A CALCULATION OF THE EXPECTATION ON THE STYLE INFORMATION

We provide details of calculating $\mathbb{E}_{\hat{s}(\tilde{X})\sim\mathcal{N}(\mu(\tilde{X}),\sigma^2\boldsymbol{I})}CE\left(g\left(\hat{s}\left(\tilde{X}\right);W_g\right),Y\right)$. We assume a normal distribution for the styles, i.e., $\hat{s}\left(\tilde{X}\right)\sim\mathcal{N}\left(\mu\left(\tilde{X}\right),\sigma^2\boldsymbol{I}\right)$. According to the definition of the cross-entropy loss, for a input pair (x,y) we have:

$$
\begin{aligned}
&\mathbb{E}_{\hat{s}(x)\sim\mathcal{N}(\mu(x),\sigma^2\boldsymbol{I})}CE\left(g\left(\hat{s}\left(x\right);W_g\right),y\right)\\
&=\mathbb{E}_{\hat{s}(x)\sim\mathcal{N}(\mu(x),\sigma^2\boldsymbol{I})}\log\frac{1}{P(Y=y|g\left(\hat{s}\left(x\right);W_g\right)))}\\
&\le\log\frac{1}{\mathbb{E}_{\hat{s}(x)\sim\mathcal{N}(\mu(x),\sigma^2\boldsymbol{I})}P(Y=y|g\left(\hat{s}\left(x\right);W_g\right)))}\\
&=\log\frac{1}{\mathbb{E}_{\hat{s}(x)\sim\mathcal{N}(\mu(x),\sigma^2\boldsymbol{I})}\frac{e^{W_{g,y}^\top\hat{s}(x)}}{\sum_j e^{W_{g,j}^\top\hat{s}(x)}}}\\
&=\log\mathbb{E}_{\hat{s}(x)\sim\mathcal{N}(\mu(x),\sigma^2\boldsymbol{I})}\sum_j e^{(W_{g,j}-W_{g,y})^\top\hat{s}(x)}\\
&=\log\sum_j e^{(W_{g,j}-W_{g,y})^\top\hat{s}(x)+\frac{1}{2}(W_{g,j}-W_{g,y})^\top\sigma^2\boldsymbol{I}(W_{g,j}-W_{g,y})}\\
&=\log\frac{\sum_j e^{W_{g,j}^\top\hat{s}(x)+\frac{\sigma^2}{2}(W_{g,j}-W_{g,y})^\top(W_{g,j}-W_{g,y})}}{W_{g,y}^\top\hat{s}(x)}\\
&\triangleq\log\frac{1}{P\left((Y=y|\overline{g}\left(\hat{s}(x);W_g\right)\right)}\\
&\triangleq CE\left(\overline{g}\left(\hat{s}\left(x\right);W_g\right),y\right),
\end{aligned}
\tag{1}
$$

where the inequality follows from the Jensen's inequality: $\mathbb{E}\log(X)\le\log\mathbb{E}X$, the expectation is calculated by leveraging the moment-generating function:

$$\mathbb{E}e^{tX}=e^{t\mu+\frac{1}{2}\sigma^2t^2},X\sim\mathcal{N}(\mu,\sigma^2). \tag{2}$$

Note that, we define the function $\overline{g}\left(\hat{s}\left(x\right);W_g\right)$ for simplicity:

$$P\left((Y=y|\overline{g}\left(\hat{s}(x);W_g\right)\right)\triangleq\frac{W_{g,y}^\top\hat{s}(x)}{\sum_{j=1}e^{W_{g,j}^\top\hat{s}(x)+\frac{\sigma^2}{2}(W_{g,j}-W_{g,y})^\top(W_{g,j}-W_{g,y})}}. \tag{3}$$

B RELATIONSHIP BETWEEN ORTHOGONALITY AND STATISTICAL INDEPENDENCE

We give the proof for the following lemma in Sec. 3.3. Note that, we use R to present the learned representation of X, and replace X with R for simplicity.

Lemma 1. *$R\in\mathbb{R}^d$ is the learned representation, where d is the number of dimension of R. Assume that R is a normal distribution with mean m and covariance matrix M. The content used for predicting labels, i.e., logits, is obtained by applying a linear functions to R, i.e., $\hat{c}\left(R\right)=W_cR$, where W_c are parameters used for mapping R to logits. The style is modeled by a normal distribution, i.e., $\hat{s}\left(R\right)=\mu\left(R;W_s\right)+\Sigma\left(Y\right)^{\frac{1}{2}}\boldsymbol{n}$, where W_s presents parameters for modeling the mean of styles, and $\boldsymbol{n}$ is sampled from a standard normal distribution. Assume that $\mu\left(R;W_s\right)$ is a linear function, i.e., $\hat{s}\left(R\right)=W_sR+\Sigma\left(Y\right)^{\frac{1}{2}}\boldsymbol{n}$. Then, setting W_s as an instantiate of the orthogonal complement of W_c leads to statistical independence, i.e., $\hat{c}\left(R\right)\perp\!\!\!\perp\hat{s}\left(R\right)$. Here, $\perp\!\!\!\perp$ denotes the statistical independence, and we define $\langle a,b\rangle_M=\langle a,Mb\rangle$ for a given semi-definite matrix M. The orthogonality $A\perp_M B$ of two subspaces A and B is defined likewise.*

Proof. Under the assumption in Lemma 1, setting W_s as an instantiate of the orthogonal complement of W_c, we have:

$$\begin{aligned}
&\ker(W_s)^{\perp} \perp_M \ker(W_c)^{\perp} \Longleftrightarrow \mathrm{im}(W_s^{\top}) \perp_M \mathrm{im}(W_c^{\top}) \Longleftrightarrow \langle W_s^{\top}\boldsymbol{a}, W_c^{\top}\boldsymbol{b}\rangle_M = 0 \forall \boldsymbol{a}, \boldsymbol{b} \\
&\Longleftrightarrow \langle W_s^{\top}\boldsymbol{a}, MW_c^{\top}\boldsymbol{b}\rangle = 0 \forall \boldsymbol{a}, \boldsymbol{b} \Longleftrightarrow W_s M W_c^{\top} = 0 \Longleftrightarrow \mathbb{E}_R W_s (R-m)(R-m)^{\top} W_c^{\top} = 0 \\
&\Longleftrightarrow \mathbb{E}_{R,\boldsymbol{n}} W_s \left(R + \Sigma^{\frac{1}{2}}\boldsymbol{n} - m\right)(R-m)^{\top} W_c^{\top} = 0 \Longleftrightarrow Cov\left(\hat{s}(R), \hat{c}(R)\right) = 0 \\
&\Longleftrightarrow \hat{c}(R) \perp\!\!\!\perp \hat{s}(R)
\end{aligned} \tag{4}$$

□

C More details about evaluation metrics and training details

Evaluation metrics. For MNIST dataset, we set the maximum perturbation bound $\epsilon = 0.3$, perturbation step size $\eta = 0.01$, and the number of iterations $K = 40$ for PGD and C&W attacks, which keeps the same as (Zhang et al., 2019). Following (Rice et al., 2020), we set perturbation bound $\epsilon = 8/255$, perturbation step size $\eta = \epsilon/10$, and the number of iterations $K = 20$ for CIFAR10 dataset.

training details. For MNIST, we use the same CNN architecture as (Carlini & Wagner, 2017; Zhang et al., 2019). Following (Zhang et al., 2019), the network is trained using SGD with 0.9 momentum for 50 epochs with an initial learning rate 0.01, and the batch size is set to 128. Hyper-parameters used to craft adversarial examples for training are the same as those used for evaluation. These two networks share the same hyper-parameters: we use SGD with 0.9 momentum, weight decay 2×10^{-4}, batch size 128, and an initial learning rate of 0.1. The maximum epoch is 120, and the learning rate is divided by 10 at epoch 60 and 90, respectively. To generate adversarial examples for training, we set the maximal perturbation $\epsilon = 8/255$, the perturbation step size $\eta = 2/255$, and the number of iterations $K = 10$, which is the same as (Rice et al., 2020).

D Experiments of WRN-34-10 on CIFAR10

Table 1: Classification accuracy (%) of WRN-34-10 on CIFAR-10 under the white-box threat model. The best-performance model and the corresponding accuracy are highlighted.

Method	Best checkpoint				Last checkpoint			
	Natural	FGSM	PGD-20	CW-20	Natural	FGSM	PGD-20	CW-20
Mardry	**86.63**	59.48	53.65	53.58	**86.60**	57.07	49.23	49.46
ADA-M	85.24	**61.22**	**55.17**	**55.68**	85.61	**60.08**	**51.76**	**52.59**
TRADES	**84.32**	60.94	56.69	54.87	**84.86**	59.94	52.04	52.39
ADA-T	84.19	**61.62**	**57.36**	**55.75**	84.35	**61.57**	**55.15**	**55.23**

In Table 1, we report the accuracy of WRN-34-10 (Zagoruyko & Komodakis, 2016) of Madry, TRADES, and the proposed method on CIFAR10 against various attacks, i.e., FGSM, PGD, and C&W attacks, which are widely used in the literature. Here, "Natural" denotes the accuracy of natural test images. We denote by PGD-20 the PGD attack with 20 iterations for generating adversarial examples, which also applies to the C&W attack. We can see that the proposed method achieves the best robustness against all three types of attacks, demonstrating that taking into account the spurious correlation can significantly improve the adversarial robustness. Note that the standard deviations of 5 runs are omitted, because they hardly affect the results.

E Ablation study

We implicitly conducted ablation studies when designing Table 1, Table 2, and Table 3. To further understand the comparative effects of different terms of the proposed method, we reorganize the robust accuracy of the best checkpoint trained on CIFAR-10 and CIFAR-100 in Table 2. Comparing Madry, TRADES, and ADA-M, we find that introducing the second (t_2) and the third term (t_3)

Table 2: Robust accuracy (%) of ResNet-18 on CIFAR-10 and CIFAR-100 under the white-box threat model. For simplicity, we use t_1, t_2, and t_3 to represent the first, second, and third terms in Eq. 11, respectively. The best-performance model and the corresponding accuracy are highlighted.

Method	t_1	t_2	t_3	CIFAR-10			CIFAR-100		
				FGSM	PGD-20	CW-20	FGSM	PGD-20	CW-20
Madry	✓			56.69	51.92	51.00	56.69	51.92	51.00
ADA-M	✓		✓	57.98	54.44	52.51	57.98	54.44	52.51
TRADES	✓	✓		57.25	53.64	51.39	57.25	53.64	51.39
ADA-T	✓	✓	✓	**58.97**	**54.55**	**52.95**	**58.97**	**54.55**	**52.95**

can improve the robustness and that the effect of these two terms is close. Similarly, comparing TRADES and ADA-T, we see that introducing the third term (t_3) can further improve the robustness.

F More details about adversarial learning

Recent work on improving adversarial robustness mainly falls into two categories: certified defense and empirical methods.

Certified defense (Raghunathan et al., 2018; Wong & Kolter, 2018; Singla & Feizi, 2020) aims to endow the model with provably adversarial robustness against norm-bounded perturbations. Although the certified defense strategy is promising, the empirical defense (Goodfellow et al., 2014; Madry et al., 2017; Zhang et al., 2019; Wang et al., 2019; Pang et al., 2020; Wong & Kolter, 2018; Xie et al., 2019; Yang et al., 2019), especially the adversarial training method (Goodfellow et al., 2014; Madry et al., 2017; Zhang et al., 2019), is currently the most effective strategy. Empirical defense firstly generates adversarial examples using a certain adversarial attack, then incorporates the generated adversarial examples into the training process.

Recently, various efforts (Najafi et al., 2019; Carmon et al., 2019; Shafahi et al., 2019; Wong et al., 2020; Wang et al., 2019; Pang et al., 2020; Zhang et al., 2020b; Rice et al., 2020) have been devoted to improving adversarial training. One line of work focuses on accelerating the training procedure (Shafahi et al., 2019; Wong et al., 2020). Another line of research (Najafi et al., 2019; Carmon et al., 2019) shows a promising direction that unlabeled training data can significantly mitigate the adversarial vulnerability. Lastly, recent work (Wang et al., 2019; Pang et al., 2020; Zhang et al., 2020b; Rice et al., 2020) provides an interesting direction where these methods rethink the adversarial training from different aspects, containing rethinking the misclassified examples (Wang et al., 2019), rethinking the importance weight of each example (Zhang et al., 2020b) and rethinking the role of normalization (Pang et al., 2020) and basic training strategies (Rice et al., 2020). However, all these methods overlook the spurious correlation between labels and the style information.

Another related work is (Ilyas et al., 2019), which provides an interesting viewpoint, i.e., adversarial examples can be viewed as a human phenomenon because the model's reliance on useful but not robust features leads to adversarial vulnerability. Our work gives a new causal perspective of adversarial vulnerability. Specifically, a) (Ilyas et al., 2019) found some features were useful but not robust, while our work explores the phenomenon's fundamental cause and provides a clear explanation of why some features are useful but not robust: Given X, labels Y are spuriously correlated with the style variables, so fitting the spurious correlation can predict labels. Thus, the style variables can be viewed as 'features'; b) (Ilyas et al., 2019) claimed that adversarial examples could be viewed as a human phenomenon, while our work shows that adversarial examples can be viewed as a model phenomenon rather than merely a human phenomenon. Specifically, the adversarial vulnerability results from fitting the correlation between labels and style variables and failing to fit the causal relations, i.e., DNNs fail to extract content variables.

G More details about causal reasoning

The most relevant work is CAMA (Zhang et al., 2020a) that aims to improve the robustness of DNNs on unseen perturbation via explicitly modeling the perturbation from a causal view. The main difference between our method and CAMA is that we focus on the adversarial vulnerability while CAMA aims to improve the robustness of unseen perturbations. In addition, CAMA assumes a hard intervention on a latent variable. It promotes robustness via modeling the perturbation in the latent space. In this paper, we employ a soft intervention and propose to penalize DNNs when the adversarial distribution is different from the natural distribution. Another related work is RELIC (Mitrovic et al., 2020), a regularizer used in self-supervised learning that uses the independence of mechanisms (Peters et al., 2017) and encourages DNNs to be invariant to different augmentations of the same instance. The self-supervised learning method (Mitrovic et al., 2020) also constructs a causal graph to model the data generation process, but the focus of RELIC is on the content invariant property, overlooking the importance of style information.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pp. 274–283. PMLR, 2018.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. *arXiv preprint arXiv:1905.13736*, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10093–10100, 2020.

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

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of science*, 74(5):981–995, 2007.

Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Ruize Gao, Feng Liu, Jingfeng Zhang, Bo Han, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Maximum mean discrepancy test is aware of adversarial attacks. In *International Conference on Machine Learning*, pp. 3564–3575. PMLR, 2021.

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.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111 (1):3, 2004.

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*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2755–2764, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Andrew Ilyas, Shibani Santurkar, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5580–5590, 2017.

Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524*, 2018.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *ICLR*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*, 2020.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

Amir Najafi, Shin-ichi Maeda, Masanori Koyama, and Takeru Miyato. Robustness to adversarial perturbations in learning from incomplete data. *arXiv preprint arXiv:1905.13021*, 2019.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, and Hang Su. Boosting adversarial training with hypersphere embedding. *arXiv preprint arXiv:2002.08619*, 2020.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. *arXiv preprint arXiv:2101.06046*, 2021.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *arXiv preprint arXiv:1904.12843*, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sahil Singla and Soheil Feizi. Second-order provable defenses against adversarial attacks. In *International Conference on Machine Learning*, pp. 8981–8991. PMLR, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33, 2020.

Takeshi Teshima, Issei Sato, and Masashi Sugiyama. Few-shot domain adaptation by causal mechanism transfer. In *International Conference on Machine Learning*, pp. 9458–9469. PMLR, 2020.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33, 2020.

Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 742–749, 2019.

Cumhur Erkan Tuncali, Georgios Fainekos, Hisahiro Ito, and James Kapinski. Simulation-based adversarial test generation for autonomous vehicles with machine learning components. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1555–1562. IEEE, 2018.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pp. 5286–5295. PMLR, 2018.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *arXiv preprint arXiv:2004.05884*, 2020.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 501–509, 2019.

Yuzhe Yang, Guo Zhang, Dina Katabi, and Zhi Xu. Me-net: Towards effective adversarial robustness with matrix estimation. In *International Conference on Machine Learning*, pp. 7025–7034. PMLR, 2019.

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

Cheng Zhang, Kun Zhang, and Yingzhen Li. A causal view on robustness of neural networks. *Advances in Neural Information Processing Systems*, 33, 2020a.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR, 2019.

Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. *arXiv preprint arXiv:2010.01736*, 2020b.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pp. 819–827. PMLR, 2013.

Yonggang Zhang, Xinmei Tian, Ya Li, Xinchao Wang, and Dacheng Tao. Principal component adversarial example. *IEEE Transactions on Image Processing*, 29:4804–4815, 2020c.