LEARNING UNFORESEEN ROBUSTNESS FROM OUT-OF-DISTRIBUTION DATA USING EQUIVARIANT DOMAIN TRANSLATOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing approaches to training robust models are typically tailored to scenarios where data variations are available in the training set. While shown effective in achieving robustness to these foreseen variations, these approaches are ineffective in learning *unforeseen* robustness, i.e., robustness to data variations with unknown characterization or without training examples reflecting them. In this work, we learn such unforeseen robustness by harnessing the variations in the abundant out-of-distribution data. As we attribute the main challenge of using these data to the domain gap, we consider using a domain translator to bridge the gap, with which we bound the intractable robustness on the target distribution. As implied by our analysis, we propose a two-step algorithm that first trains an equivariant domain translator to map out-of-distribution data to the target distribution while preserving the variation, and then regularizes a model's output consistency on the domain-translated data to improve its robustness. We empirically demonstrate the effectiveness of our method in improving both unforeseen and foreseen robustness in comparison to existing baselines. We also show that training the equivariant domain translator serves as an effective criterion for source data selection.

1 INTRODUCTION

A desirable property that trustworthy machine learning systems should have is the *robustness* to certain data variations. For example, an object classifier's prediction should be consistent under data variations that preserve the object's label, such as viewpoint changes. Despite the importance, training a model robust to *unforeseen* data variations is challenging. Prior work in training robust models is typically tailored to the scenarios where the considered data variation is foreseen, i.e., either some known transformation function characterizes it or there are pairs of training examples before and after the change to estimate it. While this is true for a few synthetic transformations, such as noise corruption (Hendrycks & Dietterich, 2019) or spatial transformations (Engstrom et al., 2019), it is rarely the case for natural variations, such as viewpoint changes (Koh et al., 2021) or temporal changes (Shankar et al., 2021), resulting in models that are robust to a limited set of data variations.

As illustrated in Figure 1, a certain type of data variation (3D viewpoint change), while unforeseen from a given training set (CIFAR-10, depicted in blue), manifests itself as pairs of transformed examples in the abundant out-of-distribution data (Objectron, a set of video clips showing viewpoint changes, depicted in orange). Based on this observation, this work proposes a new approach to learning unforeseen robustness from out-of-distribution data.

Contributions: *First*, we formulate the problem (§2) and identify the challenges in extending existing approaches to our setting: (1) Model-based data augmentation methods (MBRDL Antoniou et al. (2017); Robey et al. (2020); Zhou et al. (2022)) learn a generative model to capture the variation on source data and then apply it to augment target data. However, the generative model generalizes poorly when the domain gap is large and encounters intrinsic challenges in modeling data variations with multi-modal distributions (Salmona et al., 2022). (2) The semi-supervised consistency regularization(UDA Xie et al. (2020); Sohn et al. (2020)) directly learns robustness from source data, circumventing the challenge of modeling data variations. Nevertheless, the robustness learned from

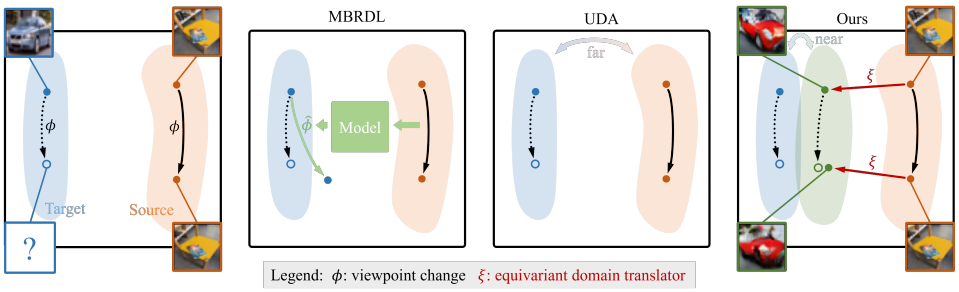


Figure 1: Illustration of our method and two existing methods extended to our setting. The arrowed line denoting the variation ϕ is solid if the variation is foreseen and dashed if otherwise. Our method trains an equivariant domain translator to translate source data to resemble the target while preserving the variation, and then learns the robustness from the translated data (depicted in green). The shown pair of images outlined in green is generated by our trained domain translator.

the source data does not necessarily generalize to the target, resulting in models with sub-optimal robustness in the presence of domain gaps.

Second, identifying the domain gap as the primary cause of the underperformance of previous methods, we analyze the problem with an auxiliary domain translator bridging the gap (§2.1). Given any domain translator, i.e., a mapping on the input space, we provide an upper-bound of the robustness loss on the target distribution which benefits from a domain translator that is both *equivariant* — data-transforming an example first and then domain-translating it gives a similar output as domain-translating the example first and then data-transforming it, and *accurate* — the domain-translated source distribution has a low Wasserstein-1 distance to the target distribution.

Third, as implied by our analysis, we propose a two-step method (§3): (1) training a domain translator. To make it accurate, we train it under the supervision of a Lipschitz-regularized domain discriminator, following WGAN (Arjovsky et al., 2017). To make it equivariant, we encourage a learnable feature extractor to extract the same transformation information from the transformed source example pairs before and after domain translation — a new heuristic method with clear intuition if we hard-engineer the feature extractor (e.g., using an optical flow estimator). (2) Using consistency regularization on the domain-translated source data to improve a model's robustness.

Fourth, we empirically evaluate our method for image classification tasks on a combination of seven source datasets, two target datasets, and two types of data variations (§4). We first verify that our method indeed learns equivariant and accurate domain translators. Then, we show the effectiveness of our method in learning unforeseen robustness compared to other baselines, and further support it by ablation studies. As a by-product, we also show that the training result of the equivariant domain translator correlates strongly (R=0.91) with the robustness benefit of a certain source dataset, indicating its usefulness as a source dataset selection criterion.

Fifth, we demonstrate the practical importance of our method by applying it to two real-world tasks. First, we learn the unforeseen 3D viewpoint change robustness on CIFAR-10 and show the improved robustness using some proxy geometric transformations. Second, we leverage out-of-distribution data to further improve the foreseen robustness on the target, achieving better robustness, in-distribution generalization, and out-of-distribution generalization.

2 Problem Analysis: Robustness from Variations on Source

This section formulates and analyzes the problem of learning unforeseen robustness from out-of-distribution data. In this problem, we are given some target examples $\{\boldsymbol{x}_i\}$ sampled from the *target data distribution* $\mathbb{P}$ on the input space $\mathcal{X}$. We consider $\mathcal{X}$ to be $\mathbb{R}^d$. In addition, we are given some source examples $\{\boldsymbol{u}_i\}$ sampled from the *source data distribution* $\mathbb{Q}$ on $\mathcal{X}$. We do learning over a family of models $\{f : \mathcal{X} \to \mathbb{R}^k\}$ which map examples in $\mathcal{X}$ to k-dimensional output vectors.

Data variation. We consider data variations that can be represented by some (possibly unknown) data transformation function $\phi : \mathcal{T} \times \mathcal{X} \to \mathcal{X}$, where $\mathcal{T}$ is the space of transformation parameters. Some

examples are group actions with $\mathcal{T}$ being some group, noise corruption, and 3D viewpoint change projected to the 2D pixel space (given that ϕ models the stochasticity). As we focus on random data transformation, we also consider some transformation parameter distribution $\mathbb{T}$ on $\mathcal{T}$. We assume that the data variation is unforeseen, i.e., we neither know the explicit data transformation function nor have transformed target example pairs $\{(\boldsymbol{x}_i, \phi_{\boldsymbol{t}_i}(\boldsymbol{x}_i))\}$. Instead, given the source examples $\{\boldsymbol{u}_i\}$, we have finite (e.g., variations extracted from a video clip) or infinite (e.g., synthetic data or transformations) transformed versions $\{\phi_{\boldsymbol{t}_{ij}}(\boldsymbol{u}_i)\}$, where $\boldsymbol{t}_{ij}$ is sampled from $\mathbb{T}$.

Robustness. We consider model robustness to random data transformations. We measure the consistency of two model outputs using some loss function $\ell : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_{\geq 0}$ that satisfies the triangle inequality $\ell(\boldsymbol{v}, \boldsymbol{v}'') \leq \ell(\boldsymbol{v}, \boldsymbol{v}') + \ell(\boldsymbol{v}', \boldsymbol{v}''), \forall \boldsymbol{v} \in \mathbb{R}^k$. Examples of such loss functions include zero-one loss $\ell_{0\text{-}1}(\boldsymbol{v}, \boldsymbol{v}') = \mathbf{1}\{\arg\max_i \boldsymbol{v}_i \neq \arg\max_i \boldsymbol{v}'_i\}$, ℓ_p loss $\ell_p(\boldsymbol{v}, \boldsymbol{v}') = \|\boldsymbol{v} - \boldsymbol{v}'\|_p$ for some $p \geq 1$, and some f-divergences such as the square root of JS-divergence (Endres & Schindelin, 2003). Given such a loss function, we define the following robustness loss.

Definition 2.1 (Robustness loss)**.** *Let ϕ be some transformation function and $\mathbb{T}$ be the distribution of transformation parameters. Then the robustness loss of a model f on the data distribution $\mathbb{P}$ is defined as $L_\phi(f, \mathbb{P}) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}, \boldsymbol{t}\sim\mathbb{T}}\left[\ell\big(f(\boldsymbol{x}), f(\phi_{\boldsymbol{t}}(\boldsymbol{x}))\big)\right]$.*

Note that the robustness loss is label-agnostic, making it well-defined on domains with different label sets. Similar robustness notions also appear in Hendrycks & Dietterich (2019) and Zhou et al. (2022).

Goal. Given target $\{\boldsymbol{x}_i\}$ and source examples $\{\boldsymbol{u}_i\}$ with their transformed versions $\{\phi_{\boldsymbol{t}_j}(\boldsymbol{u}_i)\}$, our goal is to learn a model f that minimizes the robustness loss on the target distribution $L_\phi(f, \mathbb{P})$ and some other given loss defining the primary task. For classification tasks, the significance of minimizing the robustness loss is that small robustness loss and small classification loss $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}, \boldsymbol{t}\sim\mathbb{T}}\left[\ell_{0\text{-}1}(\boldsymbol{y}, f(\boldsymbol{x}))\right]$ are sufficient to guarantee small robust classification loss $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}, \boldsymbol{t}\sim\mathbb{T}}\left[\ell_{0\text{-}1}(\boldsymbol{y}, f(\phi_{\boldsymbol{t}}(\boldsymbol{x})))\right]$, where $\ell_{0\text{-}1}$ is the zero-one loss and $\boldsymbol{y}$ is the ground-truth label of $\boldsymbol{x}$.

2.1 Bounding Robustness with Domain Translator

This section suggests using transformed source example pairs to improve the unforeseen robustness on the target distribution, given that direct optimization is infeasible. We use $\bar{\ell}_f : \mathcal{X} \to \mathbb{R}$ to denote the function $\bar{\ell}_f(\boldsymbol{x}) := \mathbb{E}_{t\sim\mathbb{T}}[\ell\big(f(\boldsymbol{x}), f(\phi_{\boldsymbol{t}}(\boldsymbol{x}))\big)]$, which intuitively measures the robustness loss of the model at a given example. Given some (measurable) function $\xi : \mathcal{X} \to \mathcal{X}$, we use $\xi_\#\mathbb{Q}$ to denote the push-forward probability distributionof $\mathbb{Q}$ on $\mathcal{X}$. We use W_1 to denote Wasserstein-1 distance. The proposition below, proved in Appendix B.1, upper-bounds the robustness loss on the target distribution by three terms illustrated in Figure 2.

Proposition 2.2. *We assume that $\bar{\ell}_f$ is Lipschitz uniformly over all models f, with a (possibly infinite) Lipschitz constant $\|\bar{\ell}\|_L$. Then for any (measurable) function $\xi : \mathcal{X} \to \mathcal{X}$, the following holds:*

$$L_\phi(f, \mathbb{P}) \leq I_1 + I_2 + I_3, \tag{2.1}$$

where $I_1 = \mathbb{E}_{u\sim\mathbb{Q}, t\sim\mathbb{T}}\left[\ell\big(f(\xi(\boldsymbol{u})), f(\xi \circ \phi_{\boldsymbol{t}}(\boldsymbol{u}))\big)\right]$,
$I_2 = \mathbb{E}_{u\sim\mathbb{Q}, t\sim\mathbb{T}}\left[\ell\big(f(\xi \circ \phi_{\boldsymbol{t}}(\boldsymbol{u})), f(\phi_{\boldsymbol{t}} \circ \xi(\boldsymbol{u}))\big)\right]$, $\quad I_3 = \|\bar{\ell}\|_L W_1(\mathbb{P}, \xi_\#\mathbb{Q})$.

We can intuitively interpret ξ as a domain translator which translates a given source example into another example that resembles target examples. I_1 measures the model's consistency loss on the domain-translated example pairs, which is estimable using source examples. I_2 measures the model's consistency loss on the ground-truth transformed example and its approximated version from the domain translator, which can be minimized if the domain translator is equivariant. I_3 measures how well the push-forward distribution approximates the target distribution.

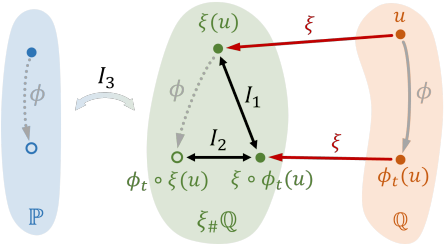


Figure 2: Illustration of our proposition.

Compared to style transfer, this domain translation is un-paired and does not need to preserve the underlying concept class. Below, we remark on two properties of the domain translator.

Equivariant domain translator minimizes I_2. Note that any domain translator ξ satisfying $\xi \circ \phi_t(\boldsymbol{u}) = \phi_t \circ \xi(\boldsymbol{u})$ (almost surely with respect to $\mathbb{P} \times \mathbb{T}$) is sufficient to minimize the term I_2 for *any* model f (assuming $\bar{\ell}_f$ has bounded range). Such ξ is said to be *equivariant* if t belongs to a group with ϕ_t being the group action. Nevertheless, we abuse the notion and refer to any ξ approximately satisfying this property (measured by some loss) as being equivariant.

Accurate domain translator minimizes I_3. Note that any domain translator ξ pushing the source distribution to match the target distribution accurately such that $W_1(\mathbb{P}, \xi_\#\mathbb{Q}) = 0$ is sufficient to minimize the term I_2 to zero for *any* model f (assuming bounded $\|\bar{\ell}\|_{\mathrm{L}}$). We refer to any ξ approximately satisfying this property as being accurate.

The above two remarks imply that we can learn an equivariant and accurate domain translator to minimize I_2 and I_3 independent of the model f, which motivates our two-step algorithm in the next section. We empirically demonstrate the existence of such domain translators for certain datasets and leave further existence discussion to Appendix B.2.

3 The Two-Step Algorithm for Learning Unforeseen Robustness

Step one: training equivariant domain translator. As a simpler case, we first propose the training objective of the equivariant domain translator when we know the transformation function characterizing the considered data variation (e.g., in the semi-supervised data augmentation setting):

$$\min_{\xi} \quad W_1(\mathbb{P}, \xi_\#\mathbb{Q}) + \lambda \mathop{\mathbb{E}}_{\boldsymbol{u}\sim\mathbb{Q}} \mathop{\mathbb{E}}_{\boldsymbol{t}\sim\mathbb{T}} [\ell\big(\xi \circ \phi_{\boldsymbol{t}}(\boldsymbol{u}), \phi_{\boldsymbol{t}} \circ \xi(\boldsymbol{u})\big)], \tag{3.1}$$

where the first term minimizes I_3, encouraging accurate domain translation, and the second term minimizes I_2, promoting equivariance. λ balances these two objectives.

To optimize the first term, we follow WGAN (Arjovsky et al., 2017) and train the domain translator ξ under the supervision of an auxiliary domain discriminator that has regularized Lipschitz constant. To estimate and optimize the second term, we sample one transformation parameter for each source example, and then do domain translation followed by transformation to get $\phi_{\boldsymbol{t}} \circ \xi(\boldsymbol{u})$, and transformation followed by domain translation to get $\xi \circ \phi_{\boldsymbol{t}}(\boldsymbol{u})$. We encourage the domain translator to generate examples such that the two terms are similar according to some loss such as ℓ_2. We use the encoder-decoder architecture from the style transfer literature to implement the domain translator.

Heuristics for learning equivariance. When learning unforeseen robustness, we only have some transformed source example pairs $\{(\boldsymbol{u}_i, \phi_{\boldsymbol{t}_i}(\boldsymbol{u}_i))\}$ without knowing the underlying data transformation function ϕ. This poses a challenge to learning equivariant domain translator ξ since we cannot transform a domain-translated example $\xi(\boldsymbol{u})$ to get $\phi_{\boldsymbol{t}} \circ \xi(\boldsymbol{u})$ in Eq. 3.1. Nevertheless, some work shows that given the transformed example pairs $\{(\boldsymbol{u}_i, \phi_{\boldsymbol{t}_i}(\boldsymbol{u}_i))\}$ and the corresponding transformation parameters $\{\boldsymbol{t}_i\}$, we can empirically encourage a model to be equivariant to the transformation $\phi_{\boldsymbol{t}}$ by predicting the transformation parameters $\{\boldsymbol{t}_i\}$ (Lenc & Vedaldi, 2019; Qi et al., 2019; Dangovski et al., 2022). Since the transformation parameters may be unknown for unforeseen variations, we propose a new heuristic method to encourage the equivariance, requiring only the transformed source example pairs and their domain-translated counterparts.

Figure 3 illustrates the method. The projector, whose architecture refers to Qi et al. (2019), inputs the original example $\boldsymbol{u}$ and its transformed version $\phi_{\boldsymbol{t}}(\boldsymbol{u})$ and outputs a vector $\boldsymbol{z}_1$. Intuition is that $\boldsymbol{z}_1$ may contain the encoded transformation parameter, which is particularly true when the projector is a hard-coded model like an optical flow estimator. An equivariant domain translator should output the domain-translated pair $\xi(\boldsymbol{u})$ and $\xi(\phi_{\boldsymbol{t}}(\boldsymbol{u}))$ that contain the same encoded transformation parameter. Thus, we encourage z_1 and z_2 to be similar, which is implemented with a predictor to prevent degeneration (referring to SimSiam (Chen & He, 2021)).

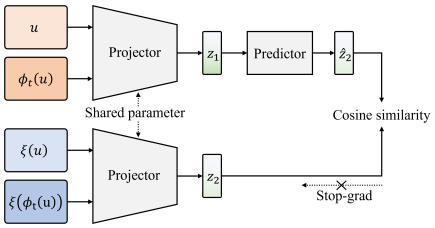


Figure 3: Our proposed heuristic method for encouraging the equivariance.

To train the domain translator, we substitute the second term in Eq. 3.1 with the cosine similarity term shown in the figure, and optimize Eq. 3.1 jointly for the domain translator, projector, and predictor.

Table 1: Results of classifiers trained using different methods and source datasets. The target dataset is CIFAR-10 and the data variation is RandAugment. The oracle method does consistency regularization directly on the target dataset.

		Robustness		Accuracy
Method	**Src**	**RC** (%)	**R** (%)	**S** (%)
ERM	/	79.1 ± 0.2	82.5 ± 0.2	89.0 ± 0.2
MBRDL	SVHN	68.7 ± 0.4	77.4 ± 0.3	78.9 ± 0.3
UDA	SVHN	82.3 ± 0.2	85.5 ± 0.3	88.2 ± 0.3
Ours	SVHN	**83.2** ± 0.3	**86.7** ± 0.3	**89.9** ± 0.2
MBRDL	STL10	72.1 ± 0.4	78.8 ± 0.3	82.9 ± 0.3
UDA	STL10	85.8 ± 0.3	89.5 ± 0.2	89.9 ± 0.3
Ours	STL10	**87.8** ± 0.2	**91.5** ± 0.3	**91.0** ± 0.3
Oracle	/	91.7 ± 0.1	94.8 ± 0.2	93.3 ± 0.1

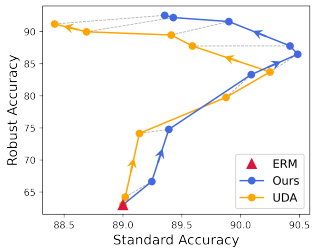


Figure 4: Robust vs. standard accuracy of classifiers trained with different consistency regularization weights. We gradually increase (denoted by the arrow) the weight from 0 to 5, producing different classifiers whose results are denoted by dots. The pair of dots connected by a dashed line have the same weight setting.

Step two: learning robust model. Our goal is to improve the robustness while doing some primary task. As an example, we consider the classification task with some given classification loss $L_{\text{classifier}}$. Based on Proposition 2.2, with the trained domain translator minimizing I_2 and I_3, we proceed to learn a robust classifier f that minimizes I_1 and $L_{\text{classifier}}$ while keeping the translator frozen.

For notation simplicity, we write I_1 as a functional of f and ξ. We use ξ^* to denote the trained domain translator, and use ξ_{id} to denote the identity domain translator which maps any example to itself (perfectly equivariant but not accurate). Then, the training objective is

$$\min_f \quad L_{\text{classifier}}(f) + \lambda_1 I_1(f, \xi^*) + \lambda_2 I_1(f, \xi_{\text{id}}), \tag{3.2}$$

where λ_1 and λ_2 are weight hyperparameters. We include the last term, which is essentially consistency regularization on source data, since we observe that additionally optimizing sometimes yields the best result. Note that UDA can be viewed as a special case of our method ($\lambda_1 = 0$, $\lambda_2 = 1$).

4 Empirical Evaluation

This section empirically evaluates our method's effectiveness to learn unforeseen robustness in image classification tasks, using two target datasets, six source datasets, and two data variations. Due to space limitations, we defer experimental details to Appendix E and more results to Appendix C.

Training equivariant domain translator. We first show that our heuristic method effectively learns accurate and equivariant domain translators. We compare three methods in training domain translators: (1) *Standard* (Std) which does not encourage equivariance ($\lambda = 0$); (2) *Equivariant-Groundtruth* (EqGt) which encourages equivariance using the groundtruth data transformation function; (3) *Equivariant-Heuristic* (EqHe) which encourages equivariance using our proposed heuristic method. We defer the results to Appendix C.1.

Learning robust classifiers. Next, we compare our method with three baselines in training robust classifiers: MBRDL (Robey et al., 2020), UDA Xie et al. (2020), and empirical risk minimization (ERM). We evaluate the trained classifiers using three metrics: (1) *Robust accuracy* (R) measures the probability of a model preserving its prediction under input variations; (2) *Robust Classification accuracy* (RC) measures the probability of a model predicting the correct label under input variations; (3) *Standard accuracy* (S) measures the probability of a model predicting the correct label. Unless otherwise specified, our method uses the EqHe-trained domain translator in all experiments.

Our method excels in learning unforeseen robustness. Despite the stark dissimilarity between SVHN and CIFAR-10, Table 1 shows that our method and UDA can harness the variations on SVHN to improve the robust classification accuracy on CIFAR-10 by 4.1% and 3.2%, respectively, indicating the feasibility of learning unforeseen robustness from out-of-distribution data. Given that the consistency regularization in UDA and our method introduces an additional weight hyperparameter, we present comparisons in Figure 4 varying this weight. For both methods, we observe two stages as the

Table 2: Ablation study of our method, varying whether to use the source (Src) dataset and the domain translator (DT).

Src	DT	SVHN		STL-10	
		RC (%)	**S** (%)	**RC** (%)	**S** (%)
✓	EqGt	83.7 (↑ 0.5)	89.5	88.1 (↑ 0.3)	91.2
✓	EqHe	83.2	89.9	87.8	91.0
✓	Std	82.8 (↓ 0.4)	88.5	86.2 (↓ 1.6)	90.6
✓	×	82.3 (↓ 0.9)	88.2	85.8 (↓ 2.0)	89.9
×	×	79.1 (↓ 4.1)	89.0	79.1 (↓ 8.7)	89.0

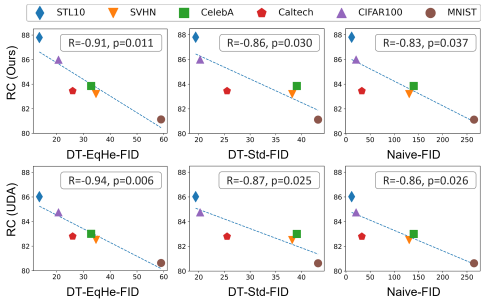


Figure 5: The correlation results for three source dataset selection criteria.

Table 3: Robust classification accuracy under six geometric data transformations, which serves as a proxy for 3D-viewpoint-change robustness.

Variations	ERM (%)	UDA (%)	Ours (%)
Affine	66.0	68.0 (↑ 2.0)	68.9 (↑ **2.9**)
Rotate	79.1	80.7 (↑ 1.6)	82.4 (↑ **3.3**)
Perspective	53.8	59.4 (↑ 5.6)	64.3 (↑ **10.5**)
Crop	83.1	85.0 (↑ 1.9)	85.6 (↑ **2.5**)
Fisheye	43.3	43.6 (↑ 0.3)	46.8 (↑ **3.5**)
Plate Spline	79.0	81.6 (↑ 2.6)	81.7 (↑ **2.7**)

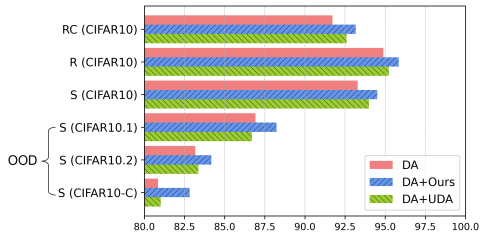


Figure 6: Our method improves robustness and generalization for foreseen variations.

regularization weight increases. In the first stage, higher weights lead to improvements in both standard and robust accuracy. In the second stage, however, raising the weight improves robust accuracy while harming standard accuracy, leading to a trade-off between the two objectives. Nevertheless, our approach outperforms UDA across all weight settings and achieves a superior Pareto-optimal in the second stage. Table 2 shows the ablation study results of our method. Both EqGt and EqHe outperform Std and the one not using the domain translator (fourth row), which underscores the significance of the equivariant domain translator.

Source dataset selection. When learning unforeseen robustness, the lack of target data variations precludes the use of cross-validation for selecting suitable source datasets. In this case, we compare three available selection criteria: (1) DT-EqHe-FID trains an EqHe domain translator and computes the FID between the target dataset and the domain-translated source dataset. (2) DT-Std-FID is the same but uses an Std-trained domain translator. (3) Naive-FID directly computes the FID between the target and the source datasets. All three criteria favor selecting source datasets with a smaller FID. Figure 6 shows that whether using our method or UDA, DT-EqHe-FID exhibits the highest correlation among the three, highlighting its effectiveness as a source dataset selection criterion.

5 Applications

Learning unforeseen robustness to natural variations. We apply our method to learn 3D-viewpoint-change robustness on CIFAR-10. To this end, we use the Objectron dataset (Ahmadyan et al., 2021) as source data, which contains video clips reflecting 3D viewpoint changes in real-world settings. We generate transformed pairs by randomly selecting an anchor frame and its adjacent frames. Since direct evaluation of 3D-viewpoint-change robustness on CIFAR-10 is infeasible, we evaluate the robustness to six common geometric transformations as a proxy. Table 3 shows that our method achieves comprehensive improvements in robustness and outperforms UDA.

Improving Foreseen Robustness and Generalization. Next, we evaluate if our method is useful when the variations are foreseen. To this end, we train classifiers on CIFAR-10 augmented with RandAugment, and use our method to learn robustness to RandAugment from STL-10. Figure 6 shows that using our method further improves robustness, in-distribution, and out-of-distribution generalization, outperforming UDA in the same setting.

REFERENCES

Adel Ahmadyan, Liangkai Zhang, Jianing Wei, Artsiom Ablavatski, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7818–7827, 2021.

Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? *Advances in Neural Information Processing Systems*, 32, 2019.

Antreas Antoniou, Amos J. Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *CoRR*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223, Fort Lauderdale, FL, USA, Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/coates11a.html.

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18613–18624. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d85b63ef0ccb114d0a3bb7b7d808028f-Paper.pdf.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljacic. Equivariant self-supervised learning: Encouraging equivariance in representations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gKLAAfiytI.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Zhun Deng, Linjun Zhang, Amirata Ghorbani, and James Zou. Improving adversarial robustness via unlabeled out-of-domain data. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2845–2853. PMLR, Apr 2021. URL https://proceedings.mlr.press/v130/deng21b.html.

Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *ICML*, 2019.

Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28, 2015.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech 256.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–189, 2018.

Zhuo Huang, Xiaobo Xia, Li Shen, Bo Han, Mingming Gong, Chen Gong, and Tongliang Liu. Harnessing out-of-distribution examples via augmenting content and style. (arXiv:2207.03162), Jul 2022. doi: 10.48550/arXiv.2207.03162. URL http://arxiv.org/abs/2207.03162. arXiv:2207.03162 [cs].

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*, pp. 1125–1134, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Leonid Koralov and Yakov G Sinai. *Theory of probability and random processes*. Springer Science & Business Media, 2007.

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

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. *International Journal of Computer Vision*, 127(5):456–476, May 2019. ISSN 1573-1405. doi: 10.1007/s11263-018-1098-y.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Hossein Vahabi, Yair Carmon, and Ludwig Schmidt. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XJk19XzGq2J.

Guo-Jun Qi, Liheng Zhang, Chang Wen Chen, and Qi Tian. Avt: Unsupervised learning of transformation equivariant representations by autoencoding variational transformations. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8129–8138, Oct 2019. doi: 10.1109/ICCV.2019.00822.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Alexander Robey, Hamed Hassani, and George J. Pappas. Model-based robust deep learning: Generalizing to natural, out-of-distribution data. *arXiv:2005.10247 [cs, stat]*, Nov 2020. URL http://arxiv.org/abs/2005.10247. arXiv: 2005.10247.

Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=77cNKCCjgw.

Antoine Salmona, Valentin De Bortoli, Julie Delon, and Agnès Desolneux. Can push-forward generative models fit multimodal distributions? In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Tsy9WCO_fK1.

Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9661–9669, 2021.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608. Curran Associates, Inc., 2020.

Ugo Tanielian, Thibaut Issenhuth, Elvis Dohmatob, and Jeremie Mary. Learning disconnected manifolds: a no gan's land. In *International Conference on Machine Learning*, pp. 9418–9427. PMLR, 2020.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=rC8sJ4i6kaH.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6256–6268. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/44feb0096faa8326192570788b38c1d1-Paper.pdf.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482. PMLR, Jun 2019. URL https://proceedings.mlr.press/v97/zhang19p.html.

Allan Zhou, Fahim Tajwar, Alexander Robey, Tom Knowles, George J. Pappas, Hamed Hassani, and Chelsea Finn. Do deep networks transfer invariances across classes? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Fn7i_r5rR0q.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A Related Work

Semi-supervised consistency regularization. A large body of work uses consistency regularization for semi-supervised learning (Sohn et al. (2020)), achieving state-of-the-art results in generalization. The key idea is to do supervised learning on the labeled data while regularizing the model to predict consistently on the unlabeled data, which potentially expands the labeled region and thus improves generalization (Wei et al., 2021). Despite the various goals previous work has, such as improving generalization (Sohn et al., 2020) or improving adversarial robustness (Zhang et al., 2019; Alayrac et al., 2019; Carmon et al., 2019; Deng et al., 2021), there is no work, to our knowledge, that learns unforeseen robustness from out-of-distribution (OOD) data. Indeed, the OOD data with potentially disjoint label sets in our setting pose a unique challenge that invalidates many common techniques such as pseudo-labeling. To harness OOD data, previous work assumes some overlaps of label sets (i.e., open-set setting, see Saito et al. (2021)) and then filters out "irrelevant" data (Xie et al., 2020; Huang et al., 2022). In contrast, overlapping label sets are not necessary for learning robustness in our setting, so we can make use of any OOD data with the desired variation.

Model-based data augmentation. Another line of work uses generative models to capture class-agnostic data variations in the dataset and then apply the trained model to do input-conditioned data augmentation for better robustness and generalization (Antoniou et al., 2017; Robey et al., 2020; Zhou et al., 2022). Modeling the variation directly from OOD data and then applying the model to the target data encounters two major difficulties. First, while the class-agnostic data variations by assumption generalize across classes and domains, the generative model capturing them may not, confining previous work to train and apply the model on the same or similar dataset. If the domain gap is large, this method can even hurt the generalization of downstream classifier. In contrast, our domain translator is trained on and applies only to the existing OOD examples, thus avoiding this issue. Second, using a GAN-based generative model to capture highly multimodal natural variations faces intrinsic challenges (Tanielian et al., 2020; Salmona et al., 2022). Indeed, prior work showed its limitation to capture geometric transformations like rotation (Zhou et al., 2022). Our method addresses this challenge by relying on the ground-truth variations from the source data, resulting in target-like rotated images as shown in the experiment.

Neural style transfer. Our approach to using a domain translator that maps source images to approximate the target distribution, is related to neural style transfer (Gatys et al., 2015; Johnson et al., 2016; Huang et al., 2018; Isola et al., 2017; Zhu et al., 2017). The similar image-to-image translation process allows us to take advantage of this rich literature and adapt various off-the-shelf network architectures to implement our domain translator. However, the goals differ. Neural style transfer aims at transferring the style of a source image to a target one while preserving some content or the underlying label. In contrast, our domain translator does not need to preserve the content or label but requires equivariance to the data variation.

B Additional Analysis

B.1 Proof of Proposition 2.2

Before giving the proof, we first state the definition of push-forward distribution, which appears in many textbooks (see, e.g., Koralov & Sinai (2007)).

Definition B.1 (Push-forward distribution)**.** *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a measurable space $(\tilde{\Omega}, \tilde{\mathcal{F}})$, and a measurable mapping $\xi : \Omega \to \tilde{\Omega}$, the push-forward distribution of $\mathbb{P}$ on the σ-algebra $\tilde{\mathcal{F}}$ is defined by*

$$\xi_{\#}\mathbb{Q}(A) = \mathbb{P}(\xi^{-1}(A)) \quad \textit{for } A \in \tilde{\mathcal{F}},$$

where $\xi^{-1}(A) := \{\omega \in \Omega : \xi(\omega) \in A\}$ denotes the pre-image of a measurable set A.

The proof follows from the assumptions that the loss ℓ satisfies the triangle inequality and $\bar{\ell}_f$ is Lipschitz uniformly over all models f. Since we are working on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with functions implemented by neural networks (with continuous activation functions) and common losses, we omit the measurability issue.

Proof. First, since ℓ is non-negative, by Tonelli's theorem, we have

$$L_\phi(f, \mathbb{P}) := \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}, \boldsymbol{t}\sim\mathbb{T}} \left[\ell\big(f(\boldsymbol{x}), f(\phi_{\boldsymbol{t}}(\boldsymbol{x}))\big)\right] = \mathop{\mathbb{E}}_{x\sim\mathbb{P}} \left[\bar{\ell}_f(x)\right],$$

where $\bar{\ell}_f(\boldsymbol{x}) := \mathbb{E}_{t\sim\mathbb{T}}[\ell\big(f(\boldsymbol{x}), f(\phi_{\boldsymbol{t}}(\boldsymbol{x}))\big)]$.

Then, since $\bar{\ell}_f$ is uniformly Lipschitz with a Lipschitz constant $\|\bar{\ell}\|_{\mathrm{L}}$, by Kantorovich-Rubenstein duality theorem (see, e.g., (Villani, 2021)), we have

$$\mathop{\mathbb{E}}_{x\sim\mathbb{P}} \left[\bar{\ell}_f(x)\right] - \mathop{\mathbb{E}}_{x\sim\xi_\#\mathbb{Q}} \left[\bar{\ell}_f(x)\right] \leq \|\bar{\ell}\|_{\mathrm{L}}\, W_1(\mathbb{P}, \xi_\#\mathbb{Q}).$$

Thirdly, since $\xi_\#\mathbb{Q}$ is the push-forward distribution of $\mathbb{Q}$ through the mapping ξ, by change of measure, we have

$$\mathop{\mathbb{E}}_{x\sim\xi_\#\mathbb{Q}} \left[\bar{\ell}_f(x)\right] = \mathop{\mathbb{E}}_{u\sim\mathbb{Q}} \left[\bar{\ell}_f(\xi(u))\right].$$

Lastly, since ℓ satisfies the triangle inequality, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{u\sim\mathbb{Q}} \left[\bar{\ell}_f(\xi(u))\right] &= \mathop{\mathbb{E}}_{u\sim\mathbb{Q}} \mathop{\mathbb{E}}_{t\sim\mathbb{T}} \left[f(\xi(\boldsymbol{u})), f(\phi_{\boldsymbol{t}} \circ \xi(\boldsymbol{u}))\right] \\ &\leq \mathop{\mathbb{E}}_{u\sim\mathbb{Q}} \mathop{\mathbb{E}}_{t\sim\mathbb{T}} \left[f(\xi(\boldsymbol{u})), f(\xi \circ \phi_{\boldsymbol{t}}(\boldsymbol{u}))\right] + \mathop{\mathbb{E}}_{u\sim\mathbb{Q}} \mathop{\mathbb{E}}_{t\sim\mathbb{T}} \left[f(\xi \circ \phi_{\boldsymbol{t}}(\boldsymbol{u})), f(\phi_{\boldsymbol{t}} \circ \xi(\boldsymbol{u}))\right]\end{aligned}$$

Rearranging terms completes the proof. □

B.2 Discussion about the Existence of Equivariant and Accurate Domain Translators

We discuss some of our conjectures about the existence here and leave the complete characterization to future work. Since we use continuous maps to instantiate ξ, we conjecture that the equivariant domain translator does not exist if the support of the source data distribution, after being expanded by the transformation, has a smaller intrinsic dimension (see, e.g., Pope et al. (2021); Salmona et al. (2022)) than that of the target. Indeed, we empirically observe that for some source and target datasets such as SVHN to CIFAR-10, training the domain translator yields a trade-off between the equivariance and the approximate performance, but such trade-off mitigates if we swap the source and target datasets. Interestingly, this existence issue seems to enable us to use the training result of an *equivariant* domain translator as the source selection criterion.

C Additional Results

C.1 Training Equivariant Domain Translator

We first show that our method learns equivariant domain translators. We compare three methods: (1) *Standard* (Std) which does not encourage equivariance ($\lambda = 0$ in Eq. 3.1), (2) *Equivariant-Groundtruth* (EqGt) which encourages equivariance using the groundtruth data transformation function, and (3) *Equivariant-Heuristic* (EqHe) which encourages equivariance using our proposed heuristic method. We use mean-squared-error (MSE) loss to evaluate and regularize equivariance, and use Fréchet Inception Distance (FID, Heusel et al. (2017)) to evaluate how well the translated source data approximate the target data (smaller values are better for both).

Figure 7 show the results when using CIFAR-10 as the target, SVHN as the source, and RandAugment (Cubuk et al., 2020) as the variation. The shown FID and MSE results are evaluated on all training data. The left-hand label of each row indicates how the images in that row are acquired. Each row's result corresponds column-wise.

Despite the stark dissimilarity between SVHN and CIFAR-10 (FID=130.7), all three domain translators successfully translate SVHN to resemble CIFAR-10 (with FIDs < 30), demonstrating that *our method learns accurate translators*.

With similar FIDs, EqGt and EqHe achieve much lower equivariance loss (0.4 and 0.8, respectively) compared to Std (3.3), demonstrating that *our method learns equivariant domain translators*. The

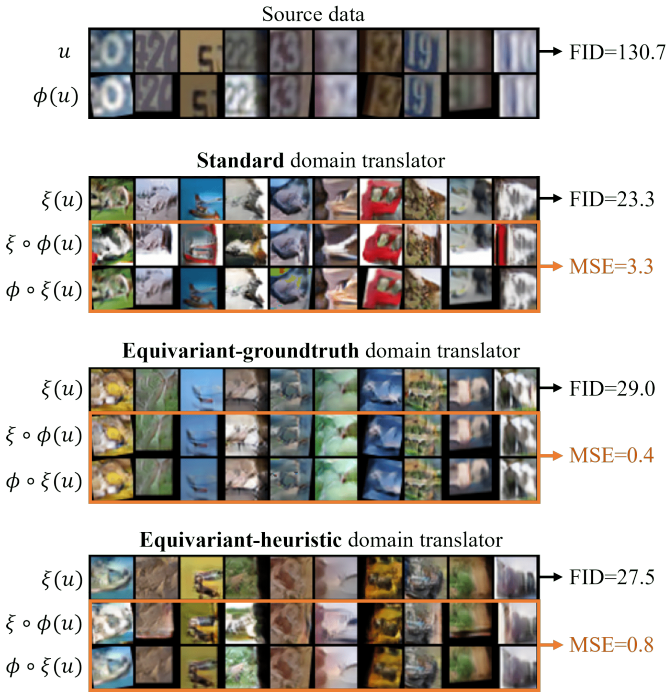


Figure 7: Training results of different domain translators ξ, including quantitative results and some input-output examples. The target dataset is CIFAR-10, the source is SVHN, and the considered variation ϕ is RandAugment. The label on the left of each row indicates how the images of that row are calculated, and the results correspond column-wise. FID measures the similarity to the target data, MSE loss measures the equivariance, both are smaller the better. While all three domain translators well-translate the source data to be target-like, only EqGt and EqHe well-preserve the variations (highlighted in orange boxes). Note that the shown FID and MSE losses are evaluated on all training data.

orange boxes highlight the images where we can visually discern improved equivariance. EqHe preserves various transformations in RandAugment, similar to EqGt, but without knowing ground-truth transformation functions or parameters, showing its effectiveness.

C.2 Visualizing the Results of Equivariant Domain Translator

We show the outputs of our domain translators in Figure 8, 9 and 10. Results demonstrate that our method can effectively translate the source data to be target-like. The trained domain translator also well-preserve the variations including random rotation, RandAugment, and 3D-viewpoint change. Therefore, we are able to do consistency regularization with the target-like images and the transformed version of them, so that to train a robust classifier under unforeseen variations. We notice that domain translators trained with different source dataset have different performances. As discussed in Section 4, the source dataset's distance to the target dataset correlates with the performance. Additionally, if the source dataset is much "simpler" than the target one, such as MNIST and SVHN, it is very difficult for the domain translator to cover the whole manifold of the target distribution, and to preserve complex variations such as RandAugment (especially the color change) on MNIST. One interesting future work is to take the intrinsic dimension of the dataset into consideration.

C.3 Results on CIFAR-100

Table 4 shows the results on CIFAR-100 where we use SVHN, STL10 and CIFAR-10 as the source data. Data variation is the RandAugment. We get consistent results where our method excels over other methods in robustness and accuracy.

(a) SVHN as the source dataset. Random rotation as the variation.

(b) STL10 as the source dataset. Random rotation as the variation.

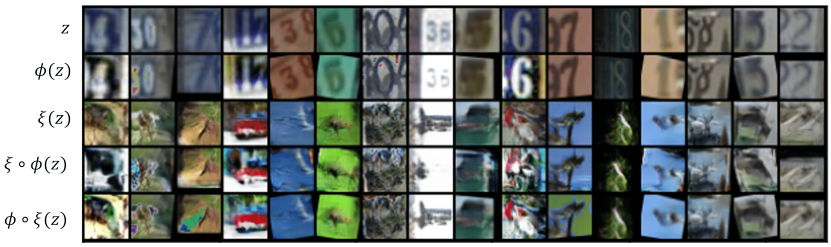
Figure 8: Results of our method with random rotation as the input variation. We use CIFAR-10 as the target dataset. z denotes the source data, ϕ denotes the variation, i.e. random rotation, and ξ denotes EqHe, the domain translator trained with the heuristic method. By comparing $\xi(z)$ with CIFAR-10 data, results indicate that our method can effectively translate the source data to be target-like. By comparing between $\xi \circ \phi(z)$ and $\phi \circ \xi(z)$, which are expected to be similar, our domain translators well-preserve the variations.

Table 4: Results of classifiers trained using different methods and source datasets. The target dataset is CIFAR-100 w/o data augmentation and the data variation is RandAugment. We show here for reference the oracle method that does consistency regularization directly on the target dataset.

		Robustness		**Accuracy**
Method	**Src**	**RC** (%)	**R** (%)	**S** (%)
ERM	/	48.8 ± 0.1	57.2 ± 0.2	62.9 ± 0.3
MBRDL	SVHN	36.9 ± 0.4	55.3 ± 0.5	52.4 ± 0.3
UDA	SVHN	51.7 ± 0.2	61.6 ± 0.2	63.2 ± 0.4
EDT (Ours)	SVHN	**53.2** ± 0.3	**63.4** ± 0.2	**64.1** ± 0.3
MBRDL	STL10	39.6 ± 0.3	56.1 ± 0.3	56.1 ± 0.2
UDA	STL10	55.9 ± 0.3	67.1 ± 0.2	64.1 ± 0.3
EDT (Ours)	STL10	**58.3** ± 0.3	**70.0** ± 0.3	**65.1** ± 0.3
MBRDL	CIFAR-10	39.6 ± 0.4	58.4 ± 0.3	56.2 ± 0.3
UDA	CIFAR-10	56.5 ± 0.2	68.3 ± 0.2	63.8 ± 0.3
EDT (Ours)	CIFAR-10	**59.0** ± 0.2	**71.2** ± 0.3	**64.5** ± 0.2
Oracle	/	70.9 ± 0.2	82.1 ± 0.2	73.6 ± 0.2

C.4 Problems of MBRDL

Figure 11 and 12 shows the performance of the variation simulator learned by MBRDL. We can see that the MBRDL suffers from two problems. Firstly, it is hard to learn a good variation simulator. As Zhou et al. (2022) observed and as shown in Figure 11, brightness change and color change are easy to learn but geometric transformations such as rotation are hard to learn. The complex variations such as RandAugment are even harder. Secondly, the learned variation simulator has poor generalization ability. Figure 12 (a) and (c) show that the variation simulator which is trained on the source data performs well on the source data. However, (b) and (d) show that the variation simulator performs



(a) SVHN as the source dataset. RandAugment as the variation.

(b) STL10 as the source dataset. RandAugment as the variation.

(c) CelebA as the source dataset. RandAugment as the variation.

(d) MNIST as the source dataset. RandAugment as the variation.

Figure 9: Results of our method with RandAugment as the input variation. We use CIFAR-10 as the target dataset. z denotes the source data, ϕ denotes the variation, i.e. RandAugment, and ξ denotes EqHe, the domain translator trained with the heuristic method. By comparing $\xi(z)$ with CIFAR-10 data, results indicate that our method can effectively translate the source data to be target-like. By comparing between $\xi \circ \phi(z)$ and $\phi \circ \xi(z)$, which are expected to be similar, our domain translators well-preserve the variations in most cases.

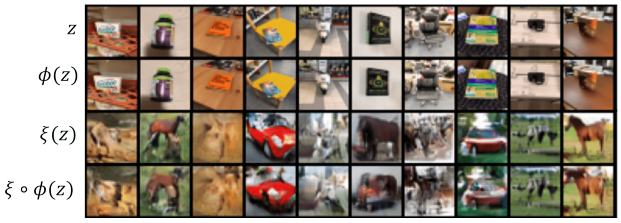
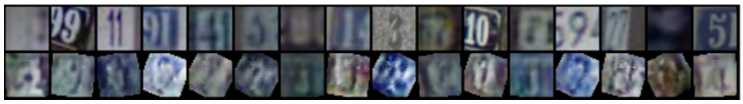


Figure 10: Results of our method with 3D-viewpoint change as the input variation. We use CIFAR-10 as the target dataset and Objectron as the source dataset. z denotes the source data, ϕ denotes the variation, i.e. 3D-viewpoint change, and ξ denotes EqHe, the domain translator trained with the heuristic method. By comparing $\xi(z)$ with CIFAR-10 data, results indicate that our method can effectively translate the source data to be target-like. $\xi \circ \phi(z)$ shows that the domain translator well-preserves the 3D-viewpoint change. For example, in the fourth column, two cars generated by $\xi(z)$ and $\xi \circ \phi(z)$ well-preserve the viewpoint change that exits in two chair images (i.e. z and $\phi(z)$).



(a) Apply rotation simulator learned on SVHN to SVHN.

(b) Apply rotation simulator learned on SVHN to CIFAR10.

(c) Apply rotation simulator learned on STL10 to STL10.

(d) Apply rotation simulator learned on STL10 to CIFAR10.

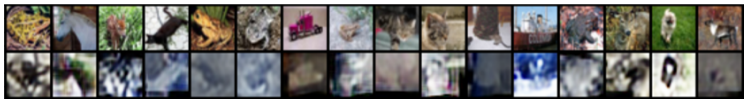
Figure 11: Results of MBRDL with random rotation as the input variation. In every subfigure, the first line shows the original images and the second line shows the transformed ones using the learned variation simulator.

badly when directly applied to the target data, resulting in blurred images or content-changed images. We suspect that it is because the variation is very hard to learn and it is even harder to learn a variation simulator that is disentangled from the source data. The problems get severe when the target domain and the source domain are far from each other. This explains why MBRDL hurts the robustness and accuracy in our experiments.

D Limitations and Societal Impact

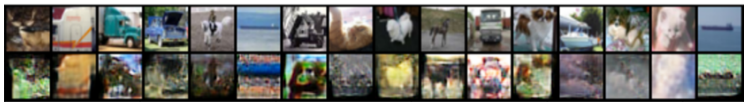
This work introduces a new approach to expanding a set of data variations that a model can learn. Unlike the prior work that trains models robust to foreseen data variations, we provide a way to expand the robustness to unforeseen data variations by harnessing out-of-distribution data (that may not have labels). This has been known as challenging due to the domain gap between the training data distribution of our interest and the source out-of-distribution data, but we introduce an approach to bridge this gap by training an accurate, equivariant domain translator. The domain translator can

(a) Apply RandAugment simulator learned on SVHN to SVHN.



(b) Apply RandAugment simulator learned on SVHN to CIFAR10.

(c) Apply RandAugment simulator learned on STL10 to STL10.



(d) Apply RandAugment simulator learned on STL10 to CIFAR10.

Figure 12: Results of MBRDL with RandAugment as the input variation. In every subfigure, the first line shows the original images and the second line shows the transformed ones using the learned variation simulator.

produce transformed source data and contribute to improving unforeseen robustness while training a classifier with consistency regularization. As (out-of-distribution) data becomes more abundant, our work is potentially practical and useful in many real-world settings.

However, the cardinality of a set of data variations available in the real world is *infinite*, and this opens-up future work opportunities. A question that our work left for future work is how we can train a robust model to cover such infinite data variations. We may find a data variation representing a group of data variations, such as 3D-viewpoint-changes, that may introduce the robustness to 2D rotations. Considering data variations during training generally leads to more computations; thus, reducing the number of variations can contribute to efficient deep-learning practices. This also suggests that evaluating the unforeseen robustness that a model learns is challenging. In our experiments, we use the surrogate transformations that (approximately) represent the original unforeseen robustness, but to precisely quantify the effectiveness of robust training methods, we encourage the community to have validation data with a broader set of data variations.

We also found that the computational demands for training an accurate, equivariant domain translator are higher than training robust models with some regularization techniques. Given the community has a great interest in sustainability, there could be a potential concern about improving robustness with generative models. However, we argue that this is not the only concern for our method, but generally, any techniques that use recent generative models, such as diffusion models, could fall into the same category. We, therefore, leave improving the computational efficiency of our method as future work.

E Detailed Experimental Setup

Setup. Our target datasets are CIFAR-10 and CIFAR-100, and we choose the source dataset from SVHN, STL-10, CIFAR-100, MNIST, CelebA, and Caltech-256. We have source datasets perceptually very different from the targets, such as MNIST or CelebA, reflecting real-world scenarios.

We consider two data variations where we know the transformation functions: (1) RandAugment (Cubuk et al., 2020), containing a random combination of 14 random transformations, spanning from geometric transformations to color space changes, and (2) random rotation that we use for verification due to its simplicity. To simulate the unforeseen robustness setting, we only use the

transformation function on source data during training. We will evaluate our algorithm against unknown transformation functions in §5.

Datasets. In Section 4, We use CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) as target datasets and SVHN (Netzer et al., 2011), STL-10 (Coates et al., 2011), CIFAR-100, MNIST (Deng, 2012), CelebA (Liu et al., 2015), and Caltech-256 (Griffin et al.). When training domain translators, we only use unlabeled images from the source and target. In Section 5, we use Objectron (Ahmadyan et al., 2021) as the source dataset to learn 3D-viewpoint-change robustness. Objectron is a collection of short, object-centric video clips. We randomly sample several frames from each clip as the anchor images and randomly sample frames in a range of 10 frames as the 3D-viewpoint changed images. We use such pairs to do 3D-viewpoint change consistency regularization. To evaluate the out-of-distribution generalization of classifiers trained on CIFAR-10, we use CIFAR-10.1 (Recht et al., 2018), CIFAR-10.2 (Lu et al., 2020), and CIFAR-10-C (Hendrycks & Dietterich, 2019) as the ood datasets. CIFAR-10.1 and CIFAR-10.2 are sampled from TinyImageNet (Le & Yang, 2015) with the same classes of CIFAR-10. CIFAR-10-C is a collection of a corrupted version of CIFAR-10 under 15 types of corruption.

Data variations. In Section 4, we use RandAugment and random rotation as the variations. RandAugment contains 14 candidate transformation functions: "ShearX", "ShearY", "TranslateX", "TranslateY", "Rotate", "Brightness", "Color", "Contrast", "Sharpness", "Posterize", "Solarize", "AutoContrast", "Equalize", and "Identity". When using RandAugment, a composition of two randomly selected functions are applied to the images. For random rotation, we use $[-30°, 30°]$ random rotation. Although the rotation is simply defined, it cannot be modeled by existing model-based methods that use MUNIT-like architectures (Zhou et al., 2022). In Section 5, we consider 3D-viewpoint change as the unforeseen variation. We randomly select two nearby frames from one video clip as the two 3D-views of one object. Since we could not evaluate the model robustness to 3D-viewpoint change on the target data (CIFAR-10), we use six proxy transformations to estimate the 3D-viewpoint robustness. Proxy transformations are geometric transformations that do warping on images, which include "Random Affine", "Random Rotate", "Random Perspective", " Random Crop", " Random Fisheye", "Random Thin Plate Spline"[1].

Evaluation Metrics. We evaluate the trained classifiers with three metrics[2]: the *robust accuracy*, denoted as R, measures the probability of a model preserving its output under input variations, the *robust classification accuracy*, denoted as RC, measures the probability of a model predicting the correct label under input variations, the *standard accuracy*, denoted as S, measures the probability of a model predicting the correct label. During testing, we randomly sample 20 transformed versions for each example to estimate the expectation of robust accuracy and robust classification accuracy.

E.1 Our Method

We use Wasserstein GAN (Arjovsky et al., 2017) to train a domain translator where the inputs of the generator (i.e. domain translator) are source images and the outputs are encouraged to be similar to the target images. We use the encoder-decoder model architecture for implementing the domain translator (i.e. generator), which consists of two convolutional layers for down-sampling, two residual blocks for latent propagation, and two other convolutional layers for up-sampling. The discriminator then distinguishes the real target data from the fake ones translated from the source data. We train generator and discriminator with adversarial training following WGAN where we use 0.01 as the clip value of the discriminator's weight. For training equivariant domain translator, we use the mean-squared-error (MSE) loss for the equivariance regularization term (the second term in Eq. 3.1). We set $\lambda = 1$ in Eq. 3.1.

For the robust classifier, we use ResNet18 as the architecture. Since the zero-one loss is difficult to optimize directly, we follow the common practice of using the surrogate loss (Bartlett et al., 2006). We use the cross-entropy loss for training the classifier, including the robustness regularization term I_1, similar to Zhang et al. (2019). The MSE loss and the L^1 norm loss are two common training

[1]Implementation follows https://kornia.readthedocs.io/en/latest/augmentation.module.html

[2]Each of them can be viewed as one minus the corresponding loss (instantiated with zero-one loss) defined in Section 2.

objectives that measure the difference between two images in the pixel space. They are used as the reconstruction loss in VAE, CycleGAN, Diffusion Model, etc. We also tried the L^1 loss for the equivariance regularization term but did not observe substantial difference. In all our experiments, we use cross-entropy loss as the surrogate loss for training and regularizing the classifier. We set $\lambda_1 = \lambda_2 = 0.5$ in Eq. 3.2. Since accurately estimating the W_1 distance for multi-dimensional non-Gaussian distributions is difficult, we use the Fréchet inception distance (FID, see Heusel et al. (2017)) to evaluate how well the domain translator pushes forward the source data to approximate the target data.

E.2 MBRDL

MBRDL (model-based robust deep learning, (Robey et al., 2020)) learns a model to simulate the natural variation. In their paper, the variation model is learned and applied to the same domain. Their method can easily extend to scenarios where variations are unforeseen in the target domain but is available in the source domain. In this paper, we first learn a variation simulator with the source data where transformed pairs are used for learning variations. We use MUNIT (Huang et al., 2018) as the variation simulator following settings in Robey et al. (2020). MUNIT is first designed for style transfer, here Robey et al. (2020) use it for input transformation. Then, we apply the variation simulator directly to the target data to do data variation and train robust classifiers with a consistency regularization loss addition to the classification loss.

E.3 UDA

UDA (unsupervised data augmentation, (Xie et al., 2020)) improves the model's robustness against variations with consistency regularization on unlabeled data. Although the unlabeled data is very similar to the target data and has foreseen variations in their paper, we can directly use their method in our case. We see source data as the unlabeled data and do consistency regularization on it while training the classifier on the target data. It's easy to see that, UDA is a simple version of our method where $\lambda_1 = 0$ in Eq. 3.2. In our experiments of UDA, we set $\lambda_2 = 1$.