example shape information is much less discriminative for distinguishing animals than artifacts in both i.i.d and o.o.d cases; texture-biased models outperform shape-biased models in corruption o.o.d scenarios such as elastic transformation and multiple blurs types where shape information is not a stable cue.

These observations inspire us to develop more robust CNNs for universal o.o.d scenarios. We design a method called Shape-Texture Adaptive Recombination (STAR) to enhance o.o.d robustness by adaptively adjusting the contributions of shape, texture and debiased features for each given instance. Specifically, a category-balanced training

dataset is firstly used to pretrain a debiased backbone and three specialized heads to pull shape, texture and debiased features out respectively. Subsequently, an instance-adaptive recombination head is trained to adaptively adjust the contributions of these distinctive features according to the specific scenario/characteristics of each instance. Benefitting from this adaptive contribution of shape and texture features, our proposed method achieves state-of-the-art o.o.d robustness across various scenarios such as image corruptions, adversarial attacks, style shifts and dataset shifts on ImageNet-C [24], ImageNet-$\overline{C}$ [35], CIFAR10-C [24], FGSM attacked [21], ImageNet-R [23], ImageNet-sketch [47], Office-Home [46] and ImageNet-V2 datasets [39] over other popular data augmentation methods.

In summary, our main contributions lie in three folds: 1) We find that the shape feature is not always superior in distinguishing different categories (i.e. animals) and the shape-biased model is not always superior to all o.o.d scenarios (i.e. elastic transformation and blurs). 2) Based on these observations, we design a method named Shape-Texture Adaptive Recombination (STAR) to adaptively adjust the contributions of shape, texture and debiased features across different instances. 3) STAR exhibits superior performance across various o.o.d scenarios including image corruption, adversarial attacks, style shifts and dataset shifts.

2. Related work

The decision rule that CNNs employ for object classification has been a debated topic for a long time. According to the shape hypothesis, CNNs combine low-level features to complex shapes during the forward propagation [2, 31, 32, 40]. However, as conflicting evidence has accumulated, local texture information seems to be a more important factor for object classification than global shape information [3, 8, 17, 18].

Geirhos *et al*. were the first to formally investigate CNNs' texture bias problem [20], and consolidated the prevalence of texture hypothesis by demonstrating humans have biases toward shape while CNNs have biases toward texture through psychophysical and computational experiments on cue-conflicting images (such as a cat shape with elephant texture, generated by style transfer). They then trained a shape-biased model which was found to have higher robustness and better transfer learning ability.

Following Geirhos *et al*., researchers started to investigate various methods to reduce texture bias, achieving impressive success in multiple research fields. Zhang *et al*. used style transfer to replace the texture information of training images with artworks and discovered improvements in domain generalization [54]. Li *et al*. employed cue-conflicting dataset to train shape-texture debiased models. They assigned image label by weighted summing its one-hot shape label and texture label [33] (similar to Mixup augmentation [50]), improving model performance on several image recognition benchmarks and adversarial robustness. Data-related methods are not the only way to reduce texture bias. Shi *et al*. discriminated texture from shape based on local self-information, and decorrelate the model output from the local texture using a dropout-like algorithm [41]. This method improves model performance on image corruption and adversarial perturbation.

There are also works analyzing the texture bias phenomenon. Hermann *et al*. investigated the origin and prevalence of texture bias in CNNs, and found that multiple factors, such as training objectives and architectures, affect the level of texture bias, while data augmentation has the largest effect [27]. More recently, Li *et al*. advocated the importance of extracting rich features for both shape and texture [33]. The idea of using both shape and texture-biased branches is also adopted by [12, 53].

Although the robustness of shape-biased models is well-recognized and analyzed **in general** by previous works, there are works challenging this notion. Mummadi *et al*. [36] investigated a **specific** method to generate shape-biased models through style transfer, and attributed its high robustness to stylization rather than shape bias. However, whether shape information is inherently robust for **specific** categories and **specific** o.o.d scenarios still remains an unexplored area.

3. Intriguing relationship between shape/ texture information and o.o.d robustness

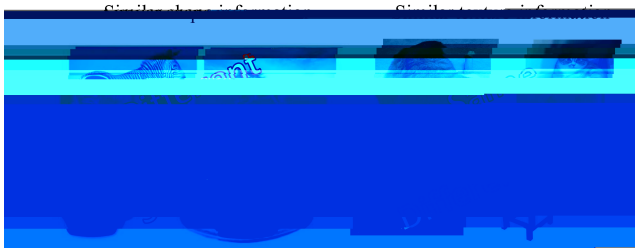

Figure 1. Daily observations supporting that shape information is not always the primary cue to define a category. (a) Using shape information alone cannot distinguish between zebra and horse. (b) Texture is the invariant information for cats. (c) Shape is the invariant information for cups. (d) Using texture information alone cannot distinguish between table and chair.

Previous research on CNNs' texture bias led to a commonly held belief: CNNs tend to be texture-biased, and increasing shape bias can improve model robustness [1, 10, 11, 20, 26, 27, 29, 37, 41, 45, 47, 49, 52, 54]. This belief holds true **in general cases** according to the verification on commonly used dataset like ImageNet [14] and CIFAR10 [30]. To further explore the relationship between shape/texture information and o.o.d robustness **in specific cases**, we investigate two questions to provide supplementary viewpoints.

Questions 1. Shape information is more robust, implying that shape information is more discriminative for o.o.d data. This motivates us to ask a fundamental question: Is shape or texture information inherently more discriminative for differentiating between categories? To answer this question, we focus on the two most general categories in the world: animals and artifacts. As shown in Figure 1, texture information seems more important to distinguish animals, while shape information seems more important to distinguish artifacts. This guess is supported by evidence from psychology and cognitive neuroscience [4, 7, 43, 44], but has not been rigorously examined in deep learning community.

Question 2. Shape-biased models are more robust in general o.o.d cases (i.e. the average of 15 corruption types in ImageNet-C dataset). It is curious to ask: Is shape information more robust in all kinds of distribution shifts? Especially, when images are corrupted by elastic transformation or blurring, will shape information be reliable?

3.1. Experimental setting for investigation

To rigorously investigate the relationship between shape/texture information and o.o.d robustness, we carefully design a category-balanced dataset, and methods to extract shape and texture features.

Shape/texture features. Following Geirhos *et al*. [20], **Shape** is defined as the set of contours that describe the 3D form of an object. In 2D case, it means the 2D projection of these 3D contours. Therefore, edge detection algorithms are natural choices for capturing shape information, and canny edge (CE) detector is adopted due to its excellent visualization results. Additionally, self-information (SI) map is used to extract shape information, as texture typically contains relatively low self-information due to its high-frequency self-repeating [41]. Inspired by Feng *et al*. [16], **Texture** is defined as the shape-unrelated portion of the image. Therefore, following the jigsaw puzzle approach [9, 38], we use permuted patches to represent the texture of the entire image, in which global shape information is largely destroyed but texture information is well preserved. Specifically, three parametric settings with shuffled 4x4, 8x8 and 16x16 patches are used to simulate different scales of texture, referred to as P4, P8 and P16. An example is illustrated in Figure 2 and more can be found in Supplementary Material 2.1.

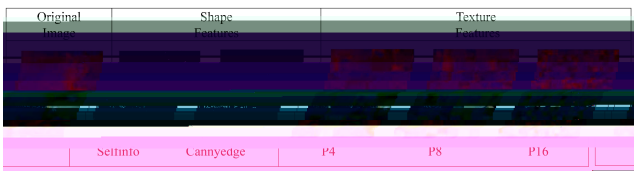

Figure 2. Illustration of shape and texture features.

Category-Balanced ImageNet dataset. 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. This dataset is a subset of ImageNet and is used to conduct control experiments. It consists 64 animal categories and 64 artifact categories, carefully selected from ImageNet 2012 dataset [14], with each containing approximately 1300 images. Detailed information regarding this dataset can be found in Supplementary Material 1. Moreover, we introduce two settings to ensure fair comparisons. In setting 1, models are trained and tested by animal and artifact sub-datasets respectively. In setting 2, the model is trained by the whole dataset, with evaluations on animal and artifact sub-datasets separately.

Training Setup. All models employ ResNet 50 architecture [22] with 100 training epochs. Other training details for all experiments in this paper can be found in the Supplementary Materials 2.2.

3.2. Is shape information more robust for all categories?

The conclusion that shape information is more robust implies that shape information is more discriminative for o.o.d data. However in conceptual science, researchers have found that natural objects are defined by their sensory features or internal structure, while man-made objects are defined by their function or usage purpose [7, 44]. From a taxonomic standpoint, the increasing use of biochemical and biophysical techniques leads to increasing determinacy in taxonomy [43]. As explored in cognitive science [4], visual shape properties' relationship with an artifact is much more articulated than their relationship with an animal, possibly because the category of an artifact is defined by its function, which is usually more related to its shape information. To thoroughly investigate this phenomenon, we ask a question: Which feature, shape or texture, is more discriminative for animal and artifact categories, in both i.i.d and o.o.d cases?

Aiming to investigate these questions, we train and test CNN models by using either shape or texture features in both i.i.d and o.o.d settings. The o.o.d test set contains 15 corrupted versions of each test image, corrupted by methods in ImageNet-C, including blurring, weather, and noise. A single evaluation metric, shape-to-texture ratio (S-T ratio), is introduced to compare the difference in discrimination ability more clearly. This ratio is calculated as the performance based on shape features divided by the performance based on texture features, and a higher value means shape information is more discriminative and vice versa.

As highlighted in Table 1, all experiments tested on animal datasets have S-T ratio smaller than 1.0, whereas all experiments tested on artifact datasets have S-T ratio larger than 1.0 in both i.i.d and o.o.d cases. These results indi-

cate that when distinguishing animals, models trained by texture features perform better than those trained by shape features. In contrast, when distinguishing artifacts, models trained by shape features perform better than those trained by texture features. This observation is consistent with our intuition that texture is the discriminative feature for animals, while shape is the discriminative feature for artifacts. Furthermore, comparing the top 2 rows with the bottom 3 rows, training with separated or combined datasets has trivial differences, indicating that the discriminative feature for a category is an intrinsic property, rather than a "relative property" compared to any other categories in the training set. A plausible explanation could be related to object definitions: the definitions of artifact categories are directly related to their functions, which are closely related to shape information, while the definitions of animal categories are directly related to their non-visual internal structures, which are much less reflected in shape information. Experiments regarding specific shape or texture features are included in Supplementary Material 3.1 with the same conclusion.

Train set	Test set	i.i.d case			o.o.d case		
		S	T	Ratio	S	T	Ratio
Animal	Animal	69	74	**0.93**	25	28	**0.90**
Artifact	Artifact	71	62	**1.15**	34	19	**1.81**
All	Animal	68	75	**0.91**	24	26	**0.92**
	Artifact	72	64	**1.12**	33	20	**1.67**
	All	70	70	1.00	28	23	1.24

Table 1. Accuracy (%) of models trained and tested on shape (S) & texture (T) features in both i.i.d and o.o.d cases. The ratio of S to T is also reported.

From these results, we notice that, **although shape features are discriminative in general cases, texture information is not necessarily always worse depending on specific categories. This conclusion encourages the use of category-balanced datasets in model training and pertaining, to fairly preserve shape and texture information.** Models should learn both shape and texture information well (discussed in section 3.3), but they may exhibit strong bias if animal-artifact distribution in the training set is strongly imbalanced.

Train set	Test set	i.i.d case	o.o.d case
Animals	Animals	0.807	0.729
Artifacts	Artifacts	0.541	0.480
All	Animals	0.791	0.738
	Artifacts	0.596	0.561
	All	0.665	0.630

Table 2. Texture bias ([0,1]) of models trained on original images and tested on cue-conflicting images in both i.i.d and o.o.d cases.

We further conduct experiments to investigate models' texture bias regarding animal and artifact categories, with details in Supplementary Material 3.2. As shown in Table 2, it is evident that models exhibit much higher texture bias on animal categories than on artifact categories. We believe that the different levels of texture bias are related to the category's intrinsic discriminative feature.

3.3. Is shape-biased model always superior in various o.o.d scenarios?

Shape information is more robust in o.o.d scenarios in general cases, since human decision rules are shape-biased. However, when analyzing models trained by shape/texture features on individual o.o.d corrupted scenarios in Section 3.2, we found models trained by shape features favor some corruption types while models trained by texture features favor others. Therefore, we are curious to investigate the robustness of shape-biased models in individual o.o.d corruption scenarios, especially considering those making shape information much less reliable such as blurring.

To investigate this question, we trained two models that are shape and texture-biased respectively. The shape-biased model is trained on all 128 categories in Category-balanced ImageNet, with each image having 50% chance replaced by either CE or SI version. The texture-biased model is trained on the same dataset, with each image having 50% chance replaced by either P4, P8 or P16 version. Both models are evaluated on the corrupted version of Category-balanced ImageNet test set, as mentioned in Section 3.2.

The results on all 15 natural corruption o.o.d scenarios are shown in Table 3. The shape-biased model performs better in general as expected, while on all types of blurs and elastic transformation, the texture-biased model shows clearly higher robustness. This observation can be explained by the fact that, when images are blurred or subjected to elastic transformation, their shape information changes more than texture, leaving their texture information as a more predictive cue. Furthermore, from the perspective of data distribution, these corrupted data have a smaller distribution shift to the augmented training dataset of the texture-biased model [35]. Results in Table 3 consolidate our intuition: **Although the shape-biased model has higher average accuracy, the texture-biased model shows noticeable advantages in certain scenarios where shape information is less reliable.**

Besides analyzing shape and texture-biased models trained by CE, SI, P4, P8, and P16 features that are extracted conventionally using simple manipulations, we further repeat our experiment using models trained by effective augmentations from Augmix variants to further support our conclusion. Augmix [25] is a well-known augmentation method that boosts robustness by applying a sequence of augmentations chosen randomly from nine basic ones, and employing consistency loss on the outputs of the original image and two augmented variants. We divide those nine

Method	avg	Noise			Blur				Weather				Digital			
		Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
S-biased model	45.6	**32.1**	**32.3**	**25.7**	42.4	37.2	45.1	44.2	**39.5**	**46.4**	**54.6**	**78.7**	**41.7**	56.6	**55.7**	**69.1**
T-biased model	41.3	23.4	21.6	15.4	**42.7**	**39.9**	**46.1**	**50.8**	28.3	33.9	43.6	72.4	39.2	**60.5**	47.2	64.9

Table 3. Accuracy (%) of shape and texture-biased models on 15 corruption scenarios. The best results are highlighted in bold.

Method	avg	Noise			Blur				Weather				Digital			
		Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
Augmix-Shape	58.1	52.7	53.6	49.7	46.4	46.4	50.9	50.9	47.6	**58.3**	60.9	**82.8**	**68.9**	64.4	**65.5**	**72.5**
Augmix-Texture	55.4	43.0	39.7	38.5	**57.1**	**48.0**	**66.9**	**64.9**	45.5	49.6	62.8	74.7	45.1	**69.6**	57.4	68.5
Augmix-Original	61.5	**56.4**	**56.6**	**54.4**	54.9	**48.0**	65.9	63.1	**52.6**	**58.3**	**69.6**	81.2	63.3	68.4	58.8	70.4

Table 4. Accuracy (%) of Augmix variants on 15 corruption scenarios. The best results are highlighted in bold.

basic augmentation methods into two groups. The texture-changing pool contains color-channel operations including solarize, equalize, auto-contrast, and posterize, which do not affect spatial information and force the model to learn robust shape features. The shape-changing pool contains spatial operations including translate xy, shear xy, and rotate which cause more changes in shape information than texture and force the model to learn robust texture features. Illustrations of these augmentations are shown in Figure 3. Following the training method of Augmix, we train 3 models, Augmix-shape, Augmix-texture, and Aumgix-original with texture-changing, shape-changing, and the original augmentation pool respectively.

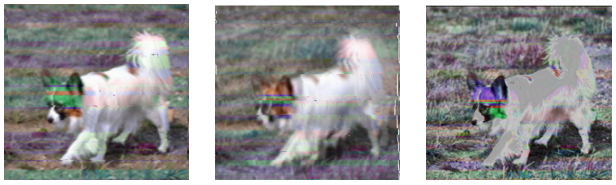

Figure 3. Illustration of shape-changing (mid) and texture-changing augmentations (right) of the original image (left).

Results in Table 4 exhibit similar trends as in Table 3, where each model favors specific types of o.o.d scenarios. In particular, Augmix-texture also favors all types of blurs and elastic as the texture-biased model in Table 3. Importantly, we noticed the results of the Augmix-original on individual cases are neither the trade-off average nor the highest value compared with its shape and texture variants. On one hand, incorporating both shape and texture information is necessary for certain o.o.d scenarios (like impulse noise and snow), where both shape and texture information remain reliable. On the other hand, incorporating both shape and texture information would weaken the predictive power under certain o.o.d (like contrast, pixelate, defocus blur, and elastic transformation), where either shape or texture information is no longer reliable. Therefore, we further analyze the working principle of Augmix, and point out its limitations to certain o.o.d scenarios where the model solely relies on shape/texture information (biased towards shape/texture) would have optimal performance.

4. Shape-Texture Adaptive Recombination

According to Section 3, the discriminative features vary across different o.o.d scenarios. Consequently, relying on fixed decision rules coupled with uniform feature extractions would be insufficient in maintaining reliability against distribution shifts. Therefore, we design a novel model named Shape-Texture Adaptive Recombination (STAR) with dynamic decision rules to enhance performance across a spectrum of o.o.d scenarios. The overall framework of the proposed method is visualized in Figure 4. A category-balanced training dataset is firstly used to pretrain a debiased backbone and three heads that specialize in extracting robust shape, texture and debiased features respectively. Then, a recombination head is trained to adaptively regulate the contributions of these distinctive features in accordance with individual instances. Compared to standard models, STAR incorporates additional parameters through three additional heads and introduces an additional training overhead of 20 epochs for the recombination head.

4.1. Unbiased backbone and specialized heads

To deal with different scenarios, the unbiased backbone (B) is designed to keep rich features ($F = B(x)$) which can be adaptively used for subsequent processes. This is mainly achieved by the objectives of the specialized heads.

Based on the backbone, three heads are designed to separate the shape, texture, and debiased features. Each head consists of a convolution layer and a fully connected layer and aims to extract one kind of feature. This is achieved by augmenting the training data using Augmix method with distinct augmentation pools. Specifically, the shape-biased head (H_s) is trained by original images and images augmented by the texture-changing pool (x_{tc}) exclusively, using cross-entropy classification loss and Jensen-Shannon divergence consistency loss as in Augmix (Equation 1). JS divergence loss is calculated by the posterior distributions of

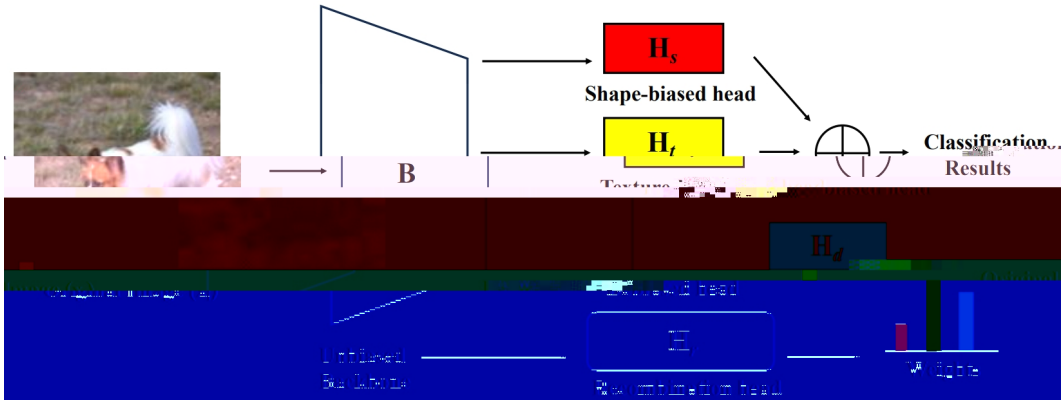

Figure 4. The overall framework of STAR. The unbiased backbone and three specialized heads first extract robust shape, texture, and debiased features. Then the recombination head provides weight adaptively across input to combine the results from specialized heads.

the original sample x_{ori} and two augmented variants. Similarly, the texture-biased head (H_t) and debiased head (H_d) are trained using images augmented by the shape-changing pool (x_{sc}) and the combined pool (x_d) exclusively.

Then, the overall loss function for the first 100 epochs is defined as Equation 2, where p^{tc}_{aug}, p^{sc}_{aug} and p^{d}_{aug} are calculated by $H_s(B(x_{tc}))$,$H_t(B(x_{sc}))$ and $H_d(B(x_d))$.

$$L_H = L_{cls}(p_{ori}; y) + \lambda JS(p_{ori}; p_{aug1}; p_{aug2}): \quad (1)$$

$$\begin{aligned} L = & L_{H_s}(p_{ori}; p^{tc}_{aug1}; p^{tc}_{aug2}; y) \\ & + L_{H_t}(p_{ori}; p^{sc}_{aug1}; p^{sc}_{aug2}; y) \\ & + L_{H_d}(p_{ori}; p^{d}_{aug1}; p^{d}_{aug2}; y): \end{aligned} \quad (2)$$

Through this approach, H_s is forced to learn similar features for the original image and images with texture-changing augmentations, enforcing its decision rule to heavily rely on shape information to minimize the consistency loss. Such a decision rule will favor o.o.d scenarios where shape information remains robust but texture is unreliable. Analogously, H_t and H_d are trained with similar underlying principles. As a result, three heads with distinct decision rules are trained to handle different o.o.d scenarios.

4.2. Recombination head

The recombination head H_r is designed to adaptively adjust the contributions of these distinctive features in accordance with each instance. Notably, images and augmented images (both noted as x_i) are fed into all heads, resulting in extracted features $H_s(Fi)$, $H_t(Fi)$ and $H_d(Fi)$. The recombination head outputs 3 values as in Equation 3, and the final prediction is the convex combination of the prediction of the specialized heads, with weights from the recombination head. The classification results $\hat{y}_i$ and training objective L_i for image x_i can be calculated as in Equation 4 and 5. In our experiment, the re8(combination)m(Tf 5.-290s 5.-291(trained)9061(by)9150(e)15(xton)]TJ 0 -11.955 Td 20in)-250(epochs)-250(with)-2

L_i

Method	i.i.d	o.o.d							
		Image Corruption		Adversarial Attack		Style Shift		Dataset Shift	Avg
	IN	IN-C(#)	IN-$\overline{C}$ (#)	FGSM0.03	FGSM0.06	IN-R	IN-S	IN-V2	
Vanilla ResNet	84.3	79.2	47.1	35.0	23.4	28.9	25.0	84.9	42.7
Cutout	84.6	79.7	46.9	24.7	15.4	29.7	24.8	85.1	40.2
Mixup	85.1	71.0	40.8	47.8	39.0	32.7	27.1	86:3	49.5
Cutmix	85.3	79.8	43.6	35.8	31.9	28.9	22.5	86:3	44.4
Patch Gaussian	84.7	74.8	46.3	24.2	16.1	30.1	24.4	84.6	40.7
Stylized IN	84.1	63.0	41.4	40.8	28.9	**38.4**	**43.4**	85.1	50.7
AutoAugment	85:4	68.9	44.1	26.4	17.3	34.5	34.6	85.2	44.4
Augmix	**85.5**	61.3	40.7	43.8	31.1	36.3	36.4	85.7	47.9
APR	85.3	58:9	**35.2**	27.9	19.0	35.7	38.1	**86.6**	50.6
STAR identi-heads	85:4	60.7	40.6	57:1	51:5	36.6	37.8	85.2	55:7
STAR(ours)	**85.5**	**57.4**	39:4	**59.3**	**54.3**	38:2	38:3	85.9	**57.2**

Table 5. I.i.d and o.o.d performances (%) for different methods on Category-balanced ImageNet (IN). 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.

distinct corruptions types, each with five severity levels per test set image in ImageNet. It is evaluated by mean corruption error, the normalized corruption error by AlexNet referred from the original paper. ImageNet-$\overline{C}$ introduces additional 10 o.o.d scenarios perceptually dissimilar to corruption types per test set image in ImageNet and evaluated by mean error. To evaluate o.o.d robustness to adversarial attacks, we obtain the test set by attacking the images in the original test set using Fast Gradient Sign Method (FGSM) [21] with ϵ = 0.03 and 0.06. To evaluate o.o.d robustness to style shift, ImageNet-R dataset [23] and ImageNet-sketch dataset [47] are used. ImageNet-R contains 30k images across 16 different rendition types (i.e. art, sculptures, sketches) and 200 ImageNet classes. ImageNet-sketch contains 50 sketch images for each ImageNet class. To evaluate o.o.d robustness to dataset shift, we adopt ImageNet-V2 dataset [39], a new test set for ImageNet with 10k images collected following the original protocol.

In our experiments on CIFAR10 dataset, a widely-used dataset with 60k 32x32 images across 10 classes, CIFAR10-C dataset [24] (CIFAR10 counterpart of ImageNet-C), is used to evaluate o.o.d robustness against image corruption. Again, we attack the test set images by FGSM with ϵ = 0.03 and 0.06 to evaluate adversarial attacks. In our experiments on Office Home, a dataset consists images of 65 objects commonly found in Office-Home settings from 4 distinct domains, we employ the real-world domain for model training and the remaining three domains (art, clip art and product) for assessing o.o.d robustness against style shift.

Comparable methods. We compare the performance of STAR with other augmentation algorithms, including Cutout [15], Mixup [50], Cutmix [48], Patch Gaussian [34], Stylized ImageNet [20], AutoAugment [13], Augmix [25] and Amplitude-phase Recombination (APR) [10]. Additionally, we introduce a model called STAR-identical heads, which mirrors the architecture of STAR, but all three heads are debiased heads trained using nine augmentations.

5.2. Results on Category-balanced ImageNet

The performance of all methods is presented in Table 5. As can be seen, our proposed method achieves the highest overall o.o.d robustness (57.2%, a notable increase of 6.5% from the second-best Stylized IN) and consistently enhances performance across all individual o.o.d scenarios. Importantly, STAR exhibits a further improvement of 1.5% over STAR-identical heads, indicating its advantage does not solely stem from the complexity of the additional heads and ensemble results. Instead, it comes from the adaptive contributions of distinctive features to suit various o.o.d cases. Meanwhile, it exhibits similar i.i.d performance as other methods, breaking the conventional trade-off between accuracy and robustness [34, 51].

Regarding o.o.d robustness against image corruptions, our proposed model achieves the lowest error in IN-C and second lowest in IN-$\overline{C}$, exhibiting improvements of 21.8% and 7.7% respectively over the ResNet baseline. The improvement from STAR-identical heads to STAR (60.7% to 57.4% in IN-C) clearly validates the effectiveness of specialized heads. To further understand the role of these heads, the performance of individual heads is evaluated and the same trend as in Table 4 is observed (i.e. texture-biased head favors blurs), indicating that H_s and H_t heads have indeed leant the intended shape and texture features. In particular, H_t shows higher accuracy for all blurs and elastic transformations over the other two heads (64.4% vs 62.0%, 63.5% on average, and details can be found in supplementary material 3.3). To further understand the role of the recombination head, we calculated the average combination weights for individual corruption types and observed

Test dataset	Architecture	Vanilla	Cutout	MixUp	CutMix	AutoA	PatchG	Augmix	APR	S-ID	ours
CIFAR10 (i.i.d scenario)	ResNet18	95.2	95.7	96.0	**96.4**	95.3	95.1	95.4	95.7	95.9	95.8
	VGG16	92.0	92.4	92.7	**93.2**	91.9	91.3	91.9	92.7	92.5	91.1
	MobileNetV3	79.0	77.9	77.0	74.2	79.1	**79.3**	**79.3**	78.6	78.8	**79.3**
	Mean	88.7	88.7	88.6	87.9	88.8	88.5	88.9	89.0	**89.1**	88.7
CIFAR10-C (image corruption)	ResNet18	74.8	75.5	79.7	71.1	82.4	83.5	88.8	88.2	88.6	**89.1**
	VGG16	76.4	74.4	77.3	72.1	80.1	78.9	83.4	83.8	84.1	**84.3**
	MobileNetV3	66.2	62.7	62.5	58.2	65.7	65.3	70.9	69.5	71.3	**71.7**
	Mean	72.5	70.9	73.2	67.1	76.1	75.9	81.0	80.5	81.5	**81.7**
CIFAR10-FGSM0.03 (adversaril attack)	ResNet18	68.5	66.5	71.3	65.1	69.7	69.6	83.1	74.9	85.1	**86.2**
	VGG16	63.1	61.4	63.1	56.5	63.6	63.6	66.3	**67.4**	67.3	67.2
	MobileNetV3	47.3	41.4	45.6	39.7	46.5	47.4	51.4	48.7	51.8	**52.1**
	Mean	59.6	56.4	60.0	53.8	59.9	60.2	67.0	63.6	68.1	**68.5**
CIFAR10-FGSM0.06 (adversaril attack)	ResNet18	59.8	57.3	66.1	57.0	61.7	61.7	79.6	63.7	82.9	**83.9**
	VGG16	47.1	44.8	51.8	44.2	46.7	48.1	51.7	50.5	51.5	**52.9**
	MobileNetV3	37.4	36.6	33.2	28.0	35.1	37.2	37.9	36.0	40.5	**41.2**
	Mean	48.1	46.2	50.4	43.1	47.8	49.0	56.4	50.1	58.3	**59.3**
OH (style shift)	ResNet18	47.4	47.2	48.0	47.3	**52.4**	47.5	50.3	48.9	51.0	51.6

Table 6. I.I.D and o.o.d performances (%) for different methods and network architectures on CIFAR10 and Office-Home (OH). S-ID means STAR-identical heads. Our proposed method consistently improves model robustness across diverse o.o.d scenarios.

the expected results. For example, the texture-biased head is assigned the highest weight for blurs (0.453), while the shape-biased head is assigned the highest weight for Pixelate (0.503). These results highlight the recombination head's capacity to allocate the most pertinent heads with the highest weights. Detailed results for individual corruption o.o.d scenarios for all methods can be found in Supplementary Material 3.4, where STAR also shows superior performance over comparable methods.

In terms of o.o.d robustness against adversarial attacks, STAR outperforms the second-best method Mixup by a substantial margin (59.3% to 47.8% in FGSM0.03, and 54.3% to 39.0% in FGSM0.06). Besides, we guess images-mixing might be an factor, since methods exhibiting strong performances largely employ this process, including Mixup, Cutmix, Augmix, Stylized IN and STAR.

For o.o.d robustness against style shifts, STAR achieves the second-best results on IN-R and IN-sketch. However, Stylized IN's top 1 accuracy might be slightly unfair, due to its utilization of an additional artworks dataset for style-transfer-based data augmentation. The various artwork styles in this additional dataset coincide with the styles in IN-R (art, paintings, sketches, etc) and IN-sketch (sketch).

Lastly, for o.o.d robustness against dataset shifts, STAR achieves a 1% improvement over the vanilla CNN, but it is less effective than Mixup, Cutmix and APR.

5.3. Results on CIFAR10 and Office-Home

As shown in Table 6, extended experiments on CIFAR10 and Office-Home prove the effectiveness of our proposed method on diverse o.o.d datasets and network architectures. For CIFAR10, STAR achieves the best performance consistently in o.o.d scenarios of image corruption and adversarial attacks, across multiple architectures like ResNet18, VGG16 and MobileNetV3. Meanwhile, the results on Office Home datasets demonstrate that STAR outperforms other methods for robustness against style shifts.

6. Conclusion

In this work, we delve deeper into the intricate relationship between shape/texture information and o.o.d robustness. We find that shape information is not always superior in distinguishing distinct categories (i.e. animals) and shape-biased model is not always superior across various o.o.d scenarios (i.e. elastic transformation and blurs). Based on these observations, we proposed Shape-Texture Adaptive Recombination to adaptively adjust the contributions of shape, texture and debiased features across different instances. The superior performance on multiple datasets under diverse o.o.d scenarios including image corruption, adversarial attacks, style shift and dataset shift proves the effectiveness and generalization of our proposed method.

Acknowledgement

This work is supported in part by National Key R&D Program of China (No.2022YFB3103800), National Natural Science Foundation of China (No.62122074, No.U1936209, and No.62002353), and National Science Fund for Distinguished Young Scholars (No.NSFC31925020).

References

[1] Nader Asadi, Amir M Sarfi, Mehrdad Hosseinzadeh, Zahra Karimpour, and Mahdi Eftekhari. Towards shape biased unsupervised representation learning for domain generalization. *arXiv preprint arXiv:1909.08245*, 2019. 1, 2

[2] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, 14(12):e1006613, 2018. 1, 2

[3] Pedro Ballester and Ricardo Matsumura Araujo. On the performance of googlenet and alexnet applied to sketches. *In AAAI Conference on Artificial Intelligence*, 2016. 2

[4] Yanchao Bi, Xiaoying Wang, and Alfonso Caramazza. Object domain and modality in the ventral visual pathway. *Trends in Cognitive Sciences*, 20(4):282–290, 2016. 3

[5] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2):115, 1987. 1

[6] Jeffrey Bisanz, Gay L Bisanz, and Robert Kail. Learning in children: Progress in cognitive development research. *Springer Science Business Media*, 2012. 1

[7] Jeannine S Bock and Charles Clifton. The role of salience in conceptual combination. *Memory & Cognition*, 28:1378–1386, 2000. 3

[8] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *In International Conference on Learning Representations*, 2019. 2

[9] Silvia Bucci, Antonio D'Innocente, Yujun Liao, Fabio Maria Carlucci, Barbara Caputo, and Tatiana Tommasi. Self-supervised learning across domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[10] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. *In IEEE/CVF International Conference on Computer Vision*, pages 458–467, 2021. 2, 7

[11] Peijie Chen, Chirag Agarwal, and Anh Nguyen. The shape and simplicity biases of adversarially robust imagenet-trained cnns. *arXiv preprint arXiv:2006.09373*, 2020. 1, 2

[12] Kenneth T Co, Luis Muñoz-González, Leslie Kanthan, Ben Glocker, and Emil C Lupu. Universal adversarial robustness of texture and shape-biased models. *In IEEE International Conference on Image Processing*, pages 799–803, 2021. 2

[13] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 7

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *In IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2, 3

[15] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 7

[16] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. *In IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

[17] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *In Advances in Neural Information Processing Systems*, 2015. 2

[18] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Texture and art with deep neural networks. *Current Opinion in Neurobiology*, 46:178–186, 2017. 2

[19] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1

[20] 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. 1, 2, 3, 7

[21] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 7

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *In IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3, 6

[23] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *In IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 2, 7

[24] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *In International Conference on Learning Representations*, 2019. 2, 6, 7

[25] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 4, 7

[26] Hwan Heo, Youngjin Oh, Jaewon Lee, and Hyunwoo J Kim. Domain generalization emerges from dreaming. *arXiv preprint arXiv:2302.00980*, 2023. 2

[27] 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. 1, 2

[28] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. *In IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019. 6

[29] Md Amirul Islam, Matthew Kowal, Patrick Esser, Sen Jia, Bjorn Ommer, Konstantinos G Derpanis, and Neil Bruce. Shape or texture: Understanding discriminative features in cnns. *arXiv preprint arXiv:2101.11604*, 2021. 2

[30] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. www.cs.utoronto.ca, 2009. 2, 6

[31] Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity.