Score-Based Generative Classifiers

Roland S. Zimmermann[1]
roland.zimmermann@uni-tuebingen.de

Lukas Schott[1]
lukas.schott@gmail.com

Yang Song[2]
yangsong@cs.stanford.edu

Benjamin A. Dunn[3]
benjamin.dunn@ntnu.no

David A. Klindt[3]
klindt.david@gmail.com

Abstract

The tremendous success of generative models in recent years raises the question whether they can also be used to perform classification. Generative models have been used as adversarially robust classifiers on simple datasets such as MNIST, but this robustness has not been observed on more complex datasets like CIFAR-10. Additionally, on natural image datasets, previous results have suggested a trade-off between the likelihood of the data and classification accuracy. In this work, we investigate score-based generative models as classifiers for natural images. We show that these models not only obtain competitive likelihood values but simultaneously achieve state-of-the-art classification accuracy for generative classifiers on CIFAR-10. Nevertheless, we find that these models are only slightly, if at all, more robust than discriminative baseline models on out-of-distribution tasks based on common image corruptions. Similarly and contrary to prior results, we find that score-based are prone to worst-case distribution shifts in the form of adversarial perturbations. Our work highlights that score-based generative models are closing the gap in classification accuracy compared to standard discriminative models. While they do not yet deliver on the promise of adversarial and out-of-domain robustness, they provide a different approach to classification that warrants further research.

1 Introduction

There exist two fundamentally distinct ways of performing classification. The standard way is to train a classifier which discriminates between classes by modeling the conditional probability $p(y|\mathbf{x})$ of labels y given some input $\mathbf{x}$. These models are called *discriminative* classifiers and have been extremely successful in supervised learning [24, 22]. The alternative is to model the conditional likelihood $p(\mathbf{x}|y)$ of an image for each label and predict the label that maximizes this likelihood [39]. This approach is much less common because it requires solving the more complicated task of learning conditional generative models of the data; these models are referred to as *generative* classifiers.

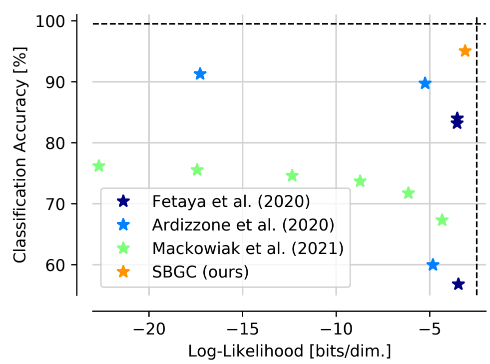

Figure 1: **Model Comparison**. Previous approaches [8, 1, 27] have demonstrated a trade-off between accuracy and likelihoods of generative classifiers on CIFAR-10; The black lines show current SOTA discriminative [horizontal, 7] and generative [vertical, 20] models.

[1]University of Tübingen, [2] Stanford University, [3] Norwegian University of Science and Technology.

A line of research has been discussing the advantages and disadvantages between using discriminative or generative models for classification [39, 33, 36, 2, 26]. They found generative models to be more data efficient, but discriminative models to be asymptotically better. However, these analyses were mostly limited to shallow models [but see 12] and performance evaluation within the same data domain. On MNIST [24] and SVHN [32], previous results have demonstrated that generative classifiers are robust even to worst-case distribution shifts in the form of adversarial perturbations [40, 17, 35]. Furthermore, it was found that generative classifiers have the highest alignment with humans on *controversial* stimuli [11].

However, so far, these results have not transferred to any natural datasets of higher visual complexity [25, 8, 1, 27]. Fetaya et al. [8] argue that 'obtaining strong classification accuracy without harming likelihood estimation is still a challenging problem'. This is empirically supported in their paper as well as a number of related works [13, 1, 27] who all demonstrate a trade-off between good likelihoods and high classification accuracy, suggesting that to better capture what is unique for a label, some performance of the generative model needs to be sacrificed [8]. We challenge these ideas and add a new datapoint to the ongoing debate around generative classifiers demonstrating that near state-of-the-art likelihoods can be combined with classification accuracy at the level of discriminative models [45].

More specifically, score-based generative models have been proposed in recent years as a promising approach towards modeling complex high-dimensional data distributions [41, 16, 42]. They are built on the idea of gradually diffusing a complex data distribution with a parameterized stochastic differential equation into a tractable noise distribution. Previous work has suggested the effectiveness of these diffusion-based models for out-of-domain (OOD) tasks such as adversarially robust classification [51] and as a general prior for natural images [48, 19].

In this work, we propose score-based generative classifiers (SBGC) and show that while these models overcome previous limitations — such as trading off likelihood for accuracy — they still do not solve the problem of worst-case distribution shifts on natural images. More precisely, using our SBGC model, we report state-of-the-art classification accuracy for generative models on CIFAR-10 [22], along with highly competitive likelihood values. Nonetheless, we find that score-based generative classifiers show only slight improvements, if at all, in OOD benchmarks and are vulnerable to adversarial examples.

2 Methods

2.1 Score-Based Generative Models

This work builds on recent advances in score-based generative models of natural images [41, 16, 42, 43, 20]. The general idea behind this approach is to define a stochastic differential equation (SDE) describing the temporal transformation of the complex data distribution (p) into a tractable noise distribution (p_T, isotropic Gaussian). This dynamical process can then be reversed by learning a time-dependent score function $f_\theta(\mathbf{x}(t), t)$ modeled by a neural network with parameters θ. Intuitively, this score function acts as an infinitesimal denoiser [19]. To generate new samples from the data distribution, one starts with random noise $\mathbf{x}(T)$ and follows the dynamics of the reverse SDE to produce a sample $\mathbf{x}(0)$ from the data distribution [see 42, for more details].

Moreover, by removing the stochasticity, the SDE becomes a (neural) ordinary differential equation [42, 5]. Using a continuous time analog of the change of variables formula, one can compute the likelihood (p_0) of an input image $\mathbf{x}(0)$ under the model [42]:

$$\log p_0(\mathbf{x}(0)) = \log p_T(\mathbf{x}(T)) + \int_0^T \nabla \cdot f_\theta(\mathbf{x}(t), t)\ \mathrm{d}t. \tag{1}$$

We use the training objective suggested in Song et al. [43] which maximizes the likelihood of the training data under the model. This is equivalent to minimizing the Kullback-Leibler divergence (KL) between the model distribution and the data distribution, i.e., $\min_\theta \mathrm{KL}(p_0|p)$ [29].

2.2 Score-Based Generative Classifiers

The most straightforward application of score-based generative models as classifiers would be to train one model per class as in Schott et al. [40]. However, this produces very unstable training

Table 1: **Model Comparison.** Accuracy (in %) and negative log-likelihoods (NLL) in bits per dimension [47] on the CIFAR-10 test set. The lower half of the table shows a baseline discriminative model [45] and the current state-of-the-art discriminative [7] and generative [20] models.

Model approach	Accuracy [%] ↑	NLL [bits/dim.] ↓
Invertible Network (Mackowiak et al. [27])	67.30	4.34
GLOW (Fetaya et al. [8])	84.00	3.53
Normalizing flow (Ardizzone et al. [1])	89.73	5.25
Energy model (Grathwohl et al. [13])	92.90	N/A
SBGC (ours)	**95.04**	**3.11**
WideResNet-28-12 (Targ et al. [45])	95.42	N/A
ViT-H/14 (Dosovitskiy et al. [7])	**99.50**	N/A
VDM (Kingma et al. [20])	N/A	**2.49**

and considerable overfitting because the dataset size is, for each model, effectively reduced by one order of magnitude. Instead, we add the image label $y \in \{1, ..., 10\}$ as a conditioning variable to the score function $f_\theta(\mathbf{x}(t), t, y)$; note that this also reduces the required amount of memory for multiple models. In terms of the network architecture, we follow concurrent work of Nichol and Dhariwal [34] and integrate this additional input by first converting it into a one-hot vector and then adding it to the model's representation in the same way as the time variable [see 42]. For classification of a given input $\mathbf{x}(0)$, we can then condition the score function on each label y

$$\log p_0(\mathbf{x}(0)|y) = \log p_T\left(\mathbf{x}(0) + \int_0^T f_\theta(\mathbf{x}(t), t, y)\,\mathrm{d}t\right) + \int_0^T \nabla \cdot f_\theta(\mathbf{x}(t), t, y)\,\mathrm{d}t, \qquad (2)$$

and predict the label that yields the highest conditional likelihood $\max_y \log p_0(\mathbf{x}(0)|y)$. To get the unconditional likelihood of samples we marginalize over the classes, i.e. $p(\mathbf{x}) = \sum_{y_i} p(\mathbf{x}|y = y_i)$.

3 Results

3.1 Classification Accuracy and Data Likelihood

First of all, we note that previous methods [8, 1, 27] have discussed a trade-off between classification accuracy and likelihoods (see Figure 1). This is further supported by observations on MNIST, where increasing the latent dimensionality of a variational autoencoder increases the likelihood at the cost of decreasing classification accuracy and robustness [6]. Older work has focused on hybrid models that combine the data efficiency of (shallow) generative models with the asymptotic performance of discriminative models. However, this comes with a cost in model likelihood [10, 36].

In contrast to these previous methods, we find that our model achieves both state-of-the-art accuracy and likelihoods for generative classifiers on CIFAR-10 (Table 1). We also test the deeper architecture proposed in Song et al. [43]: While this does not change the classification accuracy, it slightly increases the NLL to 3.08 bits/dim. We leave further exploration of this direction to future research.

3.2 Out-of-Distribution Robustness: Common Image Corruptions

Most previous approaches towards improving out-of-distribution performance on common image corruptions [14] have relied on some form of data augmentation in the form of carefully hand-crafted transformations [15] or adversarial training [4]. Here, we propose an orthogonal approach that is based on different modeling assumptions. Specifically, we are interested in seeing whether the inductive bias implicit in score-based generative classifiers improves their classification accuracy when generalizing to the image domains of CIFAR-10-C [14]. We find that our model performs better than some previous models that are solely trained on the original data (Table 2).

Since score-based models are trained like denoisers, they might work specifically well on noisy data. Thus, just comparing the mean accuracy over all corruption types, could give them an unfair advantage over the baselines. Hence, we also compare the mean accuracy for all corruptions excluding

Table 2: **Performance on Common Image Corruptions.** Accuracy (in %) on clean CIFAR-10 test set and the mean accuracy on CIFAR-10-C (considering a random subset with 10% of the original size for SBGC because of computational limitations). We refer to random flips, random crops and uniform noise as simple augmentations.

		CIFAR-10-C	
Models w/ data augm.	CIFAR-10	all corruptions	w/o noises
ResNeXt29 + AugMix (Hendrycks et al. [15])	**95.83**	89.09	90.51
ResNet-50 + adv. augm. (Calian et al. [4])	94.93	**92.17**	**92.53**
Models w/ simple data augm.			
WideResNet-28-12 (Targ et al. [45])	**95.42**	74.42	**80.21**
SBGC (ours)	95.04	**76.24**	75.71

noises [2] (Table 2). Here, we see that our SBGC model does not perform better than the baselines. We conclude that the improvements of the SBGC model in terms of out-of-distribution robustness are limited to noise corruptions.

When integrating data augmentation methods, discriminative models achieve higher performance (Table 2, top). However, we note that the approach of augmenting datasets is independent of our suggested changes, with the caveat that changing the dataset needs to be done carefully for generative models [18]. Preliminary experiments indicate that we can increase our performance on CIFAR-10-C further by leveraging more data augmentations [15] (accuracy $\sim 83\%$). However, there is still a large gap towards discriminative models with data augmentation.

3.3 Out-of-Distribution Robustness: Adversarial Perturbations

Despite their small magnitude, adversarial perturbations can cause a drastic change in the behavior of a neural network [44] and can be seen as a worst-case distribution shift. To assess the robustness of our SBGC model against such perturbations, we first generate them using a projected gradient descent (PGD) [28] on the negative cross-entropy of the model's prediction using Foolbox [37]. For this, we estimate the gradients of the model likelihoods with respect to the inputs $\mathrm{d}p_{0,y}(\mathbf{x})/\,\mathrm{d}\mathbf{x}$ using the adjoint sensitivity method [5]. This requires second-order optimization because of the Jacobian term in equation 2. We implement this efficiently with Hessian-vector-products in JAX [3] and verify the implementation by comparing the calculated gradient with numerical estimations based on finite differences. Since computing gradients is still computationally expensive (a single forward and backward pass take approximately 90 minutes on an NVIDIA Tesla V100 GPU), we follow common practice in this domain and perform the adversarial evaluation on a reduced sample size of $N = 216$.

While our approach partially improves robustness against common image corruptions, it strikingly fails in the presence of adversarial perturbations (Tab. 3). Within both standard ℓ_∞ perturbation norms ($8/255$) as well as standard ℓ_2 perturbation norms (0.5), our model miss-classifies every adversarially perturbed input image.

Table 3: **Performance on Adversarial Perturbations.** Accuracy (in %) of different models against adversarial perturbations generated by a norm-bounded ℓ_∞ and ℓ_2 PGD [28] attack (against each specific model) with a bound of $\epsilon = 8/255$ and $\epsilon = 0.5$, respectively. The bottom row shows the current state-of-the-art model for adversarially robust classification on CIFAR-10 as a reference value.

Model	Clean	ℓ_∞	ℓ_2
SBGC (ours)	**95.04**	0.00	0.00
WideResNet-70-16 + augm. (Rebuffi et al. [38])	92.23	**82.32**	**66.56**

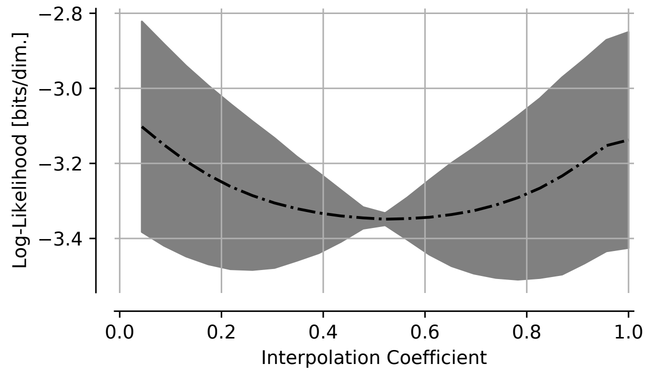

Figure 2: **Interpolation Experiment**. We linearly interpolate between dataset samples and asses the shape of the log-likelihood interpolation function (equation 3). While Fetaya et al. [8] found a concave shape, indicating high likelihoods outside the data distribution, we find a convex function suggesting that the score based generative model does assign lower likelihoods outside the training data distribution. The black line shows the mean and the shaded area indicates the standard deviation of the estimated log-likelihood over 500 image pairs.

3.4 Likelihood Estimation on Interpolated Samples

In a previous study, Fetaya et al. [8] highlighted a problem with generative classifiers. They found that the likelihood of an interpolated image on CIFAR-10 is actually higher than the likelihoods of both the start x_A and endpoint x_B of the interpolation. More precisely, they demonstrated that the function

$$f(t) = p_0\left(t \cdot x_A + (1-t) \cdot x_B\right), \quad t \in [0,1] \tag{3}$$

is concave, which is counter-intuitive because a generative model should assign low likelihood to inputs outside of its training distribution such as nonsensical interpolations in pixel space.

To test this for our SBGC model, we measure the function $f(t)$ for 24 values for $t \in [0,1]$ and average this over 500 image pairs. As depicted in Figure 2, we find that the interpolation of the likelihood is generally convex. Thus, in contrast to the analysis of previous models by [8], our SBGC model assigns lower likelihood values to interpolated images that are outside the (training) data distribution.

4 Conclusion

We have shown that the latest advances in score-based generative modeling of natural images translate into generative classifiers which have highly competitive classification accuracies as well as likelihoods. However, in the search for further benefits, we found that these models show only minor or no improvements on common image corruptions. Similarly, but in contrast to previous results [51], they spectacularly fail on worst-case distribution shifts. This highlights the necessity to perform extensive and, if at all possible, gradient-based adversarial attacks. The lack of adversarial robustness aligns with previous results [21, 25, 1, 27]. Interestingly, this vulnerability remains *despite* resolving the problem of high likelihood on interpolated images [8], and so we can exclude this as a possible explanation for the lacking robustness.

Thus, we are left wondering why generative classification of natural images does not show the same robustness to worst-case distribution shifts as observed on MNIST [40]. A possible explanation might be that the positive results on MNIST were due to the fact that Euclidean distances (on which the likelihood of variational autoencoders with a Gaussian likelihood is based) are a useful metric on MNIST but not on real images. This also led to the provable robustness on MNIST [40]. However, on natural images Euclidean distances are badly aligned with human perception [46, 50, 23] and we have no guarantee about the model behavior even in small ℓ_p epsilon balls around the training data.

[2] Gaussian noise, shot nose, impulse noise

In a nutshell, while we achieve almost the same accuracy on clean data as a baseline discriminative classifier and show high data likelihoods, we find no advantage over discriminative classifiers regarding worst-case distribution shifts and common corruptions.

References

[1] Lynton Ardizzone, Radek Mackowiak, Carsten Rother, and Ullrich Köthe. Training normalizing flows with the information bottleneck for competitive generative classification. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[2] Guillaume Bouchard and Bill Triggs. The tradeoff between generative and discriminative classifiers. In *16th IASC International Symposium on Computational Statistics (COMPSTAT'04)*, pages 721–728, 2004.

[3] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

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

[5] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6572–6583, 2018.

[6] Yanzhi Chen, Renjie Xie, and Zhanxing Zhu. On breaking deep generative model-based defenses and beyond. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1736–1745. PMLR, 2020.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Ethan Fetaya, Jörn-Henrik Jacobsen, Will Grathwohl, and Richard S. Zemel. Understanding the limitations of conditional generative models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[9] David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12):2379–2394, 1987.

[10] Akinori Fujino, Naonori Ueda, and Kazumi Saito. A hybrid generative/discriminative approach to semi-supervised classifier design. In *AAAI*, pages 764–769, 2005.

[11] Tal Golan, Prashant C Raju, and Nikolaus Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337, 2020.

[12] Ian Goodfellow, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Multi-prediction deep boltzmann machines. *Advances in Neural Information Processing Systems*, 26:548–556, 2013.

[13] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[14] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[15] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and

uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[17] An Ju and David Wagner. E-abs: extending the analysis-by-synthesis robust classification model to more complex image domains. In *Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security*, pages 25–36, 2020.

[18] Heewoo Jun, Rewon Child, Mark Chen, John Schulman, Aditya Ramesh, Alec Radford, and Ilya Sutskever. Distribution augmentation for generative modeling. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5006–5019. PMLR, 2020.

[19] Zahra Kadkhodaie and Eero P Simoncelli. Solving linear inverse problems using the prior implicit in a denoiser. *ArXiv preprint*, abs/2007.13640, 2020.

[20] Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *arXiv preprint arXiv:2107.00630*, 2021.

[21] Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *2018 ieee security and privacy workshops (spw)*, pages 36–42. IEEE, 2018.

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

[23] Valero Laparra, Johannes Ballé, Alexander Berardino, and Eero P Simoncelli. Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging*, 2016(16):1–6, 2016.

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[25] Yingzhen Li, John Bradshaw, and Yash Sharma. Are generative classifiers more robust to adversarial attacks? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3804–3814. PMLR, 2019.

[26] Percy Liang and Michael I. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 584–591. ACM, 2008. doi: 10.1145/1390156.1390230.

[27] Radek Mackowiak, Lynton Ardizzone, Ullrich Kothe, and Carsten Rother. Generative classifiers as a basis for trustworthy image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2971–2981, 2021.

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

[29] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013. ISBN 9780262018029 0262018020.

[30] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *ArXiv preprint*, abs/1906.02994, 2019.

[31] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[32] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning, 2011.

[33] Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 841–848. MIT Press, 2001.

[34] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021.

[35] Dylan M Paiton, Charles G Frye, Sheng Y Lundquist, Joel D Bowen, Ryan Zarcone, and Bruno A Olshausen. Selectivity and robustness of sparse coding networks. *Journal of Vision*, 20(12):10–10, 2020.

[36] Rajat Raina, Yirong Shen, Andrew Y. Ng, and Andrew McCallum. Classification with hybrid generative/discriminative models. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pages 545–552. MIT Press, 2003.

[37] Jonas Rauber, Roland Zimmermann, Matthias Bethge, and Wieland Brendel. Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax. *Journal of Open Source Software*, 5(53):2607, 2020.

[38] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *ArXiv preprint*, abs/2103.01946, 2021.

[39] Michael Revow, Christopher KI Williams, and Geoffrey E Hinton. Using generative models for handwritten digit recognition. *IEEE transactions on pattern analysis and machine intelligence*, 18(6):592–606, 1996.

[40] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on MNIST. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[41] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org, 2015.

[42] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ArXiv preprint*, abs/2011.13456, 2020.

[43] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *arXiv e-prints*, pages arXiv–2101, 2021.

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

[45] Sasha Targ, Diogo Almeida, and Kevin Lyman. Resnet in resnet: Generalizing residual architectures. *ArXiv preprint*, abs/1603.08029, 2016.

[46] Patrick C Teo and David J Heeger. Perceptual image distortion. In *Proceedings of 1st International Conference on Image Processing*, volume 2, pages 982–986. IEEE, 1994.

[47] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[48] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013.

[49] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[50] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.

[51] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. *ArXiv preprint*, abs/2106.06041, 2021.

A Appendix

A.1 Training and Implementation Details

We build on the architecture (baseline model with 4 residual blocks) proposed by Song et al. [43] and extend it to condition the score function also on the class label. All in all, our model has $67.19\,\mathrm{M}$ parameters. Furthermore, we use the same SDE (variance preserving) and maximum likelihood training procedure (trained for $950,000$ steps with a batch size of 128 and importance sampling) as Song et al. [43]. The training takes approximately five days on two NVIDIA Tesla V100 GPUs.

For the likelihood inference, i.e., for solving the SDE, we use the Runge-Kutta 4(5) solver of SciPy (*scipy.integrate.solve_ivp*) [49] with absolute and relative precision equal to 10^{-5}. We perform uniform dequantization (i.e., adding uniform noise with a magnitude of $1/256$) and integrate the SDE up to a minimal $\epsilon = 10^{-5}$ for numerical stability [42]. Furthermore, the final likelihood is computed by taking the average over $n = 30$ random draws for the epsilon trace estimator [42]. For the OOD tasks (CIFAR-10-C and adversarial perturbations), we lower this to $n = 10$ to reduce the compute time, thus giving a more conservative lower estimate of the performance on these datasets (see also section A.2.1).

For the WideResNet-28-12 baseline we use an existing implementation[3] in PyTorch. To make this baseline more comparable with our SBGC, we also use dequantization noise for augmenting the training data of the WideResNet.

A.2 Additional Analyses

A.2.1 Convergence Analysis

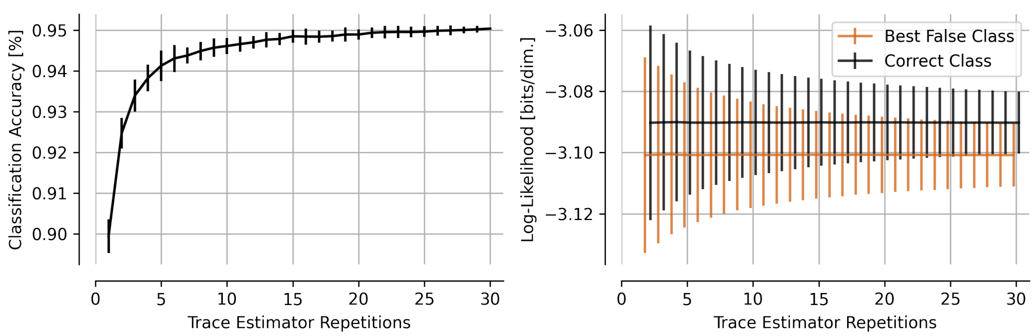

Figure 3: **Convergence Analysis.** The left plot shows the classification accuracy as a function of the number of repetitions for the trace estimator (error bars indicate two standard deviations over random subsamples). The right plot shows the estimate of the log-likelihood as a function of the number of repetitions for the trace estimator (error bars indicate standard error of the mean) for both the correct class (black) as well the wrong class with the highest likelihood (orange).

Here, we study the effect of varying the number of repetitions of the trace estimator (see previous section). In Figure. 3 on the left we see that the classification accuracy improves monotonically with the number of repetitions. A number of $n = 10$ gives nearly asymptotic performance and is, thus, a good trade-off for our OOD tasks which are computationally intensive. For asymptotic performance, we see that the accuracy rises above 95%.

On the right side of Figure. 3, we study the convergence of the trace estimator which is part of the likelihood (equation 2. We can see that the standard error of the mean (errorbars) converges very slowly. We therefore report the likelihood at $n = 30$ repetitions. Note also that the likelihood of the correct class converges (averaged across all test images) to a higher value as that of the wrong class with the highest likelihood. This is expected for a model with good classification accuracy. Thus, provided labels, our model attains slightly higher likelihoods than in the unconditioned case.

[3] `https://github.com/meliketoy/wide-resnet.pytorch`

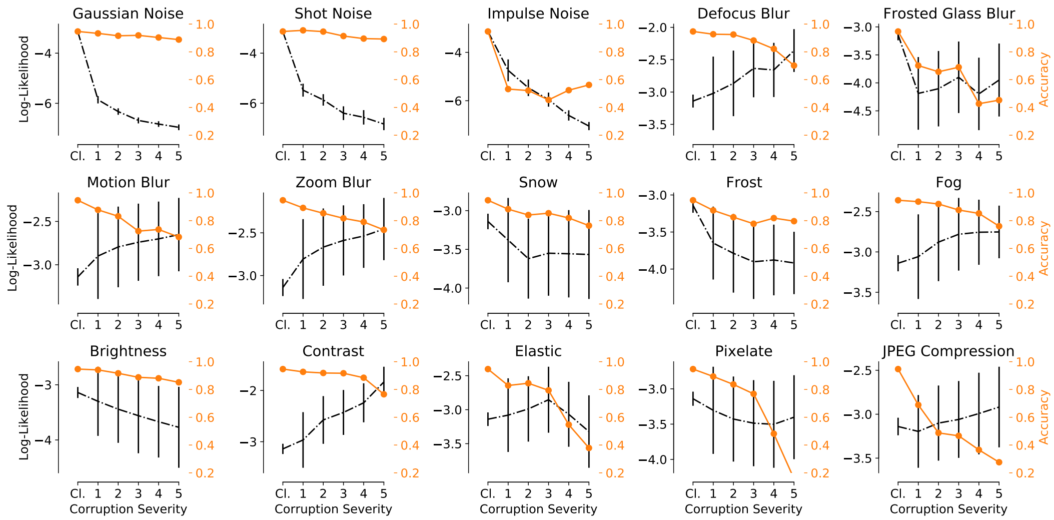

Figure 4: **Detailed Results on CIFAR-10-C.** We show the log-likelihood (orange, left axis) and the accuracy (black, right axis) across all image corruption types and severities (horizontal axis) for the CIFAR-10-C benchmark [14]. Note that the severity level *Cl.* corresponds to clean samples.

A.2.2 CIFAR-10-C Performance

Here, we further resolve the model performance on CIFAR-10-C for all different corruption types and strengths in Figure 4. Interestingly, the results highlight that likelihood and accuracy need not always be aligned. For the different corruptions based on blurring or some sort of low-pass filtering (e.g., fog or JPEG compression), we observe that while the accuracy decreases (as signal is lost) the likelihood under the model actually increases.

This observation is in line with previous studies [31]. We hypothesize that this highlights the fact that natural images follow a $1/f$ power spectrum [9] and, thus, low-pass filtering will increase the likelihood of any image. This also explains previous observations of non-natural images obtaining high likelihoods under generative models: Starting from a non-natural image we can always blur the image to obtain high likelihoods under our model (see, e.g., our Figure 4 zoom blur and Table 3 in [8]). Similarly, we can start from a natural image (e.g., from the training set) and add white noise. The high-frequency perturbations will quickly deteriorate the likelihoods (see Figure 4 Gaussian noise). Relatedly, Nalisnick et al. [30] found that a measure of dataset complexity can be leveraged in conjunction with likelihood estimates to perform OOD detection.