Improving Out-of-distribution Robustness by Adversarial Training with Structured Priors

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep models often fail to generalize well in test domains when the data distribution differs from that in the training domain. Among numerous approaches to address this Out-of-Distribution (OOD) generalization problem, there has been a growing surge of interest to exploit the input-robustness obtained by Adversarial Training (AT) to improve OOD performances. Recent works have revealed that the robust model obtained by conducting sample-wise AT also retains transferability to biased test domains. In this paper, we empirically show that sample-wise AT has limited improvement on OOD performance. Specifically, we find that AT can only maintain performance at smaller scales of perturbation while Universal AT (UAT) are more robust to larger-scale perturbations. This provides us with clues that the adversarial perturbations with universal (low dimensional) structures can enhance the robustness to large data distribution shifts which are common in OOD scenarios. Inspired by this, we propose two AT variants with low-rank structures to train OOD-robust models. Extensive experiments on DomainBed benchmark show that our proposed approaches outperform Empirical Risk Minimization (ERM) and sample-wise AT.

1 Introduction

Existing deep learning methods have achieved good performance on visual classification tasks under the same distribution of training sets and test sets. However, when the data distribution of the test set is different from that of the training set, the classification performance of the deep neural networks (DNNs) may decrease sharply [1]. This is mainly because the network may capture spurious features such as background and style information to assist the fast fitting during the training process [2]. However, in real-world scenarios, test data may differ from training data in background and style information, thus DNNs that use unstable spurious features to make predictions will fail. Solving the above problem is known as the out-of-distribution (OOD) generalization.

Another scenario where DNNs may fail is that deep models are often vulnerable to adversarial examples [3]. Adversarial training (AT) is originally proposed as an effective way to defend against adversarial attacks [4]. Moreover, there is work showing that adversarial training helps to solve the OOD generalization problem because OOD data can be seen as stronger perturbations to some extent [5]. The reason why AT can defend adversarial attacks meanwhile benefit OOD generalization is that it can make the network robust to the interference of spurious features, such as randomly injected noise (in adversarial examples) or the spurious correlation between labels and background information (in OOD generalization). In other words, AT enables the network to make predictions using intrinsic features rather than spurious features.

A potential problem, however, is that existing AT methods for solving OOD generalization ignore the specific design of perturbations. They usually simply conduct sample-wise AT [6], which only brings

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

limited performance improvement for OOD generalization. The essential reason for the failure of this type of approach is that the perturbations it uses cannot distinguish between invariant and spurious features. As a result, it improves the robustness at the expense of the standard accuracy decreasing [7]. Moreover, we empirically find that universal AT [8] , i.e. with domain-wise perturbations, shows stronger input-robustness when facing larger-scale perturbations comparing to sample-wise AT. Since the sample injected with large-scale perturbations can be regarded as OOD samples [5], we draw inspiration from this phenomenon that AT with universal (low-dimensional) structures can be the key to solving OOD generalization. To address the above problems, we propose to use structured low-rank perturbations related to domain information in AT, which can help the model to filter out background and style information, thus benefits OOD generalization. We make the following contributions in our work:

- We identify the limitations of sample-wise AT on OOD generalization through a series of experiments. To alleviate this problem, we further propose two simple but effective AT variants with structured priors to improve OOD performances.
- We theoretically prove that our proposed structured AT approach can accelerate the convergence of reliance on spurious features to 0 when using finite-time-stopped gradient descent, thus enhancing the robustness of the model against spurious correlations.
- By conducting experiments on the DomainBed benchmark [9], we demonstrate that our algorithms outperform ERM and sample-wise AT on various OOD datasets.

2 Related Work

Solving OOD Generalization with AT. According to [3], the performance of deep models are susceptible to small-scale perturbations injected in the input images, even if these perturbations are imperceptible to humans. Adversarial training (AT) [4] is an effective approach to improve the robustness of the model to input perturbations. However, many of the recent works have begun to focus on the connection between AT and OOD due to the fact that OOD data can be regard as one kind of large-scale perturbations. These works seek to exploit the robustness provided by AT to enhance the OOD generalization performance. For instance, [6] applies sample-wise AT to OOD generalization. They theoretically find that if a model is robust to input perturbation on training samples, it also generalizes well on OOD data. [5] theoretically establishes a link between the objective of AT and the OOD robustness. They reveals that the AT procedure can be regarded as a heuristic solution to the worst-case problem around the training domain distribution. Nevertheless, the discussion of [6] and [5] is restricted to the framework of using Wasserstein distance to measure the distribution shift, which is less practical for the real-world OOD setting where domain shifts are diverse. Additionally, they only study the case of sample-wise AT and did not further investigate the effect of different forms of AT (not sample-wise) on the robustness performance of OOD. Other works such as [10] focuses on the structure design of the perturbations. In [10] they use multi-scale perturbations within one picture. But they do not exploit the universal information within one training domain. In our work, we focus on real-world OOD scenarios where there are additional clues lying in the distribution shifts, i.e, the low-rank structures in the spurious features (such as background and style information) across one domain. We further design a low-rank structure in the perturbations to specifically eliminate such low-rank spurious correlations.

OOD Evaluation Benchmark. The DomainBed benchmark [9] provides a fair way of evaluating different state-of-the-art OOD methods, which has been widely accepted by the community. By conducting rigorous experiments in a consistent setting, [9] reveals that many algorithms that claim to outperform previous methods cannot even outperform ERM. Unlike previous works using AT to address OOD generalization, such as [6] and [5], in this work we adopt the Domainbed benchmark for a fair comparison of our approach with existing state-of-the-art methods.

3 Weakness of Sample-wise AT for OOD Generalization

3.1 Preliminaries

Out-of-distribution (OOD) Generalization. Let $x \in \mathcal{X}$ denote the random data in the input space $\mathcal{X}$, $y \in \mathcal{Y}$ denotes the target random data in label space $\mathcal{Y}$, and $\phi : \mathcal{X} \to \mathcal{Z}$ denotes the feature

Table 1: Test accuracy (%) on four OOD datasets on DomainBed benchmark with a fixed set of hyperparameters. The improvement of AT is marginal.

	Datasets				
Algorithm	PACS	OfficeHome	VLCS	NICO	avg
ERM	79.7 ± 0.0	59.6 ± 0.0	74.4 ± 1.0	$\mathbf{70.7 \pm 1.0}$	71.1
AT	$\mathbf{81.5 \pm 0.4}$	$\mathbf{59.9 \pm 0.4}$	$\mathbf{75.3 \pm 0.7}$	68.2 ± 2.2	**71.2**

extractor, we have the predictor as $f = w \circ \phi(x)$ where $z = \phi(x)$ is the feature variable in the feature space $\mathcal{Z}$ and $w : \mathcal{Z} \to \mathcal{Y}$ denotes the classifier.

Now we give the formal definition of the out-of-distribution generalization problem. We have a set of m training domains $\mathcal{E} = \{E_1, E_2, ..., E_m\}$, where each domain E_e is characterized by a input dataset $E_e := \{(x_i^e, y_i^e)\}_{i=1}^{n_e}$ containing n_e i.i.d input samples drawn from the distribution of $\mathcal{P}_e$, and a test domain E_{m+1} with data following the distribution of $\mathcal{P}_{te}$, where $\mathcal{P}_{te} \neq \mathcal{P}_i,\ i = 1, 2, ..., m$. $\mathcal{L} : \mathcal{X} \to \mathbb{R}^+$ denotes the loss function. The ultimate goal of OOD generalization is to find an optimal predictor f that minimizes the risk on the unseen test domain:

$$\min_f \mathbb{E}_{(x,y)\sim\mathcal{P}_{te}(x,y)}[\mathcal{L}(f(x), y)]. \tag{1}$$

Adversarial Training (AT)[1]. According to [4], AT can be expressed as the following optimization problem:

$$\min_f \mathbb{E}_{(x,y)\sim\mathcal{P}(x,y)}[\max_{\delta\in\mathcal{S}} \mathcal{L}(f(x+\delta), y] \ \text{ s.t. } \|\delta\|_p \leq \epsilon, \tag{2}$$

where $\delta \in \mathcal{S}$ is the random injected perturbation with l_p norm bounded by $\epsilon > 0$. The inner maximization problem can be optimized by fast gradient sign method (FGSM [11]), a simple one-step scheme:

$$x = x + \epsilon \text{sgn}(\nabla_x \mathcal{L}(f(x), y), \tag{3}$$

where $\text{sgn}(\cdot)$ is the sign function, or by projected gradient descent (PGD [4]), a more powerful multi-step variant:

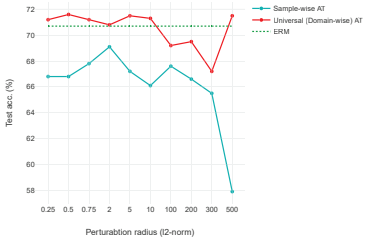
$$x^{t+1} = \prod_{x+\mathcal{S}} (x^t + \alpha \text{sgn}(\nabla_x \mathcal{L}(f(x), y))), \tag{4}$$

where $\prod$ is the projection operator onto the ϵ-ball and α is the step size.

3.2 Weakness of AT for OOD Generalization

We now highlight some weakness of sample-wise AT for OOD generalization based on series of empirical evidence. We first conduct a toy experiment on the DomainBed benchmark [9] to evaluate the OOD performance of AT. We run ERM and AT on four OOD datasets: PACS [12], OfficeHome [13], VLCS [14], and NICO [15] with a fixed set of hyperparamters (Detailed experimental setting can be found in A.1). The results are shown in Table 1. We can see that the improvement of OOD performance by AT is limited with an average improvement of only 0.1%.

We further investigate the reason behind the limitations of performance improvements on OOD datasets of AT. Although previous works have revealed that the robust features obtained by AT can improve the model generalization under OOD ([6] [5] [16]), we find that existing sample-wise AT methods only tolerate small scale perturbations. Thus, we design an experiment on NICO dataset with multiple scales of perturbations. The experiment details are shown in A.1.



we show that AT will suffer severe performance degradation when using large perturbations (shown in Figure 3.2). This provides clues to understand the failure of AT in OOD scenarios. The distribution

[1]For simplicity, we denote 'AT' for sample-wise AT by default in the rest of the paper.

shifts in OOD data usually have much larger scales than the invisible perturbations commonly used in AT [5]. Hence, AT methods designed for small perturbations cannot handle these large scale domain shifts that often appear in OOD data. However, our experiment shows that we can alleviate this problem by adapting universal AT (UAT, [8]) to the OOD setting, i.e. using domain-wise perturbations. Figure 3.2 shows that UAT will keep its generalization performance when the perturbation scale is large. There are two empirical explanations for this: first, the background and style information usually have a low-rank structure, such as the grassland and snowfield that have recurring parts. Second, similar spurious features often appear within one specific domain, such as PACS [12] and VLCS [14] datasets. As stated in [8], the universal perturbation lies in a low dimensional space. Hence using universal (domain-wise) perturbations will help to eliminate such low-rank shifts and improve model robustness.

Inspired by this, we proposed two new AT algorithms with more sophisticated low-rank structures on different dimensions to improve OOD generalization in the next section.

4 The Proposed Structured AT Method

In order to construct low-rank structured perturbations, we start by analyzing the structure of sample-wise perturbations. Assume that each input data x has a shape of $N \times N \times C$. N is the size of the input image and C is the number of the channels. For simplicity, we assume $C = 3$. We reparameterize the sample-wise perturbations as a series of 2-D matrices $\{D_1^1, D_2^1, ..., D_m^1\}$, $\{D_1^2, D_2^2, ..., D_m^2\}$, $\{D_1^3, D_2^3, ..., D_m^3\}$ where $D_e^c \in R^{n_e \times (N*N)}$ denotes the perturbations in the e-th domain for the input channel c, n_e is the number of the samples in domain E_e, and m is the number of domains. The i-th row of D_e^c represents the c-th channel of the i-th sample in domain E_e (see the first column in Figure 1 for illustration). By such reparameterization, it is natural to find that there are two orientations to reduce the rank of the perturbations:

1. **Along the dimension of the number of samples** (along the red arrow in the upper left corner of Figure 1). This corresponds to reduce the number of the perturbations used within one domain.
2. **Along the dimension of the input scale** (along the blue arrow in the upper left corner of Figure 1). This corresponds to reduce the rank of the perturbation used for a specific input sample.

In the following sections, we will propose two AT variants with structured priors that reduce the rank in these two directions.

4.1 MAT: Adversarial Training with Combinations of Multiple Perturbations

In this section, we propose domain-wise Multiple-perturbation Adversarial Training (MAT). It aims to conduct rank minimization along the dimension of the number of samples. Instead of using sample-wise perturbations, MAT constructs a combination of multiple perturbations and shares this mixed perturbation within a domain. Specifically, we choose to train the linear combination of k perturbations for each domain E_e to conduct AT. Here k is a hyperparameter and k is far less than the number of samples in domain E_e. The optimization problem can be reformulated as:

$$\min_f \sum_e E_{(x,y)\sim\mathcal{P}_e(x,y)}[\mathcal{L}(f(x+\delta^e), y], \tag{5}$$

$$\text{s.t. } \delta^e = \sum_{i=1}^{k} \alpha_i^{e*}\delta_i^{e*},\ \|\delta^e\|_p \le \epsilon,\ \sum_{i=1}^{k}\alpha_i = 1,\ \alpha_i \ge 0 \text{ for } i = 1, 2, ..., k, \tag{6}$$

where

$$\alpha_i^{e*},\ \delta_i^{e*} = \underset{\alpha_i^e,\ \delta_i^e}{\text{argmax}} E_{(x,y)\sim\mathcal{P}_e(x,y)}[\mathcal{L}(f(x+\sum_{i=1}^{k}\alpha_i^e\delta_i^e), y]. \tag{7}$$

Here e denotes a training domain and α_i^e is the weight that can be learned for each perturbation δ_i^e.

MAT works as a low-rank version of sample-wise AT. In sample-wise AT, we maintain n_e perturbations for each domain E_e, where n_e is the number of training samples in domain E_e. As for MAT,

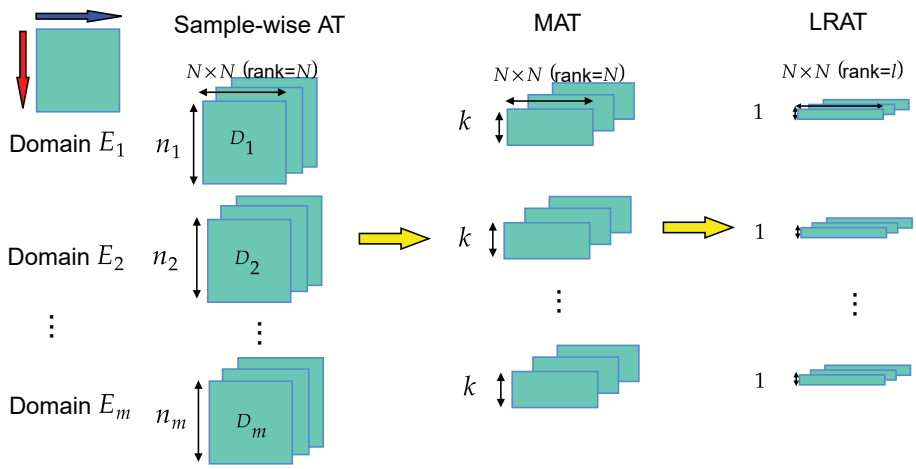


Figure 1: Illustration of how our proposed structured AT reduces the rank of the perturbations comparing to sample-wise AT. The left column shows how we reparameterize sample-wise AT. The perturbations are segmented by domains. A block represents the n_e perturbations injected in a channel of the samples in domain E_e. This figure shows the case where the input image has three channels (RGB). The red and the blue arrow in the upper left shows the two orientations to reduce the rank of the perturbations, i.e along the dimension of the total number of the perturbations and along the dimension of rank of a single perturbation, respectively. The mid column illustrates that MAT reduces the number of the perturbations used for domain E_e from n_e to k ($k \ll n_e$). The right column shows that LDAT further reduces the number of the perturbations from k to 1. Moreover, it reduces the rank of a certain perturbation from N to l ($l \ll N$).

it reduces the number of perturbations available to samples from n_e to k and to obtain low-rank structures (see the second column in Figure 1 for illustration). Therefore, it fulfills rank reduction along the sample-number dimension.

4.2 LDAT: Adversarial Training with Low-rank Decomposed Perturbations

Based on MAT, we further propose Adversarial Training with Low-rank Decomposed perturbations (LDAT). Analogous to MAT, LDAT still shares one perturbation in a specific domain. Moreover, LDAT imposes a low-rank constraint on the perturbation itself, which corresponds to the dimension of the input scale. Technically, we obtain the domain-wise low-rank perturbation matrix $\delta \in \mathcal{R}^{N\times N\times C}$ by multiplying two matrices: $\delta = AB$. Here $A \in \mathcal{R}^{N\times l\times C}$ and $B \in \mathcal{R}^{l\times N\times C}$, l is a hyperparameter and $l \ll N$. Since $\text{rank}(AB) \leq \text{rank}(A)$ and $\text{rank}(AB) \leq \text{rank}(B)$ hold for arbitrary matrices A, B, we have $\text{rank}(\delta) \leq l$. Therefore LDAT reduces the rank of the perturbation from a large value N to a relatively small value l (see the third column in Figure 1 for illustration). The formal definition of the LDAT objective is:

$$\min_f \sum_e E_{(x,y)\sim\mathcal{P}_e(x,y)}[\mathcal{L}(f(x+\delta^e), y], \text{ s.t. } \delta^e = A^{e*}B^{e*}, \|\delta^e\|_p \leq \epsilon, \tag{8}$$

where

$$A^{e*},\ B^{e*} = \underset{A^e,B^e}{\text{argmax}} E_{(x,y)\sim\mathcal{P}_e(x,y)}[\mathcal{L}(f(x+A^eB^e), y],\ A^e \in \mathcal{R}^{N\times l\times C},\ B^e \in \mathcal{R}^{l\times N\times C}. \tag{9}$$

Comparison to MAT, LDAT xxx

4.3 Theoretical Analysis

In this part, we theoretically explain why the domain-wise perturbation proposed in MAT and LDAT can help to improve the robustness of the model against spurious correlations following [2] and [17]. In general, we prove that MAT and LDAT can prevent the model from relying more on the spurious

features to make predictions as the spurious correlations in the training data increase. Consequently, the model trained with MAT or LDAT will generalize better on OOD data.

Notations. Let $x \in \mathcal{X}$ denote the random data in the input space $\mathcal{X}$ and $y \in \mathcal{Y}$ denotes the target random data in label space $\mathcal{Y}$. For simplicity, let $\mathcal{Y} \in \{1, -1\}$ in this section. Let $\mathcal{D}$ denote the distribution over $\mathcal{X} \times \mathcal{Y}$. Let x_{inv} and x_{sp} denote the invariant features and the spurious features respectively. Also for simplicity, assume that there exists an identity mapping $\Phi : \mathcal{X}_{inv} \times \mathcal{X}_{sp} \to \mathcal{X}$ such that each $D \in \mathcal{D}$ is induced by a distribution over $\mathcal{X}_{inv} \times \mathcal{X}_{sp}$ (so x can be denoted as $x = (x_{inv}, x_{sp})$). Let x_{sp} take values in $\{+\beta, -\beta\}$ for some $\beta > 0$.

A Simple OOD Task. Consider a simple OOD task where we have two training domains representing the grass and desert background respectively. Both domains have two classes: the cow class and the camel class. In the grass/desert domain, the cow/camel class predominates. During test-time, the correlation between labels and backgrounds flips. We can abstract this cow-camel dataset into the following model: a training dataset $\mathcal{S}$ with four groups of data points drawn from the four quadrants of the feature space $\{-1, +1\} \times \{-\beta, +\beta\}$ respectively. We set the invariant features $x_{inv} = y$ and the spurious features x_{sp} to be $y\beta$ with probability $p \in [0.5, 1)$ and $-y\beta$ with probability $1 - p$. For example, the data points of the cow class in the grass and the desert domain can be expressed as $x = (1, \beta)$ and $x = (1, -\beta)$ respectively. Note that p measures the intensity of spurious correlations in a certain environment. Specifically, in the desert domain, the ratio of the camel samples is p and as for the grass domain, the ratio of the cow samples is p. It is easy to know that when $p = 0.5$, there are no spurious correlations between the labels and the background in this dataset.

Consider a linear classifier $h(x) = w_{inv}x_{inv} + w_{sp}x_{sp}$. Following [17], let us consider MAT/LDAT trained with gradient descent algorithm stopped in finite time t. In order to characterize the dependence of the model on spurious features during the training process, we investigate the convergence rate of $\frac{w_{sp}\beta}{w_{inv}(t)x_{inv}}$ to 0 on the above toy dataset. The value $\frac{w_{sp}\beta}{w_{inv}(t)x_{inv}}$ denotes the ratio between the output of the spurious component to that of the invariant component. We will prove that after adding the domain-wise perturbations in finite-time-stopped gradient descent, the lower bound of the convergence rate of this ratio does not increase monotonically with p. Hence, the model will not learn a large prediction weight based on spurious features even if the spurious correlation is strong (p is large).

In the following theorem, we denote the domain-wise perturbation in MAT/LDAT as δ. Theorem 4.1 applies to both MAT and LDAT since they both use domain-wise perturbations. See appendix B for a formal statement and full proof of 4.1.

Theorem 4.1. ***(informal)*** *Let $\mathcal{H}$ be the set of linear classifiers $h(x) = w_{inv}x_{inv} + w_{sp}x_{sp}$. Consider the above 2-D OOD dataset $\mathcal{S}$. Assume that the empirical distribution of x_{inv} given $x_{sp} \cdot y > 0$ is identical to the empirical distribution of x_{inv} given $x_{sp} \cdot y < 0$. δ is the optimal perturbation obtained by optimizing (7) or (9). Let $w_{inv}(t)x_{inv} + w_{sp}(t)x_{sp}$ be initialized to the origin, and trained with MAT/LDAT to minimize the exponential loss on $\mathcal{S}$. Then, for any $(x, y) \in \mathcal{S}$, we have:*

$$\Omega\left(\frac{\frac{1}{\beta-\delta y}\ln\left[\frac{c_1+p}{c2+p^{\frac{1}{2}+\epsilon}(1-p)^{\frac{1}{2}-\epsilon}}\right]}{M\ln(t+1)}\right) \leq \frac{w_{sp}(t)\beta}{|w_{inv}(t)x_{inv}|}, \tag{10}$$

where $\epsilon := \frac{\delta y}{2\beta}$ is a real number close to 0, $c := \frac{2(2M-1)}{\beta^2}$, $c_1 := \frac{2(2M-1)}{\beta(\beta-\delta y)^2 e^{w_{inv}\delta y}}$, $c_2 := \frac{2(2M-1)}{\beta(\beta^2-\delta^2)e^{w_{inv}\delta y}(\frac{\beta-\delta y}{\beta+\delta y})^{\frac{\beta+\delta y}{2\beta}}}$. $M = \max_{x\in S} \hat{w} \cdot x$ denotes the maximum value of the margin of the max-margin classifier $\hat{w}$ on $\mathcal{S}$. Therefore, the lower bound of the convergence rate does not increase monotonically with p under the condition that when $c_2 + (\frac{c_1}{1-c_1})^{\frac{1}{2}+\epsilon}(c_1 - 2\epsilon) < 0$, the lower bound does not increase with p monotonically.

To sum up, since we can prevent this lower bound from growing monotonically with p, we accelerate the convergence rate of $\frac{w_{\text{sp}}(t)\beta}{w_{\text{inv}}(t)x_{\text{inv}}}$ to 0 when there is stronger spurious correlation (larger p). Recall that the ratio $\frac{w_{\text{sp}}(t)\beta}{w_{\text{inv}}(t)x_{\text{inv}}}$ reflects the degree of reliance on spurious features. Therefore, a faster convergence of this ratio to 0 (smaller lower bound) means that the model will end up relying less on the spurious correlations in finite time. In other words, the OOD robustness can be enhanced by using domain-wise perturbations.

Remark. Here, we demonstrate that MAT and LDAT show stronger OOD robustness comparing to ERM. We compare the result in Theorem 4.1 to that in the Theorem 2 in [2]. According to the Theorem 2 in [2], even if the max-margin classifier does not rely on the x_{sp} for any level of spurious correlation $p \in [0.5, 1)$, ERM trained by gradient descent stopped in finite time still fails to avoid using spurious features. Moreover, when conducting ERM with finite-time-stopped gradient descent, the lower bound of the convergence rate of $\frac{w_{sp}\beta}{|w_{inv}x_{inv}|}$ to 0 is

$$\Omega\left(\frac{\ln \frac{c+p}{c+\sqrt{p(1-p)}}}{M \ln t}\right) \leq \frac{w_{sp}(t)\beta}{|w_{inv}(t)x_{inv}|}, \tag{11}$$

where M and c follow the definition in Theorem 4.1. This lower bound grows monotonically with p, thus ERM will have slower convergence for larger spurious correlations. However, with domain-wise perturbations, we can slightly modify the lower bound so that it does not increase monotonically with the spurious correlation p.

5 Experiments

5.1 Experimental Setup

We conduct experiments on the DomainBed benchmark [9], a testbed for OOD generalization that implements consistent experimental protocols across various approaches to ensure fair comparisons. We evaluate on PACS [12], OfficeHome [13], VLCS [14], NICO [15], and Colored MNIST [1]. There are several changes in our experimentation setting comparing to DomainBed:

1. **Backbone Network.** We use ResNet-18 [18] as our backbone network for datasets excluding Colored MNIST instead of ResNet-50 used in [9] for efficiency.
2. **Hyperparameter Search Space.** We use a smaller hyperparameter search space than [9]. We conduct a random search of 8 trials for PACS, OfficeHome, and VLCS while 6 trials for NICO and Colored MNIST in the hyperparameter search space, instead of 20 trials adopted in [9] for feasibility. See Appendix A.2 for more details.

Model Selection Strategy. Since hyperparameter choice has a significant impact on the OOD performance, it is critical to use appropriate model selection method. For PACS, OfficeHome, and VLCS datasets, we use training-domain validation proposed in [9] since it is more in line with the OOD scenario. For NICO, we adopt OOD validation following [19]. For Colored MNIST, we use test-domain validation [9] since it can enlarge the gaps in OOD performance among the algorithms while the gap induced by training-domain validation on Colored MNIST is marginal. More details can be found in Appendix A.2.

Hyperparameters for MAT and LDAT. To retain low-rank structures in perturbations, we set the upper bound of the search space of the perturbation number k in MAT to be 20. Similarly, the upper bound of the rank of the perturbation used in LDAT l is 20. The complete setup of the hyperparameters for MAT and LDAT is provided in Appendix A.2.

5.2 OOD Performance on Benchmark datasets.

Table 2 summarizes the results on the five OOD datasets. The results of other approaches for PACS, OfficeHome, NICO, and Colored MNIST datasets are adopted from [19]. The results on VLCS of other algorithms are missing (denoted as "-") because [19] does not experiment on this dataset.

Comparison with ERM and Sample-wise AT. From Table 2, we observe that both MAT and LDAT outperform ERM (on both our runs and the results in [19]) and AT on average. In particular, MAT achieves consistently better results than ERM on all five datasets. Additionally, the average performance of AT is worse than ERM, which is consistent with our observations in section 3.2.

Comparison with Existing SOTA Approaches. Although the results from [19] uses a different training protocol from ours: they use a larger hyperparameter search space and 20 random search for the hyperparameter combinations, the comparison between ERM ([19]) and ERM (our runs) indicates that their corresponding performances are close. A similar comparison has been made in [20]. We find that MAT outperforms all previous algorithms and LDAT ranked fourth among all methods, merely after VREx [21] and IRM [1].

Table 2: Test accuracy (%) on OOD datasets within DomainBed benchmark using ResNet-18. Here 'avg^1' denote the average accuracy on xx datasets and 'avg^2 ' denotes xxx. The best results are in **bold**.

	Datasets						
Algorithm	PACS	OfficeHome	VLCS	NICO	Colored MNIST	avg^1	avg^2
ERM (Our runs)	81.7 $\pm$ 0.3	62.1 $\pm$ 0.1	74.4 $\pm$ 1.0	73.2 $\pm$ 1.9	28.1 $\pm$ 1.5	61.3	63.9
AT (Our runs)	82.6 $\pm$ 0.4	62.1 $\pm$ 0.3	**76.2** $\pm$ **0.3**	69.7 $\pm$ 1.6	29.1 $\pm$ 1.5	60.9	64.3
ERM[19]	81.5 $\pm$ 0.0	63.3 $\pm$ 0.2	-	71.4 $\pm$ 1.3	29.9 $\pm$ 0.1	61.5	-
RSC[22]	**82.8** $\pm$ 0.4	62.9 $\pm$ 0.4	-	69.7 $\pm$ 0.3	28.6 $\pm$ 1.5	61.0	-
MMD[23]	81.7 $\pm$ 0.2	63.8 $\pm$ 0.1	-	68.3 $\pm$ 1.8	50.7 $\pm$ 0.1	66.1	-
SagNet[24]	81.6 $\pm$ 0.4	62.7 $\pm$ 0.4	-	69.3 $\pm$ 1.0	30.5 $\pm$ 0.7	61.0	-
CORAL[25]	81.6 $\pm$ 0.6	63.8 $\pm$ 0.3	-	68.3 $\pm$ 1.4	30.0 $\pm$ 0.5	61.0	-
IRM[1]	81.1 $\pm$ 0.3	63.0 $\pm$ 0.2	-	67.6 $\pm$ 1.4	60.2 $\pm$ 2.4	68.0	-
VREx[21]	81.8 $\pm$ 0.1	63.5 $\pm$ 0.1	-	71.0 $\pm$ 1.3	56.3 $\pm$ 1.9	68.2	-
GroupDRO[26]	80.4 $\pm$ 0.3	63.2 $\pm$ 0.2	-	71.8 $\pm$ 0.8	32.5 $\pm$ 0.2	62.0	-
DANN[27]	81.1 $\pm$ 0.4	62.9 $\pm$ 0.6	-	68.6 $\pm$ 1.1	24.5 $\pm$ 0.8	59.3	-
MTL[28]	81.2 $\pm$ 0.4	62.9 $\pm$ 0.2	-	70.2 $\pm$ 0.6	29.3 $\pm$ 0.1	60.9	-
Mixup[29]	79.8 $\pm$ 0.6	63.3 $\pm$ 0.5	-	66.6 $\pm$ 0.9	27.6 $\pm$ 1.8	59.3	-
ANDMask[30]	79.5 $\pm$ 0.0	62.0 $\pm$ 0.3	-	72.2 $\pm$ 1.2	27.2 $\pm$ 1.4	60.2	-
MLDG[31]	73.0 $\pm$ 0.4	52.4 $\pm$ 0.2	-	51.6 $\pm$ 6.1	32.7 $\pm$ 1.1	52.4	-
MAT (Our work)	82.3 $\pm$ 0.5	**64.5** $\pm$ **2.1**	74.6 $\pm$ 0.8	74.2 $\pm$ 1.5	**65.4** $\pm$ **8.1**	**71.6**	**72.2**
LDAT (Our work)	82.6 $\pm$ 0.5	61.0 $\pm$ 0.9	75.3 $\pm$ 0.3	**74.4** $\pm$ **1.6**	52.5 $\pm$ 5.4	67.6	69.1

5.3 Empirical Understanding

Visualization. To empirically show that MAT and LDAT can reduce the reliance on spurious features, we visualize the pixel attention heatmap of ERM, AT, MAT, and LDAT on NICO dataset using [32]. The heatmap is calculated with the second-order gradients of the last layer of the network w.r.t the output for the corresponding class. It reflects the contribution of different components of the feature map to the prediction results. We pick the model with the best performance for each method. The results in 5.3 indicate that the model trained by MAT and LDAT focuses more on the object itself, while ERM and AT adopt the background information that spuriously correlates to the class to make predictions.

Parameter Analysis.

6 Conclusions

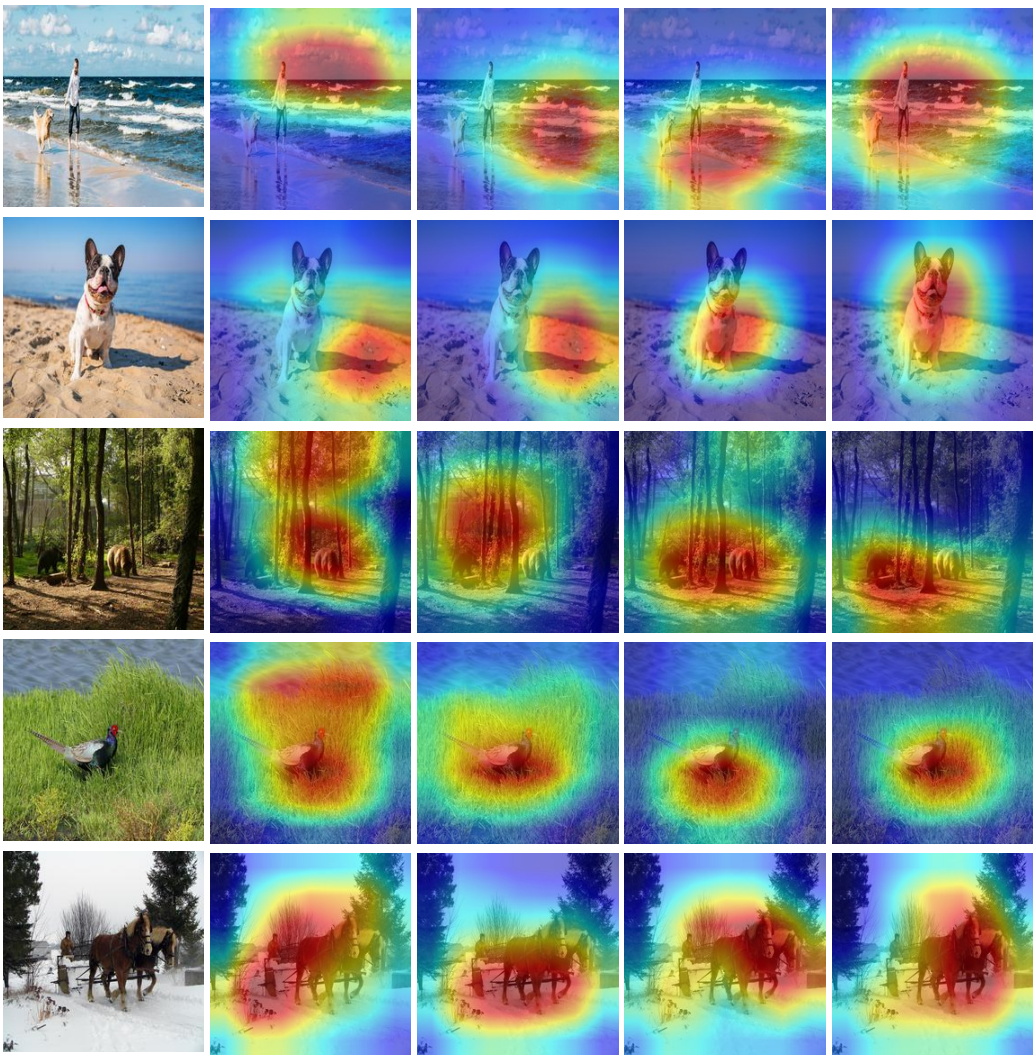
In this work, we empirically reveal the limitations of AT on OOD tasks. Due to the lack of constraints on the perturbation, AT fails to generalize well when facing large-scale perturbations which is close to the real-world OOD scenarios. Based on the observation that UAT shows a stronger robustness to large perturbations, we further proposed MAT and LDAT to exploit the low-rank structure to alleviate this issue. We validate the effectiveness of the proposed method on OOD tasks both theoretically and empirically.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

[3] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.



Origin ERM AT MAT LDAT

Figure 2: The pixel attention heatmap of ERM, AT, MAT and LDAT on NICO dataset. The redder part indicates that the model relies more on this part to make predictions.

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

[5] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*, 2018.

[6] Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhiming Ma. Improved ood generalization via adversarial training and pretraing. In *International Conference on Machine Learning*, pages 11987–11997. PMLR, 2021.

[7] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

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

[9] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[10] Charles Herrmann, Kyle Sargent, Lu Jiang, Ramin Zabih, Huiwen Chang, Ce Liu, Dilip Krishnan, and Deqing Sun. Pyramid adversarial training improves vit performance. *arXiv preprint arXiv:2111.15121*, 2021.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *Computer Science*, 2014.

[12] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[13] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[14] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

[15] Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110:107383, 2021.

[16] Klim Kireev, Maksym Andriushchenko, and Nicolas Flammarion. On the effectiveness of adversarial training against common corruptions. *arXiv preprint arXiv:2103.02325*, 2021.

[17] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 2018.

[18] Shaoqing Ren, Jian Sun, K He, and X Zhang. Deep residual learning for image recognition. In *CVPR*, volume 2, page 4, 2016.

[19] Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721*, 2021.

[20] Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *arXiv preprint arXiv:2110.10832*, 2021.

[21] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[22] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer, 2020.

[23] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.

[24] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap via style-agnostic networks. *arXiv preprint arXiv:1910.11645*, 2(7):8, 2019.

[25] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[26] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

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

[28] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017.

[29] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

[30] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*, 2020.

[31] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[32] Jacob Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section .
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[TODO]**
   (b) Did you describe the limitations of your work? **[TODO]**
   (c) Did you discuss any potential negative societal impacts of your work? **[TODO]**
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[TODO]**
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? **[TODO]**
   (b) Did you include complete proofs of all theoretical results? **[TODO]**
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[TODO]**
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[TODO]**
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[TODO]**
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[TODO]**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? **[TODO]**
   (b) Did you mention the license of the assets? **[TODO]**
   (c) Did you include any new assets either in the supplemental material or as a URL? **[TODO]**
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[TODO]**
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[TODO]**
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[TODO]**
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[TODO]**
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[TODO]**