Towards Better Robustness against Common Corruptions for Unsupervised Domain Adaptation

Zhiqiang Gao*[1,2], Kaizhu Huang†[1], Rui Zhang[2], Dawei Liu[1], and Jieming Ma[2]

[1] Duke Kunshan University, Kunshan, China
{Zhiqiang.Gao, Kaizhu.Huang, Dawei.Liu}@dukekunshan.edu.cn
[2] Xi'an Jiatong-Liverpool University, Suzhou, China
{Zhiqiang.Gao, Rui.Zhang02, Jieming.Ma}@xjtlu.edu.cn

Abstract

Recent studies have investigated how to achieve robustness for unsupervised domain adaptation (UDA). While most efforts focus on adversarial robustness, i.e. how the model performs against unseen malicious adversarial perturbations, robustness against benign common corruption (RaCC) surprisingly remains under-explored for UDA. Towards improving RaCC for UDA methods in an unsupervised manner, we propose a novel Distributionally and Discretely Adversarial Regularization (DDAR) framework in this paper. Formulated as a min-max optimization with a distribution distance, DDAR[1] is theoretically well-founded to ensure generalization over unknown common corruptions. Meanwhile, we show that our regularization scheme effectively reduces a surrogate of RaCC, i.e., the perceptual distance between natural data and common corruption. To enable a abetter adversarial regularization, the design of the optimization pipeline relies on an image discretization scheme that can transform "out-of-distribution" adversarial data into "in-distribution" data augmentation. Through extensive experiments, in terms of RaCC, our method is superior to conventional unsupervised regularization mechanisms, widely improves the robustness of existing UDA methods, and achieves state-of-the-art performance.

1. Introduction

Although Deep Neural Networks have achieved impressive performance for various applications [15, 19, 14, 39, 27, 34], they may not generalize well on new data due to the data distribution shift problem. One of such shift is called domain shift where data may come from a new target domain.Unsupervised Domain Adaptation (UDA) aims to address this domain shift with access to labeled source domain data and unlabeled target domain data. The fundamental objective is to infer the domain-invariant representations [9, 63] from data.

The vanilla UDA task assumes the data of the target domain is all clean without any corruption, which is unfortunately impractical in most cases. In other words, the target domain not only faces the label scarcity problem but also may suffer from the data corruption problem, including common corruption [16] and adversarial attack [30]. As such, recent studies have investigated how to achieve robustness for UDA. While most efforts [59, 3, 11] focus on applying Adversarial Training (AT) [29] to improve UDA's robustness against the malicious adversarial perturbations [48, 13] (i.e. small perturbations imperceptible to humans which can however easily fool the model), robustness against benign common corruption (RaCC)[16] surprisingly remains under-explored.

Common corruptions such as noise, blur, and digital defects that may occur in the real world are more relevant to practical applications [16]. For instance, autonomous cars should be able to identify pedestrians in unusual weather that would incur heavy noise from their sensors [45]. In recent studies, one of the main interests has focused on investigating data-dependent regularization methods [17, 62, 54, 6, 12, 40, 46, 22, 31].

However, conventional methods are not directly applicable to UDA. The primary reason is that most of the previous methods rely on supervised learning, while labels are not available in UDA tasks. Therefore, an unsupervised approach needs to be explored. On the other hand, data augmentation methods commonly come up with trade-offs between robustness and standard accuracy during or after

*Majority of the work was done at Duke Kunshan University.
†Corresponding author.
[1]The codes are available at https://github.com/gzqhappy/DDAR.

training. For instance, some heuristic transformations improve performance on some types of corruptions but reduce significantly the performance on clean images [42]; AT with conventional adversarial data (in pixel-space) suffers from over-smoothed decision boundary [56] and sharper training loss landscape [26], since strong "attack" introduced by augmented Out-of-distribution data in maximization stage.

Despite the importance of RaCC, only a few studies have investigated it in UDA. The pioneered study [57] reveals that adversarial data enlarging the domain shift can be a surrogate of common corruption. Since the adversarial data are generated in the pixel space, the training process has to rely on strong supervision provided by a noisy teacher, a pre-trained UDA model. As a result, an effective unsupervised regularization mechanism for UDA has not been investigated.

This paper investigates an unsupervised adversarial regularization framework to improve RaCC for UDA, named *Distributionally and Discretlly Adversarial Regularization (DDAR)*. Different from conventional classifier-tailored method [31, 51, 64, 37, 25, 24, 4], DDAR formulates a min-max optimization by using a distribution distance. The generated adversarial data are considered as the worst-case w.r.t. the latent distribution of natural data so that some other perturbed data (not only the worst ones w.r.t. classifier but those crucial for common corruptions) are also implicitly considered. As such, minimizing the distribution distance between natural data and adversarial data enables the model to generalize well on unknown types of common corruptions that share similar distribution with adversarial data. Moreover, to enable a abetter adversarial regularization, leveraging the insight from [31], we integrate an image discretization (ID) scheme into our optimization pipeline. ID transforms the pixel-space adversarial data into an "in-distribution" data augmentation which alleviates problematic regularization of AT.

To further improve the regularization performance, we update the distribution distance measurement to the Wasserstein distance on semantic space. Theoretically, we demonstrate that the proposed adversarial regularization certifies a tighter upper bound than previous methods by encouraging a better-estimated hypothesis-induced distribution distance. Empirically, compared with the previous unsupervised approach, such as the virtual adversarial training (VAT) [33] and adversarial feature desensitization (AFD) [4], our approach achieves much better generalization to unknown common corruptions by reducing the perceptual distance [32] between natural data and common corruptions, which can be considered as an effective indicator for RaCC.

In a nutshell, our key contributions are listed as follows:

1) We propose an unsupervised adversarial regularization method, working as a plug-in component, to enable robustness against common corruptions for UDA methods.

2) While it is theoretically well-founded, our novel regularization mechanism empirically certifies a better generalization over the multiple types of common corruptions, compared with previous unsupervised methods.

3) Through extensive experiments on UDA benchmarks, our method consistently improves RaCC of various UDA methods, achieving new state-of-the-art.

2. Related Work

2.1. Unsupervised Domain Adaptation

Among the current UDA methods, the adversarial domain adaptation methods not only achieve state-of-the-art performance but also offer a theoretical guarantee to reduce domain shift. The domain adversarial learning method approximates the $\mathcal{H}\Delta\mathcal{H}$-Divergence by using a domain classifier [9], which motivates a large body of heuristic learning algorithms of UDA. Recent studies devoted to learning a discriminative distribution alignment, which is to approximate a joint distribution alignment by conditioning the domain classifier on both features and corresponding predictions [35, 28, 49, 18]. The hypothesis adversarial learning method certifies a better distribution distance estimation by introducing two classifiers to measure the distribution discrepancy between two domains. The disagreement between two classifiers' predictions is used to detect the non-discriminative features which do not clearly belong to some categories. By playing the mini-max game with a feature extractor, two classifiers optimize the decision boundaries for alleviating intra-class domain discrepancy [43, 63].

2.2. Adversarial Training

The conventional adversarial training (AT) [30, 13] method has been widely studied as a promising approach to defending against adversarial attacks. It formulates a min-max optimization with the classification loss, where the adversarial examples are generated to maximize the loss to confuse the classifier, and then minimize the loss over the adversarial examples for robust learning. On the other hand, some successful methods can leverage the unlabeled data and have been applied to semi-supervised learning and UDA tasks [7, 1], such as Virtual Adversarial Training [33] that aims to smooth the output distribution around input data through AT. Moreover, generating adversarial examples by considering the inter-sample relationships, i.e. Feature-Scattering [60], or training with adversarial data by minimizing the distribution distance [4] also achieves promising performance for adversarial robustness.

2.3. Robustness against Common Corruption

Some studies focus on developing data augmentation techniques for implicit regularization. Augmix [17] is proposed to improve robustness against common corruption

and relies on a diverse set of predefined data augmentation methods. Borrowing the idea from AT, adversarial augmentation approaches propose to search augmentations adversarially, such as learning the random mixing factors for augmentation operation [54] employing augmentation policy network to produce hard augmentation policies [62], optimizing an image-to-image model to generate corrupted images [6]. Few studies investigate the supervised adversarial regularization (AR) method on intermediate feature layer [22], BatchNorm layer [46], and discrete visual coding space [31]. We further investigate an unsupervised AR method on the basis of image discretization in [31].

3. Methodology

3.1. Preliminaries

In UDA, n_s labeled examples denoted as $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ are drawn from the source distribution $\mathcal{S}$, and n_t unlabeled samples $D_t = \left\{x_j^t\right\}_{j=1}^{n_t}$ are drawn from the target distribution $\mathcal{T}$. Both $\mathcal{S}$ and $\mathcal{T}$ are defined on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the feature set and $\mathcal{Y}$ is the label in $\{1, \ldots, K\}$ in multi-class classification. The goal of a learning algorithm is to find a hypothesis $h : \mathcal{X} \mapsto \mathcal{Y}$ in the hypothesis space $\mathcal{H}$ with a low target risk $\epsilon_{\mathcal{T}}(h) = \mathbb{E}_{(x^t, y^t)\sim\mathcal{T}}\left[\ell\left(h\left(x^t\right), y^t\right)\right]$ with no access to the labels of $\mathcal{T}$, where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function. Additionally, the hypothesis is instantiated by a neural network: $h = g_\phi \circ f_\psi$ with parameter θ, where f_ψ is the feature extractor and g_ϕ is a classifier. $f = f_\psi(x)$ and $z = g_\phi(f)$ represent the produced feature and logits vector respectively.

According to the learning theories [5] of UDA, for any hypothesis $h \in \mathcal{H}$, the bound of target risk is given by

$$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda, \tag{1}$$

where $\lambda = \min\left[\epsilon_{\mathcal{S}}(h) + \epsilon_{\mathcal{T}}(h)\right]$, and

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) = 2 \sup_{(h,h')\in\mathcal{H}^2} \Big| \mathop{\mathbb{E}}_{x\sim\mathcal{S}} \mathrm{I}\left[h(x) \neq h'(x)\right] - \mathop{\mathbb{E}}_{x\sim\mathcal{T}} \mathrm{I}\left[h(x) \neq h'(x)\right] \Big|, \tag{2}$$

is $\mathcal{H}\Delta\mathcal{H}$-Divergence to measure the distribution distance. $\epsilon_{\mathcal{S}}(h)$ and $\epsilon_{\mathcal{T}}(h)$ represent the error of hypothesis h on the source and target domain respectively. $\mathrm{I}[a]$ is the indicator function, which is 1 if predicate a is true and 0 otherwise.

3.2. Distributionally Adversarial Regularization

Images destroyed by common corruptions will be significantly different from the original ones in contour and texture. These corrupted images will show significant distribution shifts compared to the original counterpart and are considered as out-of-distribution samples [42, 55]. However, the types and properties of common corruption are unknown during the training process. Therefore, it is essential to improve the generalization ability of the UDA model over unseen corruptions.

The classifier-tailored adversarial regularization (AR) method may not be the optimal strategy in our case. These methods only minimize the entropy-based loss on the adversarial data that are generated by worsening the classifier's predictions [51, 64, 37, 25, 24]. They ignore more kinds of adversarial data that might be effective surrogates for common corruption. Moreover, when a better generalization is achieved, the distribution distance between these adversarial data and the natural samples should be small.

Here, the proposed AR method takes advantage of the latent distribution information of the target domain to improve the generalization over unseen corruption. Specifically, we leverage the notion of distribution distance in UDA to define the optimization form of AR, such as $\mathcal{H}\Delta\mathcal{H}$-Divergence in Eq. 2, empirically estimated by a domain discriminator [9].

The process of generating adversarial data by maximizing a given loss function (inner optimization) can be considered as a sampling procedure. If the worst-case data of the entire distribution, not just those that fool the classifier, are sampled in this process, more kinds of adversarial data can be implicitly attained. Here, we define a distribution distance $D_\theta(\cdot, \cdot)$ on the latent space. Given the clean target domain data, i.e. $x^t \sim \mathcal{T}$, and initially sampled possible adversarial data, i.e. $x^{t,0} \sim \mathcal{T}^0$, $D_\theta\left(\mathcal{T}, \mathcal{T}^0\right)$ measures their distribution distance. By further sampling data that are more distant from $\mathcal{T}$, or dissimilar with x^t (determined by domain discriminator), the worst-case adversarial data with respect to the entire latent distribution are obtained by

$$\mathcal{T}^{adv} = \arg\max_{\mathcal{T}^0\in\mathcal{Q}} \left[D_\theta\left(\mathcal{T}, \mathcal{T}^0\right)\right], \tag{3}$$

where $\mathcal{T}^{adv}$ is the adversarial data distribution, i.e. $x^{t,adv} \sim \mathcal{T}^{adv}$, the initialized $x^{t,0}$ can be obtained by perturbing the natural data with Gaussian or uniform random noise, and $\mathcal{T}^0, \mathcal{T}^{adv} \in \mathcal{Q}$ is a feasible region of all possible perturbed adversarial data where $\mathcal{Q} \triangleq \{r : r \in B(x, \epsilon)\}$ with $B(r, \epsilon) \triangleq \{r : \|x^t - r\|_p \leq \epsilon\}$ being the l_p-ball at center x^t with radius ϵ.

Correspondingly, in order to further obtain better robust generalization over unseen corruption, the model is regularized by minimizing the distribution distance between the natural data and the adversarial data. The key insight is to enable the model to generalize well to the new distribution where the adversarial data are located, rather than only considering the error rate at a particular data point (commonly used in the classifier-tailored AR method). The optimization problem of the proposed AR is shown as follows:

$$\begin{aligned} &\min_\theta D_\theta\left(\mathcal{T}, \mathcal{T}^{adv}\right) \\ &\text{s.t. } \mathcal{T}^{adv} = \arg\max_{\mathcal{T}^0\in\mathcal{Q}} \left[D_\theta\left(\mathcal{T}, \mathcal{T}^0\right)\right], \end{aligned} \tag{4}$$

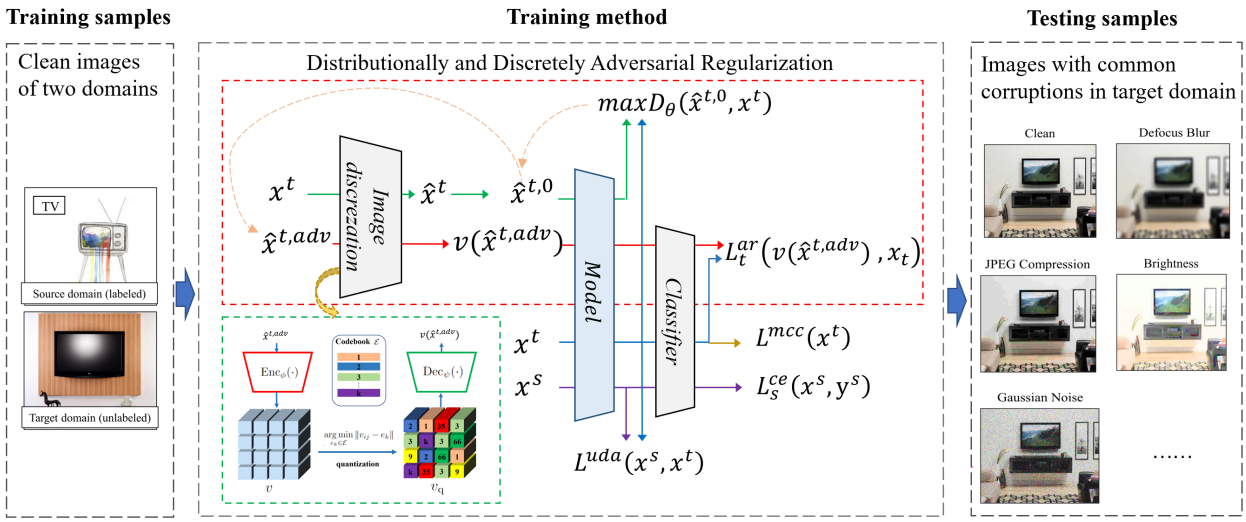

Figure 1. Overview of our *Distributionally and Discretely Adversarial regularization* (DDAR) method considering RaCC. During training, commonly occurring corruptions in the real world cannot be attained, instead only clean source and target data are available. In testing, the trained model is evaluated on common corruptions defined in [16], which includes 15 types of corruptions, and each corruption contains 5 severity. Our DDAR, working as a plug-in component, effectively improves the RaCC of UDA methods. Formulated as a min-max optimization with a distribution distance, DDAR generates adversarial data by maximizing the loss value of $D_\theta(\cdot,\cdot)$ and then regularizing the model by minimizing L_t^{ar}, which enables the model to generalize well on unknown corruptions. Meanwhile, in order to enable a better adversarial regularization method, an image discretization (ID) method [31] is leveraged to transform generated adversarial data into "in-distribution" data augmentation.

which can be used as a plug-in regularization term and jointly optimized with learning objects of UDA methods.

As such, our AR forms a min-max optimization by employing a distribution distance as a criterion. By considering the worst-case adversarial data w.r.t. natural data distribution as proxies for the unknown corruption, when the distribution distance between proxies and the natural data is reduced, our method allows the model to generalize better to the unknown corruption. We will demonstrate that our optimization goal is theoretically well-founded in Section 3.5. Empirically, this min-max optimization process can be guaranteed based on one crucial observation. Although $\mathcal{T}^0$ and $\mathcal{T}^{adv}$ are sampled from a small local region of $\mathcal{T}$ in $B(x,\epsilon)$, previous studies show that these distributions can be identified clearly. For instance, a model tends to overfit to the standard deviation of Gaussian noise used for training [22], and the adversarial attacks can be detected accurately by using maximum mean discrepancy [10].

3.3. Wasserstein Metric on Latent Space

Through the lens of UDA, widely investigating and improving the distribution distance promotes the performance of the learning algorithm. In our case, though there is an observed distribution shift between them, natural and adversarial data are paired, which is different from unpaired source and target domain samples. Furthermore, the perceptual distance [22, 32] on latent space has been shown as an effective indicator for RaCC. These observations motivate us to investigate a more advanced distance to the latent space for our adversarial regularization.

Wasserstein distance, a metric on the space of probability distributions, has been engaged to constrain the distribution distance between classifier-tailored adversarial data and natural data [47, 52]. The Wasserstein distance in the latent space considers the transportation cost c which quantifies the cost of transferring mass from (z, y) to (z^{adv}, y^{adv}):

$$c((z,y),(z^{adv},y^{adv})) = \frac{1}{2}\|z - z^{adv}\|_2^2 + \infty \cdot \mathbf{1}\{y \neq y^{adv}\}.$$

The transportation cost assumes an infinite value for data points with distinct labels, as our focus is solely on perturbation to the marginal distribution of Z. When the distance metric is defined in the semantic space by employing feature of last hidden layer, for inputs originating from the original space $\mathcal{X} \times \mathcal{Y}$, the transportation cost c_θ is denoted by

$$c_\theta((x,y),(x^{adv},y^{adv})) = c\left((g_\phi(f_\psi(x)),y),(g_\phi(f_\psi(x^{adv})),y^{adv})\right)$$

so as to assesses distance concerning the feature mapping $z = g_\phi(f_\psi(x))$.

For probability measures like $\mathcal{T}$ and $\mathcal{T}^{adv}$, both defined over $\mathcal{X} \times \mathcal{Y}$, we denote $\Pi(\mathcal{T}, \mathcal{T}^{adv})$ as their couplings, representing measures M satisfying $M(A, \mathcal{X} \times \mathcal{Y}) = \mathcal{T}(A)$ and $M(\mathcal{X} \times \mathcal{Y}, A) = \mathcal{T}^{adv}(A)$. Consequently, we define

our distance metric as follows:

$$D_\theta(\mathcal{T}, \mathcal{T}^{adv}) = \inf_{M \in \Pi(\mathcal{T}, \mathcal{T}^{adv})} \mathbb{E}_M \left[c_\theta((x, y), (x^{adv}, y^{adv})) \right].$$

3.4. Discretely Adversarial Regularization

The image discretization (ID) scheme [31] is employed to improve visual representation and alleviate the trade-off between robustness and accuracy. ID builds upon the insight that the adversarial data in Natural Language Processing tasks surprisingly will not hurt the accuracy and even improve the model generalizations [65, 20]. To achieve this goal, the perturbed images will be first transformed into discrete text-like inputs, i.e. visual words, which are then used for regularization.

VQGAN [8] is employed to perform the ID procedure. Specifically, given a continuous image $x \in \mathbb{R}^{H \times W \times 3}$, VQGAN consists of an encoder $\mathrm{Enc}_\phi(\cdot)$, a decoder $\mathrm{Dec}_\psi(\cdot)$, and a quantization function $\mathrm{q}_\mathcal{E}(\cdot)$. The encoder $\mathrm{Enc}_\phi(\cdot)$, which is a convolutional model, maps the input image x to intermediate latent vectors $v = \mathrm{Enc}_\phi(x) \in \mathbb{R}^{(h \times w) \times d}$, where h and w are the height and width of the intermediate feature map, and d is the latent dimension.

The quantization function $\mathrm{q}_\mathcal{E}(\cdot)$ learns a codebook $\mathcal{E} = \{e_k \mid e_k \in \mathbb{R}^d\}_{k=1}^K$, where each latent vector $v_{ij} \in \mathbb{R}^d$ is quantized to its nearest codebook entry e_k as follows:

$$v_\mathrm{q} = \mathrm{q}_\mathcal{E}(v) := \left(\arg\min_{e_k \in \mathcal{E}} \|v_{ij} - e_k\| \right) \in \mathbb{R}^{h \times w \times d},$$

where i, j represent each position in the feature map. The decoder $\mathrm{Dec}_\psi(\cdot)$ then reconstructs the image $\hat{x}$ from the quantized vectors v_q:

$$\hat{x} = \mathrm{Dec}_\psi(v_\mathrm{q}).$$

Training VQGAN can be conducted by minimizing the difference between the reconstructed image $\hat{x}$ and the original image x. Thus far, for a given continuous image x, we can obtain its corresponding discrete reconstruction $\hat{x}$. For simplicity, we denote the image discretization process as $\mathcal{V}$, resulting in $\hat{x} = \mathcal{V}(x)$.

Intuitively, this discretization process weakens the "attack" of the generated adversarial data. It transforms adversarial data into "in-distribution" data augmentation. In detail, when adversarial data, e.g. x^{adv}, is transformed into the latent vector v^{adv}, v^{adv} is likely to be far from v of the original sample x^0. However, when v^{adv} is further transformed into v_q^{adv} after the quantization step, the visual codes q^{adv} and q, found by v^{adv} and v separately, will not be as far apart as v^{adv} and v. As result, during training stage, employing the $\mathcal{V}(x^{adv})$ reconstruct from v^{adv} by using ID may alleviate the over-smooth and sharper training loss problem of conventional AR methods.

Leveraging the above insight, we employ ID as a predefined transformation for data augmentation. The detailed pipeline is shown in Fig. 1. First, we aim to transform the initial data distribution $\mathcal{T}^0$. This operation makes $\mathcal{T}^0$ more dissimilar from $\mathcal{T}$, leading to a better adversarial data generalization process. Consequently, $\hat{x}^{t,adv} \sim \widehat{\mathcal{T}}^{adv}$ and $\hat{x}^{t,0} \sim \widehat{\mathcal{T}}^0$ are sampled from a new discretely transformed initialization region $\hat{\mathcal{Q}}$ that is centered at $\hat{x}^t = \mathcal{V}(x^t)$. Specifically, $\hat{x}^{t,adv}$ is derived from $\hat{x}^{t,0}$, and its detail formulation is shown in Eq. 5 and 7. More importantly, discretization of all adversarial data leads to a discretely transformed adversarial data distribution $\mathcal{V}(\widehat{\mathcal{T}}^{adv})$, where $\mathcal{V}(\cdot)$ also represents a discretization function for a distribution. Here, VQGAN, as a preprocessor, will not be jointly trained with the regularization method. The detailed configuration can be seen in Section 4.1.

3.5. Model Analysis

3.5.1 Theoretical Analysis

We first establish the theoretical bound for robustness against common corruption for UDA in Theorem 1. We then analyze the relationship between our method and the previous classifier-tailored method and show that our method leads to a tighter upper bound in Proposition 1.

Theorem 1. *Let $D_\theta(\cdot, \cdot)$ be a distribution distance, $\mathcal{S}$ be a source distribution, and $\mathcal{T}$ be a target distribution, $\mathcal{T}^{adv}$ be an adversarial data distribution representing a worst-case distribution of $\mathcal{T}$ with respect to $D_\theta(\cdot, \cdot)$. For every $h \in \mathcal{H}$,*

$$\begin{aligned} \epsilon_\mathcal{T}^{cc}(h) \le \epsilon_\mathcal{S}(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda \\ + D_\theta\left(\mathcal{T}, \mathcal{V}(\widehat{\mathcal{T}}^{adv})\right) + \lambda^* \end{aligned} \tag{5}$$

$$\textit{s.t. } \widehat{\mathcal{T}}^{adv} = \arg\max_{\mathcal{T}^0 \in \hat{\mathcal{Q}}} \left[D_\theta\left(\mathcal{T}, \widehat{\mathcal{T}}^0\right) \right],$$

$$\hat{\mathcal{Q}} \triangleq \{r : r \in \hat{B}(\hat{x}, \epsilon)\}, \hat{B}(r, \epsilon) \triangleq \left\{r : \|\mathcal{V}(x^t) - r\|_p \le \epsilon\right\},$$

where $\epsilon_\mathcal{T}^{cc}(h)$ is the robust error against common corruption on the target domain, $\epsilon_\mathcal{S}(h)$ is source error, $d_{\mathcal{H}\Delta\mathcal{H}}(\cdot, \cdot)$ is $\mathcal{H}\Delta\mathcal{H}$-Divergence to measure the distribution distance, both λ and λ^* are small constants. $\hat{\mathcal{Q}}$ represents a feasible region of generated adversarial data transformed by image discretization, and $\widehat{\mathcal{T}}^0, \widehat{\mathcal{T}}^{adv} \in \hat{\mathcal{Q}}$, $\mathcal{V}(\widehat{\mathcal{T}}^{adv})$ represents the adversarial data distribution transformed by image discretization.

Proof Sketch: Borrowing the adversarial error learning bound from [4] the and generalization bound of UDA in [5], as shown in Eq. 1, we have

$$\begin{aligned} \epsilon_\mathcal{T}^{cc}(h) &\le \epsilon_\mathcal{T}(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{T}, \mathcal{T}^{aa}) + \lambda^* \\ &\le \epsilon_\mathcal{S}(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda \\ &\quad + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{T}, \mathcal{T}^{aa}) + \lambda^*, \end{aligned} \tag{6}$$

where the $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{T}, \mathcal{T}^{aa})$ measures the distribution distance between natural data and adversarial attacks. By replacing the $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{T}, \mathcal{T}^{aa})$ with our $D_\theta\left(\mathcal{T}, \mathcal{V}(\widehat{\mathcal{T}}^{adv})\right)$, a bound in Theorem 1 can be achieved.

Compared with the generalization bound of UDA in Eq. 1, Theorem 1 shows that the RaCC for UDA is upper bounded by additional two terms, a distribution distance term (the penultimate term), and a constant term (the last term). Armed with our distribution distance D and adversarial data distribution, our method certifies a tighter upper bound to improve RaCC for UDA than classifier-tailored [31] and distribution distance-based regularization [4].

Proposition 1. *Let $\mathcal{T}^{aa}$ be a data distribution of adversarial attacks of the target domain,*

$$D_\theta\left(\mathcal{T}, \mathcal{V}(\widehat{\mathcal{T}}^{adv})\right) \leq d_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{T}, \mathcal{V}(\widehat{\mathcal{T}}^{adv})\right) \leq d_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{T}, \mathcal{T}^{aa}\right),$$

where $D_\theta(\cdot,\cdot)$ is Wasserstein distance on latent space, and $\widehat{\mathcal{T}}^{adv}$ is an adversarial data distribution that is a worst-case of $\widehat{\mathcal{T}}$ with respect to $D_\theta(\cdot,\cdot)$.

Our approach enjoys two essential practices to certify a tighter bound of RaCC. First, the introduced adversarial regularization method leads to a better hypothesis-induced distribution distance $d_{\mathcal{H}\Delta\mathcal{H}}$. Since $d_{\mathcal{H}\Delta\mathcal{H}}(\cdot,\cdot)$ is induced from the difference between two hypotheses h and h', the adversarial attacks drawn from $\mathcal{T}^{aa}$ increase the sample complexity for learning good hypothesis [44] which enlarges the supremum. On the contrary, our method progressively augments the latent distribution without sacrificing the learning ability of hypotheses. As a result, the supremum in $d_{\mathcal{H}\Delta\mathcal{H}}$ becomes tighter and at the same time a more accurate distribution distance estimation is achieved. On the other hand, upgrading the distribution distance metric from $\mathcal{H}\Delta\mathcal{H}$-Divergence to Wasserstein distance on latent space also leads to a better distance estimation. A similar solution has also been investigated in Wasserstein Generative Adversarial Networks [2].

3.5.2 Empirical Analysis

To demonstrate the effectiveness of our method, we employ two widely studied methods, a classifier-tailored, i.e. VAT [33], and a distribution distance-based regularization, i.e. AFD [4], where detailed information will be given in Section 4.5. Meanwhile, we leverage the perceptual distance (PD) [32], an effective indicator for RaCC, for evaluation.

In [32], PD between data augmentation and common corruption is calculated to evaluate the heuristic data augmentation methods. In our case, we aim to evaluate the performance of the different regularization methods. More directly, we employ PD between natural data and common corruption as a metric, which is approximately estimated by

$$d_{\mathrm{PD}}\left(\mathcal{T}, \mathcal{T}^{cc}\right) \approx \left\| \mathbb{E}_{x^t \sim \mathcal{T}}[h(x^t)] - \mathbb{E}_{x^{t,cc} \sim \mathcal{T}^{cc}}[h(x^{t,cc})] \right\|_2,$$

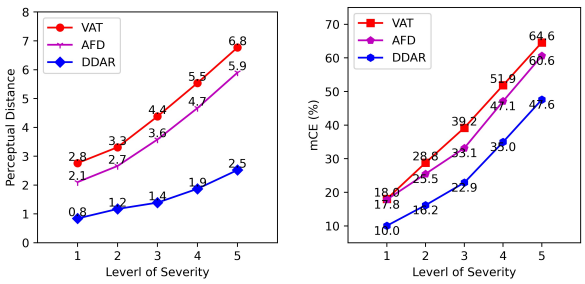

Figure 2. Illustration of Perceptual Distance (between natural and common corruption) and corruption error (mCE) of DDAR, VAT, and AFD under different levels of corruption severity. The models are evaluated on the A→D task in the Office-31 dataset.

where $x^{t,cc}$ represents the image with common corruptions in the target domain, which is drawn from the true corrupted target distribution $\mathcal{T}^{cc}$.

We compare d_{PD} and corruption error (i.e. mCE as shown in Section 4.1) among our DDAR, AFD, and VAT under different levels of corruption severity. As shown in Fig. 2, d_{PD} is positively correlated with mCE, where mCE increases with d_{PD}. Meanwhile, both DDAR and AFD reduce d_{PD} and perform better than VAT, since they regularize the model by reducing the distribution distance between natural and adversarial data. It can be observed that our method can significantly reduce d_{PD}, thus certifying a better RaCC for UDA.

3.6. Implementation and Learning Objective

Practically, we observe that the adversarial data generated by a discriminator achieve better performance than using the Wasserstein distance. Thus, we implement our method with the design shown in Fig. 3. The purpose is to ensure a min-mix optimization when a discriminator, D_d parameterized by θ_d, is involved in adversarial data generalization.

In order to engage a discriminator to generate adversarial data, the optimization for $\max_{\widehat{\mathcal{T}}^0 \in \hat{\mathcal{Q}}} D_d\left(\mathcal{T}, \widehat{\mathcal{T}}^0\right)$ is necessary. Meanwhile, two steps need to be resolved: 1) optimization for adversarial regularization (i.e., $\max_{\widehat{\mathcal{T}}^0 \in \hat{\mathcal{Q}}} D_\theta\left(\mathcal{T}, \widehat{\mathcal{T}}^0\right)$ and $\min_\theta D_\theta\left(\mathcal{T}, \mathcal{V}\left(\widehat{\mathcal{T}}^{adv}\right)\right)$ in

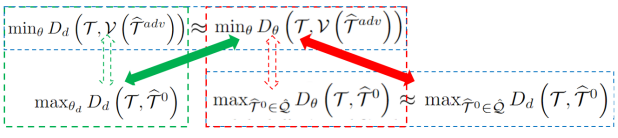

Figure 3. Illustration of the exchangeable min-max optimization objective, when a discriminator is introduced to generate adversarial data.

red dash line), and 2) adversarial learning for a discriminator (i.e., $\max_{\theta_d} D_d\left(\mathcal{T}, \widehat{\mathcal{T}}^0\right)$ and $\min_\theta D_d\left(\mathcal{T}, \mathcal{V}\left(\widehat{\mathcal{T}}^{adv}\right)\right)$ in green dash line). Assuming that Wasserstein distance is approximately equal to distribution divergence measured by a discriminator, the adversarial regularization can be done by optimizing $\max_{\widehat{\mathcal{T}}^0 \in \hat{\mathcal{Q}}} D_d\left(\mathcal{T}, \widehat{\mathcal{T}}^0\right)$ and $\min_\theta D_\theta\left(\mathcal{T}, \mathcal{V}\left(\widehat{\mathcal{T}}^{adv}\right)\right)$ (linked by red arrow), and training of discriminator can be done by optimizing $\max_{\theta_d} D_d\left(\mathcal{T}, \widehat{\mathcal{T}}^0\right)$.

As a result, the adversarial data are generated by

$$\hat{x}_j^{t,adv} = \arg \max_{\hat{x}_j^{t,0} \in \hat{B}(\hat{x}^t, \epsilon)} [\log D_d(f_\psi(x_j^t)) + (1 - \log D_d(f_\psi(\hat{x}_j^{t,0})))]. \quad (7)$$

The adversarial regularization loss L^{ar} is denoted as

$$L^{ar} = \frac{1}{n_t} \sum_{j=1}^{n_t} c_\theta \left(\left(x_j^t, y_j^t\right), \left(\mathcal{V}(\hat{x}_j^{t,adv}), y_j^{t,adv}\right)\right). \quad (8)$$

Meanwhile, the discriminator is trained with L^d loss:

$$L^d = \frac{1}{n_t} \sum_{j=1}^{n_t} [-\log D_d(f_\psi(x_j^t))) - (1 - \log D_d(f_\psi(\mathcal{V}(\hat{x}_j^{t,adv}))))]. \quad (9)$$

Additionally, since promoting discriminative feature learning is crucial for regularization, the minimum class confusion (MCC) [21] is introduced in our framework to enable discriminative learning which is crucial for unsupervised regularization. In summary, the learning object is given as:

$$L = \lambda_1 L^{src} + \mathcal{L}^{dd} + \lambda_2 L^{ar} + L^{mcc} + L^d, \quad (10)$$

where $\mathcal{L}^{src}$ is the supervised learning loss on the source domain, $\mathcal{L}^{dd}$ denotes distribution divergence loss between the source and target domain for UDA, λ_1 and λ_2 are balancing parameters of the loss.

4. Experiments

This section illustrates the experimental settings and results to evaluate our method. The **detailed results for Table 1-6**, **sensitivity analysis for λ_1 and λ_2 in Eq. 10 and ϵ in Eq. 5**, and **detail information of datasets** are shown in supplementary file.

4.1. Experimental Setup

Datasets and Corruption. For fair comparison, we adopt the commonly used benchmarks of UDA for evaluation, including **Office-31** [41], **Office-Home** [50], and **VisDa-2017** [36]. We create the corrupted datasets by using the corruption types defined by ImageNet-C [16], a widely used benchmark for corruption robustness. For each image, 15 corruption types exist with five levels of severity.

Implementation Detail. Following standard evaluation protocols for UDA, all labeled source and unlabeled target instances are used as training data. The trained classification model employs ResNet [15] as the backbone, followed by a bottleneck layer and a classifier. When combining our DDAR with given UDA methods, i.e. CDAN, the training parameter, and schedule follow the original setting of the particular UDA method. The discriminator, D_d used for generating adversarial data, is a simple three-layer neural network where each hidden layer is composed of a fully connected layer with the ReLu activation function, and the BatchNorm layer. The labels for the clean and adversarial data are set to 1 and 0 respectively. D_d will be trained 1 to 3 times in each iteration. Our method is a one-step adversarial regularization method with step size 0.5, where $p = 2$ and $\epsilon = 0.5$ in Eq. 5. The balance parameters, λ_1 and λ_2 in Eq. 10 are empirically set to 0.6 and 0.4 respectively for all datasets.

Image Discretization Configuration. As discussed in the original discrete adversarial training [31], the larger codebook size can achieve better generalization, while types of pre-training datasets have little impact. As such, we use a VQGAN model pre-trained on OpenImages [23] and will not fine-tune the model on the new dataset, where the codebook size is $K = 16384$ and the rest of the hyper-parameters of VQGAN model follow the setup in its pre-training stage [31].

Metrics for Corruption Robustness. The commonly used performance measure for robustness against common corruption is corruption error (CE) [16], which can be computed by the following:

$$\mathrm{CE}_T^f = \left(\sum_{t=1}^{5} E_{t,T}^f\right) / \left(\sum_{t=1}^{5} E_{t,T}^{\text{AlexNet}}\right),$$

where $E_{t,T}^f$ denotes the error rate of model f on the target domain data transformed by corruption T with severity t. The averaging of the CE values of the 15 corruptions, such as $\mathrm{CE}^f_{\text{Gausian Noise, ..,}}$, $\mathrm{CE}^f_{\text{Glass Blur}}$ is named as mean CE (mCE) [16] which quantifies the robustness of a model. Notably, in UDA, AlexNet is trained on a clean source domain and tested on a corrupted target domain.

4.2. Main Results

The performance over different corruptions and severity levels are shown in Table 1 and 2. These results validate the effectiveness of our method and set up a new baseline for unsupervised methods. In Table 1, models, trained with standard UDA methods, i.e. CDAN+TN and DCAN, underperform on corrupted samples as reported in [58]. By employing the pre-trained UDA model as a reference, heuristic data augmentation Augmix [17] and conventional adversarial data augmentation in pixel-space produced by DDG [58] can improve the RaCC. In contrast, our DDAR is an unsupervised adversarial regularization method that achieves better RaCC than DDG, yielding 10.8% and 3.4% improvement on Office-31 and Office-Home respectively. In Table 2, the accuracy and mCE of standard models and robust models trained by our method are shown on three datasets. Our method consistently improves the robustness of the standard model without hurting the standard accuracy significantly, especially on Office-31 and Office-Home datasets.

Method	Office-31	Office-Home
CDAN+TN	75.4	63.2
+ Augmix [17]	61.6	62.1
+ DDG [58]	59.8	56.7
DCAN	47.4	57.9
+ Augmix [17]	48.1	56.3
+ DDG [58]	41.3	53.5
MCC + DDAR	33.9	50.9
MCC + CDAN + DDAR	**30.5**	**49.9**
MCC + MDD + DDAR	34.0	54.1

Table 1. **mCE** (↓) on Office-31 and Office-home Dataset Under Common Corruptions (ResNet-50)

	Office-31		Office-Home		VisDa-2017	
	Acc	mCE	Acc	mCE	Acc	mCE
MCC	88.4	47.2	71.8	57.9	77.0	79.2
+DDAR	86.8	33.9	69.6	50.9	66.8	65.8
MCC+CDAN	**89.6**	43.7	**72.4**	55.8	81.8	65.8
+DDAR	88.9	**30.5**	71.4	**49.9**	75.1	55.6
MCC+MDD	89.3	46.8	71.6	59.6	**81.9**	63.1
+DDAR	88.6	34.0	67.4	54.1	76.8	**47.8**

Table 2. Standard accuracy (Acc) and **mCE** (↓) on Office-31, Office-Home, VisDa-2017 dataset (ResNet-50).

4.3. Ablation Study

Both training with data augmentation produced by ID (w/o adversarial perturbation) and our AR (w/o ID) improve the standard UDA model (CDAN+MCC). Integrating them together (DDAR) leads to a better performance. Additionally, as corruptions include the Gaussion noise which is commonly used for random start in AR methods [22, 20], we remove Gaussion noise to rule out the suspicion of having seen the test set. Empirically, perturbations on adversarial data are dominated by gradients generated in maximization step. As shown in Table 3, removing these noises (w/o random start) will not hurt the performances significantly.

.

Method	Avg.(↓)
CDAN+MCC	43.7
CDAN+MCC+DDAR	30.5
w/o random start	31.7
w/o ID	36.8
w/o adversarial perturbation	33.8

Table 3. **mCE** (↓) on corruptions of Office-31 dataset (ResNet-50)

4.4. Comparing with Orthogonal Methods

We further compare and combine DDAR with orthogonal methods, as shown in Table 4. Overall, our DDAR shows the superior performance to the comparison methods. Tent (finetune BN layers by entropy minimization) [53] is one representative Test-time adaptation (TTA) method which achieves comparable results with DDAR at epoch 1. However, since Tent is prone to collapse during training, its performances degrade significantly on some difficult transfer tasks. Furthermore, applying DDAR before conducting Tent produces better performances. On the other hand, DDAR outperforms AugMix (+KL) and DeepAugment (+KL) methods obviously. Moreover, combining with our regularization scheme ('W') and DDAR improves their performances significantly. Here, training hyper-parameters of Tent, AugMix and DeepAugment follows those of ImageNet in original paper.

Method	Avg.(↓)
CDAN+MCC+Tent (epoch 10)	40.2
+DDAR	32.7
CDAN+MCC+Tent (epoch 1)	30.5
+DDAR	**25.4**
CDAN+MCC+AugMix+KL	41.1
CDAN+MCC+AugMix+W	27.5
+DDAR	**26.8**
CDAN+MCC+DeepAugment+KL	38.4
CDAN+MCC+DeepAugment+W	26.0
+DDAR	**23.2**

Table 4. **mCE** (↓) on corruptions of Office-31 dataset (ResNet-50)

4.5. Comparing with Regularization Methods

By exchanging the adversarial data generation loss (Adv.) and regularization loss (Reg.), we conduct a comparison and ablation analysis which is shown in Table 5. Notably, the UDA learning method used in this section is CDAN+MCC.

Method	Adv.	Reg.	λ_2	1	2	3	4	5	**Avg. (↓)**
VAT	KL	KL	0.4	20.2	28.8	37.2	49.8	62.1	39.6
TRADES-1	KL	KL	1	19.5	27.5	36.0	49.1	61.9	38.8
TRADES-6	KL	KL	6	22.4	30.3	37.7	49.5	61.4	40.3
AFD	CE	CDAN	1	19.2	26.1	33.1	45.6	58.6	36.5
D+KL	D	KL	6	20.7	28.4	36.0	48.5	61.0	38.9
D+CDAN	D	CDAN	1	19.0	26.2	33.1	45.5	58.1	36.4
CE+W	CE	W	0.4	15.9	21.1	26.8	39.2	50.4	30.7
KL+W	KL	W	0.4	15.7	20.9	26.6	38.7	49.8	30.3
DAAR	D	W	0.4	**15.6**	**20.9**	**26.1**	**37.4**	**49.2**	**29.8**

Table 5. **mCE (↓)** of different level of severity on Office-31 Dataset (ResNet-50)

For a fair comparison, all the implemented methods are combined with our image discretization pipeline. VAT [33], TRADES-1, and TRADES-6 [61] rely on Kullback-Leibler divergence (KL) loss for adversarial data generation and regularization, where the main difference lies at the coefficients of regularization loss to control the strength of smoothness. In the original AFD [4], the adversarial data are produced by using cross-entropy (CE) loss, and the regularization follows the learning paradigms of DCGAN [38] and Wasserstein GAN [2]. In our unsupervised case, CE means the loss calculated with an online pseudo-label produced by the model itself. Meanwhile, CDAN, which is more powerful under the UDA task, is used to replace ADF's regularization loss. Additionally, λ_2 is the coefficient of regularization loss. Comparing the above implemented unsupervised methods, the performance of DDAR outperforms them significantly.

For an ablation study, 'D' and 'W' means using our adversarial data generation mechanism and regularization loss respectively. Comparing 'D'-based methods with DDAR, the results show that using 'W' in DDAR ('D'+'W') for regularization outperforms the one using KL ('D'+KL) and CDAN ('D'+CDAN) significantly. Meanwhile, a similar conclusion can be obtained from CE+'W' vs. AFD and KL+'W' vs. TRADES-6. Comparing 'W'-based methods with DDAR, we observe that our adversarial data generated by using the distribution distance in DDAR ('D'+'W') is better than the one using CE (CE+'W') and KL (KL+'W'). Meanwhile, a similar conclusion can be obtained from 'D'+KL vs. TRADES-6 and 'D'+CDAN vs. AFD.

4.6. Pixel-Space vs. Image Discretization

We compare all AR methods based on pixel-space perturbations and image discretization respectively in Table 6. These methods are implemented on the basis of the MCC+CDAN method and share the same setup with the methods in Section 4.5. The methods in 'Pixel-Space Perturbations' remove the ID and employ the generated adversarial data in pixel space for regularization directly. It can be observed that our methods both with or without using ID can outperform the other AR methods.

	Pixel-Space Perturbations		Image Discretization	
	Acc	mCE	Acc	mCE
VAT	88.2	40.8	88.9	40.4
AFD	88.4	39.0	87.5	37.3
DDAR	**89.1**	**36.8**	**88.9**	**30.5**

Table 6. Standard accuracy (Acc) and **mCE (↓)** for adversarial regularization methods based on pixel-space perturbations and image discretization on the Office-31 dataset (ResNet-50).

5. Conclusion

In this paper, we investigate how to achieve robustness against common corruption for UDA methods. A novel unsupervised method, called Distributionally and Discretely Adversarial Regularization (DDAR), is proposed to achieve a better generalization on unknown common. Our method defines a min-max optimization by leveraging a distribution distance as the criterion. It allows more types of possible adversarial data to be generated and then constrains the distribution distance between the natural and adversarial data to promote the generalization to unknown common corruptions. In addition, image discretization, which transforms adversarial data into "in-distribution" data augmentation, is introduced to enable a better adversarial regularization. Through comprehensive experiments and analysis, we show that our method is theoretically well-founded. It also empirically exceeds the previous unsupervised methods by a large margin.

Acknowledgement

The work was partially supported by the following: Jiangsu Science and Technology Programme under No. BE2020006-4, Suzhou Science and Technology Project-Key Industrial Technology Innovation (SYG202122), the XJTLU Postgraduate Research Scholarship (Grand No. PGRS1906004), and the XJTLU AI University Research Centre and Jiangsu (Provincial) Data Science and Cognitive Computational Engineering Research Centre at XJTLU.

References

[1] 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.

[2] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017.

[3] Muhammad Awais, Fengwei Zhou, Hang Xu, Lanqing Hong, Ping Luo, Sung-Ho Bae, and Zhenguo Li. Adversarial robustness for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8568–8577, 2021.

[4] Pouya Bashivan, Reza Bayat, Adam Ibrahim, Kartik Ahuja, Mojtaba Faramarzi, Touraj Laleh, Blake Richards, and Irina Rish. Adversarial feature desensitization. *Advances in Neural Information Processing Systems*, 34:10665–10677, 2021.

[5] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

[6] Dan A Calian, Florian Stimberg, Olivia Wiles, Sylvestre-Alvise Rebuffi, Andras Gyorgy, Timothy Mann, and Sven Gowal. Defending against image corruptions through adversarial augmentations. *arXiv preprint arXiv:2104.01086*, 2021.

[7] 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.

[8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[9] 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.

[10] Ruize Gao, Feng Liu, Jingfeng Zhang, Bo Han, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Maximum mean discrepancy test is aware of adversarial attacks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 3564–3575. PMLR, 2021.

[11] Zhiqiang Gao, Shufei Zhang, Kaizhu Huang, Qiufeng Wang, Rui Zhang, and Chaoliang Zhong. Certifying better robust generalization for unsupervised domain adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

[12] Chengyue Gong, Tongzheng Ren, Mao Ye, and Qiang Liu. Maxup: Lightweight adversarial training with data augmentation improves neural network training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2474–2483, 2021.

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

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

[18] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Unsupervised domain adaptation with hierarchical gradient synchronization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4043–4052, 2020.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[20] Maor Ivgi and Jonathan Berant. Achieving model robustness through discrete adversarial training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1529–1544. Association for Computational Linguistics, 2021.

[21] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 464–480. Springer, 2020.

[22] Klim Kireev, Maksym Andriushchenko, and Nicolas Flammarion. On the effectiveness of adversarial training against common corruptions. In *Uncertainty in Artificial Intelligence*, pages 1012–1021. PMLR, 2022.

[23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[24] Jingjing Li, Zhekai Du, Lei Zhu, Zhengming Ding, Ke Lu, and Heng Tao Shen. Divergence-agnostic unsupervised domain adaptation by adversarial attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8196–8211, 2021.

[25] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Pro-

gressive domain expansion network for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2021.

[26] Chen Liu, Mathieu Salzmann, Tao Lin, Ryota Tomioka, and Sabine Süsstrunk. On the loss landscape of adversarial training: Identifying challenges and how to overcome them. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[28] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in neural information processing systems*, pages 1640–1650, 2018.

[29] Chunchuan Lyu, Kaizhu Huang, and Hai-Ning Liang. A unified gradient regularization family for adversarial examples. *2015 IEEE International Conference on Data Mining*, pages 301–309, 2015.

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[31] Xiaofeng Mao, Yuefeng Chen, Ranjie Duan, Yao Zhu, Gege Qi, Shaokai Ye, Xiaodan Li, Rong Zhang, and Hui Xue. Enhance the visual representation via discrete adversarial training. 2022.

[32] Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. *Advances in Neural Information Processing Systems*, 34:3571–3583, 2021.

[33] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[34] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[35] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[36] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *CoRR*, abs/1710.06924, 2017.

[37] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020.

[38] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[40] Evgenia Rusak, Lukas Schott, Roland S Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 53–69. Springer, 2020.

[41] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, volume 6314 of *Lecture Notes in Computer Science*, pages 213–226. Springer, 2010.

[42] Tonmoy Saikia, Cordelia Schmid, and Thomas Brox. Improving robustness against common corruptions with frequency biased models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10211–10220, 2021.

[43] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.

[44] 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.

[45] Yu Shen, Laura Zheng, Manli Shu, Weizi Li, Tom Goldstein, and Ming Lin. Gradient-free adversarial training against image corruption for learning-based steering. *Advances in Neural Information Processing Systems*, 34:26250–26263, 2021.

[46] Manli Shu, Zuxuan Wu, Micah Goldblum, and Tom Goldstein. Encoding robustness to image style via adversarial feature perturbations. *Advances in Neural Information Processing Systems*, 34:28042–28053, 2021.

[47] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

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

[49] Hui Tang and Kui Jia. Discriminative adversarial domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5940–5947, 2020.

[50] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5385–5394. IEEE Computer Society, 2017.

[51] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5339–5349, 2018.

[52] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

[53] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

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

[55] Haotao Wang, Aston Zhang, Shuai Zheng, Xingjian Shi, Mu Li, and Zhangyang Wang. Removing batch normalization boosts adversarial training. In *International Conference on Machine Learning*, pages 23433–23445. PMLR, 2022.

[56] Yuxin Wen, Shuai Li, and Kui Jia. Towards understanding the regularization of adversarial robustness on neural networks. In *International Conference on Machine Learning*, pages 10225–10235. PMLR, 2020.

[57] Yifan Xu, Kekai Sheng, Weiming Dong, Baoyuan Wu, Changsheng Xu, and Bao-Gang Hu. Towards corruption-agnostic robust domain adaptation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 18(4), mar 2022.

[58] Yifan Xu, Kekai Sheng, Weiming Dong, Baoyuan Wu, Changsheng Xu, and Bao-Gang Hu. Towards corruption-agnostic robust domain adaptation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(4):1–16, 2022.

[59] Jinyu Yang, Chunyuan Li, Weizhi An, Hehuan Ma, Yuzhi Guo, Yu Rong, Peilin Zhao, and Junzhou Huang. Exploring robustness of unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9194–9203, 2021.

[60] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. *Advances in Neural Information Processing Systems*, 32, 2019.

[61] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 2019.

[62] Xinyu Zhang, Qiang Wang, Jian Zhang, and Zhao Zhong. Adversarial autoaugment. *arXiv preprint arXiv:1912.11188*, 2019.

[63] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413, 2019.

[64] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *Advances in Neural Information Processing Systems*, 33:14435–14447, 2020.

[65] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*, 2019.