STUDYING CLASSIFIER ROBUSTNESS TO ATTRIBUTE-LEVEL SHIFTS

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

ABSTRACT

While existing work in robust deep learning has focused on pixel-level ℓ_p norm-based perturbations, this class of perturbations does not account for many real-world deviations such as object-level shifts, geometric transformations, and weather-related artifacts. We consider a setup where robustness is expected over an unseen test domain that deviates from training domain in terms of attributes, specified *a* priori. We propose an adversarial training approach which learns to generate new samples so as to maximize exposure of the classifier to the attributes-space. We introduce the CLEVR-Singles dataset that allows a controlled experimental setup to study robustness of classifiers to shifts along attributes.

1 INTRODUCTION

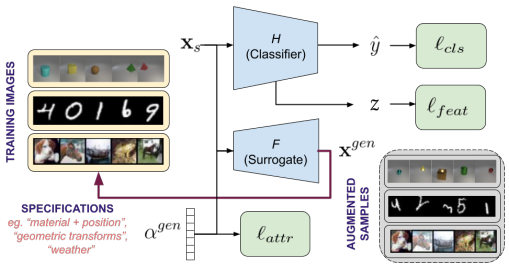

Figure 1: Overview of the problem setup and our attribute-guided adversarial training

The goal for *robust* machine learning models such as image classifiers is to make accurate predictions on *unseen* samples. In most real-world situations, the *i.i.d.* assumption (in which these unseen inputs are sampled from the same distribution as the training data) breaks down, as shown by Recht et al. (2018); Bulusu et al. (2020). To mitigate this risk, work on adversarial robustness has focused on pixel-level ℓ_p norm-bounded perturbations such as additive noise (Goodfellow et al., 2014; Sinha et al., 2018; Madry et al., 2018) to expose the model to novel input distributions. While such perturbations allow the use of tractable mathematical formulations, in practice, they do not account for many real-world deviations such as object-level shifts, geometric transformations, and weather-related artifacts.

Images can be 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 background, size, shape, or color of objects in a scene. Perturbations along some of these attributes may be irrelevant to image classification task, and do not change the true class label; 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. 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.

In this work, we propose a robust learning 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 generates images with perturbed attributes that are adversarial for the classifier, while the outer minimization solves for model parameters under these perturbations.

*Work performed during internship at LLNL. Contact Author tgokhale@asu.edu

To study robustness of classifiers to shifts in semantic attributes (such as materials and shapes of objects) under controlled settings, we create a new benchmark called the "CLEVR-Singles", based on the CLEVR dataset (Johnson et al., 2017). We find our approach to be demonstrably better than pixel-level adversarial training methods (Volpi et al., 2018; Qiao et al., 2020) for the CLEVR-Singles task, and is also flexible to support a wide-range of attribute shifts such as geometric transformations of MNIST images; and synthetic image corruptions of CIFAR-10 (Hendrycks & Dietterich, 2018).

2 Problem Setup: Robustness To Attribute-Level Shifts

Consider a classifier $H_\theta \colon \mathcal{X}_s \mapsto \mathcal{Y}$ with model parameters θ, 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. Our goal is to train H to be robust to *natural* perturbations, which are typically larger in magnitude than *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 and 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 include weather-related artifacts such as 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. 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.

3 Attribute Guided Adversarial Training (AGAT)

Our goal is to train classifiers robust to natural perturbations along attributes 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:

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} \max_{|\hat{\alpha}_i - \alpha_i| \leq \epsilon} \ell(\theta; (\mathbf{x}_{\hat{\alpha}_i}, y_i)), \quad \text{where } \ell(\cdot) \text{ is the cross-entropy loss.} \tag{1}$$

There are two fundamental issues with standard adversarial setting 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.

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 *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),
- analytical approximations (or tractable upper bound) of the natural perturbation space.

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.

Iterative Training Procedure 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, having two objectives: to minimize the classification loss over input images and to maximize the divergence between the training samples and generated perturbations. The key idea is to explore novel and hard images that *vary along the specified attributes*, by having access to a surrogate function parameterized by attributes. To achieve this, we impose a constraint that maximizes the distance between features of dataset images and perturbed images, and a similar constraint on the attributes of perturbed images.

$$\ell_{const} = \lambda_1 ||\mathbf{z} - \mathbf{z^{gen}}||_2^2 + \lambda_2 ||\alpha - \alpha^{gen}||_2^2, \quad \lambda_1, \lambda_2 \in (0, 1) \tag{2}$$

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 $\mathbf{x}$. The overall loss function is given by:

$$\ell_{AGAT} = \ell_{cls} - \beta \cdot \ell_{const}, \quad \text{where} \quad \ell_{cls} = \ell_{BCE}(\mathbf{y}, \mathbf{y^{gen}}) + \ell_{BCE}(\hat{\mathbf{y}}, \mathbf{y^{gen}}) \tag{3}$$

Algorithm 1 AGAT

Input: Source dataset $\mathcal{D}_S = \{\mathbf{x}_t, y_t\}_{i=t}^{T}$
Output: learned weights θ

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 $n \bmod N_{aug} = 0$ **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: α_t^{gen}
        for $i = 1 \dots M$ **do**
          $z_t^{gen}, \hat{y}_t^{gen} = H(\mathbf{x}_t, \alpha_t^{gen})$
          $\mathbf{x}_t^{gen} \leftarrow f(\mathbf{x}_t, \alpha_t^{gen})$
          $\alpha_t^{gen} \leftarrow \alpha_t^{gen} - \mu \nabla (\ell_{cls} - \beta \cdot \ell_{cons})$
        $\mathcal{D}_S^{aug} \leftarrow \mathcal{D}_S^{aug} \cup \mathbf{x}_t^{gen}$
    else
      for $(\mathbf{x}_t, y_t) \in \mathcal{D}_S^{aug}$ **do**
        $\theta \leftarrow \theta - \eta \nabla \ell(\theta; (\mathbf{x}_t, y_t))$

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. The pseudocode for AGAT is shown in Algorithm 1. We first pretrain the classifier only on the source samples $\mathbf{x}_s$ for N_{pre} epochs, and then initiate our augmentation process. The attribute vector is updated by minimizing Eq 3 for M gradient steps:

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

Synthetic images corresponding to α^{gen} are generated using the surrogate function $\mathbf{x}^{gen} = F(\mathbf{x}, \alpha^{gen})$, and appended to the training data. This adversarial data augmentation is performed after every N_{aug} epochs.

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 et al., 2020) and GUD (Volpi et al., 2018) perturb only in the pixel-space, thus being restricted to ℓ_p perturbations.

4 CLEVR-SINGLES 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). 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. The dataset is split along combinations of attributes as shown in Figure 2, and robustness is expected from the color classifier on unknown combinations.

5 EXPERIMENTAL RESULTS

Choice of Surrogate Function: AttGAN (He et al., 2019) is a conditional generative adversarial network that can manipulate an image along the desired attribute dimensions. We leverage this

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

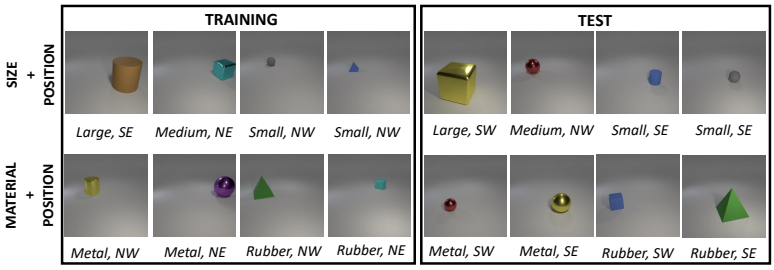

Figure 2: Sample images from the CLEVR-Singles dataset. The first row shows the train-test split on the attributes *size+position*, and the second row for *material+position*.

Method	Source	Size+Pos	Mat+Pos
Classifier Only	99.81	89.92	59.90
GUD	99.94	93.69	65.03
M-ADA	99.96	94.52	65.50
Ours (AGAT)	**99.97**	**95.22**	**69.49**

Table 1: Color-Classification accuracy with source and target sets split along *size+position* attribute for the third column, and *Material+Position* for the fourth column.

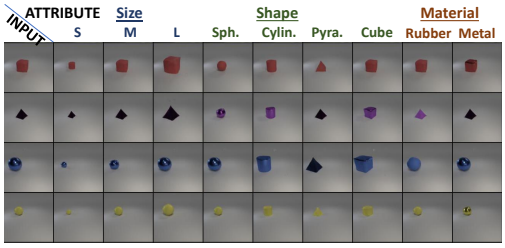

Figure 3: Images generated by AttGAN for inputs in column 1, conditioned on attributes: size(2-4), shape (5-8), material (9-10)

powerful image manipulation technique as our surrogate function $\mathbf{x}^{gen} = F_{GAN}(\mathbf{x}, \alpha)$. We define the attribute vector to be a 13-dimensional binary hash-code with 1 and 0 indicating presence or absence of an attribute. For each experiment, we train the AttGAN on the training dataset outlined in Figure 2 to generate 128x128 images, as illustrated in Figure 3.

Baselines: 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 et al., 2020) which uses a meta-learning framework to generate multiple domains; and the naive baseline of our classifier trained without any adversarial training. The same classifier architecture is used for each baseline for fair comparison, trained for 15 epochs.

Results: The test classification accuracies for different splits are reported in Table 1. 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*.

Analysis: An appropriately chosen value for β (coefficient of the constraint loss in Eq 3) encourages useful perturbations without violating the class-label consistency cost ℓ_{cls} as seen in the top row of Fig 4. On the other hand, a higher β (row 2) encourages exploration of unseen regions in attribute space at the cost of higher classification error, resulting in 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.

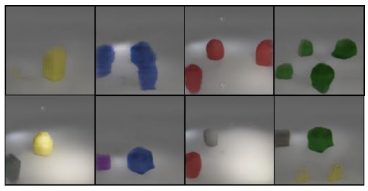

Figure 4: Effect of hyper-parameter β from Equation 3 on the novelty of generated images.

6 Outlook

In this work, we proposed AGAT – a new adversarial training strategy for robustness against large perturbations that are common in practical scenarios. AGAT learns to perturb the attribute space and utilizes surrogate functions to synthesize new images to aid training. The new CLEVR-Singles dataset that we have created can be used in future work for studying robustness to semantic shifts. In the appendix, we have shown the applicability of AGAT with different classes of surrogate functions, to geometric transforms and common image corruptions. AGAT can potentially be applied to a broad range of robustness problems not limited to classification.

REFERENCES

Saikiran Bulusu, Bhavya Kailkhura, Bo Li, Pramod K Varshney, and Dawn Song. Anomalous example detection in deep learning: A survey. *IEEE Access*, 8:132330–132347, 2020.

Taco S Cohen and Max Welling. Transformation properties of learned visual representations. *arXiv preprint arXiv:1412.7659*, 2014.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pp. 2017–2025, 2015.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4773–4783, 2019.

Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

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

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, Derek Nowrouzezahrai, and Alec Jacobson. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. *arXiv preprint arXiv:1808.02651*, 2018.

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

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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*, pp. 12556–12565, 2020.

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

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

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, 2020.

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*, pp. 5334–5344, 2018.

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

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

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

APPENDIX

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

A GEOMETRIC TRANSFORMATIONS

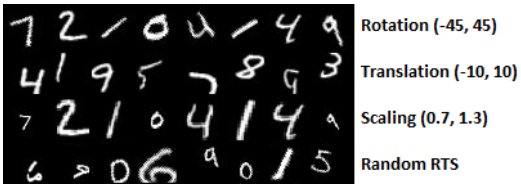

Figure 5: RTS-perturbed MNIST images.

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 & Welling (2014).

Dataset: We address this problem in the digit classification setting, with the training images belonging to the MNIST dataset LeCun (1998), and the test images that are perturbed along rotation-translation-scale (RTS), as shown in Figure 5. We use the standard RTS setup Jaderberg et al. (2015) with angle of rotation in $(-45, 45)^\circ$, translation in $(-10, 10)$ pixels in both directions, and a scale factor in the range $(0.7, 1.3)$.

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) Jaderberg et al. (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 et al. (2020) and GUD Volpi et al. (2018). Additionally, we also use the RTS perturbation sets generated by Wong & Kolter (2020) (PS) and use them as augmented training

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

Table 2: Results on the MNIST-RTS robustness benchmark for rotation (R), translation (T), scaling (S), and random combination (RTS).

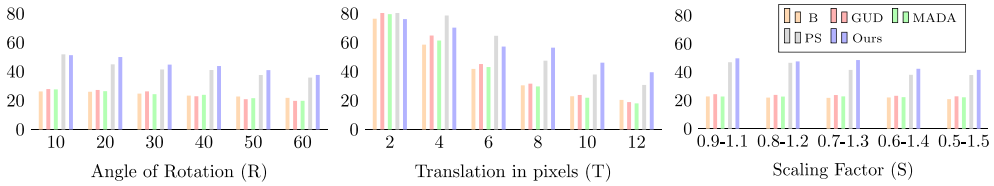

Figure 6: Comparison of random RTS accuracies when controlling each parameter to a max. value. Left: R, Center: T, Right: S

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 Equation 2, are: $\lambda_1 = 1, \lambda_2 = 1, \beta = 5$. The learning rate η for the classifier is $1\text{e}{-4}$ 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 2 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 5 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 Wong & Kolter (2020) (in gray) is competitive at lower severities.

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 2 (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 3 and show that the regularization indeed helps improve performance.

B 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.

$\mathbf{N_{aug}}$	$\mathbf{T_{aug}}$	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 3: 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 4: The effect of classification loss function at different percentages of augmented samples. GT denotes the first term and CR is consistency regularization.

Dataset The CIFAR10 dataset Krizhevsky et al. (2009) contains $50k$ training images belonging to 10 classes. Recently, CIFAR10-C Hendrycks & Dietterich (2018) which contains image corruptions for CIFAR10 images, was proposed to benchmark robustness of image classifiers. There are four main categories of these corruptions and 15 fine-grained categories: *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, of which we focus on the highest severity.

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

$$\mathbf{x}^{gen} = \frac{1}{\sqrt{2\pi\alpha_1^2}} e^{-\frac{\mathbf{x}^2}{2\alpha_1^2}} + n, \text{ where } n \sim \mathcal{N}(0, \alpha_2). \tag{4}$$

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 et al. (2018), a technique for defending against adversarial attacks, and pixel-wise domain augmentation techniques MADA Qiao et al. (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 Russakovsky et al. (2015), with group normalization Wu & 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 et al. (2018) (B+SS). Our joint-training (JT) baseline is from TTT based on Hendrycks & 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 Equation 2 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 5 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 to train the classifier, but still operates in pixel-space and does not perturb the attribute space. In Table 5 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

Method	Source	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 (blur only)	91.3	**78.4**	**75.1**	49.3	**75.4**	70.0
Ours (noise 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 5: Comparison of classification accuracies on the four corruption categories (W: Weather, B: Blur, N:Noise, D: Digital) in CIFAR-10-C. Underlined are our best scores, and **bold** are the overall best.

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.