Towards Domain Adversarial Methods to Mitigate Texture Bias

Dhruva Kashyap [*1] **Sumukh K Aithal** [*1] **Rakshith C** [*1] **Natarajan Subramanyam** [1]

Abstract

Shape-texture conflict is key to our understanding of the behavior of Convolutional Neural Networks (CNNs) and their observably good performance. This work proposes a domain adversarial training-inspired technique as a novel approach to mitigate texture bias. In our work, instead of looking at the domains as the source from which the images are from, we look at the domains as inherent features of the image. The model is trained in a method similar to Domain Adversarial training, where we define the source and target domains as the dataset and its augmented versions with minimal texture information (edge maps and stylized images), respectively. We show that using domain invariant learning to capture a prior based on the shape-texture information helps models learn robust representations. We perform extensive experiments on three subsets of ImageNet, namely, ImageNet-20, ImageNet-200, ImageNet-9. The results show that the proposed method outperforms standard Empirical Risk Minimization (ERM) in terms of test accuracy and also as evidenced by the high accuracy on the Out-Of-Distribution (OOD) datasets ImageNet-R and NICO.

1. Introduction

It is a widely held belief that the reason CNNs perform well is by first detecting low-level features and gradually moving towards higher level shapes, allowing them to capture the necessary details in images (Szegedy et al., 2014). This hypothesis was recently challenged by (Geirhos et al., 2019), which dubbed it the shape hypothesis. Although there has been extensive work to solidify this shape hypothesis, (Geirhos et al., 2019) shows that, in fact, CNNs primarily rely on object textures rather than their shapes. Although the shape is as essential to learning as texture, it must be noted that an absence of texture knowledge will undoubtedly be detrimental to the model, as shown in (Li et al., 2021). Texture plays a crucial role in fine-grained classification, which can be thought of as differentiating between German Shepherds and Poodles. Whereas shape information is at a broader level of classification, differentiating dogs from boats. This can also be observed in (Geirhos et al., 2019) when training only on the stylized version of ImageNet, which is argued to contain purely shape information, performs much poorer than training on standard ImageNet.

Though CNNs have shown that local textures are enough to achieve good performance on a diverse dataset like ImageNet, these models are not robust to corruptions or domain shifts (Hendrycks et al., 2021). The performance of these models drops significantly with minor distortions in the image, which do not change the semantics of the object class. For example, the accuracy of the ImageNet trained AlexNet model drops by more than 50% on ImageNet-Sketch (Wang et al., 2019). This drop in performance can be attributed to the over-reliance on the local textures of the object rather than the global shape. However, only the object's shape might not be enough to classify an object. Texture determines the specific fine-grained class, and shape determines the coarse-grained class. Thus, this conflict is a critical problem in visual representation learning for good OOD generalization.

In this paper, we introduce the usage of a modified Domain Adversarial training technique as an alternative to standard ERM techniques to mitigate texture bias and improve OOD generalization. We use the training dataset as the source domain and use shape agnostic representations of the dataset as the target domain. The shape agnostic representations that we use are stylized images and edge maps. We look at the target domain from the lens of a prior or a feature that can be used to learn better representations for suitable downstream tasks. We show that using augmented samples as the target domain induces a prior to mitigate texture bias. (Fig. 1) Throughout the rest of the paper, we use the term "shape texture conflict" as the conflict in the bias towards object

[*]Equal contribution [1]Department of Computer Science, PES University, Bengaluru, India. Correspondence to: Dhruva Kashyap <dhruva12kashyap@gmail.com>, Sumukh K Aithal <sumukhaithal6@gmail.com>, Rakshith C <rakshithr7r@gmail.com>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA.

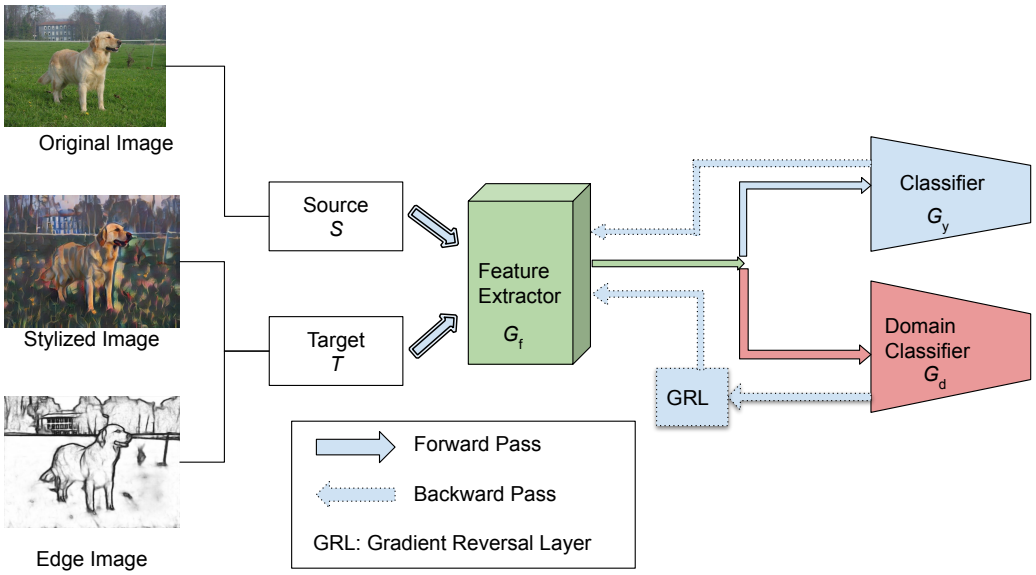

Figure 1. Overview of the proposed method. We propose the idea of using images with minimal texture information as the target domain in Domain Adversarial Training and show that it induces a shape-prior. This leads to improved OOD generalization.

textures or object shapes for deep learning trained models. We train on various flavors of ImageNet. We show results on their respective test sets and also on OOD datasets like ImageNet-Rendition (Hendrycks et al., 2021), and NICO (He et al., 2020).

In summary, our contributions are as follows:

1. We propose using the idea of Domain Adversarial training as a means to capture priors for models to learn robust representations. In this work, we use shape agnostic augmentations as the target domain.

2. We propose the interpretation of domains as features inherent to the object as opposed to originating from a different input distribution.

3. Our method outperforms baseline methods on ImageNet-R, ImageNet-9, and NICO, which are Out-of-Distribution datasets.

2. Related Work

Shape-Texture Conflict: Much work to mitigate texture bias has been studied since (Geirhos et al., 2019). This paper described using an augmented dataset that contained stylized images to combat texture bias. This approach is shown to reduce the texture bias in ResNet50 models. Another work suggested that the difference between human and ImageNet trained CNNs may stem from the training data that the model sees and not due to the internal mechanics themselves (Hermann et al., 2020). It is shown that CNNs trained with natural augmentations on training data can outperform the standard models by reducing their reliance on texture cues.

Random Convolutions (Xu et al., 2021) is a method of augmenting training data to remove local texture information. Training with random convolutions shows the effectiveness of shape-biased models in downstream tasks.

Another method, InfoDrop (Shi et al., 2020) is a Dropout inspired, a model agnostic algorithm that incorporates local self-information present in an image to increase shape-bias. They also show the interplay between robustness and shape bias. Shape texture debiased training (Li et al., 2021) shows that both shape and texture labels are required in a supervised setting to balance cue conflict correctly. They propose a method to balance these cues by inter-class image stylization and by using a label assignment strategy based on MixUp. It is also shown that shape-biased and texture-biased models rely on complementary cues. Patch-wise adversarial regularization (Wang et al., 2019) penalizes models when they can predict images using local patches and encourage the model to learn global features. It is also shown to be effective on a new proposed dataset, ImageNet-Sketch, which contains sketch images of the various class of ImageNet.

Recent works (Mummadi et al., 2021) have shown that shape bias is not correlated to the corruption robustness by training the model on the augmented dataset with edge maps and other variants. One of the reasons for the prevalence of texture over shape is shortcut learning (Geirhos et al., 2020), which shows that deep neural networks may rely on

spurious features (or shortcuts) to classify the data.

Domain Adaptation: Domain Adaptation methods have been used to learn representations of the model trained on the images from the source domain that generalizes well on the images from the target domain. Domain Adversarial Neural Networks (DANN) (Ganin et al., 2016) proposed an adversarial framework for domain adaptation by introducing a domain classifier to classify which domain a feature belongs to. DANN aims to learn domain-invariant representations that perform well on the target domain images. In this work, we use this idea to learn better representations and mitigate texture bias.

3. Proposed Method

We propose to use the idea of Domain adversarial training, a popular Domain Adaptation algorithm, to mitigate texture bias and learn more shape-related features. In previous works, the source and target domains have been used as two different sources of input. In this work, we propose to interpret the target domain as a prior to the model that learns invariant representations for specific tasks. This idea is used to model specific invariances such as shape bias, by having a shape agnostic augmentation in the target domain. We also show that this can be effectively used to avoid spurious correlations such as background (Table 4).

To remove the texture information from the images, we use the pre-trained DexiNed (Soria et al., 2020) model to generate edge maps of the image. The generated edge maps preserve the global shape of the object and contain minimal texture information. We also stylize the image using AdaIN (Huang & Belongie, 2017) similar to (Geirhos et al., 2019). Stylizing with painting changes the texture of the original image while preserving the global shape.

Domain adversarial training introduces the domain discriminator, which tries to classify the domain to which a feature map belongs. We use the original image as the source domain and the union of edge map and stylized images as the target domains. Using this choice of source and target, the model has to learn discriminative features that the model may not have learned with ERM to achieve good performance. It also introduces a prior that can generalize to other domains. We use DANN as the framework for our Domain Adaptation technique. Recent work (Musgrave et al., 2021) has shown that DANN is simple and effective for real-world Domain Adaptation problems.

We introduce a weighted target class loss term in the Domain adversarial training framework. This ensures that the model is not only robust to the augmentations but also captures the domain invariant prior. The target class loss is a standard cross-entropy loss that minimizes the error on the images in the target domain. Introducing such a loss enables the model to be robust to texture transformations. We also need the model to classify the edge map and stylized version accurately, which is a measure of shape information in the learned representation. Equation 1 captures the concept of domain adversarial training.

Notation: S refers to samples from the source domain, which refers to original images, and T refers to the samples from the target domain, which refers to the generated edge maps and the stylized images. N_S and N_T refer to the number of source samples and target samples, respectively. x_i, y_i, and d_i refer to the image, its class label, and domain label, respectively. The domain label is zero for source samples and one for target samples. L_y refers to the class loss for which we use cross-entropy loss, and L_d refers to domain loss. G_f refers to feature extractor, G_y refers to feature classifier, and G_d refers to domain classifier. $\theta_f, \theta_y, \theta_d$ refers to the trainable parameters of the feature extractor, feature classifier, and domain classifier, respectively. R refers to the gradient reversal layer. The modified learning objective is as follows:

$$\begin{aligned}\tilde{E}(\theta_f,\theta_y,\theta_d) = &\frac{1}{N_S}\sum_{x_i\in\mathcal{S}} L_y\left(G_y(G_f(\mathbf{x}_i;\theta_f);\theta_y),y_i\right)\\ &+\frac{\alpha}{N_T}\sum_{x_i\in\mathcal{T}} L_y\left(G_y(G_f(\mathbf{x}_i;\theta_f);\theta_y),y_i\right)\\ -\lambda\Big(&\frac{1}{N_S}\sum_{x_i\in\mathcal{S}} L_d\left(G_d(\mathcal{R}(G_f(\mathbf{x}_i;\theta_f));\theta_d),d_i\right)\\ &+\frac{1}{N_T}\sum_{x_i\in\mathcal{T}} L_d\left(G_d(\mathcal{R}(G_f(\mathbf{x}_i;\theta_f));\theta_d),d_i\right)\Big). \quad (1)\end{aligned}$$

The first two terms refer to the source and target class loss respectively and the next two terms refer to the source and target domain loss. The second term is the additional weighted source target loss term that we have introduced which captures domain invariant priors.

The modified domain adversarial training algorithm does not allow the model to exploit shortcuts in the dataset and thus captures more generalizable features. The proposed method leads to much better performance on OOD datasets, indicating the efficacy of the learned representations.

4. Results

We compare our result with standard baselines on three subsets of ImageNet: ImageNet-20, ImageNet-9, and ImageNet-200. We also compare our proposed method with baselines on the NICO dataset. Additional information about each dataset is reported in Appendix A and experimental details in Appendix D.

Table 1. Validation accuracies of various models on IN20 and ImageNet-R (20 classes).

Model	IN-20 Top-1 Acc	ImageNet-R
IN20	91.50	39.30
SIN	76.60	50.95
EIN	27.00	20.28
IN20 + EIN	92.30	52.81
IN20 + SIN	91.00	55.61
IN20 + EIN + SIN	92.20	60.07
ShapeTexture-Debiased Training (Li et al., 2021)	**92.65**	42.24
RandomConvolutions (Xu et al., 2021)	81.46	29.58
InfoDrop (Shi et al., 2020)	81.60	39.62
DANN - IN20, EIN+SIN (Ours)	91.90	**62.25**

Table 2. Validation accuracies of baseline and domain adversarial models on IN-200 and ImageNet-R (200 classes).

Model	IN-200 Top-1 Acc	ImageNet-R
IN200 - Baseline	**89.30**	33.02
DANN- IN200, SIN+EIN (Ours)	86.68	**44.02**

Table 3. Validation accuracies of baseline and domain adversarial models on NICO.

Model	Val Acc	Test Acc
NICO	73.84	73.00
NICO + SIN + EIN	77.38	75.23
DANN - NICO, SIN+EIN	**78.38**	**76.15**

Table 4. Performance on the ImageNet-9 dataset. BG-GAP refers to the difference in accuracy between MIXED-RAND (MR) and MIXED-SAME (MS) and indicates the impact of backgrounds on the model's prediction. (Lower BG-GAP is better)

Source	Target	Test	MS	MR	BG-GAP ↓
IN-9	-	85.95	73.80	53.58	20.22
IN-9L	-	94.61	89.90	75.60	14.30
IN-9L	SIN-9L + EIN-9L	93.43	87.46	78.69	8.77
IN-9	MR	92.81	**91.08**	84.89	6.19
IN-9	MS+MR	90.69	90.96	**87.73**	**3.23**

4.1. Datasets

ImageNet-20 (IN-20): A subset of 20 classes from ImageNet dataset which are in common with classes of ImageNet-R.

Stylized ImageNet (SIN): Stylized images of the corresponding dataset, generated using AdaIN style transfer.

EdgeMaps ImageNet (EIN): Edgemaps of the dataset generated using pretrained DexiNed model. Details about edgemap generation are available in Appendix C.

ImageNet-200 (IN-200): A subset of 20 classes from ImageNet dataset which are in common with classes of ImageNet-R.

Non-I.I.D. Image dataset with Contexts (NICO): is a dataset designed for OOD settings.

4.2. Experiments

Results on IN-20: Table 1 shows the results on the IN-20 dataset. We compare our method with other state-of-the-art methods Shape-Texture Debiased training (Li et al., 2020), Random Convolutions, and Informative Dropout. IN20 + SIN indicates that the model is trained with standard ERM on the union of the two datasets, similar to (Geirhos et al., 2019).

The proposed method outperforms a standard ERM model trained on IN-20+SIN+EIN on ImageNet-R by 1.6%. We observe that the model's performance on ImageNet-R, an OOD dataset, is heavily influenced by the training data from which the model learns. Comparing ERM models trained on standard IN20 and its stylized counterpart, we can see a significant improvement of 11% in the OOD accuracy. Furthermore, training on all flavors of IN-20 improves the OOD accuracy by 20%. Adding either only EIN or SIN to the original dataset also significantly improves OOD accuracy. However, only a slight increase can be seen on the In-distribution test set. The domain adversarial models described in our proposed method outperform the standard models in terms of In-distribution test accuracy and OOD test accuracy.

Results on IN-200: On the larger IN-200 subset of ImageNet, we compare our proposed method with standard ERM training on the dataset. Table 2 shows our results on IN-200. Although the standard model shows a slight improvement over the proposed method in the Top-1 accuracy on the validation set, it can be observed that the proposed method clearly outperforms the standard model in terms of OOD generalization, as indicated by the test accuracy on ImageNet-R by 11%.

Results on NICO: We train our baseline and domain adversarial model on seven contexts and test them on three unseen contexts. Table 3 shows the superior performance of our domain Adversarial model over the baseline model, which indicates that our Domain Adversarial model has captured better representations by focusing on the object, unlike the baseline model, which focuses more on the context of the object. It is important to note that we compare our model with a baseline consisting of a stylized version and edge maps of NICO, and yet our model surpasses the performance of the ERM model.

Results on ImageNet-9: We conduct experiments using

Table 5. Comparison of Shape-Texture cue conflict score and 4 x 4 patch accuracy of models trained on IN-200.

Model	Top-1 Acc	Shape Score	Texture Score	4x4 Patch Acc
IN200	**89.30**	135	157	61.36
DANN - IN200, EIN+SIN (Ours)	86.68	**206**	**136**	**47.21**

the source as IN-9 and targets as MIXED-RAND (MR) and MIXED-SAME (MS), and only MIXED-RAND. We also conduct experiments using IN-9L as the source domain and edgemaps and stylized images of IN-9L as the target domain. We observe in Table 4 that we obtain higher test accuracy when using stylized images and edgemaps. We also show that having target domains such as MIXED-RAND and MIXED-SAME teaches the model better background invariant representations, as shown by the background gap. It can be seen that using stylized images and edgemaps as the target domain allows the model to learn background invariant features, leading to only a 3% background gap. We also show an ablation study (Table 6) in Appendix B.

4.3. Evidence of Shape Bias

To verify that the proposed method is more shape biased, we perform experiments on Shape-texture cue-conflict dataset (Geirhos et al., 2019). For the IN-200 dataset, there are around 560 images which we evaluate on. It can be seen from Table 5 that the proposed method is more shape biased compared to the baseline model. The accuracy on the shuffled image patches is also an indication of shape bias (Luo et al., 2019). A high accuracy on the randomly shuffled patches indicates that the model is focusing more on the local patches rather than the global shape. Though Top-1 Accuracy of baseline method is higher than that of our method, the 4 x 4 patch accuracy shows that our method has a much higher patch accuracy compared to our method indicating higher texture bias of the baseline method.

5. Conclusion

We have introduced Domain Adversarial techniques as a means to mitigate shape-texture conflicts in CNNs. In this work, we show that domain adaptation methods can be effectively used to train models that are generalizable to OOD datasets. With Domain Adversarial training, the model learns both domain-specific and domain invariant features, thereby mitigating texture bias and learning generalizable representations. The results on various datasets show that the proposed method outperforms all other techniques, especially in challenging datasets like ImageNet-R. We also show results on NICO and IN-9 datasets, which evaluate the performance of the models in different contexts and background invariance, respectively.

References

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL http://dx.doi.org/10.1038/s42256-020-00257-z.

He, Y., Shen, Z., and Cui, P. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, pp. 107383, 2020.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

Hermann, K., Chen, T., and Kornblith, S. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Huang, X. and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization, 2017.

Li, Y., Yu, Q., Tan, M., Mei, J., Tang, P., Shen, W., Yuille, A., and Xie, C. Shape-texture debiased neural network training. *arXiv preprint arXiv:2010.05981*, 2020.

Li, Y., Yu, Q., Tan, M., Mei, J., Tang, P., Shen, W., Yuille, A., and cihang xie. Shape-texture debiased neural network training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Db4yerZTYkz.

Luo, T., Cai, T., Zhang, X., Chen, S., He, D., and Wang, L. Defective convolutional layers learn robust cnns. 2019.

Mummadi, C. K., Subramaniam, R., Hutmacher, R., Vitay, J., Fischer, V., and Metzen, J. H. 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.

Musgrave, K., Belongie, S., and Lim, S.-N. Unsupervised domain adaptation: A reality check. *arXiv preprint arXiv:2111.15672*, 2021.

Shi, B., Zhang, D., Dai, Q., Zhu, Z., Mu, Y., and Wang, J. Informative dropout for robust representation learning: A shape-bias perspective. In *International Conference on Machine Learning*, pp. 8828–8839. PMLR, 2020.

Soria, X., Riba, E., and Sappa, A. Dense extreme inception network: Towards a robust cnn model for edge detection. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1912–1921, Los Alamitos, CA, USA, mar 2020. IEEE Computer Society. doi: 10.1109/WACV45572.2020.9093290. URL https://doi.ieeecomputersociety.org/10.1109/WACV45572.2020.9093290.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks, 2014.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32:10506–10518, 2019.

Wang, T., Zhou, C., Sun, Q., and Zhang, H. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Xiao, K., Engstrom, L., Ilyas, A., and Madry, A. Noise or signal: The role of image backgrounds in object recognition. *ArXiv preprint arXiv:2006.09994*, 2020.

Xu, Z., Liu, D., Yang, J., Raffel, C., and Niethammer, M. Robust and generalizable visual representation learning via random convolutions. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BVSM0x3EDK6.

A. Datasets

A.1. ImageNet-20

We use a subset of 20 classes from ImageNet dataset which are in common with classes of ImageNet-R (Hendrycks et al., 2021). To ensure a balanced dataset, we construct a hierarchy of ImageNet classes of ImageNet-R. Two subclasses are randomly sampled from ten superclasses which have more than six subclasses. IN-20 comprises 20 classes with 26,000 training images and 1,000 validation images. The subset consists of birds (vulture, hen), bugs (mantis, bee), cats (tiger, leopard), dogs (golden retriever, Weimaraner), fish (anemone fish, goldfish), fruits (banana, pineapple), food (espresso, burrito), instrument(harmonica, flute), tools (candle, bucket), vehicles (jeep, tank).

A.2. ImageNet-R

ImageNet-R is collection of various renditions of images like cartoons, deviantart, graffiti, embroidery etc. consisting of 200 ImageNet classes resulting in 30,000 images. We use ImageNet-R as a primary benchmarking dataset for measuring out-of-distribution performance.

IN-200: represents the 200 classes of ImageNet which are in common with ImageNet-R. IN-200 contains 258,951 training images and 10,000 validation images.

A.3. NICO

Non-I.I.D. Image dataset with Contexts (NICO) is a dataset designed for OOD settings. NICO dataset contains 19 object classes, 188 contexts and 25,000 images in total. It simulates the real world by arbitrarily shifting image contexts. NICO comprises two superclasses, animal and vehicle. We follow the experimental settings of (Wang et al., 2021) wherein images of only the animal superclass with 10 contexts are used, out of these 10 contexts 7 are used for training and 3 are used for testing. During testing, the number of test samples across the 7 contexts is 50, and for remaining 3 contexts number of samples is 100. Unlike the original settings, long-tailed distribution is omitted during training.

A.4. ImageNet-9

We experiment with ImageNet-9 (Xiao et al., 2020), a dataset created to measure the background invariance of image classification models. The base dataset contains 9 superclasses of ImageNet, namely, dog, bird, vehicle, reptile, carnivore, insect, instrument, primate, and fish. This dataset is then used to generate other datasets with varying foreground and background information. The background gap, an indication of the model's reliance on the background is measured as the difference in test accuracy between MIXED-SAME and MIXED-RAND. We conduct experiments based on our proposed method on training on the IN-9. ONLY-BG-B contains the background of the image with the bounding box of the enclosed object blacked out. ONLY-BG-T is the background of the image with the bounding box of the object replaced with a background tile of the same image. NO-FG contains the background with the foreground object blacked out using foreground detection methods. Similarly, ONLY-FG contains the foreground object extracted using a foreground detection technique with a black background. MIXED-SAME contains images where the background has been swapped with another background of the same class, and MIXED-RAND contains images where the background is swapped with the background of a randomly chosen class.

B. Ablation

Ablation of different source-target pairs: We experiment with the source domain containing the images and their edgemaps, and the target domain having the stylized images. We observe that we obtain optimal performance when we use the original dataset as the source domain and use both of these augmented datasets as the target domain. Table 6 shows the impact of using different source and target pairs in the dataset. It can be seen that using IN20 as the source and EIN+SIN as the target works best among all the pairs. It can be seen that just using the EIN as the target significantly decreases the model's performance, and using stylized data is vital for the model to generalize well to OOD data.

Table 6. Ablation study of different source and target datasets for Domain Adversarial model.

Source	Target	IN-20 Top-1 Acc	ImageNet-R
IN20	EIN	86.30	48.24
IN20	SIN	91.10	58.65
IN20 + EIN	SIN	90.70	61.24
IN20 + SIN	EIN	89.10	57.33
IN20	EIN+SIN	**91.90**	**62.25**

C. EdgeMap Generation Technique

EdgeMaps are generated via pretrained DexiNed network (Soria et al., 2020).. DexiNed is a combination of two networks, Dense extreme inception network (Dexi), which receives an RGB image as input, and an up-sampling block (UB), which receives feature maps as input from each block of Dexi. The edge map generated from each upsampling block is combined to produce fused edge maps. DexiNed is capable of adapting to domain shifts and can outperform other state-of-the-art edge map models. DexiNed produces two variants of EdgeMaps i.e., DexiNed-averaged and DexiNed-fused. We use the authors official implementation [1] and utilize DexiNed-averaged in all our experiments.

D. Experimental details

We use ResNet50 architecture as the backbone for our Domain Adversarial model and the baseline model. The baseline refers to the model trained with ERM. The baseline model is trained for 100 epochs with a learning rate of 0.01, reduced by a factor of 10 at the 60th and 90th epochs, respectively. We use a batch size of 128 and train using SGD with momentum and weight decay of 0.01. We use the standard data augmentations used for training ImageNet models. The Domain Adversarial model is trained with the same hyperparameters as the baseline model. Unlike (Ganin et al., 2016), the same learning rate is kept across the feature extractor and feature classifier since we do not initialize the model with pre-trained weights. We use a α of 0.5 for all our experiments.

[1] https://github.com/xavysp/DexiNed