Training Robust Classifiers with Diffusion Denoised Examples

Chandramouli Shama Sastry
Dalhousie University
Vector Institute

Sri Harsha Dumpala
Dalhousie University
Vector Institute

Sageev Oore
Dalhousie University
Vector Institute

Abstract

In this paper, we explore diffusion denoised examples as augmentations to train image classifiers. In particular, we diffuse the train examples to a randomly sampled diffusion time (i.e., apply Gaussian perturbation) and then apply a single diffusion denoising step to generate an augmented train example. We provide an analysis of training classifiers with such diffusion denoised examples through comparisons with classifiers trained exclusively with (i) standard augmentations such as horizontal flips and crops and (ii) novel augmentations such as AugMix and DeepAugment. We show that classifiers trained with diffusion denoised examples are more robust than the classifiers trained using standard augmentations without sacrificing clean test accuracy. Furthermore, we demonstrate that diffusion-denoised augmentations are also useful as test-time augmentations and this allows us to introduce a simple and efficient image-adaptation method that is competitive with DDA.

1. Introduction

Image classifiers are inaccurate when presented with samples that do not lie within the train distribution. Previous works have demonstrated that image classifiers are surprisingly sensitive to a wide variety of distributional shifts leading to severely inaccurate predictions. Examples include: (a) synthetic corruptions and perturbations (e.g., Imagenet-C [12], Imagenet-P [12], Imagenet-$\bar{\text{C}}$ [20]), (b) natural and alternative renditions of in-distribution classes such as paintings, sculptures, embrodiery, etc (e.g., Imagenet-R [14]), (c) naturally occuring adversarial examples (e.g., Imagenet-A [15]), (d) stylistic alterations (e.g., Imagenet-S [10]). Collectively, these are referred to as out-of-distribution datasets.

Recent research towards improving robustness of classifiers are focused on improved training techniques or test-time adaptation. For example, AugMix[13], DeepAugment[14], PixMix [16] and Prime [21] present novel augmentation techniques that improve classifier robustness. On the other hand, test-time adaptation algorithms adapt model parameters on a batch of test images (optionally, a single image) based on a self-supervised learning objective (e.g., masked auto-encoder [8], rotation prediction [28], contrastive learning [5, 19]) or entropy minimization (e.g., TENT [29], MEMO [33], SAR [23]).

Leveraging the recent advances in diffusion generative models, Diffusion Domain Adaptation (DDA) [9] and Diffusion-TTA [25] propose test-time adaptation techniques for robust classification. Specifically, DDA utilizes an unconditional diffusion model to *transfer* an input test image into the source distribution while Diffusion-TTA utilizes a conditional diffusion model to optimize the classifier weights over a denoising loss-objective. By adapting the image instead of the model, DDA demonstrates robustness across a variety of evaluation settings (e.g., batch-sizes, batch composition, non-stationary label shifts) unlike the model-adaptation algorithms [9, 34].

In this work, we explore the application of diffusion models in training robust classifiers. Specifically, we propose to train the classifier with one-step diffusion denoised images generated from Gaussian perturbed train examples — in other words, we consider denoised examples as augmentations of the original train image. We demonstrate that classifiers finetuned on such denoised examples offer improved robustness when compared with the last checkpoint. Furthermore, we extend this to test-time and demonstrate performance comparable to or exceeding DDA using an ensemble prediction over test-time diffusion denoised images – more importantly, our method is $\sim$ 10x faster than DDA in terms of wallclock time. In summary, our contributions are as follows:

- We combine diffusion denoised augmentations with leading augmentation techniques such as AugMix and DeepAugment and demonstrate improved robustness on covairate shifts at no cost to test accuracy.
- We qualitatively analyse diffusion denoised images and provide hypotheses explaining the empirical observations.
- We extend the idea of diffusion denoised augmentations to test time to further improve the classifier robustness.

2. Background

The stochastic diffusion framework [27] consists of two key components: 1) the forward-diffusion (i.e., data to noise) stochastic process, and 2) a learnable score-function that can then be used for the reverse-diffusion (i.e., noise to data) stochastic process.

The forward diffusion stochastic process $\{\mathbf{x}_t\}_{t\in[0,T]}$ starts at data, $\mathbf{x}_0$, and ends at noise, $\mathbf{x}_T$. We let $p_t(\mathbf{x})$ denote the probability density of $\mathbf{x}$ at time t, so, e.g., $p_0(\mathbf{x})$ is the distribution of the data, and $p_T(\mathbf{x})$ is the distribution of the noise. The diffusion is structured so that $p_T(\mathbf{x})$ is independent of the starting point at $t = 0$. This process is defined with a stochastic-differential-equation (SDE):

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\, dt + g(t)\, d\mathbf{w}, \tag{1}$$

where $\mathbf{w}$ denotes a standard Wiener process, $\mathbf{f}(\mathbf{x}, t)$ is a drift coefficient, and $g(t)$ is a diffusion coefficient. The drift and diffusion coefficients are usually manually specified such the solution to the SDE with initial value $\mathbf{x}_0$ is a time-varying Gaussian distribution $p_t(\mathbf{x}|\mathbf{x}_0)$ whose mean $\mu(\mathbf{x}_0, t)$ and standard deviation $\sigma(t)$ can be exactly computed.

To sample from $p_0(\mathbf{x})$ starting with samples from $p_T(\mathbf{x})$, we have to solve the reverse diffusion SDE [1]:

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})]\, dt + g(t)\, d\bar{\mathbf{w}}, \tag{2}$$

where $d\bar{\mathbf{w}}$ is a standard Wiener process when time flows from T to 0, and dt is an infinitesimal negative timestep. In practice, the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is estimated by a neural network $s_\theta(\mathbf{x}, t)$, parameterized by θ, trained to optimize a weighted sum of denoising score-matching losses [27].

Denoised Examples. Given $(\mathbf{x}_0, y) \sim p_0(\mathbf{x})$ and $\mathbf{x} \sim p_t(\mathbf{x}|\mathbf{x}_0) = \mathcal{N}(\mathbf{x} \mid \mu(\mathbf{x}_0, t),\ \sigma^2(t)\mathbf{I})$, we can compute the denoised image $\hat{\mathbf{x}}_t$ using the pretrained score network s_θ as:

$$\hat{\mathbf{x}}_t = \mathbf{x} + \sigma^2(t) s_\theta(\mathbf{x}, t) \tag{3}$$

Intuitively, $\hat{\mathbf{x}}_t$ is an *expectation* over all possible images $\mathbf{m}_t = \mu(\mathbf{x}_0, t)$ that are *likely* to have been perturbed with $\mathcal{N}(\mathbf{0},\ \sigma^2(t)\mathbf{I})$ to generate $\mathbf{x}$ and the denoised example $\hat{\mathbf{x}}_t$ can be written as

$$\hat{\mathbf{x}}_t = \mathbb{E}[\mathbf{m}_t|\mathbf{x}] = \int_{\mathbf{m}_t} \mathbf{m}_t\, p_t(\mathbf{m}_t|\mathbf{x}) d\mathbf{m}_t \tag{4}$$

We note that the mean does not change with diffusion time t in variance-exploding SDEs while the mean decays to zero with diffusion time for variance-preserving SDEs (DDPMs).

3. Denoised Examples as Augmentations

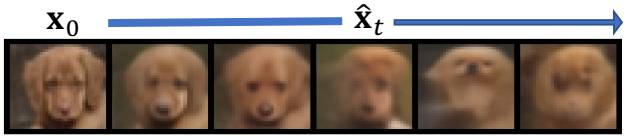


Figure 1. Denoised Examples: Given $\mathbf{x}_0$, we show $\hat{\mathbf{x}}_t$ for different values of diffusion time t arranged chronologically.

In this section, we detail the training procedure of classifiers with denoised examples as augmentations. Previous studies have mainly focused on the use of such denoised examples as approximations of clean images at inference-time to improve classifier robustness (e.g., [3]) and do not consider denoised examples for training classifiers perhaps because clean images are readily available in the training dataset. We also qualitatively analyse the denoised examples to provide a visual understanding of the generated augmentations.

To generate denoised training examples, we diffuse $\mathbf{x}_0$ to a uniformly sampled time $t \sim \mathcal{U}(0, T)$ and then, denoise it using the trained score network s_θ (Eq. 3). We illustrate some denoised examples of a CIFAR10 image in Figure 1.

To train a classifier on regular augmented examples, we optimize $\mathcal{L}(p_\phi(\mathcal{A}(\mathbf{x}_0)), y)$ where $\mathcal{A}(\cdot)$ generates a random augmentation of $\mathbf{x}_0$ and $\mathcal{L}$ is generally the cross-entropy loss. We consider the following augmentation methods to define $\mathcal{A}(\cdot)$ in this work: (i) BASE: Horizontal Flip/Crop Augmentations, (ii) AugMix (AM), (iii) DeepAugment (DA) and (iv) DeepAugment+AugMix (DAM). In the case of AugMix, the augmented examples are used to compute a Jenson-Shannon divergence in addition to the cross-entropy loss on regularly augmented examples. When additionally considering diffusion denoised augmentations, we optimize:

$$\mathcal{L}_{\text{Total}} = \mathbb{E}_{t,\mathbf{x}}[-\log p_\phi(y|\hat{\mathbf{x}}_t)] + \mathcal{L}(p_\phi(\mathcal{A}(\mathbf{x}_0)), y) \tag{5}$$

where, $t \sim \mathcal{U}(0, T)$, $\mathbf{x} \sim p_t(\mathbf{x}|\mathbf{x}_0)$, $(\mathbf{x}_0, y) \sim p_0(\mathbf{x})$ and $\hat{\mathbf{x}}_t$ is obtained using score-network (Eq. 3). We note that we do not apply the novel augmentations to the denoised images when computing the cross-entropy loss since the joint training should implicitly generalize to augmentations applied to denoised examples.

We can interpret training on denoised examples as a type of Vicinal Risk Minimization (VRM) wherein the cost-function is optimized on the *vicinal distribution* of training samples to improve generalization. For example, Chapelle et al. [4] use Gaussian perturbed examples ($\mathbf{x}$) as the vicinal distribution while MixUp [32] uses a convex sum of two random inputs (and their labels) as a vicinal distribution. A denoised example $\hat{\mathbf{x}}_t$ is a convex sum over $\mathbf{m}_t$ where each $\mathbf{m}_t$ is weighted by its likelihood of generating $\mathbf{x}$ ($p_t(\mathbf{m}_t|\mathbf{x})$) and can be considered to be *vicinal* to those examples $\mathbf{m}_t$

that have a non-trivial likelihood $p_t(\mathbf{m}_t|\mathbf{x})$. The distribution $p_t(\mathbf{m}_t|\mathbf{x})$ is concentrated around examples perceptually similar to $\mu(\mathbf{x}_0, t)$ when $\mathbf{x}$ is closer to $\mathbf{x}_0$ (i.e., smaller $\sigma(t)$) and becomes more entropic as the noise scale increases: in Figure 1, we can observe the superposition of candidate dog images with increasing diffusion time t.

Qualitative Analysis and Manifold Theory. When generating augmentations, it is important to ensure that the resulting augmentations lie on the image manifold. Recent studies [6, 24] on theoretical properties of denoised examples suggest that denoised examples can be considered to be on the data manifold under certain assumptions lending theoretical support to the idea of using denoised examples as augmentations. In addition, it is also important to preserve the class-labels upon augmentation as this can lead to manifold intrusion [11] and lead to underfitting and lower classification accuracies. Diffusion denoised augmentations generated from significantly perturbed train examples can introduce label-noise into the training since the label of the original image may not necessarily match that of the denoised image – for example, some of the diffusion denoised augmentations of the dog in Figure 2 resemble architectural buildings. In the specific case of diffusion denoised augmentations, however, the augmentations are *visually* distinct from the original image potentially allowing for the model to *learn* to be robust to noisy labels by adjusting its confidence accordingly. For example, Figure 3 shows an example augmentation that causes manifold intrusion and also does not have any visual cues that allow the model to distinguish between the original image and augmented image.

We confirm this empirically showing no degradation in classification accuracy and instead observe improved robustness to distribution shifts. We hypothesize that the diffusion denoised examples that are farther from the input image are crucial in improving classifier robustness – specifically, augmentations from larger t introduce significant changes to the input image (and possibly, the class label) requiring the classifier to observe details in input image to carefully estimate the class-membership probabilities of $\mathbf{x}_0$ based on $\hat{\mathbf{x}}_t$. For example, the diffusion denoised augmentations at $t = 999$ have minimal resemblance of the original image and we expect that the class-probabilities should be distributed evenly across the 1k classes. We evaluate the average prediction entropy as a function of the diffusion time in Figure 4 and surprisingly find that classifiers trained on state-of-the-art augmentation methods do not predict uniform class probabilities for diffusion denoised augmentations obtained at $t = 999$ while the classifiers fine-tuned with diffusion denoised images behave as expected.

Denoised Examples as Test-time Augmentations. When presented with test examples containing unknown distribution shifts, DDA [9] shows that we can project the test examples into the source distribution by first applying a forward diffusion step followed by iterative sampling. Inspired by DDA, we extend the idea of one-step diffusion denoised examples as augmentations to generate test-time augmentations of a test-example. Specifically, given a test example, we generate one-step diffusion denoised examples for various values of diffusion times t and utilize the average predictions [18] across all the augmentations to assign the class-label. We demonstrate that this technique is competitive with DDA and offers significant improvement in terms of running time as it does not involve iterative sampling.

4. Experiments

We primarily conduct our experiments on Imagenet using the Improved-DDPM [7, 22] diffusion model. In particular, we use the unconditional model open-sourced by Dhariwal and Nichol [7] as the score-network for Imagenet (256 $\times$ 256) to generate both training and test-time augmentations. In all of our experiments, we use a ResNet-50 backbone as the Imagenet classifier. We evaluate the advantages of diffusion-denoised examples as train augmentations when combined with the following augmentation-techniques: (i) BASE: Horizontal Flip/Crop Augmentations, (ii) AugMix (AM), (iii) DeepAugment (DA) and (iv) DeepAugment+AugMix (DAM). To distinguish between models trained exclusively with the above-mentioned augmentations, we use the suffix "+Diff" to denote classifiers trained additionally with diffusion-denoised examples as augmentations. We finetune the pretrained checkpoints for 5 epochs with diffusion-denoised augmentations; we also utilize the original augmentations used in pretraining in order to retain the distinct generalization benefits offered by the respective augmentations. We evaluate the pretrained models and the models finetuned with additional diffusion denoised augmentations in the following evaluation modes:

1. **DDA**: We apply the DDA algorithm to transfer a test-example with unknown distribution shift into the source distribution.
2. **DDA (SE)**: We consider both the original test example and DDA-adapted test-example by averaging the posterior probabilities following the self-ensemble (SE) strategy proposed in [9].
3. **Denoised-Ensemble (DE)**: We generate diffusion-denoised augmentations of a given test-example and then utilize the posterior probabilities averaged across the test-time augmentations. We generate 9 test-time augmentations using one-step diffusion denoising applied to images diffused to $t \in \{0, 50, \ldots, 450\}$. We follow DDA to determine the upper limit of the diffusion time $t = 450$.
4. **Default**: In the default mode, we directly evaluate the model on the test examples.

We summarize the results across all evaluation modes in Tab. 1 for Imagenet-C (severity=5) and the uncorrupted Imagenet test dataset. We also plot the effect of diffusion-time

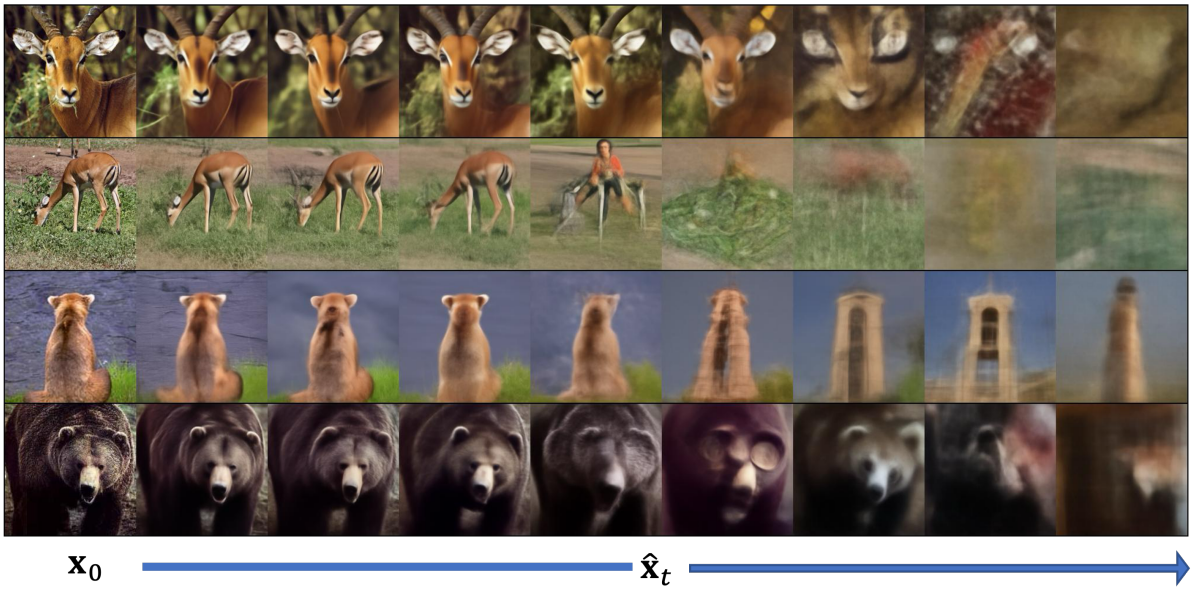


Figure 2. We show some sample diffusion denoised augmentations of samples $\mathbf{x}_0$ taken from Imagenet-train. In particular, we display 8 random augmentations for each image between $t = 350$ and $t = 700$ in steps of size 50. Augmentations generated for $t < 350$ are *closer* to the input image while the augmentations for $t > 700$ are *farther* from the input image. We observe that the diffusion denoised augmentations with larger values of t do not preserve the class label introducing noise in the training procedure. However, we find that this does not lead to empirical degradation of classification accuracy but instead leads to improved robustness to corruptions.

Figure 3. Example of Manifold Intrusion from Appendix C of Hendrycks et al. [13]. While diffusion denoised augmentations may alter class labels (Figure 2), the denoised images are visually distinguishable from the original images allowing the model to also *learn* from noisy labels without inducing manifold intrusion. On the other hand, here is an example of manifold intrusion where the augmented image does not contain any visual cues that enable the model to be robust to noisy labels.

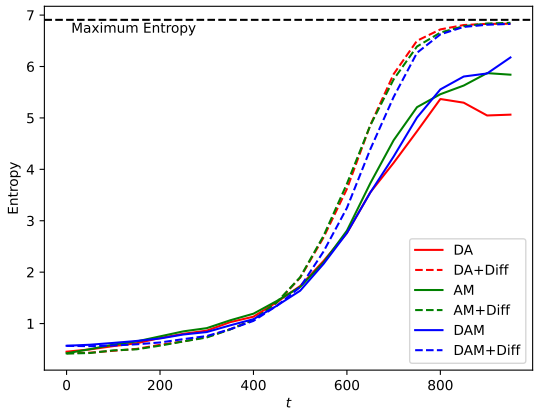


Figure 4. Average Prediction Entropy vs Diffusion Time. We compare between models trained on different augmentations by evaluating their prediction entropy over diffusion denoised images. We observe that the models trained with additional diffusion denoised augmentations (DA+Diff, AM+Diff and DAM+Diff) correctly yield predictions with higher entropies (lower confidence) for images containing imperceptible details (i.e. larger t). Surprisingly, the classifiers trained without diffusion denoised augmentations (DA, AM and DAM) do not also assign random-uniform label distribution for diffusion denoised images at $t = 999$, which have no class-related information by construction.

on the denoised-ensemble performance in Figure 5. We summarize our observations as follows:

- For Imagenet-C, models trained with diffusion denoised examples improve over models trained without these augmentations across all evaluation modes. In the evaluations

Table 1. We summarize the results for each combination of Train-augmentations and evaluation modes on Imagenet-C (severity=5). AM, DA, and DAM are short for AugMix, DeepAugment and DeepAugment + AugMix respectively. When diffusion denoised examples are additionally used to train the classifiers, we denote it with the suffix "+Diff". DDA refers to the evaluation on Denoising Diffusion Adapted samples of Imagenet-C. DE refers to the Denoised-Ensemble evaluation of the models. Def. denotes direct evaluation over Imagenet-C samples. In DDA (SE), we average the prediction probabilities obtained using the original image and DDA image. We find that diffusion denoised augmentations are effective as both train-time augmentations and test-time augmentations.

(a) Imagenet-C (severity=5)

Train Augmentations	Inference Mode				Avg
	DDA	DDA (SE)	DE	Def.	
AM	33.18	36.55	34.08	26.72	32.63
AM+Diff	34.52	38.48	38.44	29.44	**35.22**
BASE	28.35	30.62	27.2	17.87	26.01
BASE+Diff	32.1	34.13	30.66	20.44	**29.33**
DA	35.41	39.05	37.01	31.92	35.85
DA+Diff	37.59	41.35	40.65	33.74	**38.33**
DAM	40.35	44.81	41.85	39.52	41.63
DAM+Diff	41.77	46.18	44.48	41.05	**43.37**
Avg	35.41	38.90	36.80	30.09	35.30
Avg(Diff)	36.50	40.04	38.56	31.17	36.56
Avg(NonDiff)	34.32	37.76	35.04	29.01	34.03

(b) Imagenet-Test

Train Augmentations	Inference Mode				Avg
	DDA	DDA (SE)	DE	Def.	
AM	62.22	75.98	73.89	77.45	72.39
AM+Diff	63.33	76.07	75.57	77.20	**73.04**
BASE	58.09	74.38	71.20	76.10	69.94
BASE+Diff	62.83	73.93	73.65	76.00	**71.60**
DA	63.63	75.39	74.07	76.52	72.40
DA+Diff	65.33	75.57	75.17	76.40	**73.12**
DAM	65.53	74.41	73.34	75.69	72.24
DAM+Diff	66.74	74.60	74.35	75.50	**72.80**
Avg	63.46	75.04	73.91	76.36	72.19
Avg(Diff)	64.56	75.04	74.69	76.28	72.64
Avg(NonDiff)	62.37	75.04	73.13	76.44	71.74

over the uncorrupted test examples, the models trained with diffusion denoised examples preserve the accuracies of the original models.

- On average, Denoised Ensemble yields improved detection rates as compared to direct evaluation on DDA images.
- Denoised Ensemble applied to models trained with diffusion denoised images is better on average than DDA-SE applied to models trained without diffusion augmentation: for example, the non-diffusion finetuned models are 37.76% accurate on Imagenet-C in DDA-SE evaluation whereas the diffusion finetuned models are 38.56% accurate in DE evaluation. This demonstrates that we can get improved robustness by first fine-tuning the classifier over diffusion augmented images and then average predictions across diffusion denoised augmentations of a test sample to obtain competitive performance with DDA at a substantially faster ($\sim$ 10x) wallclock time (Tab. 3).

Table 2. We evaluate the models on Imagenet-R and Imagenet-S in the Default and DE evaluation modes. We find that the results on these datasets follow the same trend as observed for Imagenet-C.

Train Augmentations	Imagenet-S		Imagenet-R		Avg
	Def.	DE	Def.	DE	
AM	10.90	15.15	40.83	42.56	27.36
AM+Diff	11.10	15.79	40.93	43.22	**27.76**
BASE	7.13	11.76	36.15	38.75	23.45
BASE+Diff	7.78	12.17	37.19	41.00	**24.54**
DA	13.53	16.48	42.00	43.58	28.90
DA+Diff	13.80	17.50	42.65	44.77	**29.68**
DAM	18.93	19.44	46.72	46.44	32.88
DAM+Diff	19.31	19.88	47.08	47.38	**33.41**
Avg	12.81	16.02	41.69	43.46	28.50
Avg(Diff)	13.00	16.34	41.96	44.09	28.85
Avg(NonDiff)	12.62	15.71	41.43	42.83	28.15

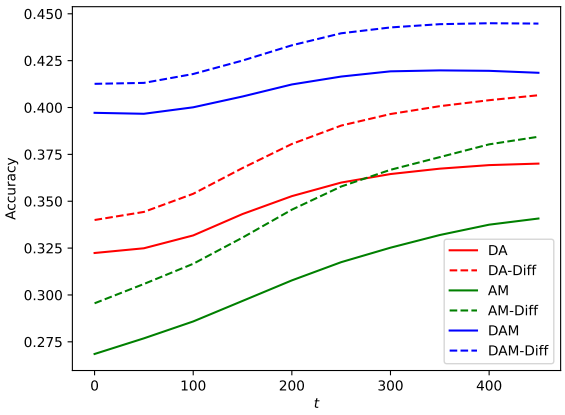
Table 3. DDA vs DE in terms of wallclock times: We use 40GB A40 GPU for determining the running time. For each method, we determine the maximum usable batch-size and report the average wallclock time for processing a single example.

Method	Wallclock Time (s)
DE	0.5
DDA	4.75

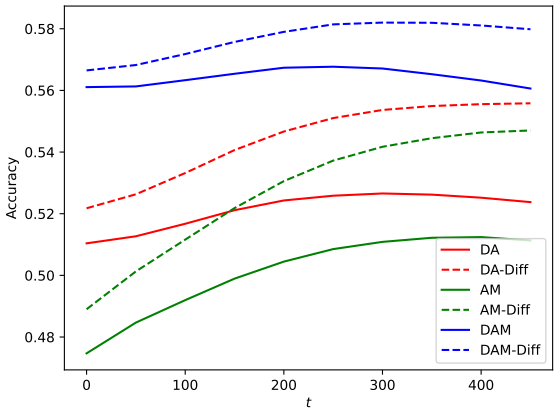
We also evaluate our method on Imagenet-R and Imagenet-S as shown in Tab. 2. Although the improvements over Imagenet-R and Imagenet-S from training over diffusion denoised augmentations are only slight, we observe some notable improvements when evaluating under the DE inference mode. Overall, the results in Tab. 2 follow the same trends as observed in Tab. 1.

5. Related Works

In this section, we discuss some related works in addition to the ones discussed in the introduction.



(a) Severity=5



(b) All Severities

Figure 5. Denoised-Ensemble: Imagenet-C Accuracy vs Diffusion Time. For each diffusion time $t \in \{0, 50, \ldots, 400, 450\}$, we plot the accuracy when considering all diffusion denoised augmentations generated up to t. For example, we consider 5 diffusion denoised augmentations of each test example to compute the accuracy corresponding to $t = 200$. We observe that classifiers trained with additional diffusion denoised augmentations not only perform better on average at $t = 0$ but can also predict more accurately when ensembling over diffusion denoised augmentations.

Synthetic Training Images. Synthetic Images from Diffusion Models have been explored to train classifiers: for example, Azizi et al. [2] fine-tune large text-to-image diffusion models to generate extra data while You et al. [31] use a three stage training process to train semi-supervised classifier wherein they generate *pseudo images* – generated with a diffusion model trained over pseudo-labels derived from first round of semi-supervised classifier training with strong augmentations – to improve standard semi-supervised training. Yamaguchi and Fukuda [30] evaluate classifiers trained on diffusion images generated from iterative diffusion sampling; in their analysis of reverse diffusion times, they expose some limitations of using synthetic images from diffusion models to train classifiers. In contrast, we train a classifier on one-step diffusion denoised images generated from the entire range of diffusion times and demonstrate improvements on robustness at no cost to original test accuracy.

Test Time Augmentations. Test time augmentations employ data augmentations at test time to improve classification accuracy [18]. In practice, some augmentations are selected to generate test time augmentations and the classifier-outputs over all augmentations are averaged to make the prediction. However, there are improved methods beyond simple averaging to generate outputs from test-time augmentations: for example, Kim et al. [17] and Shanmugam et al. [26] propose learning-based solutions for augmentation selection and aggregation respectively. We utilise diffusion denoised augmentations as test time augmentations and use simple average over outputs to improve robustness to covariate shifts. In future work, improved test time aggregation techniques could be explored to improve the performance.

6. Acknowledgements

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute https://vectorinstitute.ai/partnerships/current-partners/.

7. Conclusion

In this work, we explore diffusion denoised images as augmentations to train classifiers. When combined with leading data augmentation techniques, we find that diffusion denoised examples confer additional robustness to covariate shifts without affecting accuracy on clean examples. We qualitatively examine diffusion denoised augmentations and identify factors likely responsible for improved robustness at identical test accuracies. Furthermore, we extend diffusion denoised augmentations to test time and introduce a simple averaging technique to further improve robustness to covariate shifts.

References

[1] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. 2

[2] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *Transactions on Machine Learning Research*, 2023. 6

[3] Nicholas Carlini, Florian Tramèr, Krishnamurthy Dvijotham, Leslie Rice, Mingjie Sun, and Zico Kolter. (certified!!) ad-

versarial robustness for free! *International Conference on Learning Representations (ICLR)*, 2023. 2

[4] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000. 2

[5] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022. 1

[6] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems*, 2022. 3

[7] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *CoRR*, abs/2105.05233, 2021. 3

[8] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A Efros. Test-time training with masked autoencoders. In *Advances in Neural Information Processing Systems*. 1

[9] Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven test-time adaptation. *arXiv preprint arXiv:2207.03442*, 2022. 1, 3

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 1

[11] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization, 2018. 3

[12] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *CoRR*, abs/1903.12261, 2019. 1

[13] 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. 1, 4

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1

[15] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 1

[16] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. *CVPR*, 2022. 1

[17] Ildoo Kim, Younghoon Kim, and Sungwoong Kim. Learning loss for test-time augmentation. *Advances in Neural Information Processing Systems*, 33:4163–4174, 2020. 6

[18] Masanari Kimura. Understanding test-time augmentation, 2024. 3, 6

[19] Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. TTT++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34: 21808–21820, 2021. 1

[20] Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. In *Advances in Neural Information Processing Systems*, pages 3571–3583. Curran Associates, Inc., 2021. 1

[21] Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. PRIME: A few primitives can boost robustness to common corruptions. *CoRR*, abs/2112.13547, 2021. 1

[22] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *CoRR*, abs/2102.09672, 2021. 3

[23] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *Internetional Conference on Learning Representations*, 2023. 1

[24] Frank Permenter and Chenyang Yuan. Interpreting and improving diffusion models using the euclidean distance function. *arXiv preprint arXiv:2306.04848*, 2023. 3

[25] Mihir Prabhudesai, Tsung-Wei Ke, Alexander C. Li, Deepak Pathak, and Katerina Fragkiadaki. Test-time adaptation of discriminative models via diffusion generative feedback. In *Conference on Neural Information Processing Systems*, 2023. 1

[26] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1214–1223, 2021. 6

[27] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2

[28] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. 1

[29] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 1

[30] Shin'ya Yamaguchi and Takuma Fukuda. On the limitation of diffusion models for synthesizing training datasets, 2023. 6

[31] Zebin You, Yong Zhong, Fan Bao, Jiacheng Sun, Chongxuan Li, and Jun Zhu. Diffusion models and semi-supervised learners benefit mutually with few labels. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 6

[32] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018. 2

[33] Marvin Mengxin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*. 1

[34] Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. *arXiv preprint arXiv:2306.03536*, 2023. 1

Table 4. ImageNet-C (severity=5) accuracy for each corruption type. Relative Improvements when additionally using diffusion denoised augmentations are computed with respect to the corresponding pretrained checkpoints and averaged across all corruption types.

		Noise			Blur				Weather				Digital					
Inference Mode.	Train Aug.	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg.	Rel. Imp.
DDA	AM	50.66	52.22	50.8	18.18	25.14	18.78	24.27	22.71	33.05	5.35	40.54	11.14	40.66	54.15	50.1	33.18	
	AM+Diff	51.2	52.78	51.51	21	27.8	22.46	27.84	23.35	34.39	7.55	40.75	11.14	40.89	55.22	49.96	34.52	7.63
	BASE	45.97	47.62	46.73	12.87	17.98	12.77	20.27	17.92	27.56	5.11	35.44	5.91	36.1	47.91	45.06	28.35	
	BASE+Diff	48.91	49.99	49.57	18.39	24.15	17.47	25.48	19.49	31.37	8.31	39.39	7.09	38.79	53.87	49.28	32.1	20.17
	DA	51.48	53.37	51.15	24.11	30.07	18.07	23.22	25.31	35.46	10.69	44.11	12.33	41.26	58.68	51.82	35.41	
	DA+Diff	52.26	53.94	51.96	28.66	32.41	22.6	27.56	25.78	37.21	13.82	47.23	14.89	41.72	60.72	53.04	37.59	9.7
	DAM	53.5	55.54	54.86	31.33	37.44	28.89	28.49	30.75	39.67	12.85	49.04	21.24	45.04	61.62	54.99	40.35	
	DAM+Diff	54.29	55.82	55.44	33.12	37.91	33.13	31.58	30.96	41.04	15.21	50.85	23.33	44.72	62.93	56.19	41.77	5.01
DDA-SE	AM	49.61	51.24	50.1	20.65	23.05	24.43	32.81	25.75	36.25	20.06	54.14	14.14	39.15	54.66	52.22	36.55	
	AM+Diff	51.3	52.29	51.55	24.45	26.84	29.21	36.55	26.5	37.39	26.46	53.59	15.4	38.67	55.06	51.98	38.48	7.94
	BASE	44.85	45.59	45.17	14.33	16.2	14.23	23.94	20.54	30.4	19.54	51.65	6.61	33.19	46.47	46.54	30.62	
	BASE+Diff	48.51	49.3	48.97	21.52	22.61	20.31	30.88	20.87	32.81	23.77	50.64	7.17	34.74	51.32	48.49	34.13	16.21
	DA	53.35	54.71	53.32	25.72	28.2	20.13	26.37	30.57	40.05	28.79	59.39	13.91	39.82	59.09	52.34	39.05	
	DA+Diff	53.84	55.3	53.94	30.57	30.05	25.24	30.57	31.32	41.49	35.52	60.82	19.33	39.53	60.35	52.41	41.35	9.48
	DAM	54.17	56.33	55.21	33.59	34.12	36.29	34.89	35.82	45.12	35.52	60.9	27.84	43.33	62.99	56	44.81	
	DAM+Diff	54.6	57.06	55.85	35.74	35.5	40.09	37.22	36.34	45.96	38.72	61.88	30.91	42.59	63.41	56.85	46.18	3.75
DE	AM	32.88	36.3	32.97	21.33	30.71	25.59	32.59	24.78	39.16	17.25	55.03	6.87	43.37	54.61	57.73	34.08	
	AM+Diff	37.16	40.49	37.11	28.74	36.58	34.59	40.27	26.33	40.97	27.76	56.1	9.79	43.82	57.75	59.14	38.44	18.36
	BASE	26.88	28.92	26.99	12.11	19.62	16.06	25.01	19.52	32.22	15.28	49.72	1.44	36.55	45.45	52.25	27.2	
	BASE+Diff	30.65	34	31.74	19.27	25.78	21.49	31.24	19.13	33.64	14.63	51.03	1	38.88	51.41	56	30.66	12.99
	DA	44.53	47	46.28	21.21	30.03	22.23	29.52	28.47	40.27	24.23	57.09	4.14	43.5	57.11	59.51	37.01	
	DA+Diff	45.12	47.42	46.8	28.66	35.76	28.9	36.21	29.64	43.02	33.86	59.57	10.94	43.97	58.82	61.14	40.65	22.41
	DAM	46.63	49.23	46.88	28.87	37.4	32.61	36.16	33.47	44.43	28.72	59.79	15.09	46.47	60.82	61.22	41.85	
	DAM+Diff	47.74	50.29	47.94	33.68	40.45	37.72	39.72	34.76	45.94	36.09	60.72	21.2	46.36	62.28	62.29	44.48	9.06
Default	AM	15.01	18.37	16.64	21.48	13.69	24.88	33.66	21.54	27.13	22.91	57.91	13.09	25.16	42.33	46.99	26.72	
	AM+Diff	19.66	22.52	20.68	26	17.43	30.41	37.59	22.82	28.41	28.98	56.88	15.72	24.43	43.17	46.83	29.44	14.28
	BASE	5.68	6.49	6.45	15.04	8.24	13.29	22.86	15.59	20.43	22.21	55.64	4.23	14.31	23	34.54	17.87	
	BASE+Diff	8.29	9.26	9.47	23.36	14.43	19.47	30.84	13.86	19.91	26.75	53.29	4.1	16.06	24.14	33.45	20.44	24.03
	DA	39.6	40.78	41.91	25.47	15.74	19.02	24.58	27.41	33.58	32.03	62.62	9.55	23.69	45.41	37.47	31.92	
	DA+Diff	40.31	41.41	42.69	30.91	17.07	23.24	28.36	28	35.09	37.4	63.1	15.98	23.09	44.75	34.71	33.74	10.19
	DAM	39.61	42.75	42.13	34.47	22.95	36.57	35.58	34.05	39.85	38.74	63.95	25.6	29.62	56.45	50.51	39.52	
	DAM+Diff	41.09	44.11	43.3	36.68	25.61	40.08	37.26	34.97	40.52	40.97	64.34	29.63	28.78	57.17	51.21	41.05	4.56