Understanding of Feature Representation in Convolutional Neural Networks and Vision Transformer

Hiroaki Minoura[a], Tsubasa Hirakawa[b], Takayoshi Yamashita[c] and Hironobu Fujiyoshi[d]
Chubu University, Kasugai, Aichi, Japan

Keywords: Image Classification, Convolutional Neural Network, Vision Transformer.

Abstract: Understanding a feature representation (e.g., object shape and texture) of an image is an important clue for image classification tasks using deep learning models, it is important to us humans. Transformer-based architectures such as Vision Transformer (ViT) have outperformed higher accuracy than Convolutional Neural Networks (CNNs) on such tasks. To capture a feature representation, ViT tends to focus on the object shape more than the classic CNNs as shown in prior work. Subsequently, the derivative methods based on self-attention and those not based on self-attention have also been proposed. In this paper, we investigate the feature representations captured by the derivative methods of ViT in an image classification task. Specifically, we investigate the following using a publicly available ImageNet pre-trained model, i) a feature representation of either an object's shape or texture using the derivative methods with the SIN dataset, ii) a classification without relying on object texture using the edge image made by the edge detection network, and iii) the robustness of a different feature representation with a common perturbation and corrupted image. Our results indicate that the network which focused more on shapes had an effect captured feature representations more accurately in almost all the experiments.

1 INTRODUCTION

Following the overwhelming success of Convolutional Neural Networks (CNNs) in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 challenge, CNNs have been used in numerous computer vision tasks. However, Vision Transformer (ViT) (Dosovitskiy et al., 2021) has outperformed CNNs on image classification tasks. The ViT architecture applies a transformer (Vaswani et al., 2017) for natural language processing to computer vision. ViT first divides an input image into a sequence of patches and then obtains the correspondence between patch features by self-attention (SA) to aggregate the tokens and produce their representations. ViT achieves high recognition accuracy on ImageNet despite its simple structure. Subsequently, because of ViT's sucess in computer vision, many derivatives (Liu et al., 2021)(Touvron et al., 2021)(Wu et al., 2021) have been proposed.

Swin Transformer (Swin) (Liu et al., 2021) is a widely used derivative of ViT. Swin has improved the computational complexity of the SA within ViT and its architecture is applicable for a wide range of computer vision tasks such as object detection and semantic segmentation. Additional methods have been proposed, including CvT (Wu et al., 2021) of SA combined with the convolution and a method that use self-supervised learning (Caron et al., 2021)(Kaiming et al., 2022) have been proposed. The majority of these were derived from ViT's SA-based approach. However, ConvNeXt (Liu et al., 2022a) outperformed ViT as a result of re-designing of ResNet (He et al., 2016) on the the basis of ViT architecture. In addition, MLP-Mixer (Tolstikhin et al., 2021) was modified from the multi-head attention of ViT to simple multi-layer perceptron (MLP), and PoolFormer (Yu et al., 2022) was modified to average pooling. Non-SA-based methods also have proposed, showing that *ViT is not necessary if the spatial features of the image are captured* and highlighting the need for SA in image classification tasks.

ViT may not be required in computer vision because models such as CNN, MLP, etc. can improve recognition. Tuli *et al.* (Tuli et al., 2021) analyzed the feature representations that ViT and CNN capture in image classification. As shown in Figure 1(a), a cat image is transferred to an elephant texture in the SIN

[a] https://orcid.org/0000-0001-7373-1291
[b] https://orcid.org/0000-0003-3851-5221
[c] https://orcid.org/0000-0003-2631-9856
[d] https://orcid.org/0000-0001-7391-4725

Minoura, H., Hirakawa, T., Yamashita, T. and Fujiyoshi, H.
Understanding of Feature Representation in Convolutional Neural Networks and Vision Transformer.
DOI: 10.5220/0011621300003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 429-436
ISBN: 978-989-758-634-7; ISSN: 2184-4321

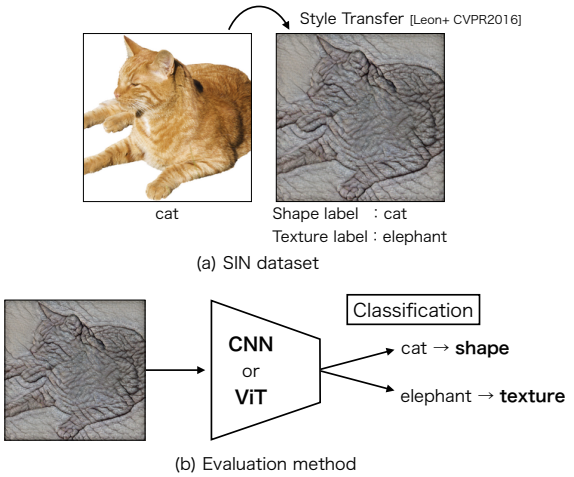

Figure 1: Evaluation method and example image from SIN dataset. The texture is generated from the original image by Style Transfer (Leon et al., 2016) includes shape and texture labels. The network takes a generated image as input and classifies it.

dataset (Robert et al., 2019). The transferred image is given the correct label *cat* and texture label *elephant*[1]. The evaluation method for the SIN dataset is shown in Figure 1(b). A network (e.g., CNN or ViT) takes the transferred image as input, and if the model classifies the image as a cat, it focuses on the object shape. If the model classifies the image as an elephant, it focuses on the object texture. This makes it possible to analyze the feature representation that the model captures and recognizes. ViT has the feature representation of the object shape, while CNN has the texture representation rather than shape.

We argue that ViT is highly effective for capturing object shapes in image classification, which has been considered difficult with conventional methods. However, it is unclear whether the methods derived from ViT can capture object shapes as well as ViT itself. In this paper, we investigate the feature representations captured by the derivative methods in an image classification task. Specifically, we investigate the following on a publicly available ImageNet (Deng et al., 2009) pre-trained model. i) Feature representation of either object shape or texture using the derivative methods. We test the cue conflict between texture and shape from the methods with the SIN dataset. ii) Classification without relying on object texture using the edge image made by the edge detection network (Xavier et al., 2020). We evaluate an image from the ImageNet validation dataset, excluding texture, using the edge detection network. The ImageNet-sketch (Wang et al., 2019) dataset is most suitable for our purpose, but it includes a shadow texture on the objects, which results in noisy of object shape. We investigate whether the model can capture the shape of the natural object by removing as much texture as possible from the natural images using the network. iii) The robustness of a different feature representation with a common perturbation and corruption image on ImageNet-C (Dan and Thomas, 2019). We examine the robustness of the shape- and texture-focused models to common noise (e.g., weather, blur) that can occur in object recognition. Experimental results show that the network focusing more on shapes has an effect on almost all the experiments.

The main contributions of this work are as follows.

- We clarify the feature representations captured by the methods derived from ViT.
- We show the potential of learning to capture shape for building robust models.

2 RELATED WORK

In this section, we review studies on image classification with ViT derivative methods.

2.1 Self-Atttention Based Approach

ViT is a model that applies a Transformer to computer vision tasks. ViT is known to outperform CNN, but in order to do so, it needs to be pre-trained on a dataset of 300 million images. Otherwise, over-training is likely to occur when there is insufficient data. Touvron *et al.* (Touvron et al., 2021) proposed DeiT, the accuracy of which is comparable to CNNs with a similar number of parameters through various data augmentations and regularization, even with insufficient data. Thus, DeiT avoids the difficulty of learning encountered in ViT. Various methods (Liu et al., 2021)(Wu et al., 2021)(Caron et al., 2021)(Kaiming et al., 2022) based on ViT have been proposed using the data augmentations and hyperparameters set up in DeiT.

Several methods have been derived from ViT, including one that reduces the amount of computation required for SA (Liu et al., 2021)(Wang et al., 2021)(Hongxu et al., 2022)(Zhuofan et al., 2022) and another that combines ViT with Convolution (Wu et al., 2021)(Zihang et al., 2021)(Zhengzhong et al., 2022). A method based on self-supervised learning has been proposed, which can be divided into two approaches: one based on contrastive learning to measure similarity between two images (Caron et al., 2021)(Xinlei et al., 2021) and one based on

[1]https://github.com/rgeirhos/texture-vs-shape

masked image modeling to reconstruct the original image using masking patch tokens (Kaiming et al., 2022)(Hangbo et al., 2022). In this paper, we investigate the feature representation of the representative method from these. We also investigate the feature representation of DeiT III (Hugo et al., 2022), which reconsiders the data expansion of DeiT to make it easier for ViT to capture the shape.

2.2 Non-Self-Atttention-Based Approach

Transformers are used in various tasks such as video recognition (Liu et al., 2022b)(Anurag et al., 2021), segmentation (Xie et al., 2021)(Zheng et al., 2021), and generative models (Yifan et al., 2021)(Kwonjoon et al., 2021) as it was found to be effective for computer vision models. These methods use self-attention based models; however, non-SA-based approach have also been proposed. For example, Liu *et al.* (Liu et al., 2022a) proposed ConvNeXt for redesigning ResNet (He et al., 2016) based on ViT architecture. Tolstikhin *et al.* (Tolstikhin et al., 2021) proposed MLP-Mixer to modify the multi-head attention of ViT to a simple MLP, in which the model is computed by MLP for each feature map. Yu *et al.* (Yu et al., 2022) proposed PoolFormer, which changes the multi-head attention of ViT to average pooling. These models have the MetaFormer structure, which captures spatial and dimensional features separately similar to ViT. Therefore, the attention mechanism is essentially unnecessary.

These findings lead us to reconsider the necessity of ViT in computer vision. On the one hand, ViT has enabled image classification which captures object shapes, which is difficult with conventional recognition models such as AlexNet, ResNet, etc. It is unclear whether the non-SA-based approaches are capable of capturing object shapes as well as ViT. Thus, we investigate the feature representations of their models to clarify the usefulness of ViT in computer vision.

2.3 Robustness of Vision Transformer

Most studies on the robustness of ViT compare ViT with ResNet. These studies (Tuli et al., 2021)(Muzammal et al., 2022) have examined the object shape and texture feature representations of ViT and CNN with the SIN dataset (Robert et al., 2019). The robustness has also been examined using natural noise images(Daquan et al., 2022)(Xiaofeng et al., 2022) and adversarial training (Sayak and Pin-Yu, 2022)(Kaleel et al., 2021)(Rulin et al., 2021)(Srinadh et al., 2021).

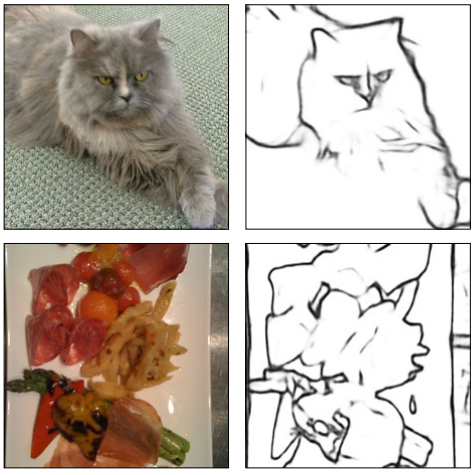

Figure 2: Example of ImageNet and Contour ImageNet.

The multi-head attention tends to capture a low-pass filter unlike CNN, as shown in (Namuk and Songkuk, 2022)(Peihao et al., 2022). While most of the prior studies compare classical CNN with ViT, our paper compares state-of-the-art methods with ViT.

3 EXPERIMENTS

In this section, we test several models on public datasets.

3.1 Evaluation Details

Datasets and Evaluation Metrics. We utilize four widely used image datasets: SIN (Robert et al., 2019), ImageNet (Deng et al., 2009), ImageNet-C (Dan and Thomas, 2019), and Contour ImageNet. The SIN dataset contains 1,280 images with style transformations between classes. In other words, there are 16 classes with 80 images per class. To compute the score in the dataset, we calculate shape bias as the proportion of correct shape decisions out of correct shape decisions and correct texture decisions. We only evaluate the subset of images for which either the shape or texture is correctly classified.

ImageNet (IN) is evaluated on validation data containing 50,000 images and 1,000 classes. The edge image was created by the edge detection network (Xavier et al., 2020) from ImageNet's validation data, which we refer to as "Contour ImageNet" (Contour). The shape of the natural object is focused on by removing as much texture as possible from the natural images using the edge detection network as shown in Figure 2. The ImageNet-C consists of 15 (+ 4 ex-

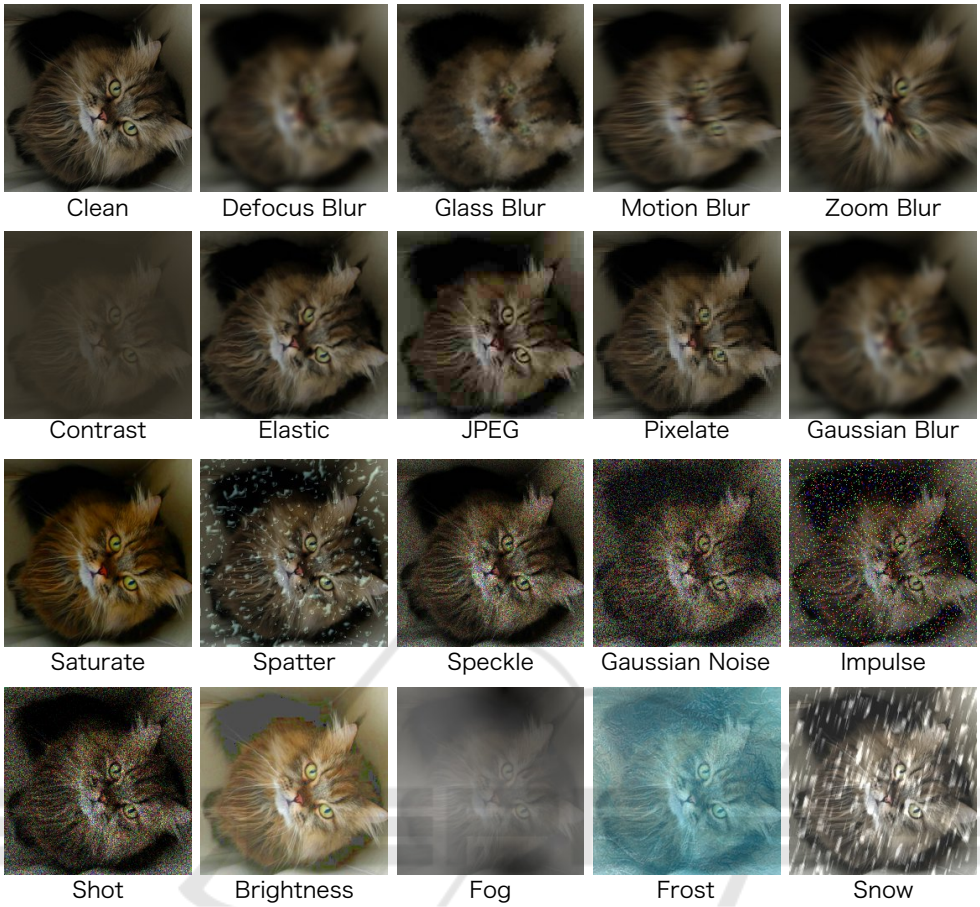

Figure 3: Example of ImageNet-C. A different corruption (blur, noise, etc.) is applied to each sample.

tra) algorithmically corruption images generated from noise, blur, weather, and digital categories as shown in Figure 3. Each type of corruption has five levels of severity. In total, there are 95 distinct collapsed images. We evaluate both the accuracy of the clean image on ImageNet and the robustness of the model on these datasets. We evaluate the *mean corruption error* (mCE) and *retention rate* (Retention) following (Dan and Thomas, 2019) and (Daquan et al., 2022) on ImageNet-C, respectively. The CE is the robustness calculated by dividing the top-1 error for each corruption by the AlexNet error, and mCE is the average CE. Retention is calculated as the ratio of ImageNet-C's robustness accuracy to ImageNet's clean accuracy.

Baseline Methods. We use the following baseline methods for comparison. We use ResNet and ConvNeXt as CNNs, ViT, DeiT, Swin, and CVT as conventional vision transformers, and PoolFormer and MLPMixer as a representative model for the Metaformer architecture. We also compare DeiT III, which is the altered data augmentation recipe for DeiT. These methods are trained by supervised learning. In addition, we use self-DIstillation with NO labels (DINO) and Masked AutoEncoder (MAE) by self-supervised learning as the method under investigation. This enables us to analyze the differences in feature representations depending on the learning method. In our experiments, we use the timm library (Wightman, 2019) or the official code for the models. Among them, we selected the large model sizes which are typically the highest performing. All methods use ImageNet pre-trained models or ImageNet21k pre-trained models fine-tuned to ImageNet.

3.2 Result of Texture vs. Shape

Figure 4 shows the fraction of shape or texture decisions of each method using the SIN dataset on ImageNet-trained models. The upper axis represents the shape and the lower axis represents the texture. The higher the value, the more the feature representation was captured. Note that ViT is compared with DeiT data augmentation applied to ViT because the

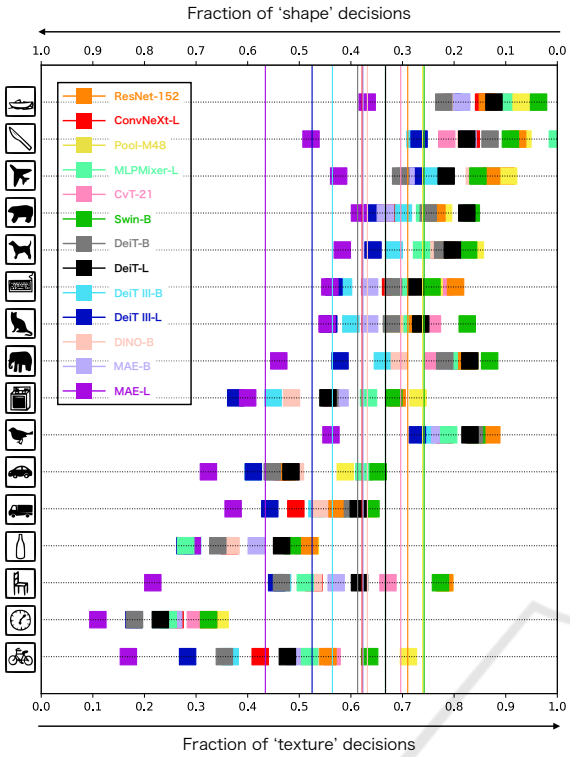

Figure 4: Classification results of shape vs. texture biases for ImageNet-trained networks. The upper axis is the fraction of shape decisions which capture the object shape. The lower axis is the fraction of texture decisions, which capture the object texture. Vertical lines indicate averages.

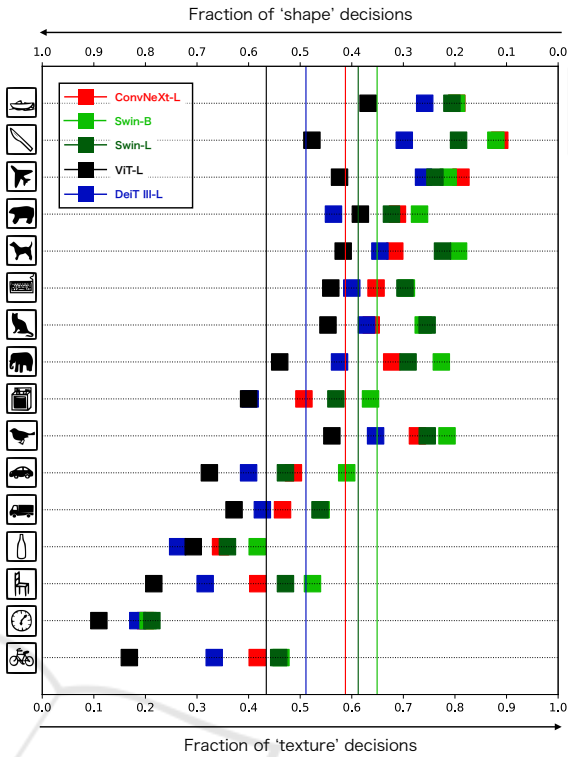

Figure 5: Classification results of shape vs. texture biases on ImageNet21k pre-trained models fine-tuned to ImageNet.

ImageNet pre-trained model is not publicly available. DeiT III, represented by dark blue in the figure, captures the shape-like features among the supervised learning. Meanwhile, Swin, Pool, and CvT are similar to ResNet in that they focus on textures for classification; however, ConvNeXt with re-designed ResNet based on ViT architecture focuses on texture and captures about 40% object shapes. Therefore, the feature representation of ConvNeXt is closer to that of ViT. MAE captures feature representations of shapes more accurately than DINO using the self-supervised learning method. This is because MAE captures spatial relationships between visualizable patch tokens at the encoder and reconstructs an object contour image to complement the masked tokens at the decoder. In addition, the larger the model size of MAE, the more of the shape it tends to capture. Only MAE-L represented by dark purple in the figure has captured a feature representation focusing on object shape.

Figure 5 shows the results of ImageNet21k pre-trained models fine-tuned to ImageNet. Overall, as the training data increases, the models are more likely to capture the object shape as a feature representation. In particular, only ViT, represented in the figure with black, has a shape bias. DeiT III captures shape and texture to the same degree as the results shown in Fig. 4. The pure ViT model tends to capture more shape than Swin and ConvNeXt.

3.3 Quantitative Results

Table 1 compares the feature representation results of several methods. MAE was the most effective on ImageNet. DeiT III had the highest score on ImageNet and ImageNet21k pre-trained under supervised learning.

Effectiveness of the Edge Object. DeiT III-L and MAE most accurately captured the feature representation of the shape rather than texture, immediately followed by ConvNeXt and DeiT III-B as shown in the Contour column of Tab. 1. Most models effectively captured the contour images if they performed well on natural images and recognized contour to some extent when pre-trained on ImageNet21k. In particular, Swin's contour result is about three times higher than that of the trained model on ImageNet. This indicates that models are more likely to capture feature representations of shapes when trained on a large dataset.

Table 1: Main results for various methods on several datasets. *PT* is the pre-training datasets and *IN-21k* is the ImageNet21k pre-trained model fine-tuned to ImageNet. *Type* is the learning type, *SL* and *SSL* are the supervised learning and self-supervised learning, respectively. *IN* is the top-1 accuracy on ImageNet. *Contour* and *IN-C* are evaluated the edge image made by the edge detection network and a common perturbation and corruption image on ImageNet-C, respectively, which affect model robustness. *Retention* and *mCE* are the evaluation metrics with respect to model robustness and mean CE, respectively.

PT	Type	Method	IN (↑)	Contour (↑)	IN-C (↑)	Retention (↑)	mCE (↓)
ImageNet	SL	ResNet-152	82.3	7.1	56.5	68.6	55.6
		ConvNeXt-L	84.2	12.6	64.4	76.5	45.8
		Pool-48M	82.4	6.6	56.5	68.6	55.7
		MLPMixer-L	68.3	2.7	34.6	50.6	83.8
		CvT-21	82.5	7.1	23.3	28.3	98.5
		Swin-B	83.2	5.0	58.7	70.6	52.9
		DeiT-B	81.1	5.2	58.8	72.5	52.8
		DeiT-L	79.7	3.6	58.6	73.4	53.0
		DeiT III-B	83.7	12.7	66.2	79.1	43.5
		DeiT III-L	**84.6**	**17.2**	**69.6**	**82.3**	**39.2**
	SSL	DINO-B	78.1	4.3	53.2	63.1	59.7
		MAE-B	83.6	8.9	59.5	71.2	52.1
		MAE-L	**85.9**	**16.5**	**67.4**	**78.5**	**42.0**
IN-21k	SL	ConvNeXt-L	86.3	18.5	68.6	79.5	40.3
		Swin-B	84.7	12.9	67.2	79.3	42.1
		Swin-L	85.5	16.6	**68.7**	80.4	**40.1**
		ViT-L	84.4	17.0	68.5	**81.1**	40.4
		DeiT III-L	**86.8**	**23.9**	67.9	78.2	41.0

Model Robustness. In-C and Retention in Tab. 1 show the top-1 accuracy and model robustness, respectively, for several methods on ImageNet-C. DeiT III was the most effective on ImageNet, followed by MAE and ConvNeXt. Interestingly, DeiT and Swin scored lower than ResNet in Contour but were more robust to noise. This result shows that SA is more robust to noise than conventional convolution. We also found that CvT of SA combined with convolution performed poorly on noisy images. The top-1 accuracy of the models pre-trained by ImageNet21k performed equally well on nearly all models, and ViT had the highest retention. Thus, ViT pre-trained on ImageNet21k was the optimal model for balancing accuracy and robustness, and the models pre-trained on ImageNet21k were robust to noise.

Table 2 shows the corruption error for each noise in ImageNet-C, and the average of these is the mCE in Tab. 1. The results on ImageNet indicated that DeiT III is robust to most noise among the supervised learning. For self-supervised learning, MAE is robust to noise than DINO. The effectiveness of DeiT III and MAE in almost all experiments shows that it is important to apply data augmentation, structures, and learning method that capture object shape, at least when training ViT on ImageNet. In addition, DeiT is much more robust to blur than ResNet, indicating that DeiT captures lower frequency components. Therefore, we believe that multi-head attention of DeiT works as a low-pass filter similarly to (Namuk and Songkuk, 2022)(Peihao et al., 2022). We also show that each model is robust to the noise on ImageNet21k pre-trained. Therefore, the model becomes more robust to noise when trained on a larger dataset.

4 CONCLUSION

In this paper, we investigated whether the methods derived from ViT capture the feature representations of object shapes and textures in an image classification task, and how they are affected by a common perturbation and corruption image with four datasets. Experimental results show that the ViT method focusing on shapes is robust to clean and noisy images on several image datasets. In self-supervised learning, we found that masked image modeling is more robust to clean and noisy images than the contrastive learning approach. Furthermore, we found that the model is robust to noise when larger data is available. Our future work will be to investigate whether learning to focus on shapes with a CNN model such as ConvNeXt enhances accuracy and robustness to noise. We also aim to identify derivative methods that are robust to adversarial training.

Table 2: Main results for several methods on ImageNet-C. The value is corruption error ; lower is desirable. *IN* is the ImageNet pre-trained model and *IN-21k* is the model pre-trained on ImageNet21k and fine-tuned on ImageNet.

PT	Type	Method	Blur				Digital				Extra				Noise			Weather			
			defocus	glass	motion	zoom	contrast	elastic	jpeg	pixelate	gaussian	saturate	spatter	speckle	gaussian	impulse	shot	brightness	fog	frost	snow
IN	SL	ResNet-152	61.9	79.3	59.7	66.3	42.2	70.3	56.8	54.0	60.8	40.3	54.2	47.9	50.8	51.9	52.4	44.7	45.5	58.6	56.0
		ConvNeXt-L	56.2	71.2	48.5	55.5	34.2	61.9	48.9	47.7	55.9	34.9	39.7	37.5	37.1	36.0	38.7	38.8	41.5	42.0	43.1
		Pool-48M	65.5	82.2	59.3	66.4	42.6	74.0	59.1	60.9	64.4	40.2	49.9	47.9	48.0	47.6	50.7	43.6	51.0	49.5	56.2
		MLPMixer-L	92.4	97.6	87.3	94.6	63.1	96.2	96.1	82.6	91.4	73.7	83.9	78.4	81.3	84.0	82.8	69.8	70.9	77.2	88.3
		CvT-21	110.8	107.1	103.8	104.6	98.9	107.2	109.9	106.2	110.8	93.0	86.0	92.0	95.1	94.9	94.7	99.8	90.9	82.5	83.9
		Swin-B	62.8	75.8	55.1	63.8	41.1	68.0	56.7	55.7	62.5	40.1	43.4	47.8	47.8	49.1	51.2	43.3	39.1	48.6	52.7
		DeiT-B	60.0	68.8	57.6	65.4	50.0	61.6	57.8	51.8	59.1	45.1	47.2	44.3	47.0	46.9	48.5	45.5	46.5	46.7	53.9
		DeiT-L	59.2	63.5	57.9	64.3	41.1	58.5	57.7	53.0	58.6	43.9	49.7	48.1	51.6	50.8	53.4	47.6	46.3	47.3	54.9
		Deit III-B	51.1	65.3	48.4	60.0	32.9	59.6	48.5	41.8	50.0	35.0	36.8	36.5	36.1	35.8	38.1	38.4	32.4	39.4	40.3
		Deit III-L	**47.4**	**59.6**	**42.2**	**51.4**	**30.1**	**54.2**	**44.1**	**36.8**	**46.4**	**32.2**	**33.4**	**32.1**	**32.0**	**31.4**	**33.6**	**35.9**	**29.2**	**35.8**	**36.6**
	SSL	Dino-B	59.5	75.9	65.2	71.7	51.0	65.3	60.5	52.9	58.1	46.2	54.8	56.6	62.1	62.6	63.1	48.1	54.5	61.3	61.5
		MAE-B	61.3	75.8	56.9	66.8	42.7	71.4	58.1	52.7	60.3	40.9	43.1	42.6	45.4	44.5	46.4	43.5	44.7	45.6	46.1
		MAE-L	**50.9**	**66.8**	**44.1**	**53.0**	**33.4**	**59.2**	**46.2**	**42.5**	**50.7**	**32.9**	**32.7**	**33.6**	**36.1**	**34.8**	**36.7**	**36.2**	**35.5**	**35.9**	**35.6**
IN-21k	SL	ConvNeXt-L	45.4	62.4	41.7	**47.3**	33.5	54.5	40.8	33.8	46.7	32.4	35.1	33.8	**35.2**	**33.3**	**36.2**	35.5	34.7	42.3	39.8
		Swin-B	49.1	65.0	44.9	53.4	34.5	57.0	44.9	37.3	49.3	34.5	34.4	35.2	37.0	36.7	37.8	35.7	32.7	39.5	40.2
		Swin-L	45.9	62.6	42.6	49.6	**32.6**	55.1	41.7	34.6	46.5	33.2	**32.8**	34.0	35.9	33.9	36.6	36.4	**31.7**	**39.2**	**36.7**
		ViT-L	45.6	**53.5**	**41.5**	50.3	31.4	**52.5**	42.8	33.5	44.2	34.4	36.9	**33.4**	36.5	36.5	37.4	35.3	35.0	44.9	40.4
		DeiT III-L	**41.3**	58.6	43.3	49.0	34.2	55.2	**38.9**	**32.1**	**41.6**	**31.7**	35.2	37.8	42.5	39.9	43.3	**34.7**	35.8	44.8	37.9

REFERENCES

Anurag, A., Mostafa, D., Georg, H., Chen, S., Mario, L., and Cordelia, S. (2021). Vivit: A video vision transformer. In *International Conference on Computer Vision*.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision*.

Dan, H. and Thomas, D. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.

Daquan, Z., Zhiding, Y., Enze, X., Chaowei, X., Anima, A., Jiashi, F., and Jose, M. A. (2022). Understanding the robustness in vision transformers. In *International Conference on Machine Learning*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai, L., and Li, F.-F. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Hangbo, B., Li, D., Songhao, P., and Furu, W. (2022). Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*.

Hongxu, Y., Arash, V., Jose, A., Arun, M., Jan, K., and Pavlo, M. (2022). A-vit: Adaptive tokens for efficient vision transformer. In *Computer Vision and Pattern Recognition*.

Hugo, T., Matthieu, C., and Herve, J. (2022). Deit iii: Revenge of the vit. *arXiv*.

Kaiming, H., Xinlei, C., Saining, X., Yanghao, L., Piotr, D., and Ross, G. (2022). Masked autoencoders are scalable vision learners. In *Computer Vision and Pattern Recognition*.

Kaleel, M., Rigel, M., and Marten, v. D. (2021). On the robustness of vision transformers to adversarial examples. In *International Conference on Computer Vision*.

Kwonjoon, L., Huiwen, C., Lu, J., Han, Z., Zhuowen, T., and Ce, L. (2021). Vitgan: Training gans with vision transformers. In *International Conference on Learning Representations*.

Leon, A. G., Alexander, S. E., and Matthias, B. (2016). Image style transfer using convolutional neural networks. In *Computer Vision and Pattern Recognition*, pages 2414–2423.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022a). A convnet for the 2020s. In *Computer Vision and Pattern Recognition*.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., and Hu, H. (2022b). Video swin transformer. In *Computer Vision and Pattern Recognition*.

Muzammal, N., Kanchana, R., Salman, K., Munawar, H., Fahad, S. K., and Ming-Hsuan, Y. (2022). Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems*.

Namuk, P. and Songkuk, K. (2022). How do vision transformers work? In *International Conference on Learning Representations*.

Peihao, W., Wenqing, Z., Tianlong, C., and Zhangyang, W. (2022). Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. In *International Conference on Learning Representations*.

Robert, G., Patricia, R., Claudio, M., Matthias, B., Felix, A. W., and Wieland, B. (2019). Imagenet-trained cnns are biased towards texture; increasing shape bias im-

proves accuracy and robustness. In *International Conference on Learning Representations*.

Rulin, S., Zhouxing, S., Jinfeng, Y., Pin-Yu, C., and Cho-Jui, H. (2021). On the adversarial robustness of vision transformers. *arXiv*.

Sayak, P. and Pin-Yu, C. (2022). Vision transformers are robust learners. In *Association for the Advancement of Artificial Intelligence*.

Srinadh, B., Ayan, C., Daniel, G., Daliang, L., Thomas, U., and Andreas, V. (2021). Understanding robustness of transformers for image classification. In *International Conference on Computer Vision*.

Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, Xiaohua Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., and Dosovitskiy, A. (2021). Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems*.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*.

Tuli, S., Dasgupta, I., Grant, E., and Griffiths, T. L. (2021). Are convolutional neural networks or transformers more like human vision? *CogSci*.

Vaswani, A., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan, N. G., Lukasz, K., and Illia, P. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. (2019). Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *International Conference on Computer Vision*, pages 568–578.

Wightman, R. (2019). Pytorch image models. In *https://github.com/rwightman/pytorch-image-models*.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., and Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. In *International Conference on Computer Vision*.

Xavier, S., Edgar, R., and Angel, D. S. (2020). Dense extreme inception network: Towards a robust cnn model for edge detection. In *Proc. WACV*, pages 1912–1921.

Xiaofeng, M., Gege, Q., Yuefeng, C., Xiaodan, L., Ranjie, D., Shaokai, Y., Yuan, H., and Hui, X. (2022). Towards robust vision transformer. In *Computer Vision and Pattern Recognition*.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*.

Xinlei, C., Saining, X., and Kaiming, H. (2021). An empirical study of training self-supervised vision transformers. In *International Conference on Computer Vision*, pages 9640–9649.

Yifan, J., Shiyu, C., and Zhangyang, W. (2021). Transgan: Two pure transformers can make one strong gan, and that can scale up. In *Advances in Neural Information Processing Systems*, pages 9640–9649.

Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., and Yan, S. (2022). Metaformer is actually what you need for vision. In *Computer Vision and Pattern Recognition*.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., and Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Computer Vision and Pattern Recognition*.

Zhengzhong, T., Hossein, T., Han, Z., Feng, Y., Peyman, M., Alan, B., and Yinxiao, L. (2022). Maxvit: Multi-axis vision transformer. In *European Conference on Computer Vision*.

Zhuofan, X., Xuran, P., Shiji, S., Li, E. L., and Gao, H. (2022). Vision transformer with deformable attention. In *Computer Vision and Pattern Recognition*.

Zihang, D., Hanxiao, L., Quoc, V. L., and Mingxing, T. (2021). Coatnet: Marrying convolution and attention for all data sizes. In *Advances in Neural Information Processing Systems*, pages 3965–3977.