공학석사 학위논문

Shape-Emphasizing Augmentation for Improving Robustness of Self-Supervised Learning

자기지도 학습의 강건성 증대를 위한 형태 강조 증강

2023 년 2 월

서울대학교 대학원

협동과정 인공지능 전공

이 상 준

Shape-Emphasizing Augmentation for Improving Robustness of Self-Supervised Learning

지도 교수 장 병 탁

이 논문을 공학석사 학위논문으로 제출함

2023 년 1 월

서울대학교 대학원

협동과정 인공지능 전공

이 상 준

이상준의 공학석사 학위논문을 인준함

2023 년 1 월

위 원 장 서 봉 원 (인)

부위원장 장 병 탁 (인)

위 원 차 지 욱 (인)

Abstract

Self-supervised learning achieved remarkable advancement comparable to supervised learning in image classification. However, its achievement is confined to test samples independently and identically distributed (IID) with a training dataset. As in supervised learning models, poor robustness to out-of-distribution (OOD) distortions still exists in self-supervised learning models. On the contrary, humans are robust to OOD distortions, and it is attributed to their shape-oriented representation with lower reliance on texture. Several previous methods were suggested to induce the image classifiers to concentrate more on shape by augmenting training images with modified textures. However, they focused on supervised learning settings rather than self-supervised ones and brought a decreased accuracy on IID test samples as a trade-off. Thus, this paper introduces shape-emphasizing augmentation, a novel data augmentation scheme for self-supervised learning. This method highlights the object's shape in an image by applying random augmentations independently to the foreground and background of the object. The self-supervised learning model learns more shape-based representation with the proposed method. Extensive experiments present its effectiveness in improving robustness to OOD distortions without sacrificing the performance on IID test samples.

Keyword : Self-supervised learning, Texture bias, Shape-based representation, Robustness to out-of-distribution distortions
Student Number : 2021-24432

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Purpose of Research

Since the first advent of a convolutional neural network (CNN)-based approach, i.e., AlexNet (Krizhevsky et al., 2012), in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015), neural networks with deep layers opened a new era in object recognition. With the consistent progress in the deep learning algorithm, ResNet (He et al., 2016) eventually surpassed the performance of humans in the challenge.

This brilliant advancement of the deep learning model in the object recognition task was accomplished in a supervised manner, requiring a bunch of labeled data per single instance to learn visual representation. However, in contrast to supervised learning, self-supervised learning does not require label information for learning visual representation. Accordingly, it has recently attracted many researchers' interest due to its efficiency in that expensive and time-consuming manual data annotation is unnecessary (Doersch et al., 2015; Pathak et al., 2016; Larsson et al., 2017; Gidaris et al., 2018; Chen et al., 2020; He et al., 2020; Grill et al., 2020).

Instead of label supervision, self-supervised learning methods learn visual representations by solving auxiliary pretext tasks predicting pseudo-labels inherently derived from training samples (Doersch et al., 2015; Pathak et al., 2016; Larsson et al., 2017; Gidaris

(a) Original Image **(b) Shape-emphasized Image**

Figure 1.1 The example of shape-emphasizing augmentation. (a) Image sample of 'wallaby' in ImageNet dataset. (b) Same image sample that shape-emphasizing augmentation is applied.

et al., 2018) or by contrastively discriminating instances with using the concept of positive and negative pairs augmented from training images (Chen et al., 2020; He et al., 2020; Grill et al., 2020).

Along with the remarkable algorithmic development in recent years, self-supervised learning models, such as SimCLR (Chen et al., 2020), have shown comparable performance to supervised learning models in the classification accuracy on test samples independently and identically distributed (IID) with the training dataset.

However, despite the stark difference in the learning mechanism between supervised and self-supervised learning, Geirhos et al. (2020) empirically showed that self-supervised learning models exhibit similar properties to supervised learning models in some aspects. First, the image classification models based on CNN architecture heavily relied on texture information in images when predicting their labels, regardless of supervised or self-supervised learning models. Besides, they were vulnerable to classifying images with out-of-distribution (OOD) distortions, e.g., modified style or texture and added synthetic noises.

In this regard, the current image classification models tend to learn a shallow correlation between the superficial attributes in an image and its label as a shortcut to classify images rather than understanding more complex visual concepts inherent in objects of the images (Beery et al., 2018; Geirhos et al., 2020). Beery et al. (2018) empirically showed that the classification model correctly recognizes cows in a common context, cows on green grass, while it fails to classify cows on the beach or seaside.

More specifically, Geirhos et al. (2019) evaluated the label predictions of the CNN models trained on ImageNet (Deng et al., 2009) and human observers on image samples whose texture and shape cues conflict with each other, e.g., an elephant's skin texture covers an image of a cat. They measured the ratio of how many times classifiers, including human observers, correctly predicted the image's label by shape cue out of the total number of their correct predictions to either shape or texture cue. In this experiment, ImageNet-trained CNN models showed a high frequency of classifying images by local texture cues. In contrast, humans predicted image labels by the global shape features of the objects. Geirhos et al. (2019) also insisted that training these CNN models to learn more shape-oriented representation helps the improvement of their robustness to OOD distortions.

In this context, previous works (Geirhos et al., 2019; Hermann et al., 2020; Sauer & Geiger, 2021) took an approach to reduce the image classification model's high dependency on local texture cues to encourage them to concentrate more on global shape features of the objects. They proposed various methods of removing existing texture cues in the original training samples by replacing them with modified ones. Geirhos et al. (2019) changed the texture in ImageNet samples to the texture of artistic paintings. Sauer and Geiger (2021) generated image samples where the texture in the foreground and background of the object are independently modified to the texture of other ImageNet object classes, respectively. Hermann et al. (2020) analyzed the effect of data augmentations changing the appearance of images in mitigating the classifier's reliance on texture information.

Though these methods substantially induced the classifier to focus more on the object's global shape features rather than local texture cues in images, however, they accompanied decreased classification accuracy on IID test samples as a trade-off (Geirhos et al., 2019; Hermann et al., 2020; Sauer & Geiger, 2021). Moreover, these previous academic efforts to mitigate a high degree of texture bias in image classification models have primarily focused on the supervised learning setting, not the self-supervised learning one.

Hence, this paper casts a research question on how a self-supervised learning model can be trained to be more robust to OOD distortions by focusing more on the object's shape rather than local texture without sacrificing accuracy on IID samples.

1.2. Research Content

As shown in Figure 1.1, this work introduces a simple yet effective novel data augmentation method, shape-emphasizing augmentation. The proposed method encourages the self-supervised contrastive learning model to learn shape-based representation for improving the model's robustness to OOD distortions without sacrificing its classification accuracy on IID test samples. [①]

The key idea of this method is that a set of random data augmentations is independently applied to the object's foreground and background, which are partitioned by the object's shape mask

[①] A preliminary version of this work was presented at CVPR 2022 workshop on Human-Centered Intelligent Services: Safe and Trustworthy.

generated from a pre-trained salient object detection model. Consequently, the object's shape in the image becomes emphasized by the disparately augmented texture on each side of the shape contour.

Shape-emphasizing augmentation can be applied to the self-supervised contrastive learning models by replacing their existing data augmentation processes for generating multi-viewed samples from each image in a mini-batch. Then, the model learns shape-based representation by contrastively discriminating the emphasized shape features common in positive samples from those in negative samples. Consequently, the proposed method induces the model to rely less on local texture cues and refer more to the global shape feature.

Earlier works (Hermann et al., 2020; Geirhos et al., 2020; Geirhos et al., 2021) revealed that SimCLR (Chen et al., 2020) has outstanding robustness to OOD distortions over other self-supervised learning models due to its particular configuration of data augmentations. This work delves into the more orthogonal add-on effect of shape-emphasizing augmentation on the self-supervised contrastive learning model by applying it to SimCLR (Chen et al., 2020) with the same kinds and hyperparameters of data augmentations of the vanilla model.

With plenty of experiments based on SimCLR (Chen et al., 2020) and another base model, this work demonstrates the effects of shape-emphasizing augmentation in encouraging the self-supervised contrastive learning model to learn shape-based representation and improving the model's robustness to OOD distortion without losing its classification accuracy on IID test samples.

1.3. Outline of Research

This section delineates the overall outline of this paper.

Chapter 1 introduces this paper's research idea, goal, background, and a brief blueprint regarding the proposed method.

Chapter 2 presents previous works related to this paper. This part explains the preceding research stream to deliver the motivation and background of this work. It also narrows vague concepts within equivocal expressions down to the specific viewpoint that this paper targets.

Chapter 3 explains the details of shape-emphasizing augmentation and its application process to a self-supervised contrastive learning model. This chapter also provides the experiment results demonstrating the proposed method's effectiveness and compares it to other baselines.

Chapter 4 provides the experiment results indicating shape-emphasizing augmentation's effect on encouraging a self-supervised contrastive learning model to learn shape-based representation.

Finally, Chapter 5 concludes this paper's research by summarizing the experiment results, describing the limitations of this work, delineating some discussion points, and providing the direction of future research.

Chapter 2. Related Works

2.1. Self-supervised contrastive learning

Self-supervised learning does not require label information when learning visual representation from an image dataset. There are two types of self-supervised approaches to substitute label supervision: auxiliary pretext tasks and contrastive learning (Albelwi, 2022).

Earlier approaches (Doersch et al., 2015; Pathak et al., 2016; Larsson et al., 2017; Gidaris et al., 2018) utilized auxiliary pretext tasks to train a model in a self-supervised manner. These tasks have self-derived pseudo-labels from the data, and the model learns visual representation by solving these tasks.

Doersch et al. (2015) made the model to predict the relative positions of each sliced patch of an image from the center of the image. In this case, relative positions around the center of the image are pseudo-labels. Gidaris et al. (2018) suggested another pretext task in which a model predicts the degree of rotation of the rotated image from its original degree. The degree of rotation, e.g., 0, 90, 180, or 270 degrees, works as a pseudo-label in this task. Pathak et al. (2016) defined an inpainting task in which a model with an encoder-decoder architecture generates an omitted part of an image when given the whole image with the missing part. They took the missing part from the original image as a pseudo-label and used reconstruction loss and adversarial loss. Also, Larsson et al. (2017) introduced a colorization

task. To solve this task, a model restores each pixel's color information from the grayscaled image. These auxiliary pretext tasks were defined differently by each research's inductive bias regarding learning visual representation (Albelwi, 2022).

However, in recent years, the approach based on contrastive learning has come to the front of self-supervised learning (Chen et al., 2020; He et al., 2020; Grill et al., 2020). This approach contrastively discriminates an instance from others by maximizing the similarity between positive samples augmented from the same image and minimizing the similarity with negative samples augmented from other images.

Chen et al. (2020) proposed SimCLR, a simple framework for self-supervised contrastive learning. SimCLR is trained by NT-Xent loss (Sohn, 2016) which pulls the representations of positive samples augmented from the same image closer in an embedding space and pushes the representations of positive samples away from those of negative samples augmented from other remaining images of a mini-batch. SimCLR adopts a projection head to give a nonlinear transformation to representation from the backbone encoder. He et al. (2020) suggested another method called MoCo, utilizing a dynamic dictionary built as a queue containing plenty of key embeddings. The dictionary is updated as a queue, which means the oldest key embeddings are dequeued when new embeddings from each mini-batch are newly enqueued. A contrastive loss between a query embedding from the backbone encoder and key embeddings from a momentum encoder is calculated by utilizing InfoNCE loss (Oord et al.,

2018). Only the backbone encoder is updated by backpropagation of the contrastive loss. The weights of the momentum encoder are updated by the exponential moving average of those of the backbone encoder.

Grill et al. (2020) presented a slightly different method called BYOL, a self-supervised contrastive learning method without negative samples. BYOL utilizes only two positive samples augmented from the same original image. Each sample is fed to an online network and a target network, respectively. Then, the mean square error between each network's final representation is calculated as a loss. Only the online network is updated with it. The weights of the target network are updated by the exponential moving average of those of the online network.

The previous works (Geirhos et al., 2020; Geirhos et al., 2021) provided the experiment results about the robustness of self-supervised learning models to OOD distortions, and SimCLR (Chen et al., 2020) showed its superiority over other self-supervised learning models. Hermann et al. (2020) and Geirhos et al. (2021) analyzed that the robustness of SimCLR came from its data augmentations utilized for generating multi-viewed samples. SimCLR (Chen et al., 2020) is set as the base framework of self-supervised contrastive learning in this work. Shape-emphasizing augmentation utilizes the same kinds of augmentations and its hyperparameter setting of the vanilla model to investigate the orthogonal effect of the proposed method while preventing the influence of data augmentations itself.

2.2. Data augmentation for improving generalization

As claimed in earlier works (Chen et al., 2020; Hermann et al., 2020; Geirhos et al., 2020; Geirhos et al., 2021), data augmentation took a critical role in the self-supervised contrastive learning model's generalization on both IID and OOD samples.

Unlike other approaches for improving a model's generalization on data samples unseen during training, such as dropout (Srivastava et al., 2014), batch normalization (Ioffe & Szegedy, 2015), transfer learning (Yosinski et al., 2014), and pre-training (Erhan et al., 2010), data augmentation provides a data-space solution dealing with a training dataset as the root of the problem (Shorten & Khoshgoftaar, 2019).

Shorten and Khoshgoftaar (2019) categorized various approaches to image data augmentation into two broad contexts: basic image manipulation without utilizing deep learning models and other approaches based on deep learning models.

Basic image manipulations are grouped into subgroups (Shorten & Khoshgoftaar, 2019). The first one is geometric transformations. A flipping augmentation flips an image by a horizontal or vertical axis. A cropping augmentation crops the center of an image or the random position of an image with a specific size. A rotation augmentation rotates an image to the right or left side with a range of degrees from 1 to 359. A translation augmentation pushes an image left, right, up, or down in the fixed view frame. Shifting an image causes the empty part in the image, and this part is filled with values of 0, 255, random numbers, or gaussian noise.

Next, the color space transformations modify the value of RGB color channels in an image data, e.g., isolating the image data by each color channel or adjusting the image's brightness by changing pixel value in RGB channels (Shorten & Khoshgoftaar, 2019). There are image processing functions with these color space transformations, e.g., color jittering that randomly alters an image's brightness, contrast, saturation, and hue and gray scaling that converts the RGB color image to grayscale. They are widely used in self-supervised contrastive learning models (Chen et al., 2020; He et al., 2020; Grill et al., 2020).

There are also manipulation methods utilizing kernel filters. The kernel filter with a specific 2-dimensional size modifies an image by sliding across the image. For instance, the Gaussian blur filter is utilized in self-supervised contrastive learning models (Chen et al., 2020; Grill et al., 2020) to blur the image.

Explicitly erasing a randomly selected part in an image is another research stream in image data augmentation (DeVries & Taylor, 2017; Zhong et al., 2020). Their approaches were motivated by dropout (Srivastava et al., 2014), which stochastically paused the neural activations in each layer of CNN. Instead, they discarded the value of a randomly selected rectangular or square part in an input image and filled it with zero or random values.

Mixing up two randomly sampled data in a mini-batch is another data augmentation approach (Zhang et al., 2018; Yun et al., 2019). Zhang et al. (2018) proposed a method called MixUp, linearly interpolating two randomly selected images from a mini-batch with the combination ratio sampled from a beta distribution. The labels of

images are also mixed by following the ratio. Yun et al. (2019) suggested another mixup strategy called CutMix. CutMix pastes the patch cropped from an image on another image. Both images are randomly chosen from a mini-batch. The labels of two images, i.e., the image where a patch was cropped and another image where the patch was pasted, are mixed by following the ratio of the patch's area to the whole image's area.

On the other hand, deep learning models are also employed as an effective tool to augment image data.

Gatys et al. (2016) introduced neural style transfer transferring a specific image's style to another image with maintaining its original contents. Geirhos et al. (2019) generated additional training samples called stylized-ImageNet, where the styles of artistic paintings are transferred to the original ImageNet samples by utilizing the neural style transfer (Gatys et al., 2016). They also proposed to train the image classification model jointly on the original ImageNet samples and stylized-ImageNet samples to improve the model's robustness to OOD distortions.

Sauer and Geiger (2021) suggested the counterfactual generative network, which utilizes BigGAN (Brock et al., 2018) as a base generative model to create counterfactual images where the texture in the foreground and background of the object are independently generated as the texture of other object classes in ImageNet, respectively. For instance, a counterfactual image can consist of an ostrich's shape, the strawberry's foreground texture, and the water's background. They suggested utilizing these counterfactual images as

the additional training samples with the original ImageNet samples to train the classifier to be invariant to the specific element in an image.

The concept of meta-learning is also utilized to optimize the best configuration of data augmentations. AutoAugment (Cubuk et al., 2019) and RandAugment (Cubuk et al., 2020) are approaches to automatically find an effective data augmentation policy for a target dataset. AutoAugment (Cubuk et al., 2019) leveraged reinforcement learning as a search algorithm to find the best composition and sequences of image processing functions with the optimal search of probabilities and magnitude of data augmentations. RandAugment (Cubuk et al., 2020) suggested the practical version of AutoAugment (Cubuk et al., 2019) by simplifying the search space, i.e., optimizing a single distortion magnitude and setting the probability of each image processing function to uniform. RandAugment (Cubuk et al., 2020) exhibited higher efficiency than AutoAugment (Cubuk et al., 2019) in terms of computational expense while achieving comparable or better performance.

In this paper, shape-emphasizing augmentation utilizes the same data augmentations with its base self-supervised contrastive learning model to observe the pure effect of the proposed method. For example, the same augmentation types and hyperparameters in the vanilla model are utilized when the proposed method is applied to SimCLR (Chen et al., 2020). Hence, the introduced method in this paper is primarily based on geometric transformations, color space transformations, and kernel filters, which are prevalently leveraged in self-supervised contrastive learning models.

2.3. Improving shape bias for robustness to distortions

Humans recognize and distinguish a particular object from others, even if they do not know how to call them. It indicates that the label with which people call the object may not be necessary for object recognition in human vision. Coincidentally, self-supervised contrastive learning models (He et al., 2020; Chen et al., 2020; Grill et al., 2020) also aim to learn visual representations without labels. This learning framework got a burgeoning interest by alleviating the inefficiency in preparing the labeled dataset necessary for the conventional supervised learning framework. It also achieved remarkable accomplishments in object recognition in recent years.

However, the previous works (Geirhos et al., 2020; Geirhos et al., 2021) empirically showed that self-supervised contrastive learning models still showed many analogous properties with supervised learning models in generalization on OOD distortions while revealing a striking difference from humans. Specifically, they reported the experiment results showing that the supervised and self-supervised contrastive learning models have a high degree of texture bias and vulnerability to OOD distortions, unlike human observers.

Geirhos et al. (2019) argued that the distinct robustness of humans to OOD distortions comes from their shape-oriented representation, compared to the image classification models heavily depending on local texture cues in images regardless of supervised or self-supervised contrastive learning models.

In this regard, Geirhos et al. (2019) measured each classifier's

degree of dependence on one of the features between the object's shape and local texture cues when classifying images by evaluating the model on the texture-shape cue conflict set. This dataset consists of test samples whose original texture is modified to the texture of another object class by utilizing neural style transfer (Gatys et al., 2016). Hence, each sample in the dataset has two labels per sample, i.e., labels for shape and texture, respectively. Then, they measured each classifier's ratio of the number of correct predictions by a shape label to the total number of correct answers on either the shape or texture label. This ratio is termed shape bias, indicating how much a specific classifier recognizes objects by their global shape features rather than by local texture cues. As a result, they argued that CNNs trained on ImageNet highly depend on texture when classifying images. On the contrary, humans predict image labels by the shape of objects in the images.

The previous research (Geirhos et al., 2019; Hermann et al., 2020; Sauer & Geiger, 2021) suggested methods to train image classification models to work with human-like characteristics in object recognition, such as a high degree of shape bias and robustness to OOD distortions. These methods commonly removed the existing texture cues in the original training samples to prohibit the classifiers from relying on texture information and encourage them to concentrate more on the object's global shape features.

Geirhos et al. (2019) and Sauer and Geiger (2021) proposed to train the classifiers jointly on ImageNet training samples and newly generated samples of modified textures from the original ImageNet

samples. In detail, Geirhos et al. (2019) introduced stylized-ImageNet, whose styles are transferred from artistic paintings by employing neural style transfer (Gatys et al., 2016). Sauer and Geiger (2021) proposed the counterfactual generative network generating counterfactual images where the foreground and background of the object are separately filled with the texture of other object classes in ImageNet.

Hermann et al. (2020) analyzed the effect of data augmentations in reducing the texture bias of image classifiers. They demonstrated that random cropping increases the texture bias while appearance-modifying augmentations, e.g., color distortion, gaussian blur, and gaussian noise, relieve the model's reliance on texture. They also figured out that the effect of these data augmentations works cumulatively.

However, unfortunately, the removal of local texture cues in the training dataset showed a trade-off between increased shape bias of the classifiers and decreased accuracy on the IID test set (Geirhos et al., 2019; Hermann et al., 2020; Sauer & Geiger, 2021; Tuli et al., 2021). In other words, the previous methods promote the image classification model to refer relatively more to global shape features by restraining the model from focusing on texture information while degrading the classifier's accuracy on the IID test samples.

Moreover, the concerns in these previous approaches were not in enhancing the self-supervised contrastive learning model’s robustness to OOD distortions.

Hermann et al. (2020), Geirhos et al. (2020), and Geirhos et al.

(2021) revealed that SimCLR's (Chen et al., 2020) outstanding robustness to OOD distortions compared to other self-supervised learning models comes from its data augmentations. However, their curiosity was targeted at checking the effect of SimCLR's data augmentations when applied to supervised learning models. In short, their interest was in verifying the importance of SimCLR's data augmentations in reducing the image classification models' texture bias rather than improving the self-supervised contrastive learning model's robustness to OOD distortions.

On the contrary, Chen et al. (2020) experimented with different combinations of data augmentations and proposed the best configuration of augmentations to increase SimCLR's performance. However, their concerns were to find a better set of transformations to increase the model's accuracy on IID test samples rather than on OOD distortions. Interestingly, they also found that AutoAugment (Cubuk et al., 2019), a more sophisticated data augmentation policy devised in a supervised learning setting, failed to facilitate SimCLR's generalization on the IID test set.

In this context, this paper proposes a novel data augmentation scheme more directly aiming at strengthening the self-supervised contrastive learning model's robustness to OOD distortions without declining the performance on IID test sets.

Chapter 3. Shape-Emphasizing Augmentation

3.1. Problem Statement

The previous methods (Geirhos et al., 2019; Hermann et al., 2020; Sauer & Geiger, 2021) struggled to mitigate the image classifier's heavy reliance on local texture cues by diversifying the existing texture in training images to overcome its poor generalization to OOD distortions. Like human observers robust to OOD distortions due to shape-oriented representation (Geirhos et al., 2019), the prior works pursued to encourage the image classification model to focus more on the object's shape. They took the approach of removing the existing texture cues in the training images to accomplish it. However, their approach deteriorated the classifier's performance on the IID test sets as a trade-off. Interestingly, it was also reported that self-supervised learning models also show similar characteristics in robustness to OOD distortions with supervised learning ones (Geirhos et al., 2020; Geirhos et al., 2021). However, unfortunately, the previous augmentation methods assumed a supervised learning setting, not a self-supervised learning one.

Accordingly, this paper proposes a novel data augmentation policy, shape-emphasizing augmentation, suitable for self-supervised contrastive learning models. Shape-emphasizing augmentation applies a set of random augmentations to training images but separately to the foreground and background of the objects in images. As displayed in

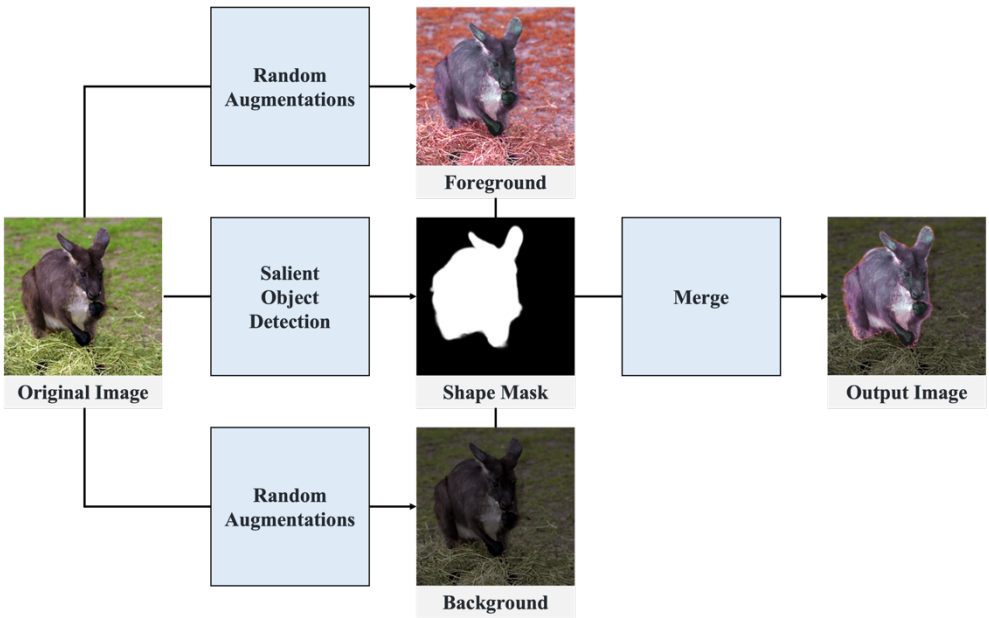

Figure 3.1 The overall processes of the shape-emphasizing augmentation.

Figure 1.1, the discrepancy of texture between the inside and outside of the object's shape mask visually emphasizes its global shape while also diversifying texture in the original training images. When applied to self-supervised contrastive learning models, they contrastively learn shape-based representation by comparing highlighted shape features common in positive samples with those of negative samples. More details about the proposed method are explained below.

3.2. Method

The detailed process of shape-emphasizing augmentation is portrayed in Figure 3.1.

Given an original image augmented by geometric transformations, e.g., cropping and flipping, the non-geometric transformations, e.g., color jittering, gray scaling, and gaussian blurring, are randomly applied to the original image two independent times. This process produces two dissimilarly augmented views: foreground and background images. Those two images are merged into the output image by the object's shape mask generated from the pre-trained salient object detection model.

The shape-emphasizing augmentation can be easily applied to the self-supervised contrastive learning model. The overall flow is depicted in Figure 3.2. Given two images differently augmented from the same image by geometric transformations, i.e., an anchor image and its positive sample, the shape-emphasizing augmentation is applied to each image respectively. Consequently, two output images

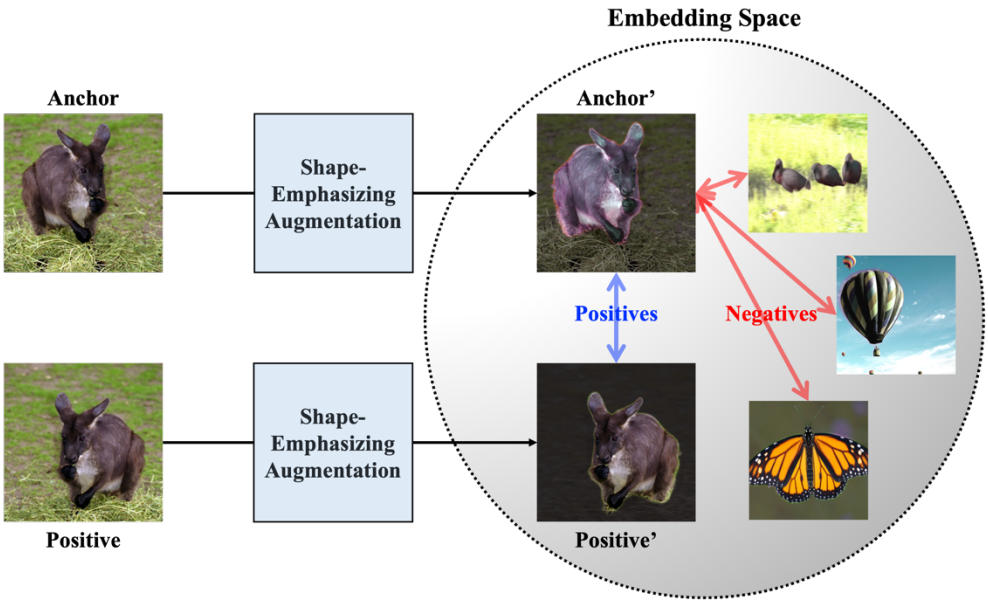

Figure 3.2 The overview of the application of shape-emphasizing augmentation to a self-supervised contrastive learning model.

with emphasized shapes are created and represented in an embedding space through the network. In the embedding space, a contrastive loss maximizes the similarity between the representations of positive samples and minimizes their similarity with negative samples. Consequently, the self-supervised contrastive learning model learns shape-based representations by contrasting accentuated shape features common in positive samples with negative samples. The additional crucial point is that shape-emphasizing augmentation still utilized the same data augmentation setting of the vanilla self-supervised contrastive learning model.

Algorithm 1 indicates more detailed procedures for applying shape-emphasizing augmentation to one of the self-supervised contrastive learning models, SimCLR (Chen et al., 2020). The gray-colored part is the same as the original SimCLR paper; only the black-colored part is changed with a few additive steps.

From a set of random data augmentations T related to geometric changes, two different transformations, t and t', are independently drawn and applied to the image x_k sampled from a minibatch, respectively. Accordingly, two differently augmented views from the image x_k with geometric transformations are generated, which are notated as $\tilde{x}_{2k-1}$ and $\tilde{x}_{2k}$.

Then, a transformation r is sampled from another set of random data augmentations R related to non-geometric modification and applied to $\tilde{x}_{2k-1}$ and $\tilde{x}_{2k}$, which are geometrically augmented images. The non-geometric transformation r is applied two independent times per image to create each pair of the foreground and background views,

Algorithm 1 Shape-Emphasizing Augmentation on SimCLR

Input: batch size N, constant τ, structure f, g, T, R, u, n
for sampled minibatch $\{x_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two geometric transformations $t \sim T$, $t' \sim T$

    # the first shape-emphasizing augmentation
    $\tilde{x}_{2k-1} = t(x_k)$
    draw a non-geometric transformation $r \sim R$
    $\tilde{x}_{2k-1}^{foreground} = r(\tilde{x}_{2k-1})$
    draw a non-geometric transformation $r' \sim R$
    $\tilde{x}_{2k-1}^{background} = r'(\tilde{x}_{2k-1})$
    $\tilde{x}_{2k-1}^{shapemask} = u(\tilde{x}_{2k-1})$
    $\tilde{m}_{2k-1} = \tilde{x}_{2k-1}^{shapemask} \odot \tilde{x}_{2k-1}^{foreground} + (1 - \tilde{x}_{2k-1}^{shapemask}) \odot \tilde{x}_{2k-1}^{background}$
    $\tilde{m}_{2k-1}^{normalized} = n(\tilde{m}_{2k-1})$
    $h_{2k-1} = f(\tilde{m}_{2k-1}^{normalized})$
    $z_{2k-1} = g(h_{2k-1})$

    # the second shape-emphasizing augmentation
    $\tilde{x}_{2k} = t'(x_k)$
    draw a non-geometric transformation $r'' \sim R$
    $\tilde{x}_{2k}^{foreground} = r''(\tilde{x}_{2k})$
    draw a non-geometric transformation $r''' \sim R$
    $\tilde{x}_{2k}^{background} = r'''(\tilde{x}_{2k})$
    $\tilde{x}_{2k}^{shapemask} = u(\tilde{x}_{2k})$
    $\tilde{m}_{2k} = \tilde{x}_{2k}^{shapemask} \odot \tilde{x}_{2k}^{foreground} + (1 - \tilde{x}_{2k}^{shapemask}) \odot \tilde{x}_{2k}^{background}$
    $\tilde{m}_{2k}^{normalized} = n(\tilde{m}_{2k})$
    $h_{2k} = f(\tilde{m}_{2k}^{normalized})$
    $z_{2k} = g(h_{2k})$
  end for

  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = z_i^{\top} z_j / \|z_i\| \|z_j\|$
  end for

  define $\ell(i,j)$ as $\ell(i,j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k}/\tau)}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$

$\tilde{x}_{2k-1}^{foreground}$ and $\tilde{x}_{2k-1}^{background}$, and $\tilde{x}_{2k}^{foreground}$ and $\tilde{x}_{2k}^{background}$. Here, the r is newly sampled from R on every implementation.

Each pair of foreground and background views is merged into the output image, $\tilde{m}_{2k-1}$ and $\tilde{m}_{2k}$, by multiplying a shape mask $\tilde{x}_{2k-1\,(or\,2k)}^{shapemask}$ to the foreground view and $1-\tilde{x}_{2k-1\,(or\,2k)}^{shapemask}$ to the background view and then summing the output values. The pre-trained salient object detection model u extracts the shape masks from each image, $\tilde{x}_{2k-1}$ and $\tilde{x}_{2k}$. The merged images are normalized and fed into the backbone encoder f(·) and projection head g(·) sequentially to obtain representations, z_{2k-1} and z_{2k}, in the embedding space.

The above procedures are done for all images in the minibatch, and pairwise similarity $s_{i,j}$ is calculated between whole embeddings of multi-viewed samples augmented from the images in the minibatch. Eventually, a contrastive loss for a positive pair of samples (i,j) is calculated by NT-Xent (the normalized temperature-scaled cross-entropy) loss with a temperature parameter τ as defined below:

$$s_{i,j} = z_i^{\mathrm{T}} z_j / \|z_i\| \|z_j\| \tag{3.1}$$

$$\ell(i,j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(s_{i,k}/\tau)} \tag{3.2}$$

The final objective function is calculated as defined below:

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\,\ell(2k-1,2k) + \ell(2k,2k-1)] \tag{3.3}$$

Accordingly, the backbone encoder f(·) and projection head g(·) are trained to minimize the objective function $\mathcal{L}$.

3.3 Experimental Setup

The shape-emphasizing augmentation is applied to SimCLR (Chen et al., 2020), and its performance is compared with the vanilla model to assess the effect of the proposed method on the self-supervised contrastive learning model. Various strategies of data augmentation methods were presented to develop the image classification model's generalization beyond training data. However, unfortunately, their interest was confined to the supervised learning setting. Then, extending the employment of these methods to a self-supervised learning setting and comparing them with shape-emphasizing augmentation is a meaningful way to demonstrate the appropriateness and necessity of the proposed method in a self-supervised learning setting.

3.3.1 Baselines

In addition to the vanilla SimCLR (Chen et al., 2020), the shape-emphasizing augmentation is compared with other types of data augmentation methods described in section 2.2: RandAugment (Cubuk et al., 2020), CutMix (Yun et al., 2019), and counterfactual generative network (Sauer & Geiger, 2021).

RandAugment (Cubuk et al., 2020) and CutMix (Yun et al., 2019)

are applied to SimCLR (Chen et al., 2020) by replacing its original data augmentation process for generating multi-viewed samples from each training image. For the employment of the counterfactual generative network (Sauer & Geiger, 2021) on SimCLR (Chen et al., 2020), it generates counterfactual images as the additional training samples and SimCLR is trained jointly on these counterfactual images and the original ImageNet training images. In a supervised learning setting, CutMix (Yun et al., 2019) and counterfactual generative network (Sauer & Geiger, 2021) utilized label information of the sampled pair of images to be mixed up or the components comprising a counterfactual image. However, they are employed just for augmenting training samples without utilizing label information when applied to SimCLR (Chen et al., 2020), learning visual representation from unlabeled data in a self-supervised manner.

3.3.2 Datasets

Each model is trained on the ImageNet (Deng et al., 2009) training set. ImageNet is one of the most prevailing datasets for visual representation learning in the deep learning era. It consists of 1,281,167 training images of 1,000 object classes and 50,000 validation samples.

After training, the models are evaluated on the validation set of ImageNet (Deng et al., 2009) and the OOD benchmark dataset (Geirhos et al., 2021), respectively. The OOD benchmark dataset comprises 17 subsets, and each subset contains distorted ImageNet samples by a

specific modification. The distortions include changes to style and texture or the addition of synthetic noises, e.g., sketch, stylized, edge, silhouette, cue conflict, colour vs. grayscale, low contrast, high-pass, low-pass (blurring), phase noise, true power spectrum vs. power equalisation, true vs. opponent colour, rotation, eidolon I, eidolon II, eidolon III, and uniform noise. The detailed source of each subset is delineated in the previous work (Geirhos et al., 2021).

However, the experiment in this paper targeted only ten subsets from a total of 17 subsets. The nine subsets assess the model's OOD distortion robustness, and the one remaining subset, cue conflict, is for measuring the degree of shape bias, indicated in section 4.2. Seven omitted subsets here are colour vs. grayscale, low-pass (blurring), true vs. opponent colour, contrast, rotation, eidolon I, and eidolon II. Out of these subsets, colour vs. grayscale, low-pass (blurring), true vs. opponent colour, and contrast are not used due to the relevance to SimCLR's data augmentation types, i.e., color jittering, gray scaling, and gaussian blurring. Geirhos et al. (2018) empirically showed that the image classifiers handle well the distortion type on which they are trained but still fail to generalize toward new types of distortion. Hence, subsets of distortions related to SimCLR's data augmentations are ignored to rule out these augmentations' influence in evaluating the OOD distortion generalization of baselines based on SimCLR. The remaining unused subsets are due to irrelevance to research scope, e.g., rotation, and duplication in types of distortions, e.g., eidolon I and eidolon II.

3.3.3 Metrics

Throughout this paper, the models are appraised by their accuracy on test sets. Accuracy is a conventional and fundamental evaluation metric for measuring an image classifier's performance. However, one metric is additionally adopted to examine more in-depth properties of baseline models toward OOD distortion generalization. It is error consistency.

Geirhos et al. (2020) suggested gauging the agreement in decision strategy between two object recognition models by measuring error consistency between them. Error consistency is the metric to measure the ratio of how many times two different models make the same decisions for each stimulus to the total number of trials in the experiment, reflecting how much consistency exceeds the expected overlap due to chance. The fundamental assumption in adopting this metric is that two systems use a similar decision strategy if they make similar errors (Geirhos et al., 2020). Error consistency between two models, i and j, is formulated by Cohen's kappa (Cohen, 1960) as defined below:

$$\kappa_{i,j} = \frac{c_{obs_{i,j}} - c_{exp_{i,j}}}{1 - c_{exp_{i,j}}} \tag{3.4}$$

$c_{obs_{i,j}}$ is calculated by $c_{obs_{i,j}} = \frac{e_{i,j}}{n}$, where $e_{i,j}$ is the number of the same decisions of two models, i.e., either both right or both incorrect, to each image in the test set and n is the total number of samples in

the test set. $c_{exp_{i,j}}$ is the expected error overlap due to chance and it is calculated by $c_{exp_{i,j}} = p_i p_j + (1 - p_i)(1 - p_j)$, where p_i and p_j are the accuracies of each observer, respectively.

Geirhos et al. (2021) provided the OOD benchmark dataset together with human observers' image-level decision data on the benchmark to measure error consistency between humans and other image classification models. This paper utilizes error consistency with human observers as one of the additional metrics to know more sophisticated characteristics of baselines regarding OOD distortions by comparing them with humans.

3.3.4 Implementation Details

The code-level implementation of SimCLR (Chen et al., 2020) is based on the GitHub repository of da Costa et al. (2022). The shape-emphasizing augmentation is also implemented on top of this code.

Each baseline model has a ResNet-50 (He et al., 2016) backbone. The backbone encoder is pre-trained for 100 epochs using SGD with an initial learning rate of 1.2, a weight decay of 0.000001, a batch size of 1,024, and layer-wise adaptive rate scaling (You et al., 2017) to adjust the learning rate during pre-training. After pre-training, a linear classifier is added to the backbone encoder. They are fine-tuned on ImageNet training samples with label information, using SGD with a learning rate of 0.4 and a batch size of 1,024. During fine-tuning, two random data augmentations are applied, i.e., random resized cropping and random horizontal flipping. Each model is fine-tuned until the

validation loss is converged to prevent the model from being overfitted to training samples, which was ten epochs here.

Shape-emphasizing augmentation utilizes the same five data augmentations as the vanilla SimCLR (Chen et al., 2020). However, they are applied differently, as described in section 3.2. Out of five data augmentations, random resized crop and random horizontal flip are from PyTorch's implementation and applied with the probability of 1 and 0.5, respectively. The remaining color jitter, random grayscale, and random gaussian blur are from the implementation of Kornia (Riba et al., 2020) and applied with the probability of 0.8, 0.2, and 0.5, respectively. Further details of hyperparameters for each data augmentation are described in the SimCLR (Chen et al., 2020). For the salient object detection model to extract a shape mask from an image, a pre-trained U2-Net (Qin et al., 2020) was employed.

For the experiment of testing RandAugment (Cubuk et al., 2020) on SimCLR (Chen et al., 2020), the existing augmentation steps in the vanilla model were replaced with RandAugment. The implementation was from the library of timm (Ross, 2019). The specific parameter of RandAugment here is 'rand-m9-mstd0.5-inc1', which means that the magnitude of 9, the number of transformation operations of 2, the standard deviation of the magnitude noise of 0.5, and the use of augmentations increasing in severity with magnitude.

The SimCLR (Chen et al., 2020) jointly trained on ImageNet samples and counterfactual images is also evaluated in this paper. The dataset configuration followed the previous work (Sauer & Geiger, 2021), i.e., half of the total samples for training are around 600,000

counterfactual images. The original paper's classifier has a shared CNN backbone and three multiple heads to independently predict each label, i.e., shape, foreground texture, and background texture in a counterfactual image. On the other hand, for SimCLR, just a backbone encoder was pre-trained in a self-supervised manner without any label information. After pre-training, a linear classifier is attached to the backbone encoder and fine-tuned by a shape label's supervision.

The implementation of CutMix (Yun et al., 2019) was adopted from the library of timm (Ross, 2019). The combination ratio between two images to be mixed up is sampled from the beta distribution, and the alpha value for it is set by 1.

Training took around seven days per baseline by multi-GPU training with distributed data-parallel module in PyTorch using 4 GeForce RTX 3090 machines.

3.4. Results and Analysis

SimCLR trained with the shape-emphasizing augmentation is compared with other baselines by testing each model on the ImageNet validation set and each subset of the OOD benchmark dataset. The classification accuracy of each model on each dataset is presented in Table 3.1. Also, the average accuracy across whole subsets of the OOD benchmark is represented in Figure 3.3.

Remarkably, the shape-emphasizing augmentation method demonstrates a superior result over the vanilla model on most OOD benchmark datasets. Especially, the proposed method boasts the most

Dataset	SimCLR	+ Shape-Augment	+ Rand-Augment	+ CGN	+ CutMix
ImageNet	67.10	**68.32** (+ 1.22)	64.57 (−2.53)	66.99 (−0.11)	65.01 (−2.09)
edge	18.13	**24.38** (+ 6.25)	18.13 (+ 0.00)	15.63 (−2.50)	16.25 (−1.88)
silhouette	38.75	40.00 (+ 1.25)	34.38 (−4.37)	**43.13** (+ 4.38)	40.63 (+ 1.88)
sketch	55.00	**58.00** (+ 3.00)	54.00 (−1.00)	55.13 (+ 0.13)	52.38 (−2.62)
stylized	29.88	**34.13** (+ 4.25)	26.75 (−3.13)	29.25 (−0.63)	29.13 (−0.75)
power-equalisation	80.98	**82.41** (+ 1.43)	74.11 (−6.87)	76.88 (−4.10)	78.48 (−2.50)
eidolon III	33.67	33.20 (−0.47)	31.88 (−1.79)	32.19 (−1.48)	**34.30** (+ 0.63)
uniform-noise	40.47	41.41 (+ 0.94)	33.05 (−7.42)	41.79 (+ 1.32)	**41.80** (+ 1.33)
high-pass	29.22	32.66 (+ 3.44)	31.41 (+ 2.19)	31.56 (+ 2.34)	**33.20** (+ 3.98)
phase-scrambling	47.32	**48.84** (+ 1.52)	44.82 (−2.50)	45.98 (−1.34)	46.88 (−0.44)

Table 3.1. Top-1 classification accuracy (%) on IID and OOD test sets. + ShapeAugment, + RandAugment, and + CutMix denote SimCLRs pre-trained with shape-emphasizing augmentation, RandAugment, and CutMix as their data augmentation methods, respectively. + CGN is SimCLR jointly pre-trained on ImageNet samples and counterfactual images. Numbers in parenthesis are the accuracy difference on each dataset between the vanilla SimCLR and each model. Bold and underlined numbers stand for the best and second-best classification accuracy, respectively.

significant boost, 6.25%p, for SimCLR's performance on the edge subset. The edge subset comprises samples where only an outline of the object remains without any texture information. It supports the hypothesis that the shape-emphasizing method induces the self-supervised contrastive learning model to focus more on global shape features than local texture cues in images.

Another encouraging point is that this enhanced generalization on OOD distortions comes without sacrificing the accuracy on IID test samples, ImageNet validation set here; instead, it comes with slightly increased accuracy.

The intensified robustness to OOD distortions from the proposed method stands out more exceptionally when compared with other baselines. Figure 3.3 indicates that the naive application of the augmentation methods devised in supervised learning settings is not extended smoothly to the self-supervised learning model. Moreover, Table 3.1 shows that they hurt the SimCLR's accuracy on IID test samples.

Specifically, RandAugment (Cubuk et al., 2020) exhibits poor accuracy on IID test samples when applied to SimCLR (Chen et al., 2020). This result is along the same line as the experiment by Chen et al. (2020). The experiment revealed a disappointing result in the accuracy on IID test samples when AutoAugment (Cubuk et al., 2019) was employed on SimCLR (Chen et al., 2020). AutoAugment is another method to automatically find the effective configuration of data augmentation policy like RandAugment, but with much more extensive search space. Both methods assume the supervised learning setting

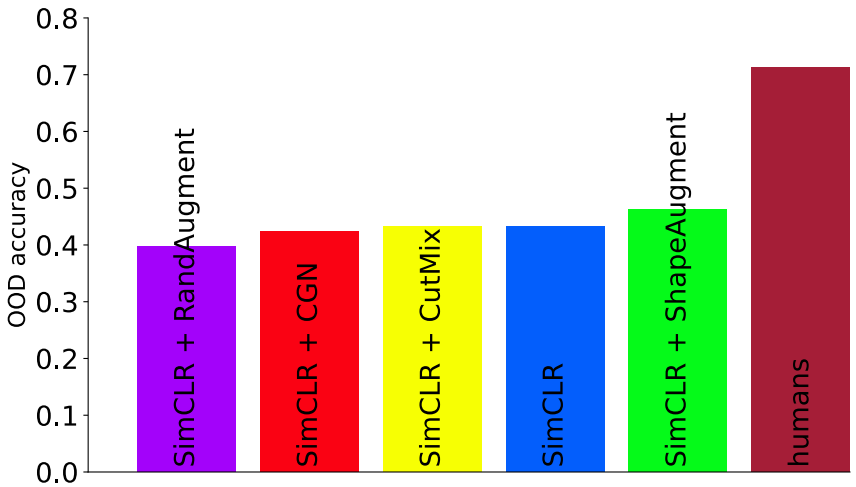

Figure 3.3 OOD accuracy averaged across datasets. The notation for each model is the same as the explanation in the caption of Table 3.1.

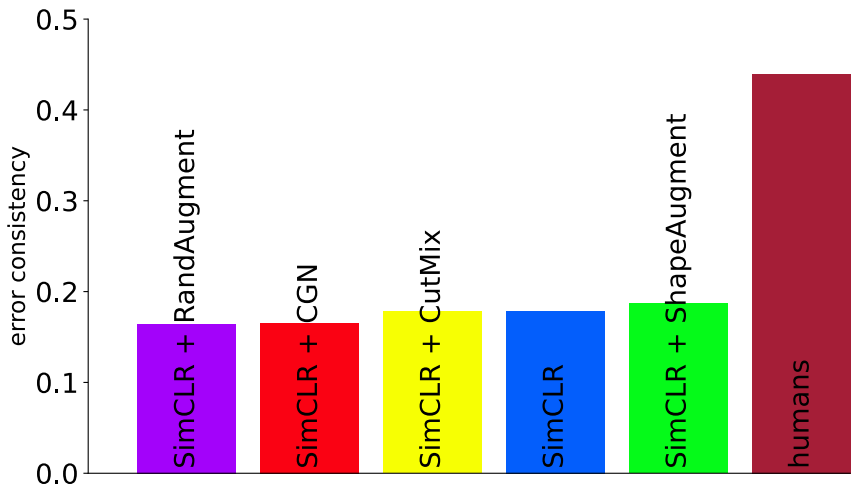

Figure 3.4 Image-level error consistency with human observers across OOD benchmark dataset.

during their search procedures for augmentation policy. The experiment results in SimCLR (Chen et al., 2020) and this paper imply that the data augmentation policies found by AutoAugment and RandAugment are not transferred well to a self-supervised learning setting. Additionally, Figure 3.3 exposes RandAugment is also ineffective in terms of generalization to OOD distortions.

The experiment results of the remaining two baselines, CGN (Sauer & Geiger, 2021) and CutMix (Yun et al., 2019), are also inferior to the shape-emphasizing augmentation and the vanilla model. These methods combine the visual components taken from different images of different object classes into the augmented image. For example, the counterfactual generative network (Sauer & Geiger, 2021) fills two sections divided by a specific object's shape mask with texture from other different object classes. CutMix (Yun et al., 2019) pastes the cropped patch from one image on top of another. When the supervised learning model is trained on these images, it is supervised by all labels of object classes that comprise the augmented image. In this regard, Table 3.1 and Figure 3.3 imply that just naively utilizing these approaches as an augmentation method for self-supervised learning without label information is not beneficial for its generalization to both IID and OOD samples.

The experiment results regarding each baseline model's error consistency with human observers are presented in Figure 3.4. Interestingly, SimCLR trained with the shape-emphasizing augmentation displays higher error consistency with human observers on OOD distortions than other baselines. It signifies that the proposed

method induces SimCLR to adopt a more similar decision strategy on each OOD sample with humans. Humans recognize objects more by global shape cues in images than local texture information and are more robust to OOD distortions than most object recognition models (Geirhos et al., 2019; Geirhos et al., 2020; Geirhos et al., 2021).

In addition to the above experiments mainly based on SimCLR (Chen et al., 2020) with ResNet-50 (He et al., 2016) backbone, the shape-emphasizing augmentation was also tested to SimCLR with the vision transformer (ViT) (Dosovitskiy et al., 2020) backbone. ViT is a novel architecture that recently came to the fore in the computer vision field by surpassing the standard CNN-based model's performance on the conventional image classification benchmark, and its operation is based on the self-attention mechanism of the transformer (Vaswani et al., 2017). After the advent of ViT and its impressive performance in a supervised learning setting, there have been trials to utilize ViT in a self-supervised learning framework as its backbone model, and it was also shown in prior work (Chen et al., 2021) that SimCLR is implemented with ViT backbone.

For the experiment in this section, the ViT backbone is implemented from the library of timm (Ross, 2019). Specifically, the ViT-Small model with the main configuration of a 16 x 16 input patch size, a hidden dimension size of 384, and 6 heads of multi-head attention is utilized. The details of training and evaluating the model are the same as in section 3.3, except for the batch size of 512.

Figure 3.5 and Figure 3.6 presents the evaluation result on the OOD benchmark dataset.

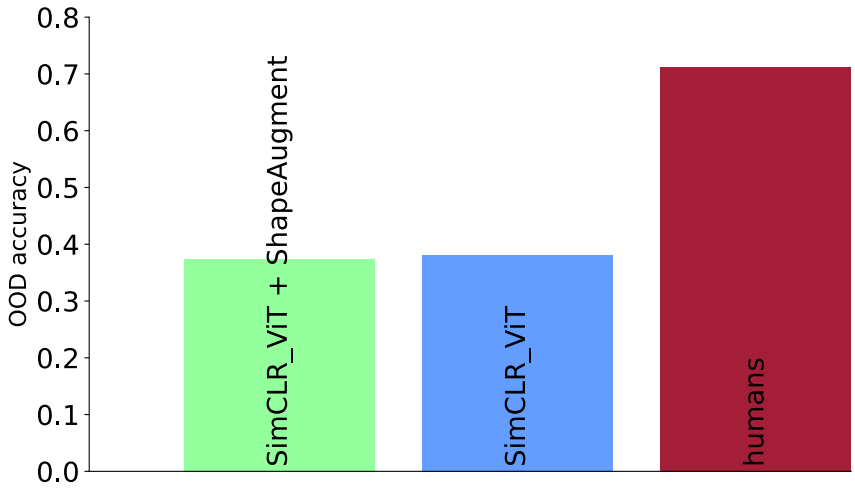

Figure 3.5 OOD accuracy averaged across datasets. SimCLR_ViT stands for a vanilla SimCLR with ViT backbone, and +ShapeAugment denotes a model pre-trained with shape-emphasizing augmentation.

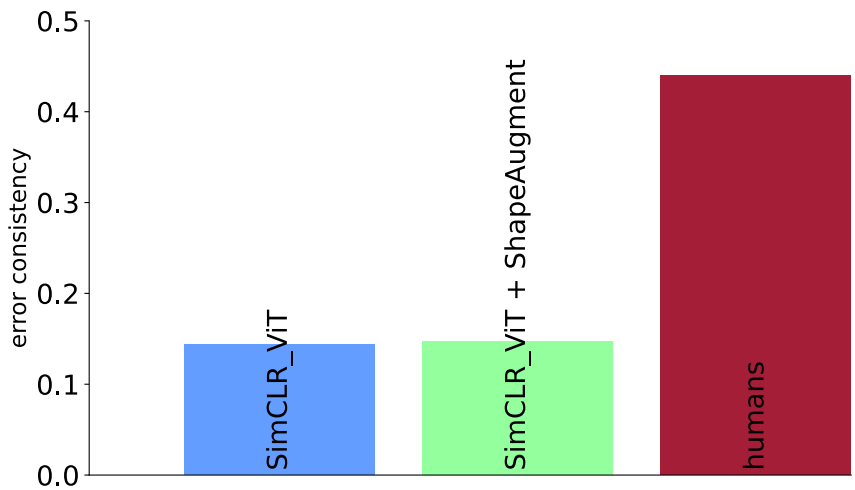

Figure 3.6 Image-level error consistency with human observers across OOD benchmark dataset.

As shown in Figure 3.5 and Figure 3.6, when the shape-emphasizing augmentation is applied to SimCLR with ViT backbone, the accuracy on the OOD benchmark is marginally decreased, while the error consistency with human observers is slightly increased. However, compared to the experiment on SimCLR with ResNet-50 backbone, the shape-emphasizing augmentation's influence on the OOD robustness of SimCLR with ViT backbone is minute.

This experiment results imply that just naively applying the suggested method to SimCLR with ViT backbone does not significantly impact the model's robustness to OOD distortions. The convolution operation in CNN models like ResNet-50 (He et al., 2016), which this paper mainly employed as a backbone model, has an entirely different learning mechanism from the self-attention of the ViT model (Dosovitskiy et al., 2020). Hence, further study is necessary to fit shape-emphasizing augmentation more suitably to a self-supervised contrastive learning model with a ViT backbone. This consideration will be deferred to future research.

Chapter 4. Shape-based Representation

4.1. Measuring the effect of the proposed method

The shape-emphasizing augmentation is expected to encourage the self-supervised contrastive learning model to learn shape-based representation by contrasting accentuated shape features common in positive pairs with those of negative samples. Hence, this chapter empirically validates the effect of the proposed method in strengthening the shape-based representation of the self-supervised contrastive learning model through relevant experiments: Measuring shape bias and employing the suggested method to supervised contrastive learning (Khosla et al., 2020). The basic setups for the experiments are the same as in section 3.3, with a few exceptions.

4.2. Shape bias

The shape bias is a metric indicating a specific classifier's degree of reliance on shape information. It is measured by testing the classifier on a texture-shape cue conflict set proposed by Geirhos et al. (2019). Each sample in this dataset has texture and shape cues conflicting with each other, i.e., a cat's image covered by an elephant's skin texture. Accordingly, there are two labels per image, a texture label, and a shape label. The degree of shape bias is defined analytically as below (Geirhos et al., 2019):

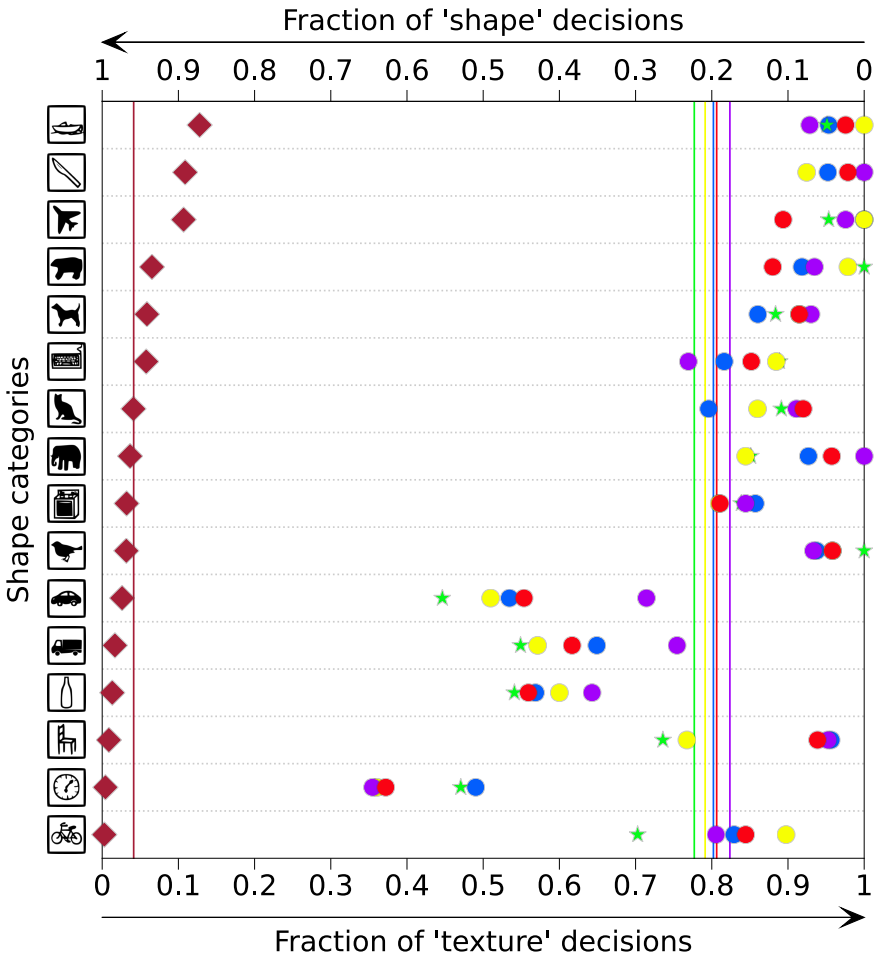

Figure 4.1 The degree of shape bias. The color code for each model is the same as in Figure 3.3. A green star and blue, yellow, red, and purple circle indicate shape-emphasizing augmentation, vanilla model, CutMix, CGN, and RandAugment, respectively. Vertical lines stand for each model's averaged value across object classes.

$$shape\ bias = \frac{the\ number\ of\ correct\ predictions\ by\ the\ shape\ label}{the\ number\ of\ correct\ predictions\ by\ shape\ or\ texture\ label} \quad (3.5)$$

Hence, the degree of shape bias can be one indicator of how much the classifier's representation is based on global shape features rather than local texture information in an image.

This section compares the influence of the shape-emphasizing augmentation in SimCLR's degree of shape bias with the same baselines as in chapter 3. Following the earlier research (Tuli et al., 2021) saying that a classifier's degree of shape bias can be affected by fine-tuning procedures, each model in this experiment is pre-trained as depicted in section 3.3.4 but fine-tuned for only one epoch. This adjustment of the length of fine-tuning is to prevent the influence of fine-tuning process from affecting the examination of each baseline method's effect on cultivating SimCLR's shape-based representation. The results of the experiment are shown in Figure 4.1.

Interestingly, human observers display a significantly higher fraction of their decisions based on shape cues than texture signals. This result is the same as the outcome from the earlier work (Geirhos et al., 2019). On the other hand, there is still a considerable gap between humans and SimCLR-based baselines, as presented in the previous research (Geirhos et al., 2020; Geirhos et al., 2021). However, as a silver lining beyond the cloud, shape-emphasizing augmentation pulls the direction of SimCLR's fraction of decisions based on one of two conflicting cues, either shape or texture, toward the side of the shape with some degree of a margin than other baselines.

4.3. Supervised contrastive learning

The supervised contrastive learning (Khosla et al., 2020) method also contrastively learns visual representations by maximizing similarities between representations of positive samples while minimizing them with negative samples. However, there is a difference with SimCLR regarding the utilization of label information when comprising a positive pair. The positive pair in SimCLR (Chen et al., 2020) consists of an anchor image and its multi-viewed sample by random data augmentations. In contrast, the supervised contrastive learning (Khosla et al., 2020) method has a group of positive samples containing the multi-viewed samples of all images in the mini-batch whose labels are the same as the anchor image. Accordingly, the supervised contrastive learning model is expected to learn highlighted shape features common in more positive samples than SimCLR when shape-emphasizing augmentation is applied.

Hence, this section delves into how the introduced method works on supervised contrastive learning's shape-based representation through several experiments. The experiments compare the effect of shape-emphasizing augmentation on supervised contrastive learning with SimCLR's case. The supervised contrastive learning models are pre-trained with the same setting of hyperparameters as SimCLR's training, except for using a supervised contrastive loss.

Each model’s accuracy on each subset of the OOD benchmark and averaged accuracy across all subsets are presented in Table 4.2 and Figure 4.2, respectively.

Dataset	SimCLR	+ Shape-Augment	SupCon	+ Shape-Augment
ImageNet	67.10	68.32 (+ 1.22)	68.09	69.25 (+ 1.16)
edge	18.13	**24.38 (+ 6.25)**	19.38	**30.00 (+ 10.62)**
silhouette	38.75	40.00 (+ 1.25)	46.25	46.25 (+ 0.00)
sketch	55.00	58.00 (+ 3.00)	55.38	55.63 (+ 0.25)
stylized	29.88	34.13 (+ 4.25)	31.00	33.63 (+ 2.63)
power-equalisation	80.98	82.41 (+ 1.43)	80.63	82.14 (+ 1.51)
eidolon III	33.67	33.20 (-0.47)	34.30	33.52 (-0.78)
uniform-noise	40.47	41.41 (+ 0.94)	44.38	39.45 (-4.93)
high-pass	29.22	32.66 (+ 3.44)	33.44	35.00 (+ 1.56)
phase-scrambling	47.32	48.84 (+ 1.52)	46.79	47.95 (+ 1.16)

Table 4.1. Top-1 classification accuracy (%) on IID and OOD test sets. SupCon stands for supervised contrastive learning model. + ShapeAugment denotes a model pre-trained with shape-emphasizing augmentation. Numbers in parenthesis are the accuracy difference on each dataset between its vanilla model. Bold values are the classification accuracy of the proposed method on the OOD benchmark subset in which the margin of improvement over the vanilla model is the largest.

Shape-emphasizing augmentation enhances the supervised contrastive learning model's classification accuracy across most subsets of the OOD benchmark dataset, as seen in SimCLR's case. However, the most impressive outcome is the accuracy improvement on the edge subset. As appeared in its name, the edge subset is composed of samples where only edges of objects exist, and the original texture information of each object is cleared away. The proposed method already showed an accuracy boost of a 6.25%p on this subset when applied to SimCLR. However, shape-emphasizing augmentation gave a 10.62%p advancement of the accuracy to the supervised contrastive learning model on the edge subset, even if the vanilla supervised contrastive learning model already has a slightly better performance on this subset than the vanilla SimCLR.

This outcome implies that the effect of shape-emphasizing augmentation could be synergized with a more extensive size of positive samples in a supervised contrastive learning model in terms of learning shape-based representation. Figure 4.3 also partly supports it. There was more boost to the supervised contrastive learning model than to SimCLR in image-level OOD error consistency with human observers recognizing objects by shape features when shape-emphasizing augmentation was applied. Additionally, Figure 4.4 shows that the introduced method also enhanced the supervised contrastive learning model's shape bias. These results are impressive, considering they occurred without degrading accuracy on IID test samples, as shown in Table 4.1.

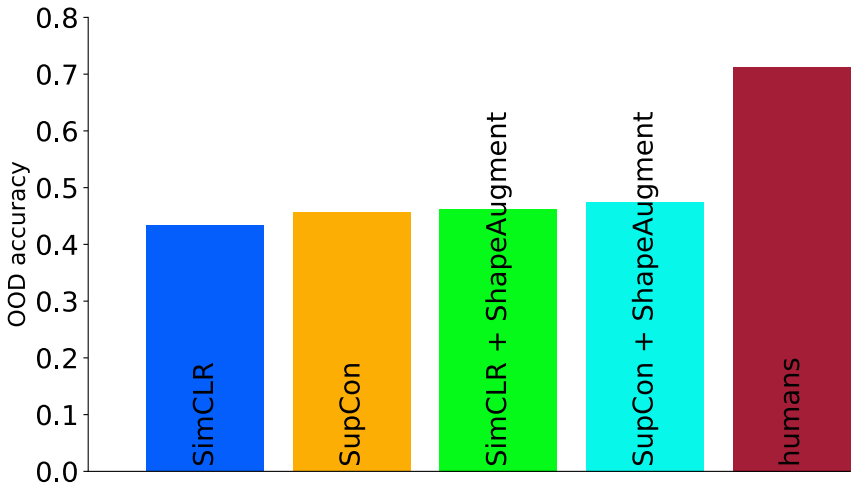

Figure 4.2 OOD accuracy averaged across datasets. The notation for each model is the same as the explanation in the caption of Table 4.1.

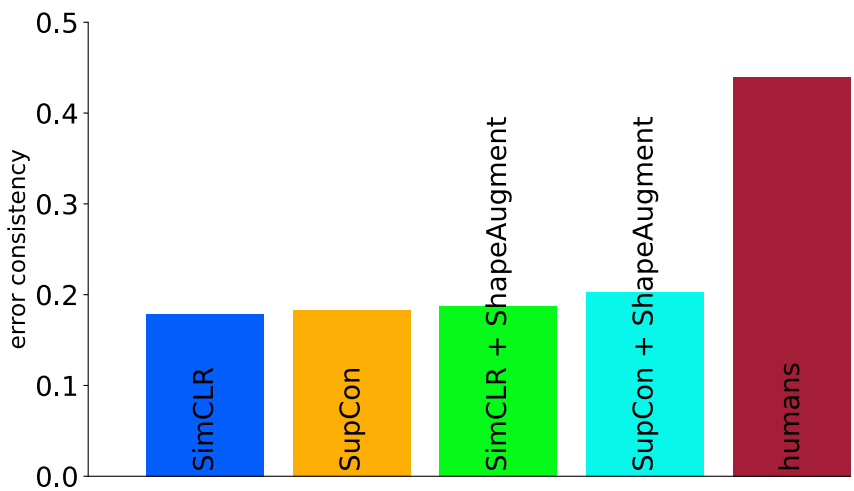

Figure 4.3 Image-level error consistency with human observers across OOD benchmark dataset.

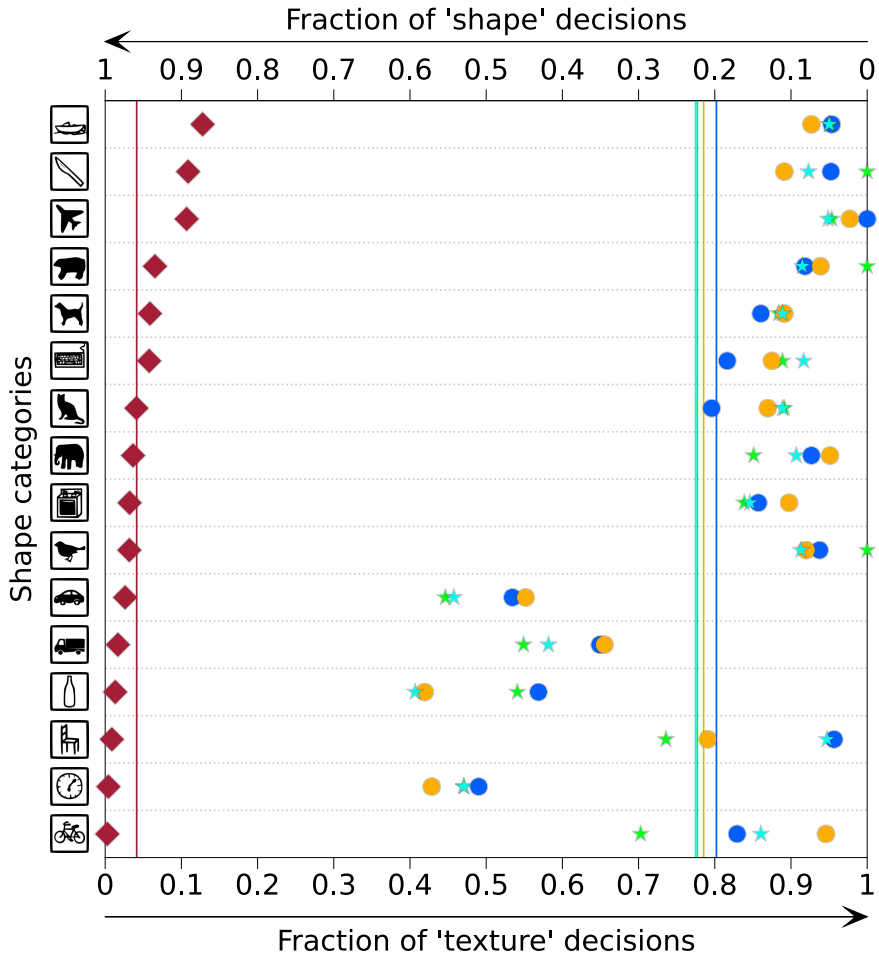

Figure 4.4 The degree of shape bias. The color code for each model is the same as in Figure 4.2. Orange and blue circles are the vanilla version of each model: SupCon and SimCLR. Light blue and green stars indicate each model with shape-emphasizing augmentation, respectively. Vertical lines stand for each model's averaged value across object classes.

Chapter 5. Conclusion

5.1. Summary of Research

This paper introduced a novel data augmentation scheme, shape-emphasizing augmentation. The suggested method applies different modifications to the texture in the inner and outer regions of the shape mask of the object in a training image by applying a set of random augmentations, respectively. When the proposed method is applied to the self-supervised contrastive learning model, it learns shape-based representation from the multi-viewed samples whose texture on each side of the object contour is differentiated by shape-emphasizing augmentation. The results of relevant experiments and the in-depth analysis demonstrated the effectiveness of the proposed method.

In detail, shape-emphasizing augmentation improved SimCLR's classification accuracy on most subsets of the OOD benchmark dataset. Specifically, on the edge subset where images do not have any local texture information of the object classes, the proposed method brought a 6.25%p of accuracy improvement. Moreover, SimCLR trained with shape-emphasizing augmentation displayed enhanced image-level error consistency with human observers on the OOD benchmark dataset. It means the model utilizes a similar decision strategy on OOD samples with humans of more shape-oriented representation and stronger robustness to OOD distortions. The increased degree of shape bias by applying shape-emphasizing augmentation also implies

it. The most encouraging thing is that this enhancement of robustness to OOD distortions occurred without the decreased accuracy on IID test samples. The employment of shape-emphasizing augmentation in the supervised contrastive learning model, which has more positive samples than SimCLR, indicates the efficacy of the introduced method in encouraging the model to learn shape-based representation.

The effectiveness and necessity of the proposed method are more remarkable when considering that previous augmentation methods in a supervised learning setting were not extended smoothly to a self-supervised learning setting. Moreover, the application of shape-emphasizing augmentation to the self-supervised contrastive learning model is more prospective in terms of the practicality in the actual use case. It is because robust OOD generalization is desirable for the self-supervised contrastive learning model's training mechanism, i.e., pre-training on the large-scale unlabeled dataset first and then finetuning with the labeled dataset that the users target specifically.

However, when the shape-emphasizing augmentation was applied to the self-supervised contrastive learning model with ViT backbone (Dosovitskiy et al., 2020), the proposed method did not show a noticeable influence on the model's robustness to OOD distortions. It is unlike the experimental results of the CNN-based self-supervised contrastive learning model, and further research would be needed to figure it out in future works.

5.2. Limitations

Shape-emphasizing augmentation presents promising results in improving the self-supervised contrastive learning model's robustness to OOD distortions without decreasing performance on IID test samples. However, the research of this paper still has limitations in some aspects described below.

First, the proposed method's effectiveness may depend on the performance of the salient object detection model. Suppose the model is terrible at segmenting the contour of the specific object due to occlusion in the scene or deformability of the object's shape. Consequently, it can hurt a self-supervised contrastive learning model's performance on related object classes when applying shape-emphasizing augmentation. However, using U2-Net as this paper's salient object detection model can be validated as a reasonable choice when considering the same model's employment in other recent papers (Sauer & Geiger, 2021; Wang et al., 2022) aiming to understand an image more structurally.

Also, the self-supervised contrastive learning models in this paper were not trained by the best training setups from their original papers due to the lack of available computing power. For example, Chen et al. (2020) and Khosla et al. (2020) reported the best performance of each vanilla model trained with much larger batch sizes and bigger backbones.

In addition, applying shape-emphasizing augmentation can increase computational and memory costs. It may come from applying

random augmentation to the image two independent times and extracting the object's shape mask from each sample of a mini-batch by utilizing the additional model for salient object detection. However, they are in the range of negligible degrees.

Moreover, the shape-emphasizing augmentation of this paper is conceptually confined in its application to 2-dimensional image data where the concept of the object or its shape can be defined visually. Hence, it is inapplicable to data with 1-dimension, such as time series data, and also data augmentation methods for this data type have different aspects from augmentations for image data (Wen et al., 2020). On the other hand, the concept of the object or its shape can also be visually defined in 3-dimensional image data with depth information. Hence, applying the concept of shape-emphasizing augmentation also can be considered, and it will be deferred to one of the promising directions for future research.

Lastly, this paper also showed that just naively applying shape-emphasizing augmentation in the same manner as the experiment on SimCLR with CNN backbone did not significantly impact the OOD robustness of the one with ViT backbone. The further discourse regarding this limitation will be a meaningful future work direction, as described more concretely in section 5.4.

5.3. Discussions

Unlike texture-shape bias and robustness to OOD distortions, which this paper dealt with, there have been various viewpoints and interests toward the definition of bias and robustness by different researchers. Some researchers assumed the image classification models have a bias toward the image's background or context where the objects are placed (Mo et al., 2021; Wang et al., 2022). Nam et al. (2020) and Lee et al. (2021) even did not predefine the specific types of bias in their works. Northcutt et al. (2021) raised the problem of robustness to label noise in the dataset. Hence, it will be an interesting point of discussion to consider how shape-emphasizing augmentation deals with other kinds of bias or robustness.

Also, this paper primarily focused on the object recognition task whose target objects are in the scope of the general domain in our everyday life. Accordingly, it may be meaningful to discuss the effect of shape-emphasizing augmentation in handling the data from more specific domains, such as medical images, or working on other computer vision tasks, such as object detection.

5.4. Future Works

Nowadays, the vision transformer (ViT) (Dosovitskiy et al., 2020) has gradually become the de facto backbone architecture of self-supervised learning models (Chen et al., 2021; Caron et al., 2021) for visual representation learning. These ViT-based self-supervised

learning models have continuously been renewing the state-of-the-art (SOTA) performance in self-supervised image classification by surpassing their counterparts based on CNN backbone.

Dosovitskiy et al. (2020) argued that ViT lacks the inductive biases that CNN architecture had, i.e., the locality, due to its learning strategy that images are sliced into patches, and the relevance between patches is measured via self-attention and position embedding. However, on the contrary, the deficiency of this inductive bias makes ViT more suitable for understanding the global context in images. In this regard, Tuli et al. (2021) presented experimental results that ViT has a higher degree of shape bias and error consistency with humans than standard CNN-based models.

However, recent research shed much less light on the ViT-based self-supervised learning model's robustness to OOD distortions. This paper also showed that the shape-emphasizing augmentation, which encouraged SimCLR with ResNet-50 backbone to be robust to OOD distortions, was not naively extended to the ViT-based SimCLR model. Hence, it would be an interesting future research direction to delve into the ViT-based self-supervised learning model's robustness to OOD distortions and develop further this paper's proposed method to be fitted to it.

Bibliography

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE international conference on computer vision (pp. 1422-1430).

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2536-2544).

Larsson, G., Maire, M., & Shakhnarovich, G. (2017). Colorization as a proxy task for visual understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition.

Gidaris, S., Singh, P., & Komodakis, N. (2018, February). Unsupervised Representation Learning by Predicting Image Rotations. In

International Conference on Learning Representations.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9729-9738).

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33, 21271-21284.

Geirhos, R., Narayanappa, K., Mitzkus, B., Bethge, M., Wichmann, F. A., & Brendel, W. (2020, October). On the surprising similarities between supervised and self-supervised models. In NeurIPS 2020 Workshop SVRHM.

Beery, S., Van Horn, G., & Perona, P. (2018). Recognition in terra incognita. In Proceedings of the European conference on computer vision (ECCV) (pp. 456-473).

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), 665-673.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.

In International Conference on Learning Representations.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

Hermann, K., Chen, T., & Kornblith, S. (2020). The origins and prevalence of texture bias in convolutional neural networks. Advances in Neural Information Processing Systems, 33, 19000-19015.

Sauer, A., & Geiger, A. (2021). Counterfactual Generative Networks. In International Conference on Learning Representations.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. Advances in Neural Information Processing Systems, 34, 23885-23899.

Albelwi, S. (2022). Survey on Self-Supervised Learning: Auxiliary Pretext Tasks and Contrastive Learning Methods in Imaging. Entropy, 24(4), 551.

Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems, 29.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural

networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning. 448-456. PMLR.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. Advances in neural information processing systems, 27.

Erhan, D., Courville, A., Bengio, Y., & Vincent, P. (2010, March). Why does unsupervised pre-training help deep learning?. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 201-208). JMLR Workshop and Conference Proceedings.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48.

DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552.

Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020, April). Random erasing data augmentation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 13001-13008).

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018, February). mixup: Beyond Empirical Risk Minimization. In International Conference on Learning Representations.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable

features. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6023-6032).

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2414-2423).

Brock, A., Donahue, J., & Simonyan, K. (2018, September). Large Scale GAN Training for High Fidelity Natural Image Synthesis. In International Conference on Learning Representations.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2019). Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 113-123).

Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 702-703).

Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. L. (2021). Are Convolutional Neural Networks or Transformers more like human vision?. arXiv preprint arXiv:2105.07197.

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. Advances in neural information processing systems, 31.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. Advances in Neural Information

Processing Systems, 33, 13890-13902.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.

da Costa, V. G. T., Fini, E., Nabi, M., Sebe, N., & Ricci, E. (2022). solo-learn: A Library of Self-supervised Methods for Visual Representation Learning. J. Mach. Learn. Res., 23, 56-1.

You, Y., Gitman, I., & Ginsburg, B. (2017). Large batch training of convolutional networks. arXiv preprint arXiv:1708.03888.

Riba, E., Mishkin, D., Ponsa, D., Rublee, E., & Bradski, G. (2020). Kornia: an open source differentiable computer vision library for pytorch. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 3674-3683).

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. Pattern recognition, 106, 107404.

Ross Wightman (2019). PyTorch Image Models, GitHub repository. https://github.com/rwightman/pytorch-image-models.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. In Proceedings of the IEEE/CVF

International Conference on Computer Vision (pp. 9640-9649).

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... & Krishnan, D. (2020). Supervised contrastive learning. Advances in Neural Information Processing Systems, 33, 18661-18673.

Wang, K., Machiraju, H., Choung, O. H., Herzog, M., & Frossard, P. (2022). CLAD: A Contrastive Learning based Approach for Background Debiasing. arXiv preprint arXiv:2210.02748.

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., & Xu, H. (2020). Time series data augmentation for deep learning: A survey. arXiv preprint arXiv:2002.12478.

Mo, S., Kang, H., Sohn, K., Li, C. L., & Shin, J. (2021). Object-aware contrastive learning for debiased scene representation. Advances in Neural Information Processing Systems, 34, 12251-12264.

Nam, J., Cha, H., Ahn, S., Lee, J., & Shin, J. (2020). Learning from failure: De-biasing classifier from biased classifier. Advances in Neural Information Processing Systems, 33, 20673-20684.

Lee, J., Kim, E., Lee, J., Lee, J., & Choo, J. (2021). Learning debiased representation via disentangled feature augmentation. Advances in Neural Information Processing Systems, 34, 25123-25133.

Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv:2103.14749.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9650-9660).

국문 초록

자기지도 학습은 이미지 분류에서 지도학습 모델에 비견되는 놀라운 발전을 이루었다. 하지만, 이러한 성과는 훈련 데이터셋과 독립적이고 동일하게 분포된 샘플들을 대상으로 한정되어 있다. 지도학습 모델과 같이, 자기지도 학습 모델은 여전히 분포 외 왜곡에 대한 낮은 강건성을 보인다. 이와 반대로, 사람은 분포 외 왜곡에 대해 강건함을 보이는데, 이는 형태 지향적인 표상과 낮은 질감 의존도에 기인한다. 이에, 최근 훈련 데이터셋의 질감을 변화시켜 이미지 분류 모델이 이미지 내 객체의 형태에 보다 집중하도록 유도하는 몇 가지 방법들이 제안되었다. 하지만, 해당 방법들은 자기지도 학습이 아닌 지도 학습에 중점을 두었을 뿐만 아니라, 독립적이고 동일하게 분포된 데이터들에 대해 오히려 성능 감소를 보였다. 이에, 본 논문에서는 자기지도 학습을 위한 새로운 데이터 증강 전략인 형태 강조 증강을 제안한다. 이 방법은 객체의 전경과 배경에 독립적으로 무작위 증강을 적용하여 이미지 내 객체의 형태를 강조한다. 다양한 실험을 통해 본 논문에서 제안하는 데이터 증강 방법이 독립적이고 동일하게 분포된 데이터들에 대한 자기지도 학습 모델의 성능 하락 없이 분포 외 왜곡에 대한 강건성을 향상시키는 데에 효과가 있음을 보인다.

주요어: 자기지도학습, 질감 편향, 형태 기반 표상, 분포 외 왜곡 강건성
학번: 2021-24432