APSIPA Transactions on Signal and Information Processing, 2023, 12, e42

Original Paper

CNN Pretrained Model with Shape Bias using Image Decomposition

Akinori Iwata* and Masahiro Okuda

Doshisha University, 1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto-fu, Japan

ABSTRACT

It is known that various implicit bias occur in Neural Networks due to their structural restrictions. Among them, texture bias caused by the convolution of CNNs has a significant impact on recognition performance. This paper shows that models with strong texture bias degrade recognition performance on datasets with large shape features, and to compensate for this characteristic of CNNs we introduce a method to increase their bias toward shapes rather than textures. Our method uses a simple image decomposition technique to create a shape-dominant dataset and then build a model with shape bias using the dataset. We experimentally show that the network can be biased towards shape without a significant loss of recognition accuracy compared to CNNs trained using conventional ImageNet. Additionally, we demonstrate that the CNN built by the proposed method obtains a higher recognition accuracy for shape-dominant images than those created using conventional methods.

Keywords: CNN, shape bias, image recognition, shape-dominant images.

*Corresponding author: Akinori Iwata, iwata@mm.doshisha.ac.jp. This manuscript is based on "Shape-Bias Evaluation of Pretrained Models Using Image Decomposition," presented at the APSIPA ASC December 2022.

Received 07 February 2023; Revised 16 August 2023
ISSN 2048-7703; DOI 10.1561/116.00000113

1 Introduction

Convolutional Neural Networks (CNNs) were originally designed to imitate the human visual system [10, 22, 23] and are utilized for numerous tasks, such as classification, semantic segmentation, and image generation. Although CNNs mimic the human visual system, there are still differences [6, 9, 13, 15, 32]. One notable difference between humans and CNNs is that humans tend to recognize objects by their global shape, whereas CNNs prioritize local texture over global shape.

The human visual cortex tends to first capture edges and points, then the shapes of objects, and finally the entire image containing the objects [24]. When humans recognize an object, they tend to focus on the global shape information of the entire object rather than on the local texture information of a part of the object. For example, when humans recognize elephants, they rarely focus on the shape of the wrinkles on their body surface but tend to recognize the body shape and positions of body parts, such as the nose and ears.

In contrast, Convolutional Neural Networks (CNNs) recognize objects primarily based on their texture, thanks to the use of convolutional layers that calculate correlations between the central pixel and its neighboring pixels to extract local texture features, rather than global shape. The pooling layer further expands the recognition range by down-sampling the feature maps. The CNN architecture gradually expands the recognition area from small to large by stacking convolutional and pooling layers. However, the local texture features obtained in the early layers tend to dominate the overall object recognition performance, as they influence the operations of the later layers. Additionally, as the kernel size per layer is usually small, the spatial receptive field is limited, and even if many layers are stacked on top of each other, the receptive field of CNNs does not become very large [5]. In other words, humans tend to focus on global shape information for object recognition, while CNNs tend to focus on local texture information for object recognition. This tendency to focus on global shapes rather than local textures in the analysis is known as shape bias, and the opposite is called texture bias.

Developing CNN with a shape bias has many advantages. One is that the CNN can grasp features more accurately for shape-dominant image data, such as cartoons or sketch images. In addition, focusing on shape also makes object judgments less susceptible to noise, blurring, and other image degradation, which may counter hostile attacks that may add perturbation and cause misclassification. Such a bias can also improve the efficiency and accuracy of CNN transfer learning and fine-tuning by selecting a model with an appropriate bias for transferring learning data. Furthermore, by analyzing the area in the image wherein the CNNs identify the shape and texture, we can improve their explainability.

In a previous study, Geirhos *et al.* [12] created Stylized-ImageNet, which is an image dataset that has different labels for local textures and global shapes, and experimentally confirmed that CNNs have a texture bias. They also used the Stylized-ImageNet to create a model with shape bias. Hermann *et al.* [20] performed various data augmentation techniques to quantitatively evaluate texture and shape biases. However, these previous studies not only changed the texture but also the shape of the object, or generated unrealistic images, which may negatively impact learning and raise questions about whether the shape is correctly recognized.

This study used image decomposition to remove texture from an image and used this decomposition to create an image dataset with information that is more meaningful for shape than texture. Thereafter, we trained the CNN on the dataset such that it had more bias toward the shape. Through experiments, we confirmed that the model had a shape bias, and fine-tuning the shape-dominant dataset resulted in the best performance compared to pretraining the model on other datasets.

The main contributions of this research can be summarized in the following two points:

- This study shows that the shape bias of a Neural Network can be increased with a shape-dominant image dataset obtained by image decomposition. The conventional dataset employed by Geirhos *et al.* [12] uses specialized techniques for dataset creation, and it is considered very time-consuming or even hard to apply to a different modality. By contrast, our method utilizes simple image decomposition techniques, providing the significant advantage of being able to create shape-biased data from any dataset and train new models with enhanced shape biases.

- While Geirhos *et al.* [12] aimed to create shape-biased models, the conventional approach did not demonstrate significant improvements in recognition performance on images from domains outside of ImageNet. In contrast, our method demonstrates enhanced recognition performance on datasets with limited texture information, which represents a significant step forward in the field.

The rest of this paper is organized as follows: Section 2 introduces related work on CNN bias and datasets. Section 3 describes the proposed method and pretraining settings. In Section 4, we explain the four experiments of the proposed method and discuss future work. Finally, Section 5 concludes the paper.

2 Related Work

2.1 Datasets

ImageNet [28] is a well-known dataset used for pertaining and includes ImageNet-1k [28] and ImageNet-21k [4], comprising 1 and 14 million images, respectively. It also has a wide variety of evaluation sets. ImageNet-C [18], where the images have been processed in various ways, such as adding noise, to verify model robustness. ImageNet-A or ImageNet-O [19], which is a collection of images originally included in ImageNet, is prone to misclassifications. ImageNet-R [17] contains a collection of images included in ImageNet, which are artifacts rather than real objects, such as pictures or dolls. ImageNet-Sketch [35]is composed of sketch images of ImageNet labels. It also contains the larger and publicly unavailable JFT-300M [31], which consists of 300 million images collected independently by Google; it is used for pretraining the recently introduced ViT [7].

2.2 Research on Implicit Bias

Studies have shown that CNN models trained using ImageNet [28] prefer local texture information to global shape information. Baker *et al.* [2] divided images into global shapes (edges and silhouettes) and local textures (parts of edges and within edges), and investigated the changes in recognition accuracy between humans and CNNs by varying each of these. Their results showed that the recognition accuracy of humans deteriorated significantly when the global shape was corrupted compared to when the local texture was corrupted. In the case of CNNs, the recognition accuracy did not deteriorate even if the global shape was corrupted, whereas it deteriorated significantly when the local texture was corrupted. Geirhos *et al.* [12] applied a style transformation to ImageNet to create image datasets with different shapes and textures and trained the CNN such that it had a shape bias. In the experiment, they measured the bias by performing image classification using two correct labels, shape and texture, for both humans and CNNs. The results showed that the human classifier focused on the global shape, whereas the original CNN focused on local texture. In contrast, models trained to have a shape bias had a bias similar to that of humans. Geirhos *et al.* [11] also compared various architectures and learning methods for humans, CNNs, and ViTs. They showed that learning on large datasets over long periods of time and learning with noise labels and data increases the bias and improves noise tolerance.

Tuli *et al.* [34] evaluated more accurate models by investigating the error consistency of deep neural networks. They found that ViTs have a more human-like bias than CNNs and a similar error consistency. Hermann *et al.* [20] investigated the impact of data augmentation and learning goals on shape bias and found that CNNs can lose shape information in their architecture as they learn. Shi *et al.* [29] proposed a vision-inspired algorithm to reduce the texture bias of CNNs and found that other algorithms, such as adversarial learning, can be used to improve recognition accuracy. Azad *et al.* [1] achieved the best results on several datasets for few-shot segmentation by reducing texture bias. Ding *et al.* [5] showed that increasing the kernel size of a CNN increases its spatial receptive field and shape bias. Raghu *et al.* [27] mapped the similarity between the layers of a CNN and ViT and showed that the CNN acquires global features from local features, whereas the ViT acquires global features from the beginning.

Many of these related studies, which attempted to obtain shape bias by modifying datasets, largely corrupted the shape of the images in the dataset. Therefore, CNNs may not have acquired the shape correctly. In this study, the image was smoothed to remove the texture while preserving the shape such that the CNN could acquire the shape from the image.

3 Proposed Method

The proposed method involves decomposing images into shape and texture components, creating and training a shape image dataset, and performing fine-tuning on datasets with shape dominance. An overview of the proposed method is shown in Figure 1. In this study, we initially decompose the pretraining images into local texture and global shape images by conventional image decomposition techniques and then reconstruct the pretraining dataset using only the global shape images. Subsequently, we pretrain the neural network using only the training images to obtain the shape bias. Finally, we quantitatively evaluate the bias through four types of experiments and demonstrate the superiority of our method for shape-dominant image data.

3.1 Image Decomposition Method

In the previous study by Geirhos *et al.* [12], images are generated using style transfer [21]. Style transfer is used to change the local texture of an image to a different local texture that is not meaningful for the correct label, thereby increasing the importance of global shape information. However, style transfer often produces unrealistic images because changing the local texture also affects the global shape, which negatively impacts learning and reduces accuracy. This is similar to learning from noisy images and has been a major concern.

In this study, we propose a method for constructing a shape-dominant dataset by using classical image decomposition methods to separate the shapes and textures of each image. We use two different methods to decompose the images: the edge-preserving smoothing algorithm proposed by Subr *et al.*

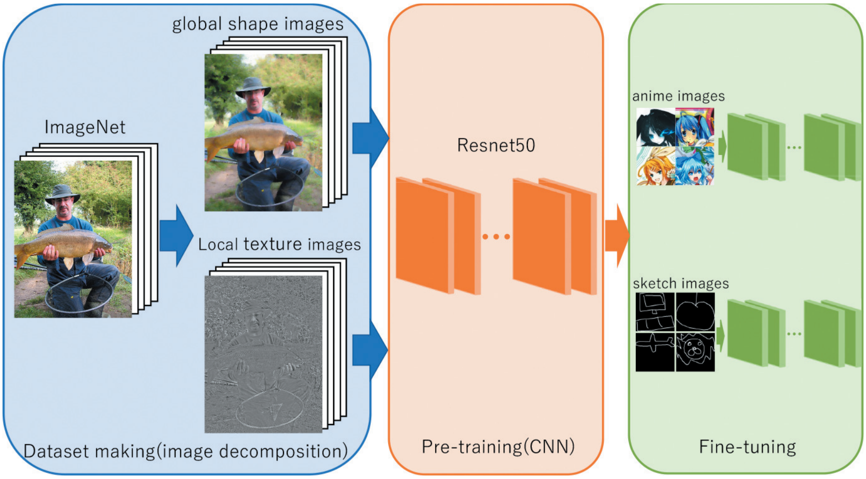


Figure 1: Overview of the proposed method. First, the dataset is decomposed into shape and texture images. Next, the shape images are used primarily to create a pretrained model. This generates a more shape-biased model. Finally, fine-tuning and transition learning are performed on the shape-biased dataset.

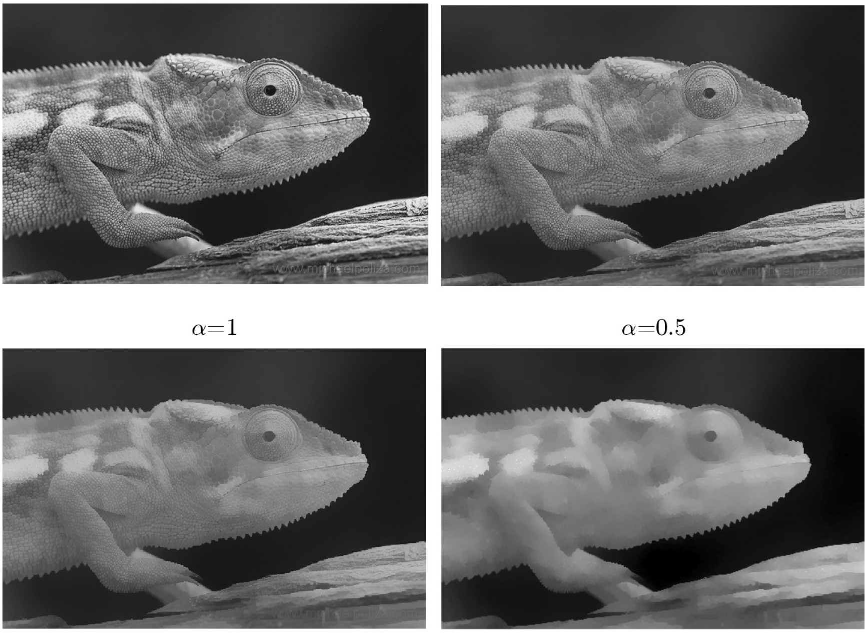
[30]and L0-Smoothing proposed by Xu *et al.* [38]. First, the edge-preserving smoothing algorithm defines the oscillation between the local minima and maxima in an image in detail (local texture). Then, the local minima and maxima in the image are calculated, and the envelope connecting each extreme value is obtained. Finally, we estimate the average of the envelopes of the local minima and maxima, and generate a smoothed image S. Next, L0-Smoothing reduces the number of pixels with non-zero differences between adjacent pixels, resulting in a smoothed image.

Using an algorithm that preserves the edges of images and maintains their shape, images can be added together. Let the original image be I, global shape image be S, and local texture image be T. We decompose I into S and T such that $I = S + T$, and reconstruct the image using the weighted combination as follows:

$$I' = S + \alpha T \ (0 \leq \alpha \leq 1). \tag{1}$$

Thus, the shape-dominant image I' is obtained. An example of images I' generated by applying the edge-preserving smoothing algorithm is shown in Figure 2. This process ($\alpha = 0$) was applied to all ImageNet images to create datasets Smooth-ImageNet (edge-preserving smoothing algorithm) and L0-ImageNet (L0-Smoothing).

Figure 3 illustrates some examples from the original ImageNet and Stylized-ImageNet datasets [12], as well as sample images from the Smooth-ImageNet and L0-ImageNet datasets. For the convenience of using style transfer [21], the



α=1

α=0.5

α=0.2

α=0

Figure 2: Sample images with some values of α. A larger α results in a texture image, while a smaller α results in a shape image. In particular, if alpha = 1, the image is the original image, and if alpha = 0, it is a Smooth-ImageNet.

images in Stylized-ImageNet were resized to (256, 256). The styles used for style transfer were based on paintings [26] primarily collected from a website called WikiArt.[1] However, due to the gaps between domains and the accuracy of the style transfer itself, the images in the Stylized-ImageNet dataset may not be interpretable to the human eye. In such cases, the image may contain a mismatch between the shape and texture, which could lead to inaccuracies in CNN's ability to acquire the shape of the original image. In contrast, the proposed datasets employ an algorithm that preserves the edges of the image while removing the texture. By training the CNN without texture information, the shape of the image is preserved, allowing humans to immediately determine the contents of the image. Comparing the Smooth-ImageNet and L0-ImageNet datasets, L0-ImageNet is smoother than Smooth-ImageNet. In terms of processing speed, creating Smooth-ImageNet required more than one month, whereas L0-ImageNet was created in approximately three days.

[1] https://www.wikiart.org/.

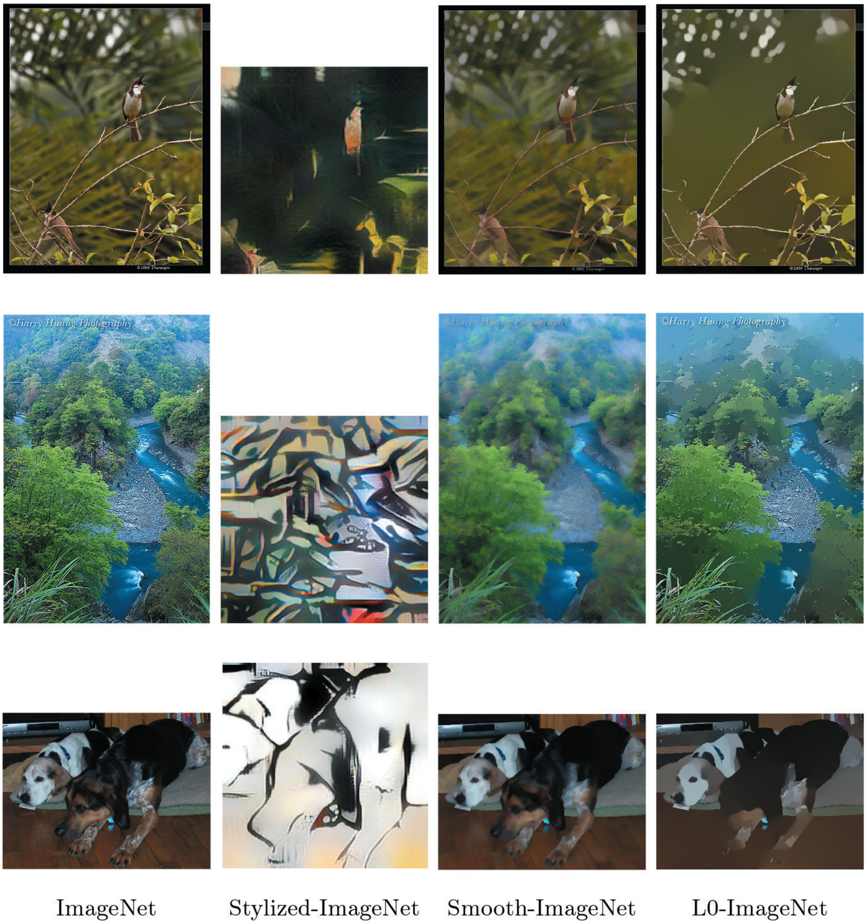


Figure 3: Partial comparison of the image datasets used in this study; Stylized-ImageNet has many images with heavily collapsed global shapes that cannot be determined by the human eye. On the other hand, the proposed dataset shows that the edges remain and the texture can be eliminated. Stylized-ImageNet is resized to (256, 256) to use style transfer.

3.2 Pretraining

We created a pretrained model of ResNet50 [16] using the proposed datasets (Smooth-ImageNet and L0-ImageNet), ImageNet as a comparison method, and Stylized-ImageNet, for a total of four types of data. For the pretraining dataset, we used only ImageNet, 1,281,167 training images and 50,000 validation images from ImageNet. We trained for 90 epochs with a batch size of 256 and an initial learning rate of 0.1. The learning rate was divided by 10 for 30, 60, and 80 epochs. For the other three datasets, both the dataset and ImageNet were used for training. For example, when training Smooth-ImageNet, we used images from both ImageNet and Smooth-ImageNet, and 2,562,334 training images. Stochastic gradient descent (SGD) and cross-entropy loss were used.

4 Experiments

In this Section, we quantitatively evaluate the shape bias proposed in previous studies. We also evaluate the efficiency of the proposed model with the shape bias by conducting experiments on three datasets characterized by shapes that are not related to either the proposed or comparative method.

4.1 Quantitative Evaluation of Shape Bias

4.1.1 Overview of the Experiment

In the first experiment, we quantitatively compared the shape bias of models trained on the proposed datasets, Smooth-ImageNet and L0-ImageNet, with those trained on the conventional datasets, ImageNet and Stylized-ImageNet. In this experiment, we tested the degree to which increasing the shape bias degrades the accuracy of conventional ImageNet classification with respect to ImageNet. Four models (Smooth-ImageNet, L0-ImageNet, original ImageNet, and Stylized-ImageNet), and three models that were fine-tuned with ImageNet for each trained model, were compared and evaluated. Pretraining was performed using the same parameter settings as those presented in Section 3.2, and the original ImageNet was used for fine-tuning. We fine-tuned for 10 epochs with a learning rate of 0.01.

We also compared and evaluated the shape bias and accuracy of the latest image recognition models, ConvNeXt-T [25] and EfficientNetV2-S [33]. The pre-training for ConvNeXt-T was performed using the official pre-training settings, and for EfficientNetV2-S, the settings in [37] were used as a reference. We fine-tuned for 30 epochs.

4.1.2 Evaluation Method

To quantitatively evaluate bias, we used the criteria proposed by Geirhos *et al.* [12]. As test data, images were style-transformed using a method similar to that used for Stylized-ImageNet to generate images with two different labels for global shapes and local textures (e.g., car for global shapes and cat for local textures). They used 16 classes consisting of airplanes, bears, bicycles, birds, boats, bottles, cars, cats, chairs, clocks, dogs, elephants, keyboards, knives, ovens, and trucks. They defined the number of shape matches ($NoSM$) as the number of model outputs that match the labels for the global shapes and the number of texture matches ($NoTM$) as the number of model outputs that match the labels for the local textures. Using these two quantities, they calculated the degree of shape bias as follows:

$$Shape\_bias = \frac{NoSM}{NoSM + NoTM}. \tag{2}$$

This metric evaluates the shape bias for each model and the accuracy of the original ImageNet validation set.

4.1.3 Results

The shape biases and recognition rates are shown in Table 1. For shape bias, Stylized-ImageNet had the highest value, followed by L0-ImageNet, Smooth-ImageNet, and ImageNet. Note that because the test data evaluated for the shape bias were created using the same method that used for creating Stylized-ImageNet, this experiment can be considered to have been conducted under favorable conditions for Stylized-ImageNet. In addition, the models trained on the proposed Smooth-ImageNet and L0-ImageNet datasets had larger shape bias values than those trained on the original ImageNet. Furthermore, when ImageNet was fine-tuned, the model trained on Stylized-ImageNet showed a slight decrease in shape bias, whereas those of models trained on Smooth-ImageNet and L0-ImageNet did not decrease.

Table 1: Shape bias and accuracy of the ImageNet validation set of ResNet50. IN: ImageNet, SIN: Stylized-ImageNet, SmIN: Smooth-ImageNet (proposed method), L0IN: L0-ImageNet (proposed method).

pretraining data	Fine-tuning data	Shape bias	TOP-1	TOP-5
IN	-	0.214	76.1	92.0
SIN + IN	-	0.370	64.6	85.1
SIN + IN	IN	0.362	74.1	91.9
SmIN + IN	-	0.266	75.4	92.5
SmIN + IN	IN	0.267	**76.8**	**93.2**
L0IN + IN	-	0.269	72.2	90.6
L0IN + IN	IN	0.272	76.3	93.1

Tables 2 and 3 shows the shape bias and accuracy of the latest image recognition models, ConvNeXt-T and EfficientNetV2-S. In both cases, the shape bias increased when trained on the proposed dataset and did not decrease significantly after fine-tuning. The accuracy is higher for the training with ImageNet alone, but the difference is not significant after fine-tuning. The proposed method has higher accuracy than Stylized-Imagenet, although there is a decrease in shape bias.

.
Table 2: Shape bias and accuracy of the ImageNet validation set of ConvNeXt-T. IN: ImageNet, SIN: Stylized-ImageNet, L0IN: L0-ImageNet (proposed method).

pretraining data	Fine-tuning data	Shape bias	TOP-1	TOP-5
IN	-	0.308	**82.0**	**95.8**
SIN + IN	-	0.460	80.2	95.0
SIN + IN	IN	0.429	81.0	95.3
L0IN + IN	-	0.372	81.2	94.1
L0IN + IN	IN	0.365	81.6	95.6

Table 3: Shape bias and accuracy of the ImageNet validation set of EfficientNetV2-S. IN: ImageNet, SIN: Stylized-ImageNet, L0IN: L0-ImageNet (proposed method).

pretraining data	Fine-tuning data	Shape bias	TOP-1	TOP-5
IN	-	0.144	**81.5**	**95.6**
SIN + IN	-	0.282	80.2	95.0
SIN + IN	IN	0.178	80.8	95.3
L0IN + IN	-	0.161	81.1	95.5
L0IN + IN	IN	0.152	81.3	**95.6**

When the proposed datasets (Smooth-ImageNet and L0-ImageNet) were used for pre-training and further fine-tuning using ImageNet, TOP-1 and TOP-5 showed almost equal or better accuracy compared to training using ImageNet. In addition, the shape bias in our method did not decrease significantly after fine-tuning compared to Stylized-ImageNet. Thus, compared to ImageNet, the proposed datasets can acquire shape bias and successfully improve or maintain accuracy without decreasing shape bias.

4.2 Evaluation with Anime Images

4.2.1 Overview of the Experiment

In the second experiment, we performed anime image recognition as an example to show that shape bias can improve classification performance. We used anime images to evaluate the effectiveness of the proposed method. Because anime images have fewer local textures than natural images, they can be considered shape-dominant. Therefore, it can be used to determine the effectiveness of a pretrained model with a shape bias for shape-dominant data.

We used the AnimeFace Character dataset;[2] 11,592 out of 14,490 images were used as training data, and 2,898 images were used as test data.

[2] http://www.nurs.or.jp/~nagadomi/animeface-character-dataset/.

Samples of images in this dataset are shown in Figure 4. The number of classes is 203, and each class contains approximately 50–100 images. For fine-tuning, we used the SGD optimizer and cross-entropy loss as the loss function. The batch size was 64 and the learning rate was 0.01. The model was trained for 40 epochs.

Figure 4: Samples of AnimeFace Character Dataset[2].

4.2.2 Results

The experimental results are shown in Table 4. The models pretrained on the proposed L0-ImageNet had the highest TOP-1 and TOP-5, followed by Smooth-ImageNet and Stylized-ImageNet; the models pretrained on the original ImageNet achieved the lowest accuracy. The model pretrained with L0-ImageNet was the most accurate for animated images with few textures. For the validation set of ImageNet, the model trained on the original ImageNet was the most accurate at the pretraining stage. However, in the fine-tuning of animated images, the accuracy of the proposed method is considered superior because of its emphasis on shape. This indicates that the proposed method is superior for animated images, which comprise shape-dominated data. Conversely, the accuracy of Stylized-ImageNet may be lower owing to the presence of different local textures resulting from its data-creation method.

Table 4: Accuracy of the AnimeFace Character Dataset. IN: ImageNet, SIN: Stylized-ImageNet, SmIN: Smooth-ImageNet (proposed method), L0IN: L0-ImageNet (Proposed method).

Pretraining data	Top-1	Top-5
IN	89.0	96.9
SIN + IN	89.6	96.8
SmIN + IN	89.3	96.6
L0IN + IN	**90.0**	**97.0**

4.3 Evaluation with Sketch Images

4.3.1 Overview of the Experiment

In the third experiment, by performing a sketch image recognition task, we showed that shape bias can improve classification performance. Sketches mainly consist of global shapes of an object and have little meaningful local texture. They have less local texture information than anime images and no color information. Therefore, the data is shape-dominant compared to natural images. We confirmed the effectiveness of the proposed method on these shape-dominant data, which are different from those of anime images.

For fine-tuning, we used a dataset comprising a collection of sketched images created by Eitz *et al.* [8]. Some samples from the dataset are shown in Figure 5. Of the 20,000 images, 16,000 were used for training and 4,000 were used for testing. The dataset comprises 250 classes, and the learning rate for fine-tuning is 0.01. SGD was used as the optimizer and cross-entropy loss was used as the loss function. The batch size was 128.

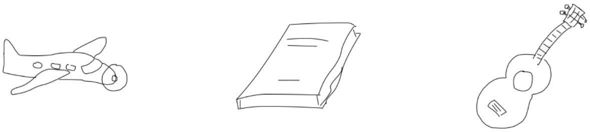


Figure 5: Samples of sketch dataset.

4.3.2 Results

The experimental results are shown in Table 5. Similar to the case of animated images, the models trained on L0-ImageNet and the proposed method showed the highest accuracy, followed by those trained on Smooth-ImageNet, Stylized-ImageNet, and original ImageNet. The group of models with low accuracy in the ImageNet validation set became more accurate for the sketched images. Thus, we created an effective pretrained model for data without color information. Furthermore, the models trained using the proposed method were found to be more effective for sketch images, which have even less local texture information than animated images.

4.4 Evaluation of Logo Images

4.4.1 Overview of the Experiment

In this experiment, we used logo images as examples of shape-dominant images. A logo is an illustration of a text string or object, and logos are applied to various items, such as organizations and products. They have a few high-frequency components similar to animated images, and some comprise different

Table 5: Accuracy of the sketch dataset. IN: ImageNet, SIN: Stylized-ImageNet, SmIN: Smooth-ImageNet (proposed method), L0IN: L0-ImageNet (Proposed method).

Pretraining data	TOP-1	Top-5
IN	78.8	94.8
SIN + IN	80.3	95.7
SmIN + IN	80.5	95.3
L0IN + IN	**81.8**	**95.9**

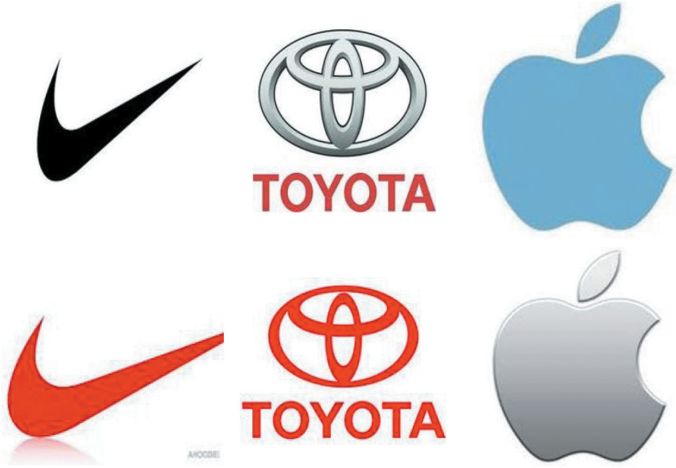


Figure 6: Example images from the logo 2k+ Dataset. The same logo can have different colors and slightly different shapes.

colors for different backgrounds or objects. Therefore, they can be considered color-independent shape images.

The Logo-2k+ dataset [36] was used in this experiment. Samples of images in the dataset are shown in Figure 6. The Logo-2k+ dataset consists of 10 parent categories and 2,341 child categories for a total of 167,140 images. Of the 167,140 images, 116,958 were used as training data and 50,182 as test data. The learning rate was set to 0.01, a cosine scheduler was used, and the warm-up time was set to five epochs. SGD was used as the optimizer and cross-entropy loss was used as the loss function. The batch size was 128.

4.4.2 Results

The experimental results are shown in Table 6. Similar to the aforementioned two datasets, the accuracy of the model trained on the proposed dataset was

Table 6: Accuracy on the Logo Dataset. IN: ImageNet, SIN: Stylized-ImageNet, SmIN: Smooth-ImageNet (proposed method), L0IN: L0-ImageNet (Proposed method).

Pretraining data	TOP-1	TOP-5
IN	58.9	69.8
SIN + IN	62.2	72.6
SmIN + IN	62.4	73.1
L0IN + IN	**63.6**	**74.5**

the highest, followed by Stylized-ImageNet and the original ImageNet. Thus, the proposed model was also effective for datasets containing logo images, where shape information is more important than color information.

These results indicate that compared to Stylized-ImageNet, Smooth-ImageNet and L0-ImageNet are shape-biased datasets and models trained on these datasets can produce effective models for shape-dominant data.

4.5 Remarks

We proposed a basis for controlling the shape bias of CNNs using image decomposition as the dataset construction method; however, there is room for further improvement. We need to conduct evaluation experiments using datasets containing other images that are considered shape-dominant, such as medical images or billboards, to further demonstrate the effectiveness of creating datasets with image decomposition. Additionally, experiments on model robustness should be conducted. CNNs are known to misclassify even small noises, such as adversarial attacks. In contrast, the same type of adversarial attacks rarely causes misclassification among humans. Therefore, it is necessary to investigate the robustness of the model trained on the proposed dataset to simple noise, such as Gaussian noise, or adversarial attacks [3].

Although optimization methods were used for shape and texture image decomposition, they require parameters to be set for each image. The proposed dataset was created without changing the parameters for each image; therefore, it was not possible to erase the texture with the optimal parameters for all images. With over one million images in ImageNet, it is practically impossible to manually set parameters manually for each image, and creating an accurate dataset for larger datasets is still problematic. Therefore, to improve texture removal, a method that automatically sets parameters can be devised or an image generation model [14, 39] that does not require parameter setting can be used.

Furthermore, it is necessary to devise an evaluation metric to measure the bias of the data to be fine-tuned. One advantage of the proposed method is that the ratio of global shape to local texture of the image can be varied by adding the decomposed images together again. By changing the ratio of shape and texture during training, a model with arbitrary bias can be created. In this study, we experimented with three types of shape-dominant images: animated images, sketches, and logo images. For example, when comparing a sketch image with an animated image, the animated image should have more local texture. Thus, we can expect more efficient learning and improved accuracy by creating a model with a bias tailored to the domain of each data set. For this purpose, it is necessary to have a measure of the bias of the data used for fine-tuning. This will enable the use of models with appropriate biases. In the future, the goal is to provide pretrained models with biases that match the biases of the fine-tuned data.

5 Conclusion

In this study, we created a dataset to train a CNN with a shape bias, which was fine-tuned and evaluated using data predominantly containing shapes. To achieve a shape bias, we employed an image decomposition method that separates images into their shape and texture components. This method overcomes the limitations of previous studies by preventing shape collapse during texture transformation, thus enabling us to reliably produce a shape-focused CNN. Our proposed method outperformed several baseline methods on test datasets where shape predominates. This dataset creation approach can be used to develop models with arbitrary biases by adjusting the proportions of shape and texture in the image through their combination. Moreover, the edges obtained through our method closely resemble those of the original images. By training on image datasets with varying proportions of shape and texture, CNNs with different biases can be created.

Acknowledgements

This work was supported in part by JSPS KAKENHI Grant Number 23K11174 and MEXT Promotion of Distinctive Joint Research Center Program Grant Number JPMXP 0621467946.

References

[1] R. Azad, A. R. Fayjie, C. Kauffmann, I. B. Ayed, M. Pedersoli, and J. Dolz, "On the Texture Bias for Few-Shot CNN Segmentation," *2021 IEEE Winter Conference on Applications of Computer Vision*, 2021, 2673–82.

[2] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman, "Deep convolutional networks do not classify based on global object shape," *PLoS Computational Biology*, 14, 2018.

[3] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*, Ieee, 2017, 39–57.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, 248–55, DOI: 10.1109/CVPR.2009.5206848.

[5] X. Ding, X. Zhang, Y. Zhou, J. Han, G. Ding, and J. Sun, "Scaling Up Your Kernels to 31x31: Revisiting Large Kernel Design in CNNs," *arXiv preprint arXiv:2203.06717*, 2022.

[6] S. F. Dodge and L. Karam, "A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions," *2017 26th International Conference on Computer Communication and Networks*, 2017, 1–7.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[8] M. Eitz, J. Hays, and M. Alexa, "How Do Humans Sketch Objects?" *ACM Trans. Graph.*, 31(4), 2012, DOI: 10.1145/2185520.2185540.

[9] N. Ford, J. Gilmer, N. Carlini, and E. D. Cubuk, "Adversarial Examples Are a Natural Consequence of Test Error in Noise," in *International Conference on Machine Learning*, 2019.

[10] K. Fukushima, S. Miyake, and T. Ito, "Neocognitron: A neural network model for a mechanism of visual pattern recognition," *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5), 1983, 826–34, DOI: 10.1109/TSMC.1983.6313076.

[11] 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 34*, 2021.

[12] 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 *International Conference on Learning Representations*, 2019, https://openreview.net/forum?id=Bygh9j09KX.

[13] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in Humans and Deep Neural Networks," in, *NIPS'18*, Montréal, Canada: Curran Associates Inc., 2018, 7549–61.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, ed. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Vol. 27, Curran Associates, Inc., 2014, https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[15] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *International Conference on Learning Representations*, 2015, http://arxiv.org/abs/1412.6572.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 770–8.

[17] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization," *2021 IEEE International Conference on Computer Vision (ICCV)*, 2021.

[18] D. Hendrycks and T. Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," *Proceedings of the International Conference on Learning Representations*, 2019.

[19] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. X. Song, "Natural Adversarial Examples," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 15257–66.

[20] K. Hermann, T. Chen, and S. Kornblith, "The origins and prevalence of texture bias in convolutional neural networks," *Advances in Neural Information Processing Systems*, 33, 2020, 19000–15.

[21] X. Huang and S. Belongie, "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization," in *ICCV*, 2017.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, *NIPS'12*, Lake Tahoe, Nevada: Curran Associates Inc., 2012, 1097–105.

[23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 86(11), 1998, 2278–324, DOI: 10.1109/5.726791.

[24] G. W. Lindsay, "Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future," *Journal of Cognitive Neuroscience*, 33, 2020, 2017–31.

[25] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 11976–86.

[26] K. Nichol, "Painter by Numbers," https://www.kaggle.com/competitions/painter-by-numbers, 2016.

[27] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do Vision Transformers See Like Convolutional Neural Networks?" In *Advances in Neural Information Processing Systems*, ed. A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, 2021, https://openreview.net/forum?id=Gl8FHfMVTZu.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, 115(3), 2015, 211–52, DOI: 10.1007/s11263-015-0816-y.

[29] B. Shi, D. Zhang, Q. Dai, Z. Zhu, Y. Mu, and J. Wang, "Informative Dropout for Robust Representation Learning: A Shape-bias Perspective," in *International Conference on Machine Learning*, 2020, 8828–39, http://proceedings.mlr.press/v119/shi20e.html.

[30] K. Subr, C. Soler, and F. Durand, "Edge-Preserving Multiscale Image Decomposition Based on Local Extrema," *ACM Trans. Graph.*, 28(5), 2009, 1–9, DOI: 10.1145/1618452.1618493.

[31] C. Sun, A. Shrivastava, S. Singh, and A. K. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, 843–52.

[32] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014, http://arxiv.org/abs/1312.6199.

[33] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*, PMLR, 2021, 10096–106.

[34] S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths, "Are Convolutional Neural Networks or Transformers more like human vision?" *ArXiv*, abs/2105.07197, 2021.

[35] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning Robust Global Representations by Penalizing Local Predictive Power," in *Advances in Neural Information Processing Systems*, 2019, 10506–18.

[36] J. Wang, W. Min, S. Hou, S. Ma, Y. Zheng, H. Wang, and S. Jiang, "Logo-2K+: A Large-Scale Logo Dataset for Scalable Logo Classification," in *AAAI Conference on Artificial Intelligence. Accepted*, 2020.

[37] R. Wightman, H. Touvron, and H. Jégou, "Resnet strikes back: An improved training procedure in timm," *arXiv preprint arXiv:2110.00476*, 2021.

[38] L. Xu, C. Lu, Y. Xu, and J. Jia, “Image Smoothing via L0 Gradient Minimization,” *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2011.

[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” in *2017 IEEE International Conference on Computer Vision*, 2017.