Learning visual representations for transfer learning by suppressing texture

Shlok Mishra[1]
shlokm@umd.edu
Anshul Shah[2]
ashah95@jhu.edu
Ankan Bansal[1]
ankan@umd.edu
Janit Anjaria[1]
janit@cs.umd.edu
Jonghyun Choi[3]
jc@yonsei.ac.kr
Abhinav Shrivastava[1]
abhinav@cs.umd.edu
Abhishek Sharma[4]
abhisharayiya@gmail.com
David Jacobs[1]
djacobs@umiacs.umd.edu

[1] University of Maryland, College Park
[2] Johns Hopkins University
[3] Yonsei University
[4] Axogyan AI

Abstract

Recent literature has shown that features obtained from supervised training of CNNs may over-emphasize texture rather than encoding high-level information. In self-supervised learning, in particular, texture as a low-level cue may provide shortcuts that prevent the network from learning higher-level representations. We hypothesize that retaining more edge information and suppressing texture can help in alleviating these problems. To this end, we propose to use a simple classical idea based on anisotropic diffusion to augment training using images with suppressed texture. We empirically show that our method achieves improved results on image classification with five diverse datasets in both supervised or self-supervised learning tasks such as MoCoV2 and Dense-CL. Our method is particularly effective for transfer learning tasks, and we observed improved performance on twelve transfer learning datasets. The large improvements (up to 11.49%) on the Sketch-ImageNet and Synthetic-DTD datasets, and additional visual analyses of saliency maps suggest that our approach helps in learning better representations that transfer well to downstream tasks. We show that our method is simple to implement and can be integrated into various computer vision tasks easily.

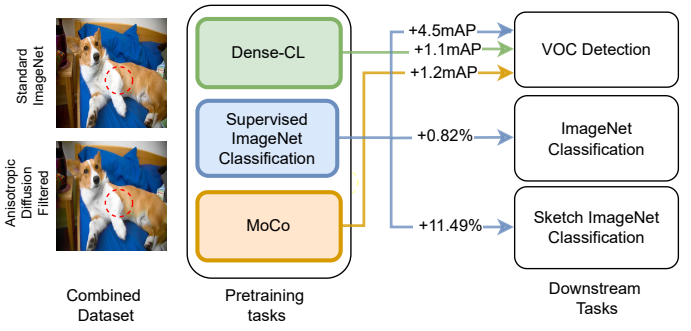

Figure 1: An overview of our approach. We propose to augment the ImageNet dataset by adding with Anisotropic diffused images. The use of this augmentation helps the network rely less on texture information and increases performance in several settings.

1 Introduction

Deep convolutional neural networks (CNNs) learn powerful visual features that have resulted in significant improvements on many computer vision tasks such as semantic segmentation [52], object recognition [32], and object detection [50]. However, CNNs often fail to generalize well across datasets under domain-shift due to varied lighting, sensor resolution, spectral-response etc. One of the reasons for this poor generalization is CNNs' over-reliance on low-level cues like texture [18].

These low-level cues and texture biases have been identified as grave challenges to various learning paradigms ranging from supervised learning [2, 18, 51] to self-supervised learning (SSL) [3, 14, 15, 46, 47]. We propose to use classical tools to suppress texture in images as a form of data augmentation to encourage deep neural networks to focus more on learning representations that are less dependent on textural cues. We use the Perona-Malik nonlinear diffusion method [48], robust Anisotropic diffusion [1], and Bilateral filtering [56] to augment our training data. These methods suppress texture while retaining structure by preserving boundaries.

Our work is inspired by the observation that ImageNet pre-trained models fail to generalize well across datasets [18, 49], due to over-reliance on texture and low-level features. Stylized-ImageNet [18] attempted to modify the texture of images using style-transfer to render images in the style of randomly selected paintings from the Kaggle paintings dataset. However, this approach offers little control over exactly which cues are removed from the image. The resulting images sometimes retain texture and distort the original shape. Stylized-ImageNet especially doesn't work well in the case of SSL since the network learns the texture and uses those textures to solve the SSL tasks (see Tab 2 Row1. In our approach (Fig. 1), we suppress the texture instead of modifying it. We empirically show that this helps in learning better higher-level representations and works better than CNN-based stylized augmentation. We compare our approach with Gaussian blur augmentation, recently used in [7, 8], and show that Anisotropic-filtering for texture suppression is better because isotropic Gaussian blur can potentially suppress edges and other higher-level semantic information as well. Our proposed method works well in self-supervised and supervised learning tasks, and we outperform both Gaussian blurring and Stylized-ImageNet in both settings.

Anisotropic-filtering is simple to implement and can be integrated easily in various com-

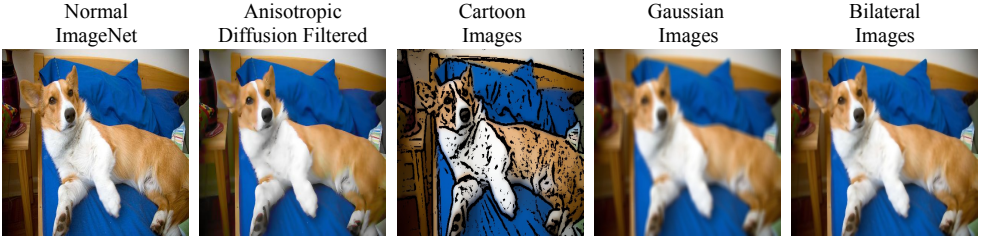

Figure 2: Four different methods for reducing texture in images.

puter vision-based methods. In the case of supervised learning, we pre-train on ImageNet, and test on twelve different datasets including ImageNet, Pascal VOC [16], Synthetic-DTD [45], CIFAR 100 [23], Sketch ImageNet [59], etc. For self-supervised setting, we use two learning frameworks: Dense-CL [60], and MoCoV2 [8] and pre-train on ImageNet and COCO. Our texture-suppressing augmentation consistently outperforms MoCoV2 and Dense-CL, which uses Gaussian blurring, on transfer learning experiments on VOC classification, detection, segmentation benchmarks, and also on classification tasks for other transfer learning datasets, including DTD [9], Cars [31], Aircraft [41], etc. With the help of qualitative and quantitative analysis, we show that our model is less reliant on high-frequency information in images and is more robust to common corruptions on datasets like ImageNet-C [22], and CIFAR-100 [23]. Our model also learns better shape bias than a Standard-ImageNet pretrained model and is also more confident in making correct predictions. Overall, we achieve significant improvements on several benchmarks:

- In a set of **twelve** diverse datasets, our method exhibits substantial improvements (as high as $+11.49\%$ on Sketch ImageNet and $+10.41\%$ on the DTD dataset) in learning visual representations across domains.
- We also get improvements in the same domain visual recognition tasks on ImageNet validation (+0.82%), and on label corruption task [23].
- We achieve improved results in self-supervised learning on image classification transfer learning tasks and on VOC detection.

2 Related Work

In this section, we review relevant methods that aim to remove texture cues from images to reduce the dependency of CNNs on low-level cues. Since we also experiment with the application of our method in self-supervised learning, we review recent work in this area as well.

Reliance on Low-Level Texture Cues. Recent studies have highlighted that deep CNNs can leverage low-level texture information for classification on the ImageNet dataset. Contrary to popular belief that CNNs capture shape information of objects using hierarchical representations [33], the work in [18] revealed that CNNs trained on ImageNet are more biased towards texture than shape information. This dependency on texture not only affects generalization, but it can also limit the performance of CNNs on emerging real-world use-cases like few-shot image classification [51]. [2] showed that a bag of CNNs with limited

receptive field in the original image can *still* lead to excellent image classification performance. Intuitively, a small receptive field forces the CNNs to heavily rely on local cues vs learning hierarchical shape representations. This evidence strongly suggests that texture alone can yield competitive performance on ImageNet, and the fact that it's relatively easier to learn vs hierarchical features may explain deep CNNs' bias towards texture.

To reduce reliance on texture, Stylized-ImageNet [18] modified the ImageNet images into different styles taken from the Kaggle Painter by Numbers dataset. While trying to remove texture, this approach could also significantly affect the shape. Also, there isn't an explicit control over the amount of removed texture. Moreover, this method may not be directly applicable to self-supervised learning because the fixed number of possible texture patterns result in images with strong low-level visual cues resulting in shortcuts. We show that the accuracy on downstream tasks, when MoCoV2 and Jigsaw are trained with Stylized-ImageNet, decreases dramatically (Table 1 Supplementary). Some of the recent methods inspired by Stylized-ImageNet [18] that try to learn better shape representation [26, 35, 44] face similar issues. Info-Dropout[53] tries to learn shape information by using dropout based methods which zeros out neurons by a high probability if the input patch contains less self-information.

On the other hand, we use Perona-Malik's anisotropic diffusion [48] and bilateral filtering [56] as ways of suppressing texture in images. These methods remove texture without degrading the edge information. Consequently, the shape information of the objects is better preserved. Also, these methods provide finer control over the level of texture suppression. Suppressing the texture in training images forces CNN to build representations that put less emphasis on texture. We show that such data augmentation can lead to performance improvements in both supervised and self-supervised settings. We also distinguish our work from other data augmentation strategies like Auto-Augment [12] which uses Reinforcement Learning to automatically search for improved data augmentation policies and introduces Patch Gaussian Augmentation, which allows the network to interpolate between robustness and accuracy [39].

Self-Supervised Learning. To demonstrate the importance of removing texture in the self-supervised setting, we consider two pretext tasks. The first pretext task is Jigsaw [46], a patch-based self-supervised learning method that falls under the umbrella of visual permutation learning [10, 11] . Some of the most recent self-supervised methods are contrastive learning based methods [3, 4, 5, 7, 8, 17, 21, 24, 25, 28, 34, 42, 43, 52, 60]. In [3], the authors have proposed using contrastive losses on patches, where they learn representations by predicting representations of one patch from another. In MoCo [21], a dynamic dictionary is built as a queue along with a moving average encoder. Every image will be used as a positive sample for a query based on a jittered version of the image. The queue will contain a batch of negative samples for the contrastive losses. MoCo has two encoder networks. The momentum encoder has weights updated through backpropagation on the contrastive loss and a momentum update. In MoCoV2, Gaussian blur and linear projection layers were added that further improve the representations. MoCo and MoCoV2 have shown competitive results on ImageNet classification and have outperformed supervised pre-trained counterparts on seven detection/segmentation tasks, including PASCAL VOC [16], and COCO [36].

Transfer Learning. Transfer learning is one of the most important problems in computer vision due to difficulty in collecting large datasets across all domains. In this work, we discuss transfer learning in the context of ImageNet. A lot of early datasets were shown to be too small to generalize well to other datasets [57]. Following this, many new large-scale datasets were released [13, 36], which are believed to transfer better. However, recent

results show that these datasets do not generalize well in all cases [30, 49]. [30] showed that ImageNet features generally transfer well, but not to fine-grained tasks. We show results of transfer learning on some of the datasets that were used by [30].

3 Methods

CNN-based classifiers have been shown to exploit textures rather than shapes for classification [2, 18]. We aim to reduce the prominence of texture in images and thus encourage networks trained to learn representations that capture better higher-level representations.

Gaussian Blur. Gaussian blurring is one of the most popular smoothing methods in computer vision, and it has been recently proposed as data augmentation for SSL [7, 8]. However, along with low-level texture, Gaussian filtering also blurs across boundaries, diminishing edges and structural information.

3.1 Anisotropic diffusion

We propose to use Anisotropic Diffusion Filters (ADF) [48], which keep the shape information coherent and only alter low-level texture. Specifically, we use Perona-Malik diffusion [48]. These filters smooth the texture without degrading the edges and boundaries. Intuitively, this will encourage the network to extract high-level semantic features from the input patches.

Perona-Malik diffusion smooths the image using the differential diffusion equation:

$$\frac{\partial I}{\partial t} = c(x,y,t)\Delta I + \nabla c \cdot \nabla I \tag{1}$$

$$c(x,y,t) = e^{-(\|\nabla I(x,y,t)\|/K)^2} \tag{2}$$

where I is the image, t is the time of evolution, Δ is the Laplacian operator, K controls sensitivity to edges, ∇ is gradient and (x,y) is a location in the image. The amount of smoothing is modulated by the magnitude of the gradient in the image through c the diffusion coefficient. The larger the gradient, the smaller the smoothing at that location. Therefore, after applying Anisotropic diffusion, we obtain images with blurred regions, but edges are still prominent. Fig. 2 shows some examples of the application of the filter. Note that ADF reduces the texture in the image without replacing it, the domain gap between images is not large, while in the case of Stylized ImageNet, the domain shift will be large. Recently, there has been some work on removing textures using deep learning as well [37, 40, 51]. We find, though, that fast and simple classical methods work well on most tasks. For all our experiments, we create a dataset 'Anisotropic ImageNet' by adding ADF filtered ImageNet images to the standard ImageNet dataset. We also experiment with training an Image-to-Image translation model Pix2Pix [27] to suppress texture. We train the model to produce images that are similar in style to anisotropic diffusion. Pix2pix model also helps in capturing different variations of texture, e.g., the amount of smoothing, according to the target task. Details of the Pix2Pix based approach are mentioned in the supplementary.

3.2 Other texture suppressing methods

We also experiment with a few other texture removing methods like robust Anisotropic diffusion [1], Bilateral filtering [56], and Cartoonization. However, empirically we find that the

most simple Anisotropic diffusion method has the best results as discussed in Section 4.2. We will discuss these other texture removing methods briefly next.

Bilateral Filtering: [55] is an efficient method of anisotropic diffusion. In Gaussian filtering, each pixel is replaced by an average of neighbouring pixels, weighted by their spatial distance. Bilateral Filtering is its extension in which weights also depend on photometric distance. This also limits smoothing across edges, in which nearby pixels have quite different intensities.

Cartoonization: A more extreme method of limiting texture is to create cartoon images. To convert an image into a cartoonish image, we first apply bilateral filtering to reduce the image's colour palette. In the second step, we convert the actual image to grayscale and apply a median filter to reduce noise in the grayscale image. After this, we create an edge mask from the greyscale image using adaptive thresholding. Finally, we combine these two images to produce cartoonish looking images (see Fig. 2).

4 Experiments

We start by briefly describing the datasets used in our experiments. We then show the effectiveness of ADF for supervised and self-supervised learning. We find that ADF is particularly effective when there is a domain shift, supporting our hypothesis that variation in texture is a significant effect of domain shift. The effect is larger when we transfer from ImageNet to datasets such as Sketch Imagenet [59], and Synthetic-DTD [45], where the domain shift is larger. Our method is also able to outperform Stylized-ImageNet [18] and Gaussian Blur [7].

Datasets. In all experiments, we use the ImageNet training set as the source of our training data. For object detection and semantic segmentation, we evaluate on Pascal VOC 2007 and VOC 2012. For label corruption, we evaluate on CIFAR100. When the downstream task is classification we evaluate on Synthetic-DTD [45], Sketch-ImageNet [59], Birds [53], Aircraft [41], Stanford Dogs[29], Stanford Cars [31],DTD [9], ImageNet-C [22], and the ImageNet validation dataset.

Experimental Details. For SSL we build on MoCoV2 [8] and Dense-CL [60]. For supervised learning, we use the ResNet50 [20] model, closely following [18]. After training on Anisotropic ImageNet, we fine-tune our model on the standard ImageNet training set following the procedure of [18]. We set the conduction coefficient (K) of Anisotropic Diffusion to 20, and a total of 20 iterations are used. We use MedPy implementation. All other hyper-parameters are described in the supplementary material.

4.1 Self-Supervision for Transfer Learning

We first experiment with Anisotropic ImageNet on Self-Supverised methods. We have doubled the number of images (Anisotropic images + normal images) as compared to normal ImageNet. So for a fair comparison, we only train our methods for half the number of epochs compared to training with just ImageNet. We then fine-tune the network pre-trained on the Anisotropic ImageNet for the downstream tasks, including image classification, object detection, and semantic segmentation on PASCAL VOC, and other transfer learning datasets. Since, we are removing low-level cues from the images, we expect to see better results when transferring to different datasets.

MoCo V2. We evaluate our method with MoCo V2 [8] and Dense-CL [60], which are the state-of-the-art methods in SSL. MoCoV2 and Dense-CL [60] used Gaussian blurring

Table 1: Comparison with MoCoV2 and Dense-CL in SSL. We note that using Anisotropic diffusion with improves performance on VOC detection and Semantic Segmentation (SS). We test on COCO-based metrics as used in [8]. We also improve performance over the baseline on the semantic segmentation (SS) task [38]. Although we have only focussed on MoCoV2 Dense-CL, our technique can potentially be extended to other state-of-art methods.

Methods	Dataset	AP_{50}	AP	AP_{75}	mIoU (SS)
Stylized ImageNet		43.5	28.80	33.7	-
Supervised ImageNet		81.6	54.2	59.8	59.8
MoCo V2 [8]	ImageNet	82.4	57.0	63.6	67.5
MoCo V2 Anistropic (Ours)	ImageNet	**83.7**	**58.2**	**64.8**	**67.8**
Dense-CL [60]	ImageNet	82.8	58.7	65.2	69.4
Dense-CL [60] Anistropic (Ours)	ImageNet	**83.5**	**59.6**	**66.4**	**70.5**
Dense-CL CC [60]	COCO	81.7	56.7	63.0	67.5
Dense-CL CC Anistropic (Ours) [60]	COCO	**83.1**	**57.9**	**64.2**	**68.6**

Table 2: Transfer learning across different datasets. Note that our approach leads to improvements in both supervised and SSL set-up.

Dataset	Aircraft [41]	Birds [53]	Dogs [29]	Cars [31]	DTD [9]
Supervised (Reproduced)	90.88	90.3	85.35	92.1	72.66
SimCLR [7]	88.1	_	_	92.1	73.2
BYOL [19]	88.1	_	_	91.7	76.2
MoCo V2 [8]	91.57	92.13	87.13	92.8	74.7
MoCo V2 Anistropic (Ours)	**92.71**	**93.29**	**88.81**	**94.3**	**76.3**

with 0.5 probability as data augmentation. In our case, we add Anisotropic diffusion on the images with 0.5 probability, and for the remaining 50% of the images, we apply Gaussian blurring with 0.5 probability. So, in our setup every image has 0.5 probability of coming from Anisotropic ImageNet, 0.25 of Gaussian blurring, and 0.25 of being normal ImageNet. Also, the number of iterations on Anisotropic filtering is chosen randomly between 10 to 20. For object detection starting from a MoCoV2 initialization, we train a Faster R-CNN [50] with C4-backbone, which is fine-tuned end-to-end.

We show improvements over MoCoV2 and Dense-CL for object detection on the VOC Dataset. In the first setup, we show improvements on COCO-based evaluation metrics (i.e., AP_{50}, $AP_{0.05:0.05:0.95}$, AP_{75}) as shown in the first three columns of Table 1, achieve competitive results. We also observe an improvement of 1.3 mean IoU on semantic segmentation [38] over MoCo V2 baseline and 1.1 over Dense-CL baseline. We also see improved performances on Dense-CL when we pre-train on COCO dataset as well. These results show that in the case of transfer learning, we improve across different datasets. More details can be found in the supplementary material. Our method is not bound to a particular pretext task and can be potentially added to any state-of-art method to achieve even further improvements. In the supplementary material, we show that our method leads to improvements with the Jigsaw [47] task.

Table 3: Experiments with Sketch-ImageNet. Use of Anisotropic ImageNet shows that our method is better at capturing representation that are less dependent on texture.

Method	Top-1 Acc	Top-5 Acc
ImageNet Baseline	13.00	26.24
Stylized Baseline	16.36	31.56
Pix2Pix-Anisotropic (Ours)	**24.49**	**41.81**

These results suggest that training the network on the Anisotropic ImageNet dataset forces it to learn better representations. This is consistent with our hypothesis that Anisotropic diffusion leads to smoothing of texture in images. This forces the network to be less reliant on lower-level information to solve the pretext task and, hence, learn representations that focus on higher-level concepts.

Experiments with Stylized ImageNet on MoCoV2 and Jigsaw. We now show experiments that indicate that, while effective in a supervised setting, Stylized ImageNet does not help with SSL. We train a model with MoCoV2 and Jigsaw as pretext tasks on the Stylized-ImageNet (SIN) dataset [18] and fine-tune on the downstream tasks of object detection and image classification on PASCAL VOC. In Table 2 (and Table 2 in supplementary), we show that there is a huge drop in performance. One reason for this failure using the SIN dataset could be that the model is able to memorize the textures in the stylized images since it only has 79,434 styles. This is not a problem in the original fully-supervised setting where the authors used SIN for supervised image classification. In that case, the network can learn to ignore texture to discriminate between classes.

4.2 Transfer Learning for Supervised Learning

As shown in the last section, suppressing texture leads to performance improvements in the case of domain transfer with SSL. In this section, we also show improvements in supervised learning and domain transfer. In the case of supervised learning we also show results using Pix2Pix-Anistropic model.

4.2.1 Across Domains

We hypothesize that learning with texture bias is most harmful to domain transfer. Thus, we first describe a challenging experimental setup for learning visual representation across domains.

Sketch-ImageNet. For a cross-domain supervised learning setup, we chose to use the Sketch-ImageNet dataset [59]. Sketch-ImageNet contains sketches collected by making Google image queries "sketch of X", where "X" is chosen from the standard class names of ImageNet. The sketches have very little to no texture, so performance on Sketch-ImageNet is a strong indicator of how well the model can perform when much less texture is present. As shown in Table 3, the difference between the Pix2Pix-Anisotropic model and the baseline model is 11.49% for Top-1 accuracy. This result implies that our model captures less dependent representations on texture than standard ImageNet and Stylized ImageNet.

Other Datasets - Aircraft, Birds, Dogs, and Cars. We further evaluate our method on image classification tasks using four different fine-grained classification datasets. We also observe improvement on image classification across five datasets in Table 2. This shows that

Table 4: Comparison using different texture removing methods, with different hyper-parameters for Anisotropic diffusion methods. We observe that the most simple [43] performs the best, and removing more texture from images does not improve performance.

Method	# Iterations	Top-1 Acc	Top-5 Acc	Object Detection
Baseline Supervised	-	76.13	92.98	70.7
Stylized ImageNet [18]	-	76.72	93.27	75.1
Perona Malik with Pix2Pix [43]	20	**76.95**	**93.36**	**75.21**
Perona Malik [43]	20	76.71	93.26	74.37
Perona Malik [43]	50	76.32	92.96	73.80
Robust AD [1]	20	76.58	92.96	73.33
Robust AD [1]	50	76.64	93.09	73.57
Gaussian Blur	-	76.21	92.64	73.26
Cartoon ImageNet	-	76.22	93.12	72.31
Bilateral ImageNet	-	75.99	92.90	71.34

in the case of domain shift, capturing higher level semantics helps in better transfer learning performance.

Object Detection. The biggest improvement we observe on transfer learning is on object detection on Faster-RCNN [50] as shown in Table 4. This improvement suggests that we are able to attend to more high-level semantics, which helps in transfer learning performance on object detection.

4.2.2 Same Domain

ImageNet: In Table 4, we show results using Anisotropic ImageNet for supervised classification. We observe that Anisotropic ImageNet improves performance in both ImageNet classification and object detection. We also use Pix2Pix model for learning to suppress texture. Pix2Pix model converts a normal image to texture suppressed image. For Gaussian blurring experiments, we closely follow [8] and add a Gaussian blur operator with variance from 10 to 20 and train in a similar manner to Stylized ImageNet [18]. We can see that our proposed Pix2Pix Anisotropic ImageNet performs better than both Stylized ImageNet and Gaussian blurring. Hence, blurring the image completely without respecting boundaries and edges, or distorting the shape information by style transfer does not improve performance.

Different Texture Removing Methods. We also provide results using different texture removing methods and different hyper-parameters for Anisotropic diffusion in Table 4. We observe that as we increase the number of iterations and remove more and more texture from images, performance starts to degrade, possibly due to the difference that comes in the data distribution after removing texture information. The most simple texture removing method [43] has the best results.

5 Analysis

We now show qualitatively and quantitatively how our model is less dependent on texture information.

5.1 Synthetic-DTD Dataset:

To better demonstrate the effectiveness of less texture dependent representations, we used the dataset introduced by [45]. This dataset provides four variations in images: texture, color, lighting, and viewpoint. This dataset is created by taking 47 different textures from DTD dataset[45] and applying them to a 3D dataset of 10 classes, called ShapeNet[6] to yield the same object rendered with different view-points and multiple textures. It contains 480,000 training images and 72,000 testing images. In this dataset, we made sure that texture information during training and testing are completely different. So, the texture is not a cue when we use this dataset. We evaluate our Pix2Pix-Anisotropic model on this dataset and compare against the baseline normal ImageNet model. The Pix2Pix-Anisotropic model achieves a performance boost of 10.41% in classification which suggests that we are indeed able to learn texture agnostic feature representations.

5.2 Experiments on testing shape bias:

To show the low reliance of our model on texture and greater reliance on high-level features like shape, we use the Geirhos Style-Transfer (GST) dataset [18]. It consists of 1,248 images of 16 classes from ImageNet with shape and texture coming from different classes. This dataset teases apart the importance of shape vs texture for CNNs. We observe an improvement of 1.02% on identifying the class representing the shape over a ImageNet pretrained Resent-50. This shows that we can successfully reduce the reliance on texture.

6 Conclusion

We propose to help a CNN focus on high-level cues instead of relying on texture by augmenting the ImageNet dataset with images filtered with Anisotropic diffusion, in which texture information is suppressed. Empirical results suggest that using the proposed data augmentation for pretraining self-supervised models and for training supervised models gives improvements across ten diverse datasets. Noticeably, the 11.4% improvement while testing the supervised model on Sketch ImageNet suggests that the network is capturing more higher-level representations than the models trained on ImageNet alone.

Acknowledgement

SM was supported in part by, by the US Defense Advanced Research Projects Agency (DARPA) Semantic Forensics (SemaFor) Program under grant HR001120C0124. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the DARPA. AS was supported by an ONR MURI grant N00014-20-1-2787. JC was partly supported by the NRF grant (No.2022R1A2C4002300), IITP grant (No.2020-0-01361-003, AI Graduate School Program (Yonsei University), 10%, No.2021-0-02068, AI Innovation Hub, 10%) funded by the Korea government (MSIT).

References

[1] Michael J. Black, G. Sapiro, D. Marimont, and D. Heeger. Robust anisotropic diffusion. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 7 3:421–32, 1998.

[2] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet, 2019.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

[4] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. 2019.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021.

[6] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository, 2015.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.

[9] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[10] R Santa Cruz, Basura Fernando, Anoop Cherian, and Stephen Gould. Visual permutation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41: 3100–3114, 2018.

[11] Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, and Stephen Gould. Deeppermnet: Visual permutation learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6044–6052, 2017.

[12] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *ArXiv*, abs/1805.09501, 2018.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[15] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015.

[16] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2009.

[17] Songwei Ge, Shlok Kumar Mishra, Chun-Liang Li, Haohan Wang, and David Jacobs. Robust contrastive learning using negative samples with diminished semantics. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=xLExSzfIDmo.

[18] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv*, abs/1811.12231, 2018.

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2019.

[22] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ArXiv*, abs/1903.12261, 2019.

[23] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[24] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ArXiv*, abs/1808.06670, 2018.

[25] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding, 2019.

[26] Md Amirul Islam, Matthew Kowal, Patrick Esser, Sen Jia, Björn Ommer, Konstantinos G. Derpanis, and Neil Bruce. Shape or texture: Understanding discriminative features in {cnn}s. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=NcFEZOi-rLa.

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

[28] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21798–21809. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f7cade80b7cc92b991cf4d2806d6bd78-Paper.pdf.

[29] A. Khosla, Nityananda Jayadevaprakash, B. Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization : Stanford dogs. 2012.

[30] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2656–2666, 2019.

[31] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[33] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

[34] Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. i-mix: A strategy for regularizing contrastive representation learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=T6AxtOaWydQ.

[35] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and cihang xie. Shape-texture debiased neural network training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Db4yerZTYkz.

[36] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[37] Sifei Liu, Jin shan Pan, and Ming-Hsuan Yang. Learning recursive filters for low-level vision via a hybrid neural network. In *ECCV*, 2016.

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2015.

[39] Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin Dogus Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *ArXiv*, abs/1906.02611, 2019.

[40] Kaiyue Lu, Shaodi You, and Nick Barnes. Deep texture and structure aware filtering network for image smoothing, 2018.

[41] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[42] Shlok Mishra, Anshul Shah, Ankan Bansal, Abhyuday Jagannatha, Abhishek Sharma, David Jacobs, and Dilip Krishnan. Object-aware cropping for self-supervised learning, 2021.

[43] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations, 2019.

[44] Chaithanya Kumar Mummadi, Ranjitha Subramaniam, Robin Hutmacher, Julien Vitay, Volker Fischer, and Jan Hendrik Metzen. Does enhanced shape bias improve neural network robustness to common corruptions? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=yUxUNaj2Sl.

[45] Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. doi: 10.1109/cvpr42600.2020.00737. URL http://dx.doi.org/10.1109/CVPR42600.2020.00737.

[46] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. *ArXiv*, abs/1603.09246, 2016.

[47] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[48] Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12:629–639, 1990.

[49] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019.

[50] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[51] Sam Ringer, Will Williams, Tom Ash, Remi Francis, and David MacLeod. Texture bias of cnns limits few-shot classification performance, 2019.

[52] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, Apr 2017. ISSN 2160-9292. doi: 10.1109/tpami.2016.2572683. URL http://dx.doi.org/10.1109/TPAMI.2016.2572683.

[53] Baifeng Shi, Dinghuai Zhang, Qi Dai, Zhanxing Zhu, Yadong Mu, and Jingdong Wang. Informative dropout for robust representation learning: A shape-bias perspective, 2020.

[54] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6827–6839. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c2e5eaae9152079b9e95845750bb9ab-Paper.pdf.

[55] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 839–846, 1998.

[56] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Iccv*, volume 98, page 2, 1998.

[57] A. Torralba and Alexei A. Efros. Unbiased look at dataset bias. *CVPR 2011*, pages 1521–1528, 2011.

[58] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[59] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary Chase Lipton. Learning robust global representations by penalizing local predictive power. *ArXiv*, abs/1905.13549, 2019.

[60] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021.

[61] Yichong Xu, Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, and Zheng Zhang. Scale-invariant convolutional neural networks. *ArXiv*, abs/1411.6369, 2014.