Model-Based Robust Deep Learning: Generalizing to Natural, Out-of-Distribution Data

Alexander Robey [1] **Hamed Hassani** [1] **George J. Pappas** [1]

Abstract

Despite success in many applications, deep learning is known to be fragile to unbounded shifts in the data distribution due to many forms of natural variation, including changes in weather or lighting in images. However, there are remarkably few techniques that consistently improve robustness to natural shifts in the data distribution. To address this gap, we propose a new approach called *model-based robustness*. Critical to our approach is to first use unlabeled data to learn *models of natural variation*, which vary data over a range of natural conditions. We then introduce a novel min-max optimization-based formulation and a family of algorithms which enforce invariance to these learned models of natural variation. Our extensive experiments show that across a variety of natural conditions in twelve distinct datasets, classifiers trained with our algorithms significantly outperform standard data augmentation, domain adaptation, and adversarial training baselines. Specifically, when training on ImageNet and testing on various subsets of ImageNet-c, our algorithms improve over baseline methods by up to 30 percentage points in top-1 accuracy. Further, we show that our methods provide robustness against multiple simultaneous distributional shifts and to domains entirely unseen during training.

The last decade has seen remarkable progress in deep learning (DL), which has prompted integration of DL frameworks into myriad application domains (LeCun et al., 2015), including medical diagnostics and robotics (Melis et al., 2017; Papangelou et al., 2018). Importantly, many of these domains are *safety-critical*, meaning that the detections, recommendations, or decisions made by DL systems can directly impact the well-being of humans (Dreossi et al., 2019). To this end, it has been repeatedly shown that DL is fragile to seemingly innocuous distributional shifts in the input data (Hendrycks & Dietterich, 2019; Szegedy et al., 2013). Indeed, these observations have culminated in a string of very recent papers, which have rigorously studied the fragility of DL to unbounded shifts in the data distribution due to a wide range of naturally-occurring phenomena (Djolonga et al., 2020; Taori et al., 2020; Hendrycks et al., 2020), showing conclusively that the DL lacks robustness to these out-of-distribution shifts. However, despite the pressing nature of this challenge, there are remarkably few principled techniques that have been shown to provide robustness against these forms of out-of-distribution, natural variation (Hendrycks et al., 2019a).

In response to this challenge, in this paper we propose a a principled framework that can be used to train neural networks to generalize well to data that has undergone natural, out-of-distribution shifts. The key idea in our approach is to first learn unsupervised *models of natural variation*, which use unlabeled data to capture natural distribution shifts such as changes in the weather conditions in images. Then, given these models, we propose a family of algorithms, which are designed to *enforce invariance* to the shifts captured by learned models of natural variation. Our experiments show that across a variety of challenging, naturally-occurring conditions, such as variation in lighting, haze, rain, and snow, and across various datasets, including CURE-TSR, ImageNet, and ImageNet-c, classifiers trained with our algorithms significantly outperform standard DL baselines, including empirical risk minimization (ERM), data augmentation, perturbation-based adversarial training, and, when applicable, domain adaptation methods.

Contributions. Our contributions are as follows:

- We formulate a novel robust optimization procedure that leverages models of natural variation to search for challenging shifts in the data distribution.
- We propose a family of efficient algorithms to train classifiers to be invariant with respect to the transformations captured by models of natural variation and accordingly improve the robustness of DL against challenging out-of-distribution shifts.
- We show experimentally that our algorithms consis-

[*]Equal contribution [1]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, Pennsylvania, USA. Correspondence to: Alexander Robey <arobey1@seas.upenn.edu>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.,

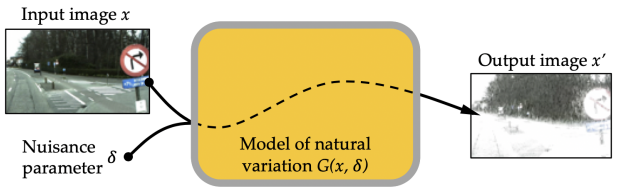

Figure 1. **Models of natural variation.** Throughout this paper, we will use *models of natural variation* to describe a wide variety of natural transformations that data are often subjected to in natural, real-world environments. In our formulation, models of natural variation take the form $G(x, \delta)$, where x is an input datum such as an image and δ is a *nuisance parameter* that characterizes the extent to which the output datum $x' := G(x, \delta)$ is varied.

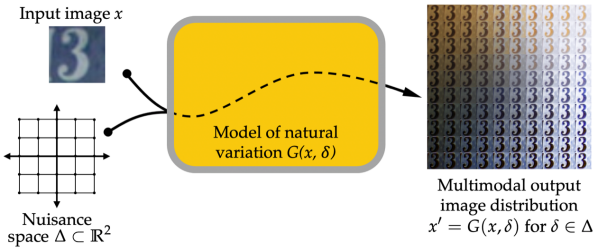

Figure 2. **Capturing the many modes of natural variation.** The nuisance parameter $\delta \in \Delta$ of a model of natural variation is designed to capture the fact that distributional shifts can change images in diverse, multi-modal ways. In this example, by gridding the nuisance space Δ, we show the range of images produced by a learned model of natural variation trained to capture shifts in brightness on SVHN.

tently and uniformly improve robustness to commonly-occuring distributions shifts that frequently occur in real-world environments, including snow, rain, fog, and brightness on numerous datasets, including CURE-TSR, ImageNet, and ImageNet-c.

1. Model-based robustness

Underlying the task of improving the robustness of neural networks to natural, out-of-distribution data are two fundamental challenges. Firstly, unlike in the perturbation-based robustness community, in real-world, safety-critical environments, data can vary in unknown, unbounded, and highly nonlinear ways. Thus, the first step toward building a training procedure that can generalize to out-of-distribution data should be to design a mechanism that accurately describes how data varies in such environments. Next, assuming a suitable model that captures natural variation in real-world environments, the second challenge is to formulate a training procedure that exploits this model toward improving robustness against out-of-distribution shifts. In this section, we present novel solutions to each of these challenges.

1.1. Models of natural variation

In order to effectively model shifts in natural variation in a domain-agnostic setting, we abstractly define *models of natural variation*. Concretely, a model of natural variation $G(x, \delta)$ is a map that describes how an input datum x can be naturally varied by a *nuisance parameter* δ resulting in a new datum $x' = G(x, \delta)$. Ideally, for a fixed datum x, varying the nuisance parameter δ should vary the severity of the natural conditions in the generated datum x'. An example of such a model is shown in Figure 1, where an image x on the left (in this case, in sunny weather) can be naturally varied by δ and consequently transformed into the image x' on the right (in snowy weather). Furthermore, we note that in real-world environments, natural variation can often produce a range of outputs corresponding to different levels of natural variation. For example, in Figure 2, we show the diverse output distribution over shifts in brightness on SVHN that is captured by varying the nuisance parameter.

Learning models of natural variation from data. In many situations, models of natural variation are not known a priori or are too costly to obtain. For example, consider Figure 1 in which a model $G(x, \delta)$ takes an image x of a street sign in sunny weather and maps it to an image $x' := G(x, \delta)$ in snowy weather. Even though there is a relationship between the two images, obtaining a model of natural variation G relating these two domains is extremely challenging if we resort to geometric structure. For such problems we advocate for *learning* an unsupervised model G from data.

Throughout our experiments, we assume that we have access to two unpaired datasets from domains A and B. For example, domain A might contain images in sunny weather and domain B might contain images in snowy weather. To train a model of natural variation from this data, we rely on the Multimodal Unsupervised Image-to-Image Translation (MUNIT) framework (Huang et al., 2018), which combines two autoencoders and two generative adversarial networks (Goodfellow et al., 2014a), to learn unsupervised maps between the image domains A and B. We note that many choices of unconditional image-to-image translation networks satisfy our criteria for G, and in future work we plan to investigate the efficacy of these architectures.

1.2. Model-based robust training formulation

We now describe our model-based training framework. In particular, we consider a setting in which the training data $(x, y) \sim \mathcal{D}$ are distributed according to a joint distribution $\mathcal{D}$ over instances $x \in \mathbb{R}^d$ distributed according to the marginal distribution $\mathbb{P}_A$ and labels $y \in [k] := \{1, 2, \ldots, k\}$. In our framework, we additionally assume that the instances x can be transformed according to a model of natural variation

$G(x, \delta)$ by choosing different values of δ from a given *nuisance space* Δ. The goal of the model-based approach is to train a classifier f_w parameterized by weights $w \in \mathbb{R}^p$ that achieves high accuracy on an out-of-distribution test set drawn from $\mathbb{P}_B$ that has been subjected to the same source of natural variation that $G(x, \delta)$ models. This perspective can be captured by the following robust optimization problem:

$$\min_{w} \mathbb{E}_{(x,y)\sim\mathcal{D}} \Big[\max_{\delta\in\Delta} \ell(G(x, \delta), y; w) \Big]. \tag{1}$$

In this notation, $\ell : \mathbb{R}^d \times [k] \times \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ denotes an appropriately-chosen loss function. The robust optimization problem in (1) comprises two coupled optimization problems: an inner maximization problem and an outer minimization problem. In particular, the inner maximization problem can be written in the following way:

$$\delta^\star \in \arg\max_{\delta\in\Delta} \ell(G(x, \delta), y; w). \tag{2}$$

In this problem, given an instance-label pair (x, y) and a fixed weight $w \in \mathbb{R}^p$, we seek a nuisance parameter $\delta^* \in \Delta$ that produces a corresponding instance $x' := G(x, \delta^*)$ which gives rise to high loss values $\ell(G(x, \delta^*), y; w)$ under the current weight w. One can think of this vector δ^* as characterizing the *most-challenging* distributional shift that can be captured by the model $G(x, \delta^*)$ for the original instance x. After solving this inner problem, we can rewrite the outer minimization problem in the following way:

$$w^\star \in \arg\min_{w\in\mathbb{R}^p} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell(G(x, \delta^\star), y; w)\right] \tag{3}$$

In this outer problem, we seek the weight $w \in \mathbb{R}^p$ that minimizes the risk against challenging instances of the form $G(x, \delta^*)$. By training the neural network to correctly classify these challenging data points, ideally the classifier should become invariant to the model $G(x, \delta)$ for any $\delta \in \Delta$ and consequently to the original source of natural variation.

2. Model-based training algorithms

We now assume that we have access to a suitable model of natural variation $G(x, \delta)$ and shift our attention toward exploiting G in the development of novel robust training algorithms. Following (1), our goal is to design algorithms that search over the nuisance space Δ of a fixed model of natural variation $G(x, \delta)$ to find out-of-distribution instances which are difficult to correctly classify. Considering the geometric perspective introduced in the previous section, the algorithms we propose employ different techniques for searching over the learned image manifolds $B(x)$ to enforce invariance to variation captured by $G(x, \delta)$. To this end, rather than assuming access to the full joint distribution $\mathcal{D}$, we now consider the empirical version of (1) wherein we assume that we are given given a set of i.i.d. samples $\mathcal{D}_n :=$ $\{(x_j, y_j)\}_{j=1}^n$ drawn from $\mathcal{D}$. In this way, the empirical counterpart of (1) can be expressed in the following way:

$$w^\star \in \arg\min_{w\in\mathbb{R}^p} \sum_{j=1}^{n} \Big[\max_{\delta\in\Delta} \ell\left(G\left(x_j, \delta\right), y_j; w\right) \Big]. \tag{4}$$

Note that when w parameterizes a neural network, (4) is a nonconvex-nonconcave min-max problem, which is difficult to solve exactly. Thus, we resort to approximate optimization techniques for solving (4). Specifically, we propose three algorithmic variants: *Model-based Adversarial Training* (MAT), *Model-based Robust Training* (MRT), and *Model-based Data Augmentation* (MDA). Pseudocode for MAT is given in Algorithm 1.

Algorithm 1 Model-based Adversarial Training (MAT)

1: **Input:** Number of steps $k \in \mathbb{Z}_+$, step size $\alpha > 0$
2: **Output:** Learned weight w
3: **repeat**
4: **for** minibatch $\{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \mathcal{D}_n$ **do**
5: $\delta := (\delta_1, \delta_2, \ldots, \delta_m) \leftarrow (0_q, 0_q, \ldots, 0_q)$
6: **for** k steps **do**
7: $g \leftarrow \nabla_\delta \sum_{j=1}^m \ell(G(x_j, \delta_j), y_j; w)$
8: $\delta \leftarrow \Pi_\Delta[\delta + \alpha g]$
9: **end for**
10: $g(w) \leftarrow \sum_{j=1}^m [\ell(G(x_j, \delta_j), y_j; w)$
11: $+\lambda \cdot \ell(x_j, y_j; w)]$
12: $w \leftarrow w - \eta \nabla_w g$
13: **end for**
14: **until** convergence

At a high level, each of these algorithms alternates between solving the outer problem and the inner problem. To this end, each of these algorithms uses SGD to solve the outer problem; the methods differ in how they seek solutions to the inner problem, and in what follows, we describe each of these procedures in more detail.

Model-based Adversarial Training. In MAT, we seek an exact solution to the inner maximization problem by performing k steps of gradient ascent in the nuisance parameter δ on the objective $\ell(G(x, \delta), y; w)$. This procedure is illustrated in lines 5-9 of Algorithm 1. The resulting nuisance parameter $\delta^\star \in \Delta$ found by gradient ascent is one that causes $\ell(G(x, \delta^\star), y; w)$ to have high loss under the current weight w. Among the three algorithms we propose, MAT most-strongly enforces invariance to G.

Model-based Robust Training. In MRT, we consider a sampling-based approach to solving the inner maximization problem. In this way, we first randomly sample $\delta_j \in \Delta$ for $j \in [k]$. We then select the $j^\star \in [k]$ such that $\ell(G(x, \delta_{j^\star}), y; w)$ is maximized. Thus, whereas MAT seeks to exactly solving the inner problem, via a first-order gradient-based scheme, in MRT we use a zeroth-order,

Table 1. **Out-of-distribution robustness.** In each experiment, we train a model to map from challenge-level 0 to challenge-level 2 data from different subsets of CURE-TSR. We then perform model-based training using challenge-level 0 data and test on challenge-levels 3-5.

CURE-TSR subset	Test accuracy (top-1) on levels 3, 4, and 5								
	ERM + Aug			PGD + Aug			MAT		
	3	4	5	3	4	5	3	4	5
Snow	86.5	74.8	60.9	82.9	77.3	61.8	**88.0**	**77.8**	**70.7**
Haze	55.2	54.0	47.5	83.8	63.1	53.4	**83.9**	**79.1**	**70.1**
Decolorization	87.9	85.1	78.8	84.7	75.2	64.9	**90.5**	**89.6**	**89.4**
Rain	72.7	71.7	66.9	68.9	66.4	60.5	**80.7**	**78.7**	**74.8**

Table 2. **ImageNet to ImageNet-c robustness.** In each experiment, we train a model of natural variation to map from classes 0-9 of ImageNet to the same classes from a subset of ImageNet-c. Next, we use this model to perform model-based training on classes 10-59 of ImageNet, and we test each network on classes 10-59 from the same subset ImageNet-c on which the model was trained.

Model dataset (classes 0-9)	Training dataset (classes 10-59)	Test dataset (classes 10-59)	Test accuracy (top-1/top-5)					
			ERM		AugMix		MAT	
Snow	ImageNet	Snow	20.9	49.9	1.10	8.3	**31.1**	**61.2**
Contrast		Contrast	41.1	73.4	0.72	6.76	**50.0**	**79.5**
Brightness		Brightness	26.9	59.2	0.56	5.20	**53.0**	**81.7**
Frost		Frost	16.3	39.0	29.5	58.4	**36.0**	**67.2**

sampling-based approach to finding challenging model-based data $G(x, \delta_{i^\star})$.

Model-based Data Augmentation. In MDA, rather than explicitly trying to solve the inner problem, we seek a *diverse* collection of naturally-varying data rather than "worst-case" model-based variation. In this way, MDA samples $\delta_i \in \Delta$ for $i \in [k]$ and then appends $\{G(x, \delta_i), y\}_{i=1}^k$ to the training dataset. Therefore, this method seeks to enforce invariance to G by enlarging the training set with data transformed WRT different levels of natural variation, such as the variation in brightness shown in Figure 2.

3. Experiments

Out-of-distribution robustness. In many applications, one might have data corresponding to low levels of natural variation, such as a dusting of snow in images of street signs. However, it is often difficult to collect data corresponding to high levels of natural variation, such as images taken during a blizzard. In such cases, we show that our algorithms can be used to provide significant out-of-distribution robustness against data with high levels of natural variation by training on data with relatively low levels of the same source of natural variation. To do so, we use data from the CURE-TSR dataset (Temel et al., 2019), which contains images of street signs divided into subsets according to various sources of natural variation and corresponding severity levels. For example, for images in the "snow" subset, level 0 corresponds to no snow, whereas level 5 corresponds to a full blizzard. Thus, for each row of Table 1, we use unlabeled data from levels 0 and 2 to learn a model of natural variation corresponding to a given source of natural variation in CURE-TSR. We then train classifiers using MRT with labeled level 0 data. We also train classifiers using ERM and PGD using the labeled data from levels 0 and 2. We then test all classifiers on data from levels 3-5. *Note that while this is an unfair comparison for our methods, given that the model-based algorithms are not given access to labeled level 2 data, our algorithms still outperform the baselines by as much as 20 percentage points on level 5 data.*

Robustness on the shift from ImageNet to ImageNet-c. To demonstrate the scalability of our approach, we perform experiments on ImageNet (Deng et al., 2009) and the recently-curated ImageNet-c dataset (Hendrycks & Dietterich, 2019). ImageNet-c contains images from the ImageNet test set that are corrupted according to artificial transformations, such as snow, rain, and fog, and are labeled from 1-5 depending on the severity of the corruption. For numerous challenging corruptions, we train models to map from the classes 0-9 of ImageNet to the corresponding classes of ImageNet-c. We then train all networks on classes 10-59 of ImageNet, and test on the corresponding classes for various subsets of ImageNet-c. Note that in this setting, the ImageNet classes used to train the model of natural variation are disjoint from those that are used to train the classifier, so many techniques, including most domain adaptation methods, do not apply; to offer a point of comparison, we include the baseline AugMix (Hendrycks et al., 2019a), which performs data augmentation using fixed transformations such as rotations and scalings. *In all of the settings we considered,*

our algorithms significantly outperformed both baselines.

References

Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., and Marchand, M. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

Albuquerque, I., Monteiro, J., Falk, T. H., and Mitliagkas, I. Adversarial target-invariant representation learning for domain generalization. *arXiv preprint arXiv:1911.00804*, 2019.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Blanchard, G., Lee, G., and Scott, C. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in neural information processing systems*, pp. 2178–2186, 2011.

Blanchard, G., Deshmukh, A. A., Dogan, U., Lee, G., and Scott, C. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017.

Daumé III, H. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Djolonga, J., Yung, J., Tschannen, M., Romijnders, R., Beyer, L., Kolesnikov, A., Puigcerver, J., Minderer, M., D'Amour, A., Moldovan, D., et al. On robustness and transferability of convolutional neural networks. *arXiv preprint arXiv:2007.08558*, 2020.

Dobriban, E., Hassani, H., Hong, D., and Robey, A. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

Dreossi, T., Donzé, A., and Seshia, S. A. Compositional falsification of cyber-physical systems with machine learning components. *Journal of Automated Reasoning*, 63(4): 1031–1053, 2019.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1625–1634, 2018.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014a.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

Gowal, S., Qin, C., Huang, P.-S., Cemgil, T., Dvijotham, K., Mann, T., and Kohli, P. Achieving robustness in the wild via adversarial mixing with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1211–1220, 2020.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019b.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., and Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pp. 1989–1998. PMLR, 2018.

Hosseini, H. and Poovendran, R. Semantic adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1614–1619, 2018.

Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172–189, 2018.

Jalal, A., Ilyas, A., Daskalakis, C., and Dimakis, A. G. The robust manifold defense: Adversarial training using generative models. *arXiv preprint arXiv:1712.09196*, 2017.

Javanmard, A., Soltanolkotabi, M., and Hassani, H. Precise tradeoffs in adversarial training for linear regression. *arXiv preprint arXiv:2002.10477*, 2020.

Karianakis, N., Dong, J., and Soatto, S. An empirical evaluation of current convolutional architectures' ability to manage nuisance location and scale variability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4442–4451, 2016.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015.

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., and Tao, D. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Melis, M., Demontis, A., Biggio, B., Brown, G., Fumera, G., and Roli, F. Is deep learning safe for robot vision? adversarial examples against the icub humanoid. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 751–759, 2017.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18, 2013.

Papangelou, K., Sechidis, K., Weatherall, J., and Brown, G. Toward an understanding of adversarial examples in clinical trials. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–51. Springer, 2018.

Pei, Z., Cao, Z., Long, M., and Wang, J. Multi-adversarial domain adaptation. *arXiv preprint arXiv:1809.02176*, 2018.

Robey, A., Hassani, H., and Pappas, G. J. Model-based robust deep learning. *arXiv preprint arXiv:2005.10247*, 2020.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. *arXiv preprint arXiv:1712.02560*, 2017.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pp. 5014–5026, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *arXiv preprint arXiv:2007.00644*, 2020.

Temel, D., Chen, M.-H., and AlRegib, G. Traffic sign detection under challenging conditions: A deeper look into performance variations and spectral characteristics. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

Tramer, F., Carlini, N., Brendel, W., and Madry, A. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

Wong, E. and Kolter, J. Z. Provable Defenses Against Adversarial Examples Via the Convex Outer Adversarial Polytope. *arXiv preprint arXiv:1711.00851*, 2017.

Wong, E. and Kolter, J. Z. Learning perturbation sets for robust machine learning. *arXiv preprint arXiv:2007.08450*, 2020.

Xiao, C., Zhu, J.-Y., Li, B., He, W., Liu, M., and Song, D. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018.

For more details, please see our arXiv paper: (Robey et al., 2020).

A. Related work

Fragility of DL to out-of-distribution shifts. Deep learning is known to be fragile to many forms of *natural variation* in data (Hendrycks & Dietterich, 2019; Djolonga et al., 2020; Taori et al., 2020; Hendrycks et al., 2020). In image classification, a growing literature has conclusively demonstrated that DL is fragile to shifts such as changes in weather or background color (Eykholt et al., 2018; Hendrycks et al., 2019b; Hosseini & Poovendran, 2018), rotation and scaling (Xiao et al., 2018; Karianakis et al., 2016), and sensor-based attacks (Kurakin et al., 2016). Because such transformations frequently arise in safety-critical domains, it is critically important for the DL community to develop algorithms that generalize to natural out-of-distributions data. In this paper, we specifically address this challenge by proposing a principled, optimization-based framework which can be used to improve robustness to arbitrary forms of natural variation.

To this end, two concurrent works have sought to address out-of-distribution robustness under the assumption that data is corrupted according to a fixed generative architecture. Gowal et al. (2020) exploit properties specific to the StyleGAN architecture to formulate a training algorithm that provides robustness against color-based shifts on MNIST and CelebA. In our work, we propose a more general framework and three novel robust training algorithms that can exploit any suitable generative network, and we show improvements on more challenging, naturally-occurring shifts across twelve distinct datasets. Wong & Kolter (2020) use conditional VAEs to learn perturbation sets corresponding to simple corruptions from pairs of images. In our framework we improve robustness against more challenging, natural shifts by learning from *unpaired* datasets and we do not rely on class-conditioning to generate realistic images.

Perturbation-based robustness. Recently, there has been particular interest in the DL community surrounding robustness to artificial, norm-bounded, adversarially-chosen perturbations of the input data (Goodfellow et al., 2014b; Szegedy et al., 2013). This problem has motivated an arms-race-like amalgamation of proposed adversarial attacks and defenses within the scope of norm-bounded perturbations (Athalye et al., 2018; Tramer et al., 2020) and has prompted researchers to closely study the theoretical properties of adversarial robustness (Dobriban et al., 2020; Schmidt et al., 2018; Jalal et al., 2017; Tsipras et al., 2018; Javanmard et al., 2020). In particular, the dominant paradigm toward improving robustness against norm-bounded, additive perturbations relies on a robust optimization perspective wherein neural networks are trained to correctly classify worst-case perturbations of data (Madry et al., 2017; Wong & Kolter, 2017). In this paper, while we also employ a robust optimization perspective in our formulation, we study the separate problem of out-of-distribution generalization to *naturally-occurring* distributional shifts.

Domain adaptation and domain generalization. In the domain adaptation literature, various methods have been proposed which rely on the assumption that unlabeled data corresponding to a fixed distributional shift is available during training (Ajakan et al., 2014; Ganin & Lempitsky, 2015; Saito et al., 2017; Daumé III, 2009). Several works in this vein use an adversarial formulation to adapt the feature representations of classifiers trained on a source domain to perform well on a related target domain (Tzeng et al., 2017; Hoffman et al., 2018; Long et al., 2018; Pei et al., 2018). We note that the main difference between domain adaptation techniques and our framework is that our solution does not assume access to unlabeled data from a fixed shift and can be applied to datasets that are entirely unseen during training. Moreover, rather than adapting the feature space of classifier to perform well on a new domain, we directly attack the end-to-end problem of training classifiers to generalize to multiple arbitrary out-of-distribution shifts.

Also related is the field of domain generalization (Blanchard et al., 2011; Muandet et al., 2013; Li et al., 2018), in which one assumes access to a variety of training domains, all of which are related to an unseen target domain on which the trained classifier is ultimately evaluated. Such works often rely on transfer learning (Blanchard et al., 2017) and feature space alignment (Albuquerque et al., 2019) to improve classification accuracy on the unseen domain. While some of our experiments tackle a similar problem, in which knowledge from one domain is used to learn a classifier that performs well on an unseen test domain, the experiments in the other subsections of our experiments show that our model-based paradigm is much more broadly applicable to settings that fall outside of the scope of domain generalization.

B. Further experiments

B.1. Robustness to simultaneous distributional shifts

In practice, it is common to encounter multiple simultaneous distributional shifts. For example, in image classification, there may be shifts in both brightness and contrast; yet while there may be examples corresponding to shifts in either brightness or contrast in the data, there may not be any examples of both shifts occurring simultaneously. To address this robustness challenge, for each row of Table 3, we learn two models of natural variation G_1 and G_2 using unlabeled training data corresponding to two separate shifts, which map domains $A_1 \to B_1$ (e.g. low- to high-brightness) and $A_2 \to B_2$

Table 3. **Composing models of natural variation.** We consider shifts in two distinct and simultaneous sources of natural variation. To perform model-based training, we compose two models of natural variation trained separately on each of the two corruptions.

Dataset	Challenge 1 (dom. A_1→dom. B_1)	Challenge 2 (dom. A_2→dom. B_2)	Test acc. (top-1)	
			ERM	MDA
SVHN	Brightness (low→high)	Contrast (low→high)	54.9	**67.2**
ImageNet	IN-c brightness (low→high)	IN-c contrast (high→low)	13.6	**49.9**
ImageNet	IN-c brightness (low→high)	IN-c fog (no→yes)	50.3	**58.8**
ImageNet	IN-c contrast (high→low)	IN-c fog (no→yes)	8.40	**23.2**

Table 4. **Transferability of model-based robustness.** In each experiment, we train a model of natural variation on a given training dataset $\mathcal{D}_1$. Then, we use this model to train on a new dataset $\mathcal{D}_2$ entirely unseen during the training of the model.

Training dataset $\mathcal{D}_1$	Test dataset $\mathcal{D}_2$	Challenge (dom. A→dom. B)	Test accuracy (top-1)				
			ERM	PGD	MRT	MDA	MAT
MNIST	Fashion-MNIST	Background color (blue→red)	69.3	67.7	**81.4**	80.1	76.1
	Q-MNIST		87.0	79.9	**98.0**	**98.0**	**98.0**
	E-MNIST		63.5	49.3	**86.1**	85.9	84.1
	K-MNIST		47.9	47.7	89.1	**89.3**	86.8
	USPS		89.9	87.4	93.3	**93.4**	91.9
GTSRB	CURE	Brightness (high→low)	47.6	43.6	**73.0**	72.4	67.8
ImageNet & ImageNet-c	CURE	Snow (no→yes)	52.0	53.0	59.4	**62.2**	59.4
		Brightness (low→high)	41.5	40.2	46.6	46.7	**47.5**

(e.g. low- to high-contrast). We then compose these models to form a new model $G(x, \delta) = G_1(G_2(x, \delta), \delta)$ which can be used to provide robustness against both shifts simultaneously. We then train classifiers on labeled data from $A_1 \cup A_2$ and test on data from $B_1 \cap B_2$. To create the data from $B_1 \cap B_2$ for the ImageNet experiments, we apply pairs of transformations that were originally used to create the ImageNet-c datasets; more details are in the appendix. *In each row of Table 3, MDA outperformed baseline methods by as much as 35 percentage points.*

B.2. Transferability of model-based robustness

Because we learn models of natural variation offline before training a classifier, our paradigm can be applied to domains that are *entirely unseen* while training the model. In particular, we show that models can be reused on similar yet unseen datasets to provide robustness against a common source of natural variation. For example, one might have access to two domains corresponding to the shift from images of European street signs taken during the day to images taken at night. However, one might wish to provide robustness against the same shift from daytime to nighttime on a new dataset of American street signs without access to any nighttime images in this new dataset. Whereas many techniques, including most domain adaptation methods, do not apply in this scenario, in the MBRDL paradigm, we can simply learn a model corresponding to the changes in lighting for the European street signs and then apply this model to the dataset of the American signs. Table 4 shows several experiments of this stripe in which a model G is learned on one dataset $\mathcal{D}_1$ and then applied on another $\mathcal{D}_2$; *we improve robustness on unseen domains by up to 40 percentage points.*

B.3. Unsupervised domain adaptation

While our approach does not require labeled data from domain B, when such data is available, it is of interest to evaluate how our approach compares to relevant methods such as domain adaptation. In Table 5, for each shift from domain A to B, we assume access to labeled data from domain A and unlabeled data from domain B. In each row, we use unlabeled data from both domains to train a model of natural variation. We then train classifiers using our algorithms, as well with ERM and PGD, using data from domain A and test on data from the test set for domain B. Furthermore, we compare to ADDA, which is a well-known domain adaptation method (Tzeng et al., 2017). *In every scenario, our model-based algorithms significantly outperform the baselines, often by 10-20 percentage points.* Note that while this is one of the most commonly studied settings in domain adaptation, it represents only one particular setting to which our framework can be applied.

Table 5. **Unsupervised domain adaptation.** In each experiment, we assume access to unlabeled data from domain B, which we use to train a model of natural variation. We compare to suitable baselines, including domain adaptation.

Dataset	Challenge (dom. A→dom. B)	Test accuracy (top-1)					
		ERM	PGD	ADDA	MRT	MDA	MAT
SVHN	Brightness (low→high)	30.5	36.2	60.1	**70.9**	69.5	52.2
SVHN	Contrast (low→high)	55.9	57.9	54.6	**74.3**	74.1	55.2
GTSRB	Brightness (low→high)	40.3	34.7	27.6	50.4	48.3	**64.8**
GTSRB	Contrast (low→high)	44.5	41.9	14.7	68.4	**69.4**	55.1
CURE	Snow (no→yes)	52.0	53.0	16.1	74.0	**74.5**	72.3
CURE	Haze (no→yes)	57.2	50.9	49.2	72.5	70.0	**74.6**
CURE	Rain (no→yes)	62.6	62.3	16.5	75.2	73.7	**75.3**