Improving Robustness to Texture Bias via Shape-focused Augmentation

Sangjun Lee, Inwoo Hwang, Gi-Cheon Kang, Byoung-Tak Zhang
AI Institute (AIIS), Seoul National University
Seoul, Republic of Korea
{sjlee, iwhwang, gckang, btzhang}@bi.snu.ac.kr

Abstract

Despite significant progress of deep neural networks in image classification, it has been reported that CNNs trained on ImageNet have heavily focused on local texture information, rather than capturing complex visual concepts of the objects. To delve into this phenomenon, recent studies proposed to generate images with modified texture information for training the model. However, these methods largely sacrifice the classification accuracy on the in-domain dataset while achieving improved performance on the out-of-distribution dataset. Motivated by the fact that human tends to focus on shape information, we aim to resolve this issue by proposing a shape-focused augmentation where the texture in the object's foreground and background are separately changed. Key idea is that by applying different modifications to the inside and outside of an object, not only the bias toward texture is reduced but also the model is induced to focus on shape. Experiments show that the proposed method successfully reduces texture bias and also improves the classification performance on the original dataset.

1. Introduction

With the rapid development of deep learning algorithms, current deep neural networks (DNN) achieved remarkable performance in image classification [8, 10, 15, 18]. However, recent works discovered that DNNs often learn only shallow correlations between images and their labels, instead of learning intrinsic visual concepts [5, 11, 16]. For example, convolutional neural networks (CNN) trained on ImageNet [15] rely on texture information when classifying images, i.e. biased toward texture [5]. This induces the model to be vulnerable to out-of-distribution samples and results in poor generalization capability [14].

To reduce the model's heavy reliance on the image's texture information in classifying images, numerous works [7, 17] proposed to generate images whose texture is modified and to train the model on the generated images along with the original dataset. This method improves the model's robustness by preventing the model from relying on the spurious correlation between the image's texture and its label [7]. For example, Geirhos *et al*. [7] proposed Stylized-ImageNet (SIN) where they utilized neural style transfer [4] to change the texture of the samples in ImageNet to that of artistic paintings. Sauer-Geiger [17] proposed counterfactual generative network (CGN) changing the texture of the foreground and background of an object in the image from ImageNet separately to the texture of other classes. However, while these methods force the model to be less reliant on texture information and successfully improve the model's robustness to out-of-distribution samples, it is also shown that the model's classification accuracy decreases when tested on the original dataset. This trade-off indicates the drawback of these methods which only focus on diminishing the model's reliance on texture.

(a) Original Image (b) Shape-focused augmentation

Figure 1. Shape-focused augmentation. (a) Image sample of label 'spoonbills' in ImageNet-100. (b) Same image sample that different augmentations are applied to the foreground and background of 'spoonbills' in the image.

Our hypothesis is that the radical modification of texture information e.g., replacing with another image's texture as done in prior works [7, 17] hinders the model from learning semantically meaningful representations of images in the original dataset and thus negatively affects the model's accuracy on the in-domain samples. Motivated by the prior work [7] empirically showing that humans tend to classify images by their shape rather than by their texture and in-

creasing this shape bias improves the model's robustness to out-of-distribution samples, we propose a novel data augmentation scheme, shape-focused augmentation. In our method, the augmentation is applied to reduce the spurious correlation between the image's texture and its label, but separately on each side of the shape boundary, i.e., the inside and outside of an object. Among the set of data augmentations, i.e., `ColorJitter`, `RandomGrayscale` and `RandomGaussianBlur`, subsets are randomly sampled twice and each is applied to the foreground and background of the object respectively as shown in Fig. 1. This simple technique not only mitigates the model's reliance on texture information but also emphasizes shape feature in the image and induces the model to focus on it, as humans do.

We implemented the experiment to validate the effectiveness of shape-focused augmentation by comparing it with the prior method [17] in terms of the accuracy on both the original dataset (ImageNet-100) and out-of-distribution samples (OOD dataset [6]). The experiment results show that the proposed method outperforms baselines on both datasets which implies that our method effectively alleviates the model's texture bias and enables the model to learn robust representations by focusing on shape without sacrificing the test accuracy on the original dataset.

Our contributions are summarized as follows:

- Motivated by the human behavior that focuses on shape, our method successfully improves the classifier's robustness to out-of-distribution samples without degrading the accuracy on the original dataset.
- Our method is simple yet effective, and it can be equipped with various models and applied to both supervised and self-supervised learning tasks.

2. Related works

Reducing the texture bias of CNN. As one of the critical pitfalls of current deep neural networks, there have been several approaches to reduce the model's reliance on local texture information. Bahng *et al*. [1] trained the model which is intentionally biased to the image's texture by extremely reducing the size of receptive fields of the CNN model. With the biased model, they train a de-biased model by enforcing the learned representation to be independent of the biased model's representation.

Another approach is to generate images whose texture is modified from the original dataset and train the model with the generated dataset. For example, Geirhos *et al*. [7] utilizes neural style transfer [4] to generate a dataset whose texture is changed to that of artistic paintings. Sauer-Geiger [17] proposed the generative model to create counterfactual images where the texture in the foreground and background of an object are independently altered to that of other classes in ImageNet. The classifier to be trained on the generated images consists of 3 multiple heads with a shared backbone to predict each label of shape, foreground, and background in counterfactual images and use the average of output from 3 heads when predicting the label of samples in ImageNet.

Hermann *et al*. [9] examined the effectiveness of data augmentations in preventing the model from preferentially classifying images by their texture feature. They showed that data augmentation is as effective as the methods using the additional set of generated images e.g., Stylized-ImageNet [7], in terms of reducing the model's texture bias.

Our work in this paper falls into the same category as the previous work by Hermann *et al*. [9] in the point that both works utilize data augmentations as the tool to relieve the correlation between an image's label and its texture. However, our work also partly shares the same intuition with CGN [17] in that defining visual features in an image more structurally as the foreground and background of the object and giving different variations to each part.

Importance of shape-bias for OOD robustness. It is known that humans classify images based on their shape rather than their texture [7] and there exists a significant gap between the robustness of humans and neural networks on out-of-distribution samples [6]. In this regard, it has been argued that training the model to focus on the shape attribute of an image, i.e., shape bias, is essential for improving its robustness to out-of-distribution samples [7].

Geirhos *et al*. [6, 7] emphasized the importance of improving the model's shape bias by comparing the model's vulnerability to out-of-distribution samples to the robustness of humans. To measure the robustness, they proposed `OOD benchmark` dataset [6]. This dataset is generated from ImageNet by applying 17 kinds of adjustments to its texture. It is divided into 2 groups and we denote these groups as `OOD-noise` and `OOD-style`. `OOD-noise` contains images changed by 12 kinds of noise-related modifications to their texture. `OOD-style` consists of images adjusted by 5 kinds of style-related alterations to the texture of images.

Our work is along the same line as previous works [6, 7] in the standpoint that the model needs to be biased toward shape to obtain the robustness to out-of-distribution samples. However, being deviated from former research streams usually concentrating on relatively increasing the model's shape bias by reducing its texture bias, our work looks into the method directly emphasizing the shape feature of the image during training.

3. Method

In this section, we first describe the overall process of our proposed method, shape-focused augmentation. To demonstrate the necessity and effectiveness of our method, we sequentially show 1) the influence of drastic adjustment of texture in the images from the original dataset

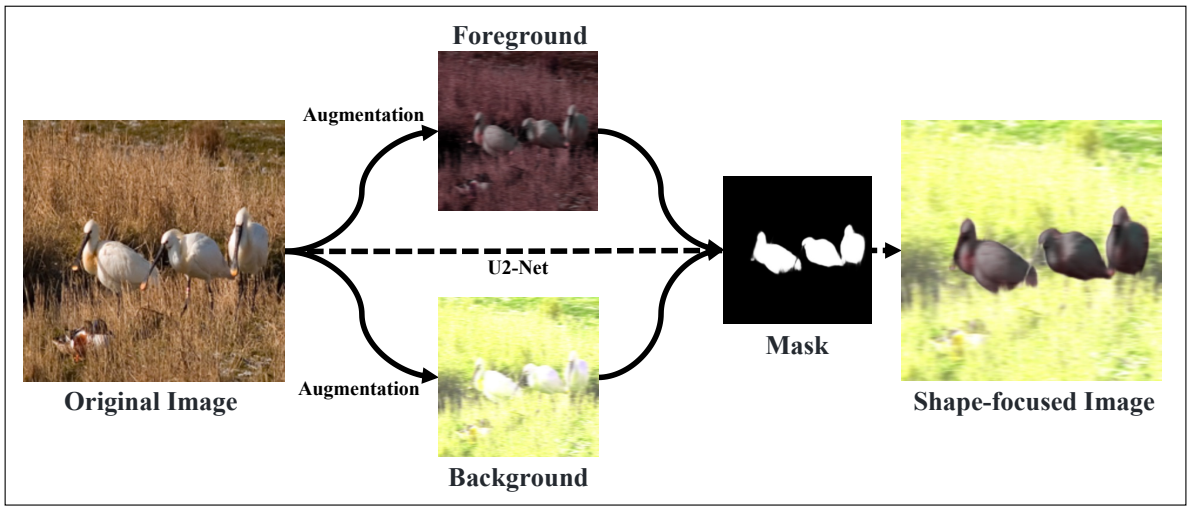

Figure 2. Overall process of shape-focused augementation. Pretrained U2-Net [13] is used for segmenting the mask of shape.

on the model's test accuracy across out-of-dstribution and in-domain samples and 2) the efficacy of data augmentations when applied separately to the object's foreground and background in an image. Lastly, we suggest the necessity of the learning algorithm directly emphasizing the highlighted shape features in the output images of the shape-focused augmentation.

3.1. Shape-focused Augmentation

The overall process of shape-focused augmentation is shown in Fig. 2. The original image has been augmented via `RandomResizedCrop` and `RandomHorizontalFlip`. We compose two independent sets of data augmentations by randomly sampling them from `ColorJitter`, `RandomGrayscale`, and `RandomGaussianBlur`. Each of two sets of indenpendently sampled data augmentations is applied to the original image respectively and it consequently creates two images which are augmented differently from one another. These two dissimilarly augmented images are combined into one image by the shape mask generated from pretrained U2-Net [13]. In other words, two disparately augmented images fill in the inside and outside of the shape mask, respectively.

Algorithm1 summarizes how the shape-focused augmentation can be plugged into the training process of a vanilla model in the supervised learning setting. First, we get two independently augmented images, a foreground image f and a background image b, with the random data augmentation module A. The data augmentations in module A are arbitrarily chosen from `ColorJitter`, `RandomGrayscale`, and `RandomGaussianBlur` on every implementation. Then, we extract the shape mask m of the object in the given image by utilizing pretrained U2-Net [13], denoted as U. Finally, the foreground image f and the background image b are combined into the shape-focused image x_{aug} by summing the element-wise multiplications of the shape mask m and $1-m$ with f and b, respectively. The augmented image x_{aug} is fed into the ResNet-50 model R and the cross-entropy(CE) loss is minimized during its training.

Algorithm1: Shape-focused Augmentation

Input: image x, label y, iteration t, end of iteration T
U : Pretrained U2-Net
R : ResNet-50
CE : Cross-entropy loss
A : Random data augmentation
for $t = 1$ **to** T **do**
  $f \leftarrow A(x)$
  $b \leftarrow A(x)$
  $m \leftarrow U(x)$
  $x_{aug} \leftarrow m \odot f + (1 - m) \odot b$
  Train R with $CE(R(x_{aug}), y)$
end

3.2. Effects of Modifying Texture on the Accuracy

In spite of the neural network's improved robustness to out-of-distribution samples by training it on generated images with modified textures, the model often shows a decline in the accuracy on the original dataset as a trade-off [9]. To figure out the factors affecting this degradation, we scrutinize the influence of the degree of changing the image's texture from the original dataset. We hypothesized that the model's decreasing accuracy on the original dataset is influenced by the adjustment of the training image's texture to the extent of losing the semantic similarity with the images with the same label in the original dataset.

(a) Shape ≠ Foreground

Shape: 'golden retriever'
Foreground: 'broccoli'
Background: 'hay'

(b) Shape = Foreground

Shape: 'golden retriever'
Foreground: 'golden retriever'
Background: 'hay'

Figure 3. Samples of counterfactual images. (a) The sample of counterfactual images that texture in both foreground and background of an object is altered. (b) The sample of counterfactual images that texture only in the object's background is adjusted.

In this regard, we compare the two extremely different cases of texture modification. By utilizing CGN [17]'s method, we generate a set of the counterfactual images where the alteration of texture makes them semantically not resembled with the images of the identical label from the original dataset as shown in Fig. 3 (a). On the contrary, we also create another set of counterfactual images explicitly discernible as the images with the same label from the original dataset by adjusting the texture only in the background of the object like Fig. 3 (b). Then, we train the model jointly on each set of the counterfactual images and the original dataset, i.e., ImageNet-100.

Model	ImageNet-100	OOD-noise	OOD-style
ResNet-50	**77.64**	39.41	35
CGN(a)	69.54 (-8.1)	**53.96**	**37**
CGN(b)	76.44 (-1.2)	43.83	33

Table 1. Top-1 classification accuracy(%). CGN(a) denotes the ResNet-50 model trained on the set of counterfactual images in Figure 3 (a) along with ImageNet-100. CGN(b) denotes the ResNet-50 model trained on the set of counterfactual images in Figure 3 (b) along with ImageNet-100. Numbers in parenthesis is the accuracy difference on ImageNet-100 between the vanilla model and each model.

In Tab. 1, the considerable improvement in the test accuracy on OOD datasets is shown by the model trained on the counterfactual images like Fig. 3 (a) which is unrecognizable as 'golden retriever' of the original dataset anymore. However, the model shows the sharp drop of the accuracy on ImageNet-100 when compared to the vanilla model. On the other hand, the model trained on the counterfactual images like Fig. 3 (b) which is still perceptible as 'golden retriever' in the original dataset exhibits much less difference from the vanilla model in the accuracy across all datasets.

Through this result, we empirically show that losing the semantic representation of the images in the original dataset by exhaustively changing their texture leads the model trained on the modified images to misclassify the in-domain samples often than before.

3.3. Effectiveness of Shape-focused Augmentation

Based on the caveat from the previous experiment result, we assumed that the image's texture needs to be modified within the range of retaining the semantic closeness to the images of the same label in the original dataset in order to prevent the model trained on the adjusted images from missing the capability of classifying in-domain images.

In this respect, Hermann *et al*. [9] empirically show the comparable effectiveness of data augmentations to generating new images with the altered textures in improving the model's robustness to out-of-distribution samples. We assume that data augmentations give variations to the image's texture while sufficiently retaining recognizable semantic attributes of the original images. Additionally, Sauer-Geiger [17] proposed the method differentiating an image into two parts by the shape silhouette of the object and independently altering texture in each divided part to the texture of other classes. Consequently, this method brings a more diversified texture to the image.

Motivated by the findings from the prior works [9, 17], we designed the shape-focused augmentation introduced in Sec. 3.1. As shown in Fig. 1, shape-focused augmentation emphasizes the shape attribute in the output image by applying data augmentations separately to the inside and outside of the object in an image.

We hypothesize that the model is trained to focus more on the image's shape feature while alleviating its inclination to classify images by their texture information through the application of shape-focused augmentation in its learning process as elaborated in Sec. 3.1. In this aspect, we expect the model becomes less susceptible to out-of-distribution samples without compromising the capacity of classifying the images from the original dataset. To validate our hypothesis, we examine the test accuracy of the model trained with the shape-focused augmentation on both out-of-distribution samples and in-domain samples. Then, we demonstrate the efficiency of our method by comparing it with the results of other baselines.

In Tab. 2, the model trained with the shape-foucsed augmentation shows a substantial improvement over the vanilla model in the accuracy on both two types of OOD datasets. The enhancement of the test accuracy on OOD datasets is even larger than that of the model trained on counterfactual images. The model trained with our method also shows the encouraging result of reducing the gap with the vanilla model in the accuracy on the original dataset when compared to the model trained on counterfactual images.

Model	ImageNet-100	OOD-noise	OOD-style
ResNet-50	**77.64**	39.41	35
+ShapeAug	74.56 (-3.08)	**62.62**	**40**
CGN(a)	69.54 (-8.1)	53.96	37

Table 2. Top-1 classification accuracy (%). +ShapeAug denotes the ResNet-50 model trained with shape-foucsed augmentation in Figure 2. CGN(a) denotes the ResNet-50 model trained on counterfactual images in Figure 3 (a) along with ImageNet-100. Numbers in parenthesis is the accuracy difference on ImageNet-100 between the vanilla model and each model.

Regarding the superior performance of the model trained with shape-focused augmentation over the model trained on counterfactual images across all OOD datasets, we supposed that it results from the inherent limitations that the model architecture in CGN [17] has. The model is formed with a shared backbone and three multiple heads. On the counterfactual images which have shape, foreground, and background labels for one image, each of the three heads output a logit for the respective labels, and it is trained to be invariant to all but one feature out of the three features of counterfactual images. However, on the normal images with one label per image, the model calculates the average of three logits respectively came from each of the three heads to predict the image's label. In this respect, we presumed that this averaging operation might have decreased the maximum upper bound of the model's predictability on OOD dataset's images with one label even if the specific head's logit might have shown the maximal accuracy in classifying the image.

Accordingly, it also displays the convenience of the shape-focused augmentation which is able to be utilized by simply plugging it into the existing learning steps of the vanilla model without the additional modifications to the model architecture.

The experiment results show that the shape-focused augmentation simultaneously improves the model's robustness to OOD datasets and reduces the gap with the vanilla model in the accuracy on the original dataset. Furthermore, our approach also has uncomplicated processes in the usage than the method generating the additional set of images with the adjusted texture. Namely, the shape-focused augmentation can be utilized by simply adding it to the model's current learning procedures without changing the model architecture. However, the generation-based models need to generate additional images in advance to train the model on it, and some of the models even needs to alter the model structure which can affect their classification performance.

3.4. The Necessity of Shape-focused Learning

Our proposed method applies different subsets of randomly sampled data augmentations to the image's two separated parts which are segmented by the object's shape boundary. It highlights the shape of the object in the image more explicitly, e.g., the example in Fig. 1. This approach brought a noticeable enhancement in the model's robustness to out-of-distribution samples and also in the reduction of the amount of falling accuracy on the in-domain samples. However, the vanilla CNN does not have the learning properties of directly accentuating the highlighted shape features of augmented images during their training. We make an assumption that the vanilla CNN's lack of learning processes immediately emphasizing the shape attribute results in the still remaining reduction of the accuracy on the original dataset despite the application of the shape-focused augmentation during training.

To jump over the constraints in the vanilla CNN, we apply the shape-focused augmentation to contrastive learning models [2, 12] by replacing their existing augmentation procedures with ours. The fundamental idea of contrastive learning is to pull together an anchor image and its augmented sample, i.e., a positive sample, in the embedding space, and to push apart the anchor image from all the other samples, i.e., negative samples. In this way, the model learns the common meaningful representations between the anchor image and its positive sample by contrasting them with negative samples without utilizing the supervision by the image's label. We expect this contrastive characteristic contributes the model to learn the emphasized shape features between the images augmented by shape-focused augmentation.

In our work, we choose the supervised contrastive learning [12] as the main model where the shape-focused augmentation is employed on. In this method, not only a single positive sample augmented from an anchor image but also all images with the same label as the anchor image are included in its positives. We presume that this aspect enables the model to learn the common shape attributes among images with the same label as the anchor image as well as the accentuated shape features in output images of the shape-focused augmentation.

The general outline of the application of shape-focused augmentation in supervised contrastive learning is shown in Fig. 4. The positive sample is augmented from an anchor image via `RandomResizedCrop` and `RandomHorizontalFlip` augmentations. The anchor image and the positive are passed through the shape-focused augmentation module respectively and the two output images are represented into the embedding space. The details of the implementation of shape-focused augmentation are the same as depicted in Sec. 3.1. In the embedding space, the anchor image and positives are pulled closer and nega-

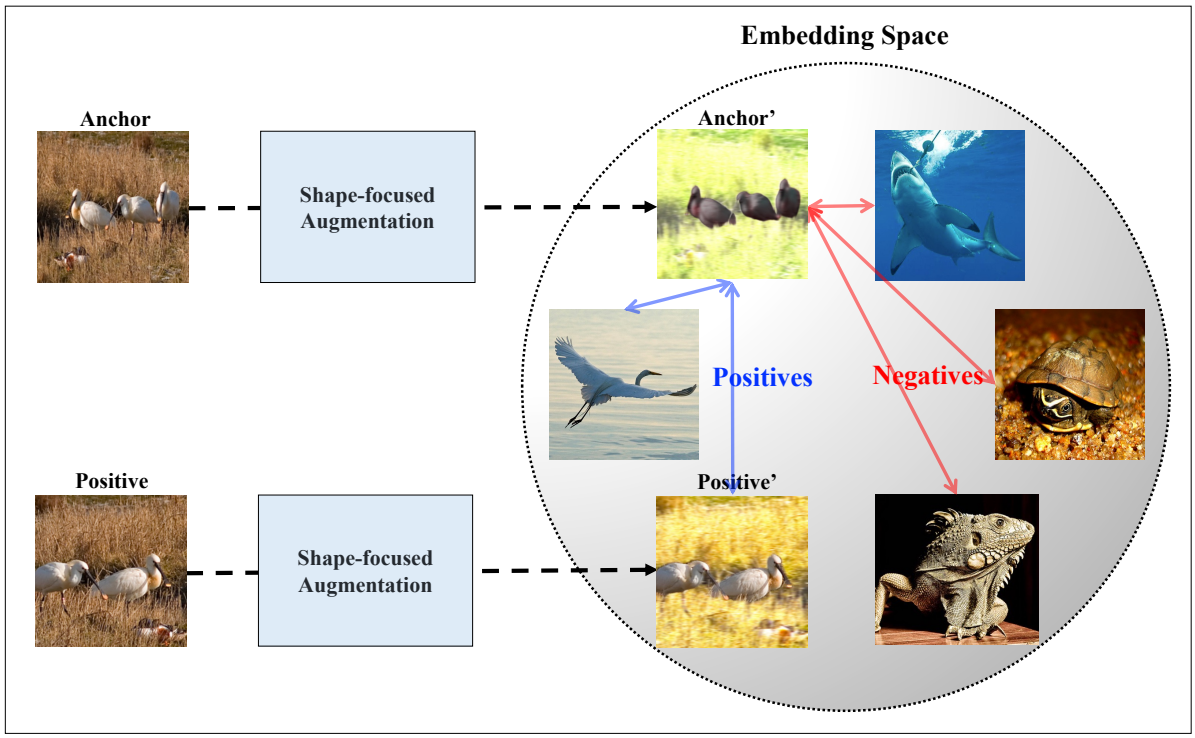

Figure 4. The overall overview of the application of shape-focused augmentation in supervised contrastive learning.

tives are pushed away from the anchor image. The positives include the image augmented from the anchor image and other images with the same label as the anchor image, and the negatives contain all of the remaining images in a batch.

We expect that the inherent learning property of supervised contrastive learning encourages the model to learn the common representations of shape attributes in the output images from shape-focused augmentation and the images with the same label. Accordingly, in the next section, we empirically show that the employment of shape-focused augmentation with contrastive learning methods leads us to achieve our conflicting goal, i.e. increasing the model's robustness to out-of-distribution samples without the decline of accuracy on the original dataset.

4. Experiment

In this section, we demonstrate the effectiveness of our method in both reducing the classifier's texture bias and increasing its accuracy on the in-domain dataset. We apply the shape-focused augmentation to two contrastive learning models, i.e., supervised contrastive learning (SupCon) [12] and SimCLR [2], each in supervised and self-supervised settings respectively. In Sec. 4.2, we verify the efficacy of our method under supervised contrastive learning model, and in Sec. 4.3 we show our method's efficiency in self-supervised learning setting, which is SimCLR.

4.1. Implementation details

Baselines. We compare our method with vanilla contrastive models, i.e. SupCon and SimCLR, and the model pretrained on counterfactual images generated by CGN [17].

Datasets. We utilized ImageNet-100 dataset for pretraining the vanilla model and the model with our method. As the comparison model, we pretrained the model jointly on ImageNet-100 dataset and counterfactual images generated from CGN as shown in Fig. 3 (a). We used OOD dataset to measure the model's robustness to texture bias and also used the validation set of ImageNet-100 to measure the accuracy on the in-domain samples.

Training details. We pretrained contrastive learning models with ResNet-50 backbone. All the implementations of the models are from the prior work [3] and five types of data augmentations are randomly applied to the images during the vanilla model's pretraining, which are `RandomResizedCrop`, `RandomHorizontalFlip`, `RandomGaussianBlur` `RandomGrayscale`, and `ColorJitter`. For our method, we employ the shape-focused augmentation to the models as illustrated in Sec. 3.4 during their pretraining. For the comparison model, we pretrained the model jointly on ImageNet-100 and counterfactual images. By following the linear evaluation protocol in both contrastive learning models, we trained a linear classifier on ImageNet-100 dataset on top of the frozen version of pretrained models.

4.2. Results on Supervised Contrastive Learning

Model	ImageNet-100	OOD-noise	OOD-style
SupCon	80.06	69.61	49
+CGN	77.42 (-2.64)	68.324	47
+ShapeAug	**82.24** (+2.18)	**71.82**	**54**

Table 3. Top-1 classification accuracy (%). SupCon denotes the vanilla supervised contrastive learning model. +ShapeAug denotes SupCon pretrained with shape-foucsed augmentation as shown in Fig. 4 and +CGN denotes SupCon jointly pretrained on ImageNet-100 and counterfactual images in Fig. 3 (a). Numbers in parenthesis is the accuracy difference on ImageNet-100 between the vanilla model and each model.

We show the efficiency of our method when it is applied on supervised contrastive learning. We compare the model pretrained with our method with the vanilla model and the model pretrained on the counterfactual images by measuring each model's accuracy on OOD datasets and ImageNet-100 dataset. Tab. 3 shows the comparison of each model's accuracy on each dataset.

Remarkably, our approach demonstrates the enhanced performance over the vanilla model on OOD datasets and even on the original dataset by large gaps. It shows the proposed method successfully reduces the model's propensity to predict the image's label by its texture information while not losing, even advancing, its ability to classify images in the original dataset. We suppose that it resulted from the supervised contrastive learning model's learning properties which reinforce our method's efficacy. In other words, the model is encouraged to learn the common representations in the images of the same label as well as the highlighted shape features in the output images of shape-focused augmentation by contrastively comparing them with images with different labels.

On the other hand, the model pretrained on counterfactual images along with ImageNet-100 shows the decreased accuracy on not only ImageNet-100 dataset but also OOD dataset. It respectively indicates the importance of preserving the semantic closeness to the original dataset when modifying the image's texture, and the efficaciousness of data augmentations in improving the model's robustness to texture bias.

4.3. Results on SimCLR

We show the effectuality of our method when it is applied on a self-supervised learning setting, i.e SimCLR which has a single positive augmented from an anchor image without utilizing label information in the dataset. We compare the performance of our method with that of the vanilla model and the model pretrained jointly on counterfactual images and ImageNet-100.

Model	ImageNet-100	OOD-noise	OOD-style
SimCLR	74.98	66.67	49.5
+CGN	69.74 (-5.24)	**67.59**	44
+ShapeAug	**77.64** (+2.66)	67.22	**53.5**

Table 4. Top-1 classification accuracy (%). +ShapeAug denotes SimCLR pretrained with the shape-foucsed augmentation in Fig. 2 and +CGN denotes SimCLR jointly pretrained on ImageNet-100 and counterfactual images in Fig. 3 (a). Numbers in parenthesis is the accuracy difference on ImageNet-100 between the vanilla model and each model.

Tab. 4 shows the comparisons of each model's accuracy on ImageNet-100 and OOD datasets. Our method also shows the enhanced accuracy over the vanilla model across all validation sets. Contrary to our method, the model pretrained on counterfactual images shows the lower accuracy on ImageNet-100 than the vanilla model. However, it shows higher accuracy on one of OOD datasets, `OOD-noise`, than the vanilla model and our method. This results were not shown in the experiment on supervised contrastive learning setting. We presume it is caused by the difference of the learning protocol between supervised contrastive learning and SimCLR, which means whether they incorporate images with the same label as the anchor image in positives or not.

4.4. Ablation studies

Model	ImageNet-100	OOD-noise	OOD-style
SupCon	80.06	69.61	49
+Aug-Aug	78.8 (-1.26)	70.90	**58.5**
+ShapeAug	**82.24** (+2.18)	**71.82**	54

Table 5. Top-1 classification accuracy (%). +Aug-Aug denotes SupCon pretrained by applying data augmentations twice. +ShapeAug denotes SupCon pretrained with shape-foucsed augmentation as shown in Fig. 4. Numbers in parenthesis is the accuracy difference on ImageNet-100 between the vanilla model and each model.

As an ablation study, we show that the trade-off between the model's robustness to out-of-distribution samples and the accuracy on the original dataset are not able to be overcome by merely applying multiple times of data augmentations during training. We compare the accuracy of our method with that of the supervised contrastive learning model which is pretrained by consecutively applying data augmentations twice.

Tab. 5 shows the accuracy on OOD datasets and ImageNet-100 by each model. Only our approach shows the higher accuracy across all test sets than the vanilla model.

Model	Mixed-Same	Mixed-Rand	BG-Gap ⇓
SupCon	76.59	69.04	7.55
+CGN	75.93	66.52	9.41
+ShapeAug	**79.19**	**72.30**	**6.89**

Table 6. Top-1 classification accuracy (%) on two subsets of Backgrounds Challenge dataset [19], i.e., Mixed-Same and Mixed-Rand, and BG-Gap score. +CGN denotes SupCon jointly pretrained on ImageNet-100 and counterfactual images as shown in Fig. 3 (a). +ShapeAug denotes SupCon pretrained with shape-foucsed augmentation as show in Fig. 4. Downwards Arrow ⇓ implies the lower measure the less dependence of the classifier on the image's background information.

The experiment results show the importance of differentiating the foreground and background of the object in an image when we give variety to the image's texture.

As an additional ablation study, we evaluate our method's accuracy on two subsets of Backgrounds Challenge dataset [19] and measure the gap between the accuracies on the two subsets.

The two subsets are Mixed-Same and Mixed-Rand dataset. Mixed-Same dataset contains the images with the background which is modified to the background of relevant classes of images in ImageNet. By contrast, Mixed-Rand dataset includes the images whose backgrounds are randomized and have no more correlation with the labels of the original images. Xiao *et al*. [19] also proposed BG-Gap score which is the gap in the model's accuracies on Mixed-Same and Mixed-Rand as the measure of the model's reliance on the background information when classifying the images. We compare the BG-Gap score of our method with that of other baselines.

Tab. 6 exhibits the accuracy on Mixed-Same and Mixed-Rand by each model and consequently its BG-Gap score. Our method not only displays the highest accuracy on both Mixed-Same and Mixed-Rand dataset but also the lowest BG-Gap score. It means that there is less drop in the model's classification performance even if the object in an image from the original dataset is located on out-of-distribution backgrounds. This outcome shows that our method enables the model to be trained to perceive images more structurally by differentiating the foreground and background of the object in an image while increasing the model's robustness to out-of-distribution samples.

5. Conclusion

In this paper, our work aims to overcome the trade-off between improving the classifier's robustness to out-of-distribution samples and increasing its accuracy on the in-domain samples. For this challenging issue, we proposed a novel data augmentation scheme called shape-focused augmentation. Our method differently modifies the texture of the foreground and background of the object in an image. The main intuition in shape-focused augmentation is training the model to concentrate on the global shape feature in the augmented image while reducing the model's dependence on the image's local texture feature when classifying images. We experiment our method on both the supervised learning model and the self-supervised learning model and demonstrated the effectiveness of our method by comparing its accuracy on various datasets with multiple baselines. As a result, we showed that our aims to overcome the conflicting trade-off issues in the classification model can be achieved when the shape-focused augmentation is applied to contrastive learning methods.

Acknowledgement

This work was partly supported by the Institute of Information & Communications Technology Planning Evaluation (2015-0-00310-SW.StarLab/20%, 2017-0-01772-VTT/10%, 2018-0-00622-RMI/20%, 2019-0-01371-BabyMind/20%, 2021-0-02068-AIHub/20%, 2021-0-01343-GSAI/10%) grant funded by the Korean government.

References

[1] H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh. Learning de-biased representations with biased representations. In *In International Conference on Machine Learning.*, pages 528–539, 2020. 2

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *In International conference on machine learning*, pages 1597–1607, 2020. 5, 6

[3] Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23(56):1–6, 2022. 6

[4] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *ICCV*, pages 2414–2423, 2016. 1, 2

[5] R. Geirhos, J. H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. In *Nature Machine Intelligence*, pages 665–673, 2020. 1

[6] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, and W. Brendel. Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems*, 2021. 2

[7] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *arXiv preprint arXiv:1811.12231*, 2018. 1, 2

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *In Proceedings of the IEEE conference on computer vision and pattern recognition.*, pages 770–778, 2016. 1

[9] K. Hermann, T. Chen, and S. Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 19000–19015, 2020. 2, 3, 4

[10] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1

[11] J. Jo and Y. Bengio. Measuring the tendency of cnns to learn surface statistical regularities. In *arXiv preprint arXiv:1711.11561.*, 2017. 1

[12] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, ..., and D. Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, pages 18661–18673, 2020. 5, 6

[13] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. In *Pattern Recognition*, volume 106, page 107404, 2020. 3

[14] S. Ringer, W. Williams, T. Ash, R. Francis, and D. MacLeod. Texture bias of cnns limits few-shot classification performance. In *arXiv preprint arXiv:1910.08519.*, 2019. 1

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, ..., and L. Fei-Fei. Imagenet large scale visual recognition challenge. In *International journal of computer vision 115(3)*, pages 211–252, 2015. 1

[16] Beery S., Van Horn G., and Perona P. Recognition in terra incognita. In *ECCV*, pages 456–473, 2018. 1

[17] A. Sauer and A. Geiger. Counterfactual generative networks. In *ICLR*, 2021. 1, 2, 4, 5, 6

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, ..., and A. Rabinovich. Going deeper with convolutions. In *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[19] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry. Noise or signal: The role of image backgrounds in object recognition. In *In International Conference on Learning Representations.*, 2021. 8