Supplementary Material for "Shift from texture-bias to shape-bias: edge deformation-based augmentation for robust object recognition"

Xilin He[1,2,3], Qinliang Lin[1,2,3], Cheng Luo[1,2,3] Weicheng Xie[1,2,3,*], Siyang Song[4], Feng Liu[1,2,3], Linlin Shen[1,2,3]
[1]Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University
[2]Shenzhen Institue of Artificial Intelligence & Robotics for Society
[3]Guangdong Key Laboratory of Intelligent Information Processing
[4]University of Leicester
{2020152115, linqinliang2021, luocheng2020}@email.szu.edu.cn
{wcxie,feng.liu,llshen}@szu.edu.cn, ss1535@leicester.ac.uk

1. Online vs offline augmentation

To investigate the performance of the online fashion of the data augmentation, we compare it with offline augmentation in terms of adversarial robustness and runtime cost on CIFAR-10. The offline augmentation refers to implementing TSD multi-times to expand the training set. For the online augmentation, we augment the shape the same fold as offline augmentation with TSD, and use them repeatedly for the following training. Experiment results are shown in Fig. S1, where the proposed online augmentation achieves much better adversarial robustness than the baseline, and results in only slightly lower robustness than the offline augmentation. Meanwhile, our algorithm needs only the similar runtime cost as the vanilla training, which is significantly lower than that of offline augmentation.

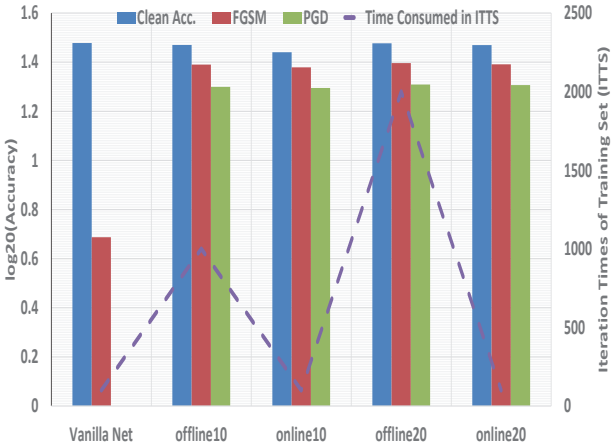

Figure S1. Accuracy and iteration times of training set (ITTS) of offline and online augmentation on CIFAR-10. 'online10' denotes a tenfold increase in training data by TSD.

2. Additional visualization comparisons

In this section, we shed light on how our edge deformation-based generation images differ from directly deformed images.

Comparison with direct image deformation: As shown in Fig. S2, as the deformation intensity λ increases, producing directly deformed images with *hard constraint* [23] needs to strictly follow the deformation strategy, leading to ignorance of semantic information, shape rationality or texture details. By contrast, we perform TPS deformation on the edge map of an input image, then we inpaint the texture via a generator trained with edge guidance *soft constraint* loss (Eq. (7) in the manuscript). With this loss, generated images on the deformed edge map can be restricted in a reasonable range, leading to diversity and rationality of deformed shapes.

The performance of EMSE: To shed light on how the introduced self-information edge map works, the produced edge maps with different shape encoding are presented in Fig. S3.

Fig. S3 shows that the edge maps of object boundaries obtained by our method can encode richer shape cues to better against perturbation noises, compared with the Robust Canny in the 2nd row. Meanwhile, our shape encoding enhances the importance of object edges, thus representing shapes more accurately than those with the self-information guided map in the 3rd row.

The performance of TSG: To study the performance of TSG, we visualized the generated samples with or without the proposed generator in Fig. S4. Specifically, compared with the pix2pix approach [9], the proposed TSG generates images with well-preserved sharpness of object shape and realistic texture, e.g. the generated image in the 4th column, i.e. it better builds up the connection between shape and the

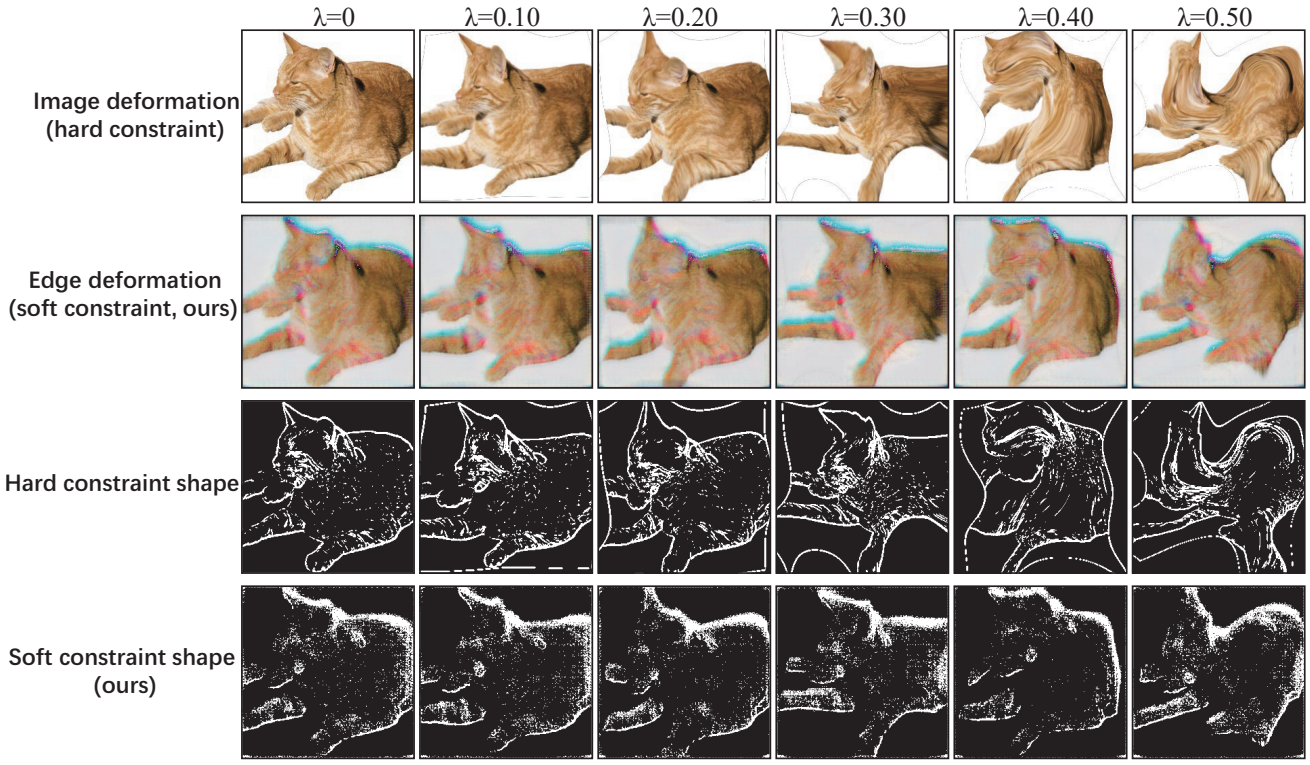

Figure S2. Visualizations of image deformation and our edge deformation-based generation images and their self-information guided maps.

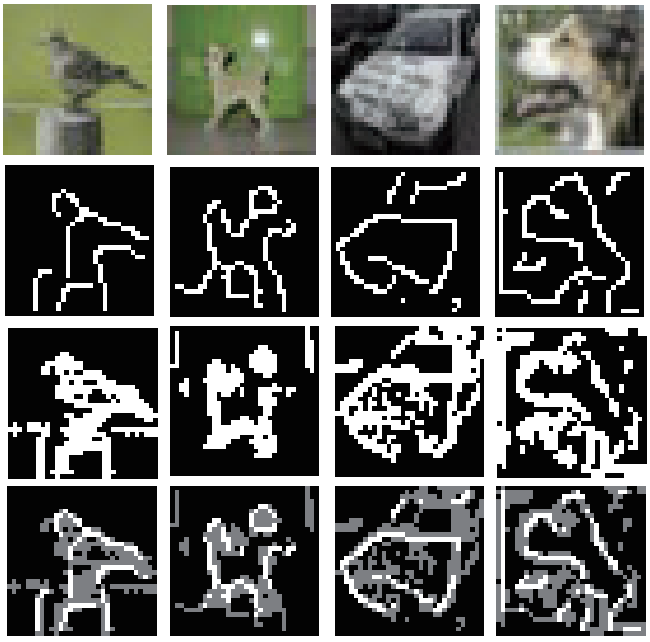

Figure S3. From the 1st to 4th rows, it represents raw images from CIFAR-10, edges extracted by Robust Canny [19], self-information guided map [18], the proposed extended edge map.

texture label.

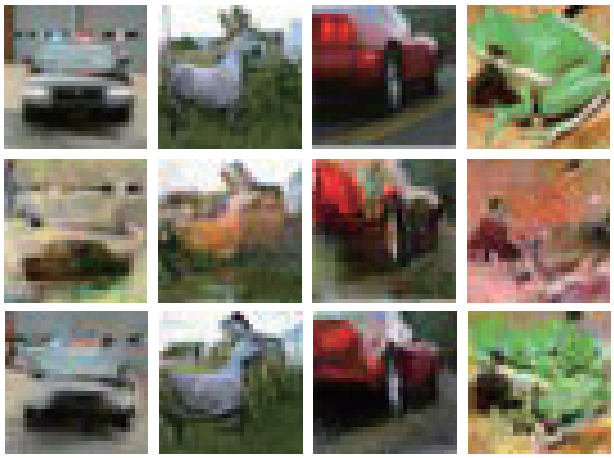

Figure S4. From the 1st to 3rd rows, it represents raw images from the testing set of CIFAR-10, generated images by vanilla pix2pix [9] together with EMSE and TSD, and our SDbOA.

3. Hyperparameter analysis

Deformation intensity λ: While Fig. S5 shows the sensitivity of shape-bias against the deformation intensity λ, we further present example samples with different λ in Fig. S6. It is clear that *different samples can withstand varying degrees of deformation intensity before being misclassified*. More precisely, when λ is relatively small, the global shape structure maintains similar to the benign image, while

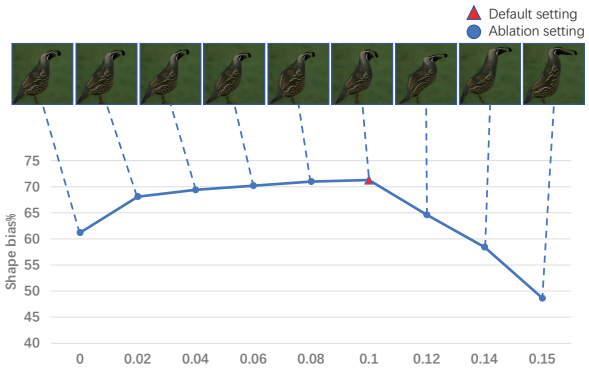

Figure S5. Shape-bias sensitivity against different deformation intensities (λ) on the texture-shape cue conflict dataset [2] in terms of Geirhos' shape-bias metric [3].

it could be damaged when the images are highly twisted under a large λ. As labeled in the red rectangles of Fig. S6, these generations are no longer easily classified correctly, which may hence impair the generalization ability of the learned feature representation when they are used for training.

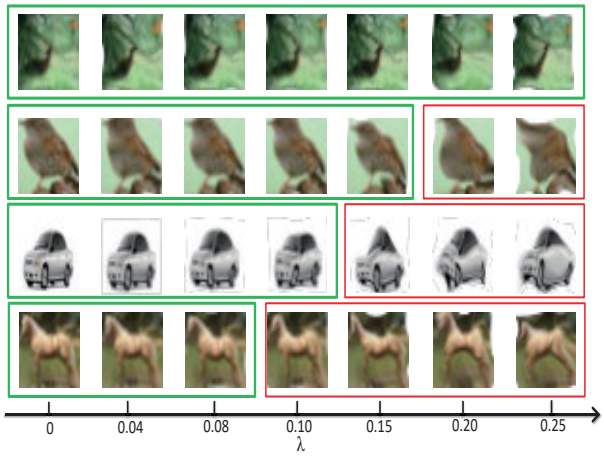

Figure S6. Synthesized images with different deformation degrees (λ). The red rectangles label the deformed images that are wrongly classified.

Number of grids n: In this section, we analyze how the number of grids n during the meshing TPS procedure affects the model's shape bias. As shown in Fig. S7, we evaluate shape bias of models trained with different number of grids n and a fixed deformation intensity $\lambda = 0.1$.

4. More quantitative results

Results in terms of shape-bias metrics: In this section, we provide more quantitative results in terms of two shape-bias metrics in Tab. S1, it shows that our SDbOA achieves larger shape-bias values compared with SOTA data augmentation techniques [6, 7, 21] in terms of both metrics.

Robustness against common corruptions: To study the robustness of the proposed algorithm under common corruptions [5], we follow the protocol in Sun et al. [19], and present the robustness performances in Tab. S2. In Tab. S2, the mean corruption error (mCE) [5] across all 15 corruptions and 5 severities for each corruption is used as the corruptions metric. Tab. S2 shows that our SDbOA can consistently improve the model robustness of the vanilla net under common corruptions, and achieves an improvement of 14.77% on ImageNet-C.

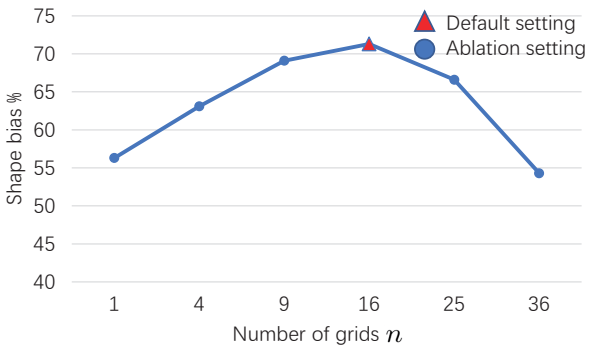

Figure S7. Shape-bias sensitivity against different number of grids (n) on the texture-shape cue conflict dataset [2] in terms of Geirhos' shape-bias metric [3].

Method	Dataset	sb_{GE} [3]	sb_{IS} [8]
Standard	IN	21.39	17.0
AugMix [6]	IN	26.85	17.4
AugMax [21]	IN	29.51	17.4
PixMix [7]	IN	32.79	20.1
SDbOA	IN	**71.28**	**31.2**

Table S1. Quantitative results (%) in terms of two shape-bias metrics with SOTA data augmentation methods, i.e. AugMix [6] (ICLR'19), AugMax [21] (NeurIPS'21), PixMix [7] (CVPR'22) and our SDbOA.

Models	FM-C	CA-C*	C10-C	IN-C
Vanilla Net	67.52	65.37	65.83	38.64
SDbOA	**72.05**	**71.80**	**78.46**	**53.41**

Table S2. Robustness performance (%) under common corruptions [5]. * denotes that the corruptions are generated by us. 'FM-C' is the abbreviation of 'Corrupted Fashion MNIST'.

Performance under adversarial training: Since SDbOA is a non-adversarial-training method, to further study the adversarial robustness of SDbOA incorporated with adversarial training [14], we further compare our algorithm with two generative-model-based adversarial training methods [16, 17]. As shown in Tab. S3, despite SDbOA incorporated with adversarial training can achieve stronger robustness, standard SDbOA can better trade off the accuracy and robustness with less training time cost.

Method	Clean	Robust	Time	AT
FDA_{28} [16]	**85.97**	60.73	Several Days	✓
$SDbOA_{28}$+AT	83.04	**72.46**	One Day	✓
$SDbOA_{28}$	85.27	68.15	3-5 Hours	✗
$PORT_{34}$ [17]	87.00	60.60	Several Days	✓
$SDbOA_{34}$+AT	85.76	**75.08**	One Day	✓
$SDbOA_{34}$	**87.12**	73.36	3-5 Hours	✗

Table S3. Adversarial robustness (%) of ours and relative generation-based data augmentation SOTAs, i.e. FDA [16] (NeurIPS'21) and PORT [17] (ICLR'22) employing adversarial training. We follow the network architecture set up and the robust metric in [16] measured by AutoAttack [1]. AT represents PGD adversarial training [14]. '28' or '34' stands for using WideResNet-28-10 or WideResNet-34-10 [24] as the classifier network, following [16, 17].

5. Hyper parameter settings

In this section, the detailed experimental settings common across most experiments are provided.

Training setup: Network architectures from the ResNet family, namely ResNet-18, ResNet-50 [4] and variants of WideResNet [24] following previous works [7, 16, 17] are used. For the proposed TSG module, we implement it as a two-branch variant of Pix2Pix [9]. Each network is trained based on Adam optimizer [11]. In the first stage for training the TSG module, the learning rate, batch size and iteration epochs are set as 0.001, 64 and 100, respectively. In the second stage of the training, we jointly train the TSG module and the classifier network, where their learning rates are initialized as 0.0001 and 0.001, where a decay ratio of 0.1 for every 60 epochs is adopted in both two stages of training. We use the ResNet-18 model for Fashion MNIST [22], CelebA [13] and CIFAR-10 [12], and pretrained ResNet-50 for ImageNet in the adversarial robustness benchmark (Table 2 in the manuscript). ResNet-18 is used in the backdoor attack benchmark (Table 5 in the manuscript). We use variants of WideResNet (specified in the manuscript) in generative-model-based adversarial robustness benchmark (Table 3 in the manuscript) and corruptions robustness benchmark (Table 4 in the manuscript). Runtime evaluation and comparison (Table 3 in the manuscript) are conducted based on a NVIDIA A5000.

Evaluation setup: For adversarial robustness, adversarial samples are generated based on the PyTorch library Torchattacks [10]. For corruption robustness benchmark (Table 4 in manuscript), the CIFAR-10-C dataset with 15 kinds of image corruptions and 5 serverities is used. Results reported in Tab. S2 are evaluated on corresponding corruption databases for Fashion MNIST, CIFAR-10 and ImageNet [5], while the corruption database for CelebA is generated by us with the Python library Imagecorruptions [15]. The protocol and setup in [19, 20] are followed for evaluating backdoor attack robustness.

Pipeline setup: For the Robust Canny [19] in the proposed EMSE, the same hyper-parameter setting in [19] is employed. For representing self-information edge map, the average self-information value is used as the threshold. For the TSD module, we set the number of grids n as 16, the deformation intensity λ as 0.1. The setting of these hyper-parameters is fixed for EMSE and TSD on all the datasets.

6. More Visualization Results of Deformed Edge Maps and Synthesised Images

In this section, more edge maps acquired by our TSD, together with their synthesized images with online data augmentation are visualized in Fig. S8.

References

[1] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 4

[2] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021. 3

[3] 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. In *International Conference on Learning Representations*, 2018. 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[5] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018. 3, 4

[6] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 3

[7] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16783–16792, 2022. 3, 4

[8] Md Amirul Islam, Matthew Kowal, Patrick Esser, Sen Jia, Bjorn Ommer, Konstantinos G Derpanis, and Neil Bruce. Shape or texture: Understanding discriminative features in cnns. In *International Conference on Learning Representations*, 2021. 3

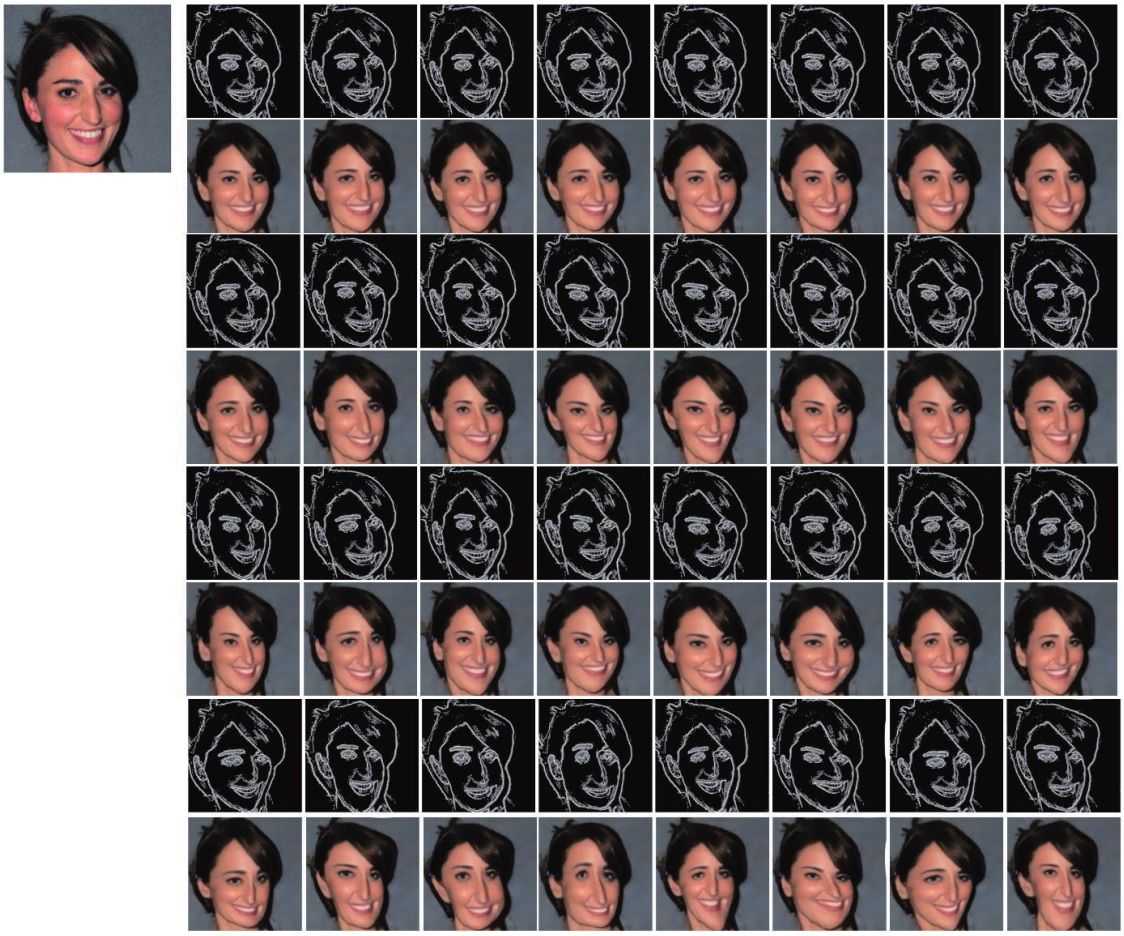

Figure S8. Deformed edge maps and the specific synthesized images. The image on the upper left corner is a raw image from CelebA [13].

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 2, 4

[10] Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020. 4

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[13] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 4, 5

[14] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 3, 4

[15] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 4

[16] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. In *Advances in Neural Information Processing Systems*, 2021. 3, 4

[17] Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *International Conference on Learning Representations*, 2022. 3, 4

[18] Baifeng Shi, Dinghuai Zhang, Qi Dai, Zhanxing Zhu, Yadong Mu, and Jingdong Wang. Informative dropout for robust representation learning: A shape-bias perspective. In *International Conference on Machine Learning*, pages 8828–8839. PMLR, 2020. 2

[19] Mingjie Sun, Zichao Li, Chaowei Xiao, Haonan Qiu, Bhavya Kailkhura, Mingyan Liu, and Bo Li. Can shape structure features improve model robustness under diverse adversarial settings? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7526–7535, 2021. 2, 3, 4

[20] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018. 4

[21] Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. Augmax: Adversarial composition of random augmentations for robust training. *Advances in neural information processing systems*, 34:237–250, 2021. 3

[22] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 4

[23] Jun Yu, Guochen Xie, Zhongpeng Cai, Peng He, Fang Gao, and Qiang Ling. Micro expression generation with thin-plate spline motion model and face parsing. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7210–7214, 2022. 1

[24] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016. 4