Attribute-Guided Adversarial Training for Robustness to Natural Perturbations

Tejas Gokhale[1]*, Rushil Anirudh[2], Bhavya Kailkhura[2], Jayaraman J. Thiagarajan[2], Chitta Baral[1], Yezhou Yang[1]

[1] Arizona State University, [2] Lawrence Livermore National Laboratory
{tgokhale, chitta, yz.yang}@asu.edu, {anirudh1, kailkhura1, jjayaram}@llnl.gov

Abstract

While existing work in robust deep learning has focused on small pixel-level norm-based perturbations, this may not account for perturbations encountered in several real-world settings. In many such cases although test data might not be available, broad specifications about the types of perturbations (such as an unknown degree of rotation) may be known. We consider a setup where robustness is expected over an unseen test domain that is not i.i.d. but deviates from the training domain. While this deviation may not be exactly known, its broad characterization is specified *a* priori, in terms of attributes. We propose an adversarial training approach which learns to generate new samples so as to maximize exposure of the classifier to the attributes-space, without having access to the data from the test domain. Our adversarial training solves a min-max optimization problem, with the inner maximization generating adversarial perturbations, and the outer minimization finding model parameters by optimizing the loss on adversarial perturbations generated from the inner maximization. We demonstrate the applicability of our approach on three types of naturally occurring perturbations — object-related shifts, geometric transformations, and common image corruptions. Our approach enables deep neural networks to be robust against a wide range of naturally occurring perturbations. We demonstrate the usefulness of the proposed approach by showing the robustness gains of deep neural networks trained using our adversarial training on MNIST, CIFAR-10, and a new variant of the CLEVR dataset.

1 Introduction

The goal of *robust* machine learning models for tasks such as image classification is to make accurate predictions on *unseen* samples. The i.i.d. assumption is the simplest case in which unseen samples come from the same distribution as the training dataset. However, in most real-world situations, this assumption breaks down and so do models trained under the i.i.d. paradigm (Recht et al. 2018; Bulusu et al. 2020).

Most work on adversarial robustness has focused on pixel-level ℓ_p norm-bounded perturbations such as additive noise (2014; 2018; 2018; 2018). While such perturbations allow the use of tractable mathematical formulations, in practice, they are not the only perturbations that might be encountered at test time. For example, geometric transforms such as rotation, translation, or scaling of images, that are commonly encountered in the real world are not accounted for by pixel-wise ℓ_p bounded perturbations.

As such, images are parameterized by several unique attributes ranging from low-level information responsible for image formation like lighting, camera angle and resolution; to high-level semantic information like changes in background, size, shape, or color of objects in a scene. Perturbations along many of these attributes are irrelevant to tasks like image classification and are thus "semantics-preserving" perturbations. For instance, translating a digit inside an image in a digit classification task, or manipulating the shape of an object in a color classification task, will not result in a change in the true class-label. Yet, perturbations along these attributes are likely to cause models to fail when they are changed intentionally or otherwise (Xiao et al. 2020; Joshi et al. 2019; Liu et al. 2018). Shifts in such "nuisance attributes" typically result in large ℓ_p perturbations, posing significant challenges for existing pixel-level perturbation models. On the other hand, it is impractical to sample the entire attribute space effectively in order to guard against potential failures at test time.

In this work, we propose a robust modeling technique for image classification problems, which learns to generate new samples so as to maximize the exposure of the classifier to variations in the attribute space. Our approach falls under the broad category of adversarial training (Madry et al. 2017), and utilizes a min-max optimization setup, wherein the inner maximization step generates adversarial attribute perturbations while the outer minimization step identifies model parameters that reduce the task-specific loss (e.g., categorical cross entropy) under these perturbations. We find that the attribute-based specification produces models that can more effectively handle challenging real-world distribution shifts than standard ℓ_p norm-bounded perturbations (Qiao, Zhao, and Peng 2020). Furthermore, our proposed approach is flexible to support a wide-range of attribute specifications, which we demonstrate with three different use-cases:

1. Object-level shifts from a conditional GAN for adversarial training on a new variant of the CLEVR dataset;
2. Geometric transformations implemented using a spatial transformer for MNIST data; and
3. Synthetic image corruptions on CIFAR-10 data.

*Work performed during internship at LLNL.

Our contributions can be summarized as follows:

- We consider the problem of robustness under a set of specified attributes, that go beyond typically considered ℓ_p robustness in the pixel space.
- We present Attribute-Guided Adversarial Training (AGAT), a robust modeling technique that solves a min-max optimization problem and learns to explore the attribute space and to manipulate images in novel ways without access to any test samples.
- We create a new benchmark called "CLEVR-Singles" to evaluate robustness to semantic shifts. The dataset consists of images with a single block having variable colors, shapes, sizes, materials, and position.
- We demonstrate the efficacy of our method on three classes of semantics-preserving perturbations: object-level shifts, geometric transformations, and common image corruptions.
- Our method outperforms competitive baselines on three robustness benchmarks: CLEVR-Singles, MNIST-RTS, and CIFAR10-C.

2 Related Work

Most existing work on robustness deals with the problem of finding ℓ_p perturbations which focus on additive noise, with tractable mathematical guarantees of performance when test data falls within an ϵ-ball of the training distribution (Goodfellow, Shlens, and Szegedy 2014; Sinha, Namkoong, and Duchi 2018; Madry et al. 2018; Raghunathan, Steinhardt, and Liang 2018). Such perturbation are typically *imperceptible* to the human eye. As a result, there is an increasing interest in addressing challenges that arise from natural corruptions or perturbations (Hendrycks and Dietterich 2018) that are *perceptible* shifts in the data, more likely to be encountered in the real world. For example, (Liu et al. 2018) use a differentiable renderer to design adversarial perturbations sensitive to semantic concepts like lighting and geometry in a scene; (Joshi et al. 2019) design perturbations only along certain pre-specified attributes by optimizing over the range-space of a conditional generator. Our work focuses on building robust models against semantic, or more generally attribute guided concepts that may or may not exist in the training distribution, using a surrogate function.

ℓ_p-norm based robustness methods make no assumptions about the test distribution, except that the methods are guaranteed to be robust only inside the ϵ-ball of the training distribution (Volpi et al. 2018; Qiao, Zhao, and Peng 2020). Some recent approaches extend this notion to assume some access to data from the test distribution such as TTT (Sun et al. 2020) that achieves robustness for a test example by minimizing the cost of an auxiliary task for each test sample; and (Wong and Kolter 2020) learn a CVAE using possible corruptions one might encounter, to then guarantee robustness of a classifier within the learned perturbation set. For comparison, our method assumes access to no data from the test distribution, but only knowledge of a specification, which is the intended functionality of the system specified in human-understandable attributes. Under this challenging set-up, we show our method still outperforms existing robustness techniques on popular and standard benchmarks.

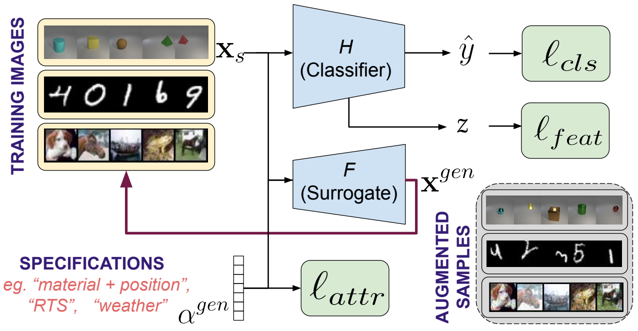

Figure 1: Overview of the problem setup and our attribute-guided adversarial training method.

3 Problem Setup

We begin by defining the classifier parameterized by a set of neural network weights, θ, as $H_\theta : \mathcal{X}_s \mapsto \mathcal{Y}$, where $\mathcal{X}_s$ denotes the space of the observed image data (or source) and $\mathcal{Y}$ denotes the label space for the task of interest.

Robustness to natural perturbations Our goal is to train an H_θ that is robust to *natural* perturbations, which are typically larger in magnitude than the *imperceptible* ℓ_p-bounded pixel-space perturbations, considered in the literature.

We consider a broad range of natural semantics-preserving perturbations that will not affect the predictions for the task under consideration – (a) **Object-level shifts**, where attributes of the object are manipulated so as to considerably change the appearance of the object, without changing the task label; such as changing the shape or size of an object in a color classification task. (b) **Geometric transformations**, where the test image may be scaled, rotated, and shifted in arbitrary ways; and (c) **Common image corruptions**, which may occur in the real world like fog, image compression artifacts, blurs, and other forms of noise.

Most of these perturbations do not naturally fall within small ℓ_p-norm ball deviations ($|\mathbf{x}-\tilde{\mathbf{x}}|_p \leq \epsilon$), for which most existing robustness methods are designed, and are bound to fail when the classifier encounters such data in the wild. However, making ϵ arbitrarily large in robustness formulations does not work in practice, since the image quality degrades significantly. Hence, we propose a new framework to design models that are robust to such natural perturbations.

4 Attribute Guided Adversarial Training

Let us denote an image by $\mathbf{x}_\alpha$ parameterized by a set of attributes α related to image formation (lighting, viewing angle, position) as well as abstract semantic information (color, shape, size, etc.). Manipulating images with new combinations of attributes that are not seen in the training set, requires access to the underlying physical generative processes, which is unrealistic. We do not assume direct access to such deterministic mechanisms.

Our goal is to train classifiers robust to natural perturbations along attributes in α that are specified *a priori*. Inspired by recent developments in robust optimization and adversarial training (Madry et al. 2018), we consider the following worst-case problem around N attributes of the training data:

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} \max_{|\hat{\alpha}_i - \alpha_i| \leq \epsilon} \ell(\theta; (\mathbf{x}_{\hat{\alpha}_i}, y_i)), \tag{1}$$

where $\ell(\cdot)$ is the cross-entropy loss.

The solution to worst-case optimization in Equation 1 guarantees good performance against test data that is distance ϵ away from the training data in the attribute space. In other words, we expect the model learned using (1) to be robust against ϵ-bounded natural perturbations. Interestingly, as we will empirically show later, models learnt using (1) perform better than existing pixel-level techniques even on ℓ_p-bounded imperceptible perturbations.

Although the structure of the attribute-guided adversarial training problem may look similar to standard adversarial training, we explain next why solving Eq (1) is significantly more challenging and requires us to make several algorithmic innovations. Note that (1) solves a min-max optimization problem, with the inner maximization generating natural perturbations by maximizing the classification loss over attribute space, and the outer minimization finding model parameters by minimizing the loss on natural perturbations of the training data generated from the inner maximization. The success of this method crucially relies on solving the inner optimization problem. Motivated by the standard adversarial training, one might be temped to approximately solve the re-parameterized inner optimization problem

$$\max_{|\delta|_p \leq \epsilon} l(\theta; (\mathbf{x}_{\alpha_i + \delta}, y_i)), \tag{2}$$

and generate the natural perturbations $\mathbf{x}^*_{\hat{\alpha}_i}$ using projected gradient descent (PGD) as:

$$\delta_i{}^* := \mathcal{P}_\epsilon(\delta_i - \lambda \nabla_\delta l(\theta; (\mathbf{x}_{\alpha_i + \delta}, y_i))), \tag{3}$$

where λ is the gradient step and $\mathcal{P}_\epsilon$ is projection on l_p ball of radius ϵ. However, there are two fundamental issues with this approach making it infeasible in practice: first, we cannot compute gradients as we do not have access to the attribute space; and second, we do not have access to the true generative mechanism conditioned on the attributes.

4.1 Proposed Approach

Surrogate Functions We propose to use differentiable surrogate functions parameterized by attributes to overcome the limitation described above. In other words, we have $\mathbf{x}_{\alpha+\delta} \approx F_\delta(\mathbf{x}_\alpha)$, where F_δ is differentiable. Typically, exact perturbations $\mathbf{x}_{\alpha+\delta} = F_\delta(\mathbf{x}_\alpha)$ can be performed for PGD attacks or other ℓ_p norm bounded attacks. However, in our case accessing the true generative process to manipulate images along α is not feasible. For example, we cannot rely on deterministic functions to manipulate semantic features in the image like size, shape or texture of an object. As a result, we resort to using *approximate* image manipulators in the form of surrogate functions which act as proxies to the true generative process. Depending on the type of attributes against which we wish to train for robustness, the surrogate function can take different forms:

- generative editing models for semantic perturbation that is learned from the training data itself,
- analytical functions for geometric transformations in the form of spatial transformers (STNs), or
- an analytical approximation (or tractable upper bound) of the natural perturbation space.

For example, if we want a classifier robust to unknown affine transforms then F is the spatial transformer layer parameterized by α which now represents 6 parameters controlling rotation, scale, and shift of the image.

Note that we do not assume access to any additional data other than the clean training dataset $\mathcal{X}_s$, and specification of the class of functions against which robustness is desired. While such surrogate functions only approximate the natural perturbations, we show that they are sufficient for enabling us to make classifiers more robust to natural perturbations.

Iterative Training Procedure Having access to attribute parameterized surrogate function, we aim to solve (1). Note, the success of the adversarial training is dependent on the quality of the generated perturbations. Thus, we aim to generate natural perturbations that have a larger coverage over the specified attribute space than the training samples images $\mathbf{x}_s$. Consider the classifier H_θ which outputs the predicted class $\hat{y}$ and intermediate features z, let the surrogate function F be parameterized by the attribute vector α. We propose an iterative training procedure called Attribute-Guided Adversarial Training (**AGAT**) detailed in Algorithm 1. Our algorithm has two objectives: to minimize the classification loss over input images and to maximize the divergence between the training samples and generated perturbations. Thus, the key idea here is to explore novel and hard images that only *vary along the specified attributes*. To achieve this, we impose a constraint that maximizes the distance between features of dataset images and perturbed images. Additionally, since we would also like to explore new regions in the attribute space, we impose a similar constraint on the attributes of perturbed images.

We express this constraint as the loss function given by:

$$\begin{gathered} \ell_{const} = \lambda_1 \ell_{feat} + \lambda_2 \ell_{attr}, \;\; \lambda_1, \lambda_2 \in (0, 1) \\ \text{where } \;\; \ell_{feat} = ||\mathbf{z} - \mathbf{z^{gen}}||_2^2, \text{ and } \; \ell_{attr} = ||\alpha - \alpha^{gen}||_2^2 \end{gathered} \tag{4}$$

To ensure that the generated images belong to the same class as the input image, we combine classification loss with respect to the ground truth label $\mathbf{y}$ with consistency regularization with respect to the predicted label $\hat{\mathbf{y}}$ of $\mathbf{x}$.

$$\ell_{cls} = \ell_{BCE}(\mathbf{y}, \mathbf{y^{gen}}) + \ell_{BCE}(\hat{\mathbf{y}}, \mathbf{y^{gen}}) \tag{5}$$

The overall loss function is computed as the Lagrangian:

$$\ell_{AGAT} = \ell_{cls} - \beta \cdot \ell_{const} \tag{6}$$

Intuitively, ℓ_{cls} encourages the augmented images to belong to the same class-label as the input image, while the constraint ℓ_{const} encourages the adversarial learning algorithm to perturb the image features as well as the attributes away from the input features and attributes. We first pretrain the classifier only on the source samples $\mathbf{x}_s$ for N_{pre} epochs. Then, we initiate our augmentation process. To generate new samples, we minimize Equation 6 and update the attribute vector for M update steps as:

$$\alpha^{gen} \leftarrow \alpha^{gen} - \mu \nabla \ell_{AGAT}. \tag{7}$$

Algorithm 1 Attribute-Guided Adversarial Training

```
Input: Source dataset $\mathcal{D}_S = \{\mathbf{x}_t, y_t\}_{i=t}^{T}$
Output: learned weights $\theta$
Initialize: $\theta \leftarrow \theta_0$, $\mathcal{D}_S^{aug} \leftarrow \mathcal{D}_S$
for $n = 1 \dots N_{epochs}$ do
    if $n < N_{pre}$ then
        for $t = 1 : T$ do
            $\theta \leftarrow \theta - \eta \nabla \ell_{cls}(\theta; (\mathbf{x}_t, y_t))$
    else
        if $n$ mod $N_{aug} = 0$ then
            for $t = 1 \dots T_{aug}$ do
                sample $(\mathbf{x}_t, y_t)_{t=1}^{T_{aug}}$ from $\mathcal{D}_S$
                $z_t, \hat{y}_t = H(\mathbf{x_t})$
                Initialize: $\alpha_t^{gen}$
                for $i = 1 \dots M$ do
                    $z_t^{gen}, \hat{y}_t^{gen} = H(\mathbf{x}_t, \alpha_t^{gen})$
                    $\mathbf{x}_t^{gen} \leftarrow f(\mathbf{x}_t, \alpha_t^{gen})$
                    $\alpha_t^{gen} \leftarrow \alpha_t^{gen} - \mu \nabla(\cdot \ell_{cls} - \beta \cdot \ell_{cons})$
                $\mathcal{D}_S^{aug} \leftarrow \mathcal{D}_S^{aug} \cup \mathbf{x}_t^{gen}$
        else
            for $(\mathbf{x}_t, y_t) \in \mathcal{D}_S^{aug}$ do
                $\theta \leftarrow \theta - \eta \nabla \ell(\theta; (\mathbf{x}_t, y_t))$
```

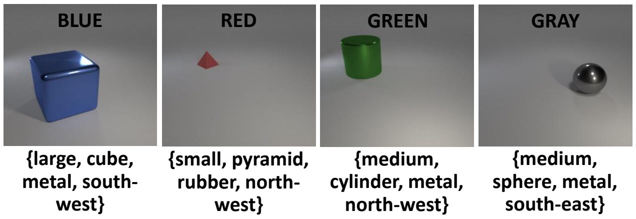
BLUE
RED
GREEN
GRAY
{large, cube, metal, south-west}
{small, pyramid, rubber, north-west}
{medium, cylinder, metal, north-west}
{medium, sphere, metal, south-east}

Figure 2: Examples of images from CLEVR-Singles and the color labels and (size, shape, material, position) attributes.

Finally, synthetic images are generated using the surrogate function $\mathbf{x}^{gen} \leftarrow F(\mathbf{x}, \alpha^{gen})$, and appended to the training data. This adversarial data augmentation is performed after every N_{aug} epochs during which T_{aug} images are generated. The total number of augmented samples is expressed as a percentage of the number of training samples so as to allow fair comparison across datasets and types of perturbations. The pseudocode for AGAT is shown in Algorithm 1.

The distinguishing factor for **AGAT** is that we perturb the attribute space and use surrogate functions to synthesize images, while previous adversarial augmentation protocols such as M-ADA (Qiao, Zhao, and Peng 2020) and GUD (Volpi et al. 2018) perturb only in the pixel-space, thus being restricted to ℓ_p perturbations. It is important to note that our method is agnostic to the choice of surrogate functions, which can take the form of additive noise, affine transformation in pixel-space, or conditional generative adversarial networks (Mirza and Osindero 2014) trained to transform an input image according to an input attribute vector.

5 Experiments

In this section, we introduce the three types of robustness specifications that we experiment with, along with details about the datasets, baselines, and metrics used for each.

Attribute	Train	Test
Size and Position	(small, NW), (medium, NE), (large, SE)	(small, SE), (medium, NW), (large, SW)
Material and Position	(metal, NW), (rubber, NW), (metal, NE), (rubber, NE)	(rubber, SW), (metal, SW), (rubber, SE), (metal, SE)

Table 1: The train and test splits for our experiments with semantic object-level perturbations for CLEVR-Singles.

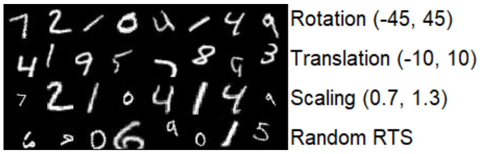

Figure 3: RTS-perturbed MNIST images.

5.1 Semantic Object-Level Perturbations

One class of real-world perturbations is when properties or attributes of images or objects in images change at test time. These changes do not affect the classification label, but significantly change the appearance of the image. For instance, consider the task of color classification in objects of varied shapes and textures. Here, *red metallic spheres* and *red rubber cubes* both belong to the class label "red", however may appear very different in their shapes and textures. Thus, if only *red metallic cubes* are seen during training, conventional classifier predictions for test images consisting of *red rubber cubes* can fail to generalize. Thus, although the class label is invariant to such semantic factors, robustness to perturbations along these factors is desirable.

Dataset: To study the problem of such object-level shifts along semantic factors of an image in a controlled fashion, we create a new benchmark called CLEVR-Singles[1] by modifying the data generation process from CLEVR (Johnson et al. 2017). We create images of single objects having one of eight colors, and use color classification as our task in this paper. Each object has four variable attributes that do not affect the color class of the image; these are: *shape* (cube, sphere, pyramid, or cylinder), *size* (small, medium, or large), *material* (rubber or metal), and *position* (northwest, southwest, northeast, southeast). While the objects are generated at continuous (X, Y, Z) world coordinates, we assign them a discrete position class for our experiments. Object-level perturbations can be made over these four attributes for our robustness experiments. In other words, it is known that one or more of {*shape, size, material, position*} of the image may change at test-time without knowing the magnitude or combinations of the change. We split the dataset based on a combination of attributes as shown in Table 1; for instance only certain combinations of size and position are observed in the training set, but robustness is expected from the color classifier on unknown combinations.

[1] Dataset: https://github.com/tejas-gokhale/CLEVR-Singles

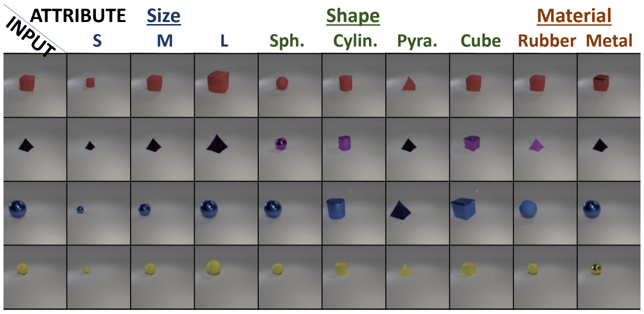

Figure 4: Images generated by AttGAN for the images in column 1, conditioned on attributes.

AttGAN as the Surrogate Function: Conditional generative adversarial networks (cGANs) have been shown to perform exceptionally well on image-to-image translation in various domains (Isola et al. 2017; Zhang et al. 2017; Karras, Laine, and Aila 2019). AttGAN (He et al. 2019) is one such conditional GAN which is trained to manipulate attributes of input face images. Thus given an image and a vector of desired attributes, AttGAN can manipulate the face image along the desired attribute dimensions. We leverage this powerful image manipulation technique as our surrogate function $\mathbf{x}^{gen} = F_{GAN}(\mathbf{x}, \alpha)$. Formally, we define the attribute vector to be a 13-dimensional binary hashcode with 1 and 0 indicating presence or absence of an attribute. For each experiment, we train the AttGAN on the training dataset outlined in Table 1 to generate 128x128 images, with a learning rate of $2\text{e}{-4}$ for 100 epochs on a single 16GB GPU. Examples of images generated by AttGAN are shown in Figure 4, when manipulating certain attributes such as size, shape, and material of the object.

Baselines: For the color classification task on CLEVR-Singles images, we compare against two pixel-level domain augmentation baselines: GUD (Volpi et al. 2018) which performs adversarial data augmentation to generate fictitious target domains, and M-ADA (Qiao, Zhao, and Peng 2020) which uses a meta-learning framework to generate multiple domains of samples. We also report the performance of a classifier directly without any adversarial training as a naive baseline. The same classifier architecture is used for each baseline for fair comparison. All models are trained for 15 epochs including pre-training epochs $N_{pre} = 5$, batch-size 64, and $M = 15$ update steps for adversarial augmentation. The number of augmented samples T_{aug} is 30% of the original source data, and augmentation interval N_{aug} is fixed at 2 epochs. For our model the coefficients in Equations 4, and 5 are: $\lambda_1 = 0.5, \lambda_2 = 0.5, \beta = 0.25$. The learning rates η, μ for the classifier and adversarial augmentation are both 5e-5.

Results: The test classification accuracies for different splits are reported in Table 2. We observe that our model is better than all baselines considered here, with a boost of 5 percentage points in accuracy on the harder experiment along *Material+Position*.

Method	Source	Size+Pos.	Mat.+Pos.
B	99.81	89.92	59.90
GUD (2018)	99.94	93.69	65.03
M-ADA (2020)	99.96	94.52	65.50
Ours	99.97	95.22	69.49

Table 2: Classification accuracy for color-classification on CLEVR-Singles. Source and target sets are split on *size+position* attribute for the third column, and *Material+Position* for the fourth column.

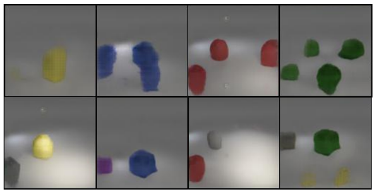

Figure 5: Visualization of the effect of weight β of the constraint loss ℓ_{const} on the generated images. Row 2 has higher β than Row 1. Illustration also shows that AttGAN is able to generate multiple objects (of same color for Row 1 and different colors for Row 2), though absent in training data.

Analysis: In Figure 5, we show 8 different examples generated by AttGAN during adversarial training. We can see the effect of the coefficient β, from the constraint loss ℓ_{const} in eq (6), in exploring the attribute space. An appropriately chosen value for β encourages useful perturbations without violating the class-label consistency cost ℓ_{cls} as seen in the top row of Figure 5. On the other hand, a higher β would mean a higher weight for exploring the regions (or combinations) in attribute space not seen in training. In the bottom row we see that a high β encourages novel attribute exploration at the cost of higher classification error as a result of generating objects with different colors within the same image. It is noteworthy that AttGAN is able to generate images with multiple objects, even when it trained on images with only a single object, thus demonstrating its suitability to explore novel attributes using the proposed AGAT training.

5.2 Geometric Transformations

Another common class of perturbations is geometric transformations, i.e. a composition of rotation, translation, and scaling of an image. These perturbations are common since cameras may capture a scene from different orientations, distances, and inclinations. It is well known that standard image classifiers are not robust to these common perturbations (Cohen and Welling 2014).

Dataset: We address this problem in the digit classification setting, with the training images from MNIST (LeCun et al. 1998), and the test images that are perturbed along rotation-translation-scale (RTS), as shown in Figure 3. We use the standard RTS setup (Jaderberg et al. 2015) with angle of rotation in $(-45, 45)°$, translation in $(-10, 10)$ pixels in both directions, and a scale factor in the range $(0.7, 1.3)$.

Method	**R**	**T**	**S**	**RTS**
B	84.44	27.67	95.76	21.91
GUD (2018)	86.08	29.09	97.89	23.10
MADA (2020)	87.37	29.25	**98.32**	22.68
PS (2020)	**87.86**	45.36	96.00	39.38
Ours (T_{aug}=30%)	84.93	**52.95**	96.11	**41.43**

Table 3: Results on the MNIST-RTS robustness benchmark for rotation (R), translation (T), scaling (S), and random combination (RTS).

Surrogate Function: The attributes of interest, α, consist of a 2×3 affine matrix that controls rotation, translation, and scale. To perform affine transformations on the image with a perturbed α, we use Spatial Transformer Networks (STN) (2015) which allow differentiable spatial manipulation of input images in a convolutional neural network, such as RTS and or general warping. The perturbed images are generated as: $\mathbf{x}^{gen} = F_{STN}(\mathbf{x}_s, \alpha)$.

Baselines: We compare the robustness performance to RTS perturbations with a naive baseline, denoted by (B), that is only trained on the standard MNIST dataset, and pixel-level perturbation methods MADA (Qiao, Zhao, and Peng 2020) and GUD (Volpi et al. 2018). Additionally, we also use the RTS perturbation sets generated by (Wong and Kolter 2020) (PS) and use them as augmented training samples. All models are trained for 12 epochs including pre-training epochs $N_{pre} = 5$, with a batch-size 64, and $M = 10$ update steps for adversarial augmentation. The number of augmented samples T_{aug} is 30% of the original source data, and augmentation interval N_{aug} is fixed at 10 epochs. Our model the coefficients in Equations 4, and 5 are: $\lambda_1 = 1, \lambda_2 = 1, \beta = 5$. The learning rate η for the classifier is 1e-4 and μ for the adversarial augmentation is 0.1.

Results: We report digit classification accuracies on the target test set containing only rotations (R), only translations (T), only skew (S), as well as a random combination of RTS. Our model performs well on all four metrics, and beats the perturbation sets (PS) even though their augmentation model has access to RTS perturbations during training. In particular, we observe a significant improvement compared with MADA and GUD, in the robustness on the translation experiment, which is the hardest task among the three.

Analysis The pixel-level perturbation methods still perform reasonably well on rotation and scale experiments in Table 3 because in each case the rotations/translations/scale are randomly sampled, resulting in several test examples that are very close to the training examples (with no RTS). In order to resolve this further, we study the performance by controlling the magnitude of R, T, and S in the test set. Figure 3 shows the bar-plots when the range of rotation is varied from $(-10, 10)$ to $(-60, 60)$, translation from $(-2, 2)$ to $(-12, 12)$ pixels, and scaling factor from $(0.9, 1.1)$ to $(0.5, 1.5)$. It can be observed that at higher severity of perturbation, our model (in blue) significantly outperforms all baselines. The model trained with Perturbations Sets (2020) (in gray) is competitive at lower severities.

$\mathbf{N_{aug}}$	$\mathbf{T_{aug}}$	**R**	**T**	**S**	**RTS**
1	10	84.12	43.33	96.65	34.12
	30	83.80	54.17	95.89	40.46
	50	84.49	59.97	96.29	47.62
	70	84.35	62.76	96.24	51.13
2	10	84.97	47.21	96.41	36.84
	30	84.93	52.95	96.11	41.43
	50	86.35	61.07	95.76	47.59
	70	84.59	62.75	95.79	50.28

Table 4: The effect of augmentation interval (N_{aug}) at different percentages of augmented samples (T_{aug}).

Loss	$\mathbf{T_{aug}}$	**R**	**T**	**S**	**RTS**
GT	10	85.41	29.48	96.74	23.57
	30	84.82	48.46	96.75	37.80
	50	84.17	52.44	95.86	41.82
	70	84.07	55.14	95.70	44.20
GT + CR	10	84.97	47.21	96.41	36.84
	30	84.93	52.95	96.11	41.43
	50	86.35	61.07	95.76	47.59
	70	84.59	62.75	95.79	50.28

Table 5: The effect of classification loss function at different percentages of augmented samples. GT denotes the first term and CR is consistency regularization in Eq (5).

We also analyze the effect of the number of augmented samples (T_{aug}) expressed as a percentage of the size of the training data, while controlling for the augmentation interval N_{aug}. As expected, larger number of augmented samples improve robustness even higher than in table 3 (which fixes number of additional augmented examples at 30% for all baselines). Larger augmentation intervals contribute positively at lower percentages of augmented samples.

Finally, we perform an ablation study with and without the consistency regularization defined in Eq (5) and show that the regularization indeed helps improve performance.

5.3 Common Image Corruptions

Image corruptions are another common class of perturbations. These can occur due to image digitization artifacts, weather, camera calibration, and other sources of noise.

Dataset The CIFAR10 dataset (Krizhevsky 2009) contains $50k$ training images belonging to 10 classes. Recently, CIFAR10-C (Hendrycks and Dietterich 2018) which contains image corruptions for CIFAR10 images, was proposed to benchmark robustness of image classifiers, with 4 major and 15 fine-grained categories of corruption: *Weather* (fog, snow, frost), *Blur* (zoom, defocus, glass, motion), *Noise* (shot, impulse, Gaussian), and *Digital* (JPEG, pixelation, elastic transform, brightness, contrast). There are five levels of severity of corruptions; we focus on the highest severity.

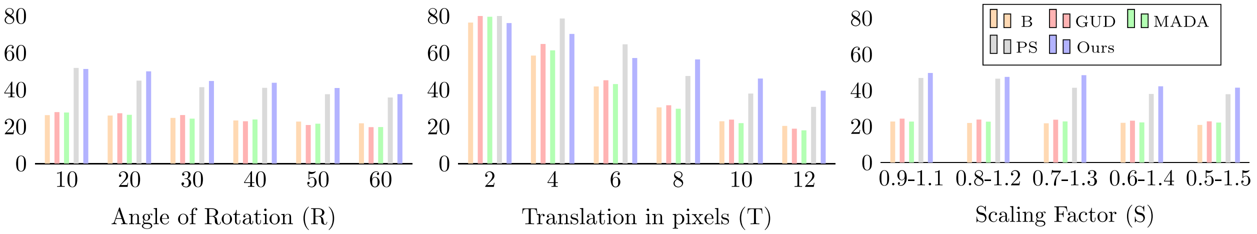

Figure 6: Comparison of random RTS accuracies when controlling each parameter to a max. value. Left: R, Center: T, Right: S

Surrogate Function: We use a general surrogate function – a composition of additive Gaussian noise and Gaussian blur filter parameterized by $\alpha = \{\alpha_1, \alpha_2\}$:

$$\mathbf{x}^{gen} = \frac{1}{\sqrt{2\pi\alpha_1^2}} e^{-\frac{\mathbf{x}^2}{2\alpha_1^2}} + n, \text{ where } n \sim \mathcal{N}(0, \alpha_2). \qquad (8)$$

We evaluate the performance gains using this surrogate function with the proposed AGAT training on the challenging CIFAR-10-C dataset.

Baselines: Test-Time Training (TTT) (Sun et al. 2020) is a recent approach in which a classifier is trained only on source data, but the test sample is utilized to update the classifier during inference. Adversarial Logit Pairing (ALP) (Kannan, Kurakin, and Goodfellow 2018), a technique for defending against adversarial attacks, and pixel-wise domain augmentation techniques MADA (Qiao, Zhao, and Peng 2020) and GUD (Volpi et al. 2018) are also considered as baselines. We use ResNet-26 (He et al. 2016) specially designed for CIFAR-10 (2015), with group normalization (Wu and He 2018) which is stable with different batch sizes. This acts as the naive classifier-only baseline (B). We also consider the classifier trained with an auxiliary self-supervised task of angle prediction (Gidaris, Singh, and Komodakis 2018) (B+SS). Our joint-training (JT) baseline is from TTT based on (Hendrycks and Dietterich 2018).

We compare three versions of our model: with additive noise only, with Gaussian filtering, and with a composition of Gaussian filter and noise. Our models are trained for 150 epochs including pre-training epochs N_{pre}=100, batch-size 128, and M=15 update steps for adversarial augmentation. The number of augmented samples is 30% of the original source data, and augmentation interval N_{aug} is fixed at 2 epochs. For our model the coefficients in Equations 4, and 5 are: $\lambda_1 = 0.5, \lambda_2 = 0.5, \beta = 0.25$. The learning rates η, μ for the classifier and adversarial augmentation are both 5e-5.

Results In Table 6 we show the classification accuracies on CIFAR10-C. It can be seen that our method consistently outperforms all baselines overall, and also on three of the four categories of corruptions (weather, blur, and digital). It is interesting to note that the ALP performance on the Noise category is distinctly greater than all previous methods, potentially because it is designed to defend against projected gradient descent adversarial attacks (Madry et al. 2017). ALP uses a similar loss function as Equation 5 to train the classifier, but still operates in pixel-space and does not perturb the attribute space. In Table 6 we also demonstrate that our models which uses only blur or only noise as surrogate are also better than previous state-of-the art. Note that the “noise only” model is in essence a pixel-level perturbation achieved by only perturbing along the variance parameter using AGAT training, and yet we see a significant boost in performance over all other pixel-level additive noise methods. Similarly, the “blur only” model also gives performance boosts on weather and digital categories, further indicating the general applicability of our AGAT training approach.

Method	**Src.**	**W**	**B**	**N**	**D**	**Avg.**
B	90.6	70.6	69.0	45.5	71.6	66.4
B+SS	91.1	70.6	68.5	48.7	69.7	67.0
GUD	-	71.7	59.2	30.5	64.7	58.3
MADA	-	75.6	63.8	54.2	65.1	65.6
JT	91.9	71.7	69.0	50.6	71.6	68.3
ALP	83.5	60.9	74.7	**75.4**	68.5	70.0
TTT	**92.1**	73.7	71.3	54.2	73.4	70.5
Ours (b. only)	91.3	**78.4**	**75.1**	49.3	**75.4**	70.0
Ours (n. only)	90.3	75.0	73.3	62.4	73.1	71.3
Ours	89.3	77.8	74.1	65.8	71.6	**72.3**

Table 6: Comparison of classification accuracies on CIFAR-10-C corruption categories (Weather, Blur, Noise, Digital). Our best scores are underlined; overall best are bold.

6 Conclusion

In this paper, we propose a new adversarial training strategy for robustness against large perturbations that are common in practical settings. Our adversarial training algorithm perturbs the attribute space to synthesize new images instead of pixel-level perturbations which are common to the robustness literature. The new CLEVR-Singles dataset that we have created can be used in future work for studying robustness to semantic shifts. We extensively evaluate AGAT training on three benchmarks and achieve state-of-the-art performance. We empirically show that AGAT is applicable to a three types of naturally occurring perturbations, and can be used with different classes of surrogate functions. AGAT can potentially be applied to a broad range of robustness problems not limited to classification.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, Lawrence Livermore National Security, LLC. This document was prepared as an account of the work sponsored by an agency of the United States Government. Neither the United States Government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or Lawrence Livermore National Security, LLC. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes. This work was supported by LLNL Laboratory Directed Research and Development project 20-ER-014 and released with LLNL tracking number LLNL-JRNL-814425.

Broader Impact

The concept of robustness is critical when it comes to deploying machine learning systems in practical settings where input signals may undergo perturbations due to weather (such as fog, smog, rain), digital corruptions in transmission, or changes in camera inclinations causing geometric transformations or artifacts such as defocusing, or motion blur. Our method for developing robust classifiers is broadly applicable if such classes of perturbations are known *a priori*.

Robustness research is also crucial for avoiding or removing unintended biases that may percolate from the training data into the classification model. Recent studies (Bolukbasi et al. 2016; Zhao et al. 2017; Hendricks et al. 2018) have shown that models trained on biased data can in fact amplify this bias when performing inference on test samples. We believe that work in the lines of AGAT could be potentially used for mitigating social biases due to biased training data, such as gender or racial biases.

References

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, 4349–4357.

Bulusu, S.; Kailkhura, B.; Li, B.; Varshney, P. K.; and Song, D. 2020. Anomalous Example Detection in Deep Learning: A Survey. *IEEE Access* 8: 132330–132347.

Cohen, T. S.; and Welling, M. 2014. Transformation properties of learned visual representations. *arXiv preprint arXiv:1412.7659* .

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations*.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, Z.; Zuo, W.; Kan, M.; Shan, S.; and Chen, X. 2019. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing* 28(11): 5464–5478.

Hendricks, L. A.; Burns, K.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, 793–811. Springer.

Hendrycks, D.; and Dietterich, T. 2018. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Advances in neural information processing systems*, 2017–2025.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2901–2910.

Joshi, A.; Mukherjee, A.; Sarkar, S.; and Hegde, C. 2019. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE International Conference on Computer Vision*, 4773–4783.

Kannan, H.; Kurakin, A.; and Goodfellow, I. 2018. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373* .

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4401–4410.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, University of Toronto* .

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11): 2278–2324.

Liu, H.-T. D.; Tao, M.; Li, C.-L.; Nowrouzezahrai, D.; and Jacobson, A. 2018. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. *arXiv preprint arXiv:1808.02651* .

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

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* .

Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12556–12565.

Raghunathan, A.; Steinhardt, J.; and Liang, P. 2018. Certified Defenses against Adversarial Examples. In *International Conference on Learning Representations*.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2018. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451* .

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3).

Sinha, A.; Namkoong, H.; and Duchi, J. 2018. Certifying Some Distributional Robustness with Principled Adversarial Training. In *International Conference on Learning Representations*.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A. A.; and Hardt, M. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*.

Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J. C.; Murino, V.; and Savarese, S. 2018. Generalizing to unseen domains via adversarial data augmentation. In *Advances in neural information processing systems*, 5334–5344.

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

Wu, Y.; and He, K. 2018. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xiao, K.; Engstrom, L.; Ilyas, A.; and Madry, A. 2020. Noise or Signal: The Role of Image Backgrounds in Object Recognition. *arXiv preprint arXiv:2006.09994* .

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 5907–5915.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2979–2989.