ON THE EFFICACY OF GENERALIZATION ERROR PREDICTION SCORING FUNCTIONS

Puja Trivedi[⋆] *Danai Koutra*[⋆] *Jayaraman J. Thiagarajan*[†]

[⋆] University of Michigan
[†]Lawrence Livermore National Laboratory

ABSTRACT

Generalization error predictors (GEPs) aim to predict model performance on unseen distributions by deriving dataset-level error estimates from sample-level scores. However, GEPs often utilize disparate mechanisms (e.g., regressors, thresholding functions, calibration datasets, etc), to derive such error estimates, which can obfuscate the benefits of a particular scoring function. Therefore, in this work, we rigorously study the effectiveness of popular scoring functions (confidence, local manifold smoothness, model agreement), independent of mechanism choice. We find, absent complex mechanisms, that state-of-the-art confidence- and smoothness- based scores fail to outperform simple model-agreement scores when estimating error under distribution shifts and corruptions. Furthermore, on realistic settings where the training data has been compromised (e.g., label noise, measurement noise, undersampling), we find that model-agreement scores continue to perform well and that ensemble diversity is important for improving its performance. Finally, to better understand the limitations of scoring functions, we demonstrate that simplicity bias, or the propensity of deep neural networks to rely upon simple but brittle features, can adversely affect GEP performance. Overall, our work carefully studies the effectiveness of popular scoring functions in realistic settings and helps to better understand their limitations.

Index Terms— Generalization, Data Augmentation, Out-of-Distribution

1. INTRODUCTION

Safe deployment of machine learning models requires suitable failure indicators so that models whose performance falls below an acceptable tolerance can be temporarily pulled from production. While learning-theoretic complexity measures can be used to estimate model performance under *i.i.d* assumptions [1, 2], they are currently insufficient for estimating model generalization on *out of distribution (o.o.d)* data.

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344. Supported by the ASCR Co-Design project. It was also partially supported by the National Science Foundation under CAREER Grant No. IIS 1845491. Code available: `https://github.com/pujacomputes/icassp23-gengap.git`

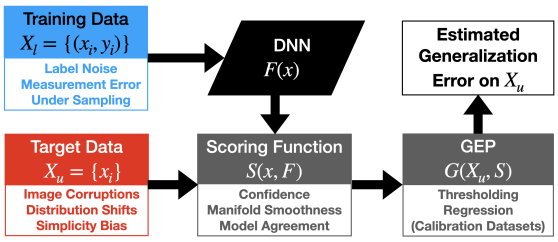

Fig. 1. **Generalization Error Prediction.** We focus on the problem of generalization error prediction with classifiers and study the design of scoring functions under various distribution shifts, corruptions, and data fidelity issues.

To this end, generalization error predictors (GEPs), which are designed to estimate performance on *arbitrary* target datasets, have become popular [3, 4, 5, 6, 7].

In brief, GEPs *aggregate* sample-level *scores* to predict generalization error on unlabeled target datasets (see Fig. 1). Popular scoring functions, such as manifold proximity [6], confidence estimates [4], local manifold smoothness [3], and agreement between independently trained models [8, 9, 10], attempt to measure the likelihood that the predicted label of a sample is correct. GEPs then use different mechanisms (thresholding functions, regressors, calibration datasets) to create dataset-level error estimates from the provided scores. However, since these mechanisms can vary in complexity, the efficacy of a particular scoring functions can obfuscated. For example, state-of-the-art GEPs often rely upon multiple, labeled calibrated datasets [6, 11, 4, 12], which can provide additional information that bolsters the performance of otherwise subpar scoring functions. Therefore, in this paper, we use a simple, fixed GEP, and rigorously study the effectiveness of popular scoring functions (confidence, local manifold smoothness, model agreement) in several realistic settings and identify a potential cause of poor GEP performance.

GEP performance under distribution shifts (Sec. 3.1): Using a large family of image-level corruptions and distribution shifts, we benchmark the three scoring functions and provide key insights on their efficacy in practice.

Impact of training data fidelity on GEP performance (Sec. 3.2): Scoring functions directly depend on data and model properties. Therefore, we consider common data fidelity is-

sues (label noise, measurement errors and sampling discrepancies) to study what role data quality plays on the GEP performance.

Effect of simplicity bias on GEP performance (Sec. 4): Deep neural networks are susceptible to relying upon simple, spurious features [13] at the expensive of robust generalization. We study the impact of this behavior on the efficacy of different scoring functions [13, 14].

2. PRELIMINARIES

We begin by formally introducing the problem setting and scoring functions. Let $\mathcal{X}_u = \{\bar{\mathrm{x}}_i\}$ be an unlabeled, target dataset and $\mathcal{X}_\ell = \{(\mathrm{x}_i, \mathrm{y}_i)\}$ be a labeled training dataset where y_i is one of C classes. Further, let $F : x \rightarrow [0,1]^C$ be a model (e.g., DNN) trained on $\mathcal{X}_\ell$ that outputs softmax probabilities over C classes. GEPs utilize scores computed from model features on the target data distribution, which are expected to be correlated with dataset performance. We focus on popular sample-level scoring functions, $\mathrm{S}(\mathrm{x}; \mathrm{F}) \rightarrow \mathbb{R}$, and define them below.

- **Confidence** (`Conf`) [4]. We can directly obtain a sample-level score from F by using the maximum softmax probability (e.g., the predicted class's confidence):

$$\mathrm{S}(\mathrm{x}; \mathrm{F}) = \max \mathrm{F}(\mathrm{x}). \tag{1}$$

- **Local Manifold Smoothness** (`LMS`) [3]. Let $q(\mathrm{x}'|\mathrm{x})$ be a local probability distribution over augmented samples, x', that can be generated from a given natural sample, x. Then, the LMS score is defined as $\mathrm{S}(\mathrm{x}) = \mathbb{E}_{\mathrm{x}' \sim q(\mathrm{x}'|\mathrm{x})} \quad \mathbb{I}[\mathrm{F}(\mathrm{x}') = \mathrm{F}(\mathrm{x})]$, where $\mathbb{I}$ is the indicator function. Note, in practice, the expectation over q is approximated by sampling set of k augmented samples $\mathrm{X}' := \{\mathrm{x_j}' \sim q(\mathrm{x}'|\mathrm{x})\}_{j=1}^{k}$ and we define q using RandAug [15].

$$\mathrm{S}(\mathrm{x}; \mathrm{F}) \approx \frac{1}{k} \sum_{j=1}^{k} \quad \mathbb{I}[\mathrm{F}(\mathrm{x_j}') = \mathrm{F}(\mathrm{x})]. \tag{2}$$

- **Model Agreement** (`MA`) [16, 8, 9]. Let $F_{0 \ldots r}$ be a set of r independently trained models (e.g., models are trained using r different seeds). WLOG, let F_0 be the base model for which the accuracy is estimated. Then the score can be computed as:

$$\mathrm{S}(\mathrm{x}; \mathrm{F}) = \frac{1}{r-1} \sum_{i=1}^{r} \mathbb{I}[\mathrm{F}_0(\mathrm{x}) = \mathrm{F}_r(\mathrm{x})] \tag{3}$$

Given these sample-level scores, GEPs then estimate the performance of deep neural networks (DNNs) under a wide-variety of distribution shifts and are defined as follows.

Definition 2.1. For an unlabeled dataset $\mathcal{X}_u$, a generalization error predictor $\mathrm{G}(\mathcal{X}_u; \mathrm{S})$ returns the estimated error of the pretrained classifier F, based on the scores from $\mathrm{S}(\mathcal{X}_u; F)$.

State-of-the-art approaches propose to curate multiple (labeled) calibration datasets or train multiple models (with different hyper-parameters) in order to construct a well-calibrated GEP [6, 11, 4, 12]. However, it can be difficult to obtain such calibration datasets in practice and training multiple models is expensive. Moreover, using such strategies can obfuscate the effectiveness of a given scoring function. Therefore, we focus on a simple, popular *thresholding-based* GEP. This GEP simply aggregates thresholded sample-level scores to obtain a dataset-level estimate: $\frac{1}{|X|} \sum_i \mathbb{I}(\mathrm{S}(\bar{\mathrm{x}}_i; \mathrm{F}) > \tau)$, where the threshold hyperparameter τ is identified by regressing G to recover the true accuracy on a pre-defined, validation dataset. Given this fixed GEP, we are ready to assess the performance of different scoring functions in a fair setting.

3. A CLOSER LOOK AT SCORING FUNCTIONS

As discussed in Sec. 1, it is critical to disentangle the scoring function from the GEP mechanism to understand the former's effectiveness. Using a fixed thresholding-based GEP, we evaluate the ability of different scoring functions to accurately predict generalization over various distribution shifts (Sec. 3.1) and under the realistic setting of training on low fidelity data (Sec. 3.2).

Experimental Setup. For all experiments, CIFAR10 is the source distribution on which we train ResNet-18 for 200 epochs with lr=0.05. STL10, CIFAR10.1 and CIFAR-10-C [17] are the target distributions, for which we estimate the generalization performance. CIFAR10.1 and STL10 represent near and far distribution shifts respectively. CIFAR-10-C contains samples generated from 15 different naturalistic corruptions, such as "fog" or "blur", applied at five severity levels. Increased severity corresponds to increased shift from the training data. In all experiments, we report the mean absolute error (MAE) between the true target accuracy and the predicted target accuracy. The threshold, τ, is determined by optimizing the thresholding function to minimize prediction error on the CIFAR10 validation dataset. Note that all results are averaged over 10 seeds. We compute `MA` using a 10-member ensemble and use 10 augmentations (RandAug [15]) to compute `LMS`.

3.1. GEP Performance under Distribution Shifts

Results are shown in Table 1. We make the following observations. Across all target datasets, and corresponding levels of distribution shifts, `MA` is by far the most effective at predicting generalization. For example, on the challenging STL10 benchmark, `MA` achieves improvements of 10%

Table 1. Assessing Scoring Functions Under Distribution Shifts. We report the mean absolute error between the true and estimated accuracies obtained with different scoring functions. We average CIFAR-10-C corruptions by severity. The top two methods are **bolded** and underlined respectively.

S(x; F)	STL10	CIFAR10.1	Sev. 1	Sev. 2	Sev. 3	Sev. 4	Sev. 5
Conf	0.1488	0.0117	0.0246	0.0430	0.0476	0.0596	0.0661
LMS	0.3438	0.0961	0.0942	0.1350	0.1704	0.2147	0.2739
MA	**0.0420**	**0.0054**	**0.0115**	**0.0189**	**0.0154**	**0.0294**	**0.0519**

and 30% over Conf and LMS scores respectively. Similarly, MA provides consistent gains at all corruption levels on CIFAR-10-C. Notably, while LMS was originally proposed in the context of sample-level scoring, it trails behind Conf, which was shown to more effective as a distribution-level scoring function by [4]. While we used RandAug as it was shown to be effective by [3], our results here indicate that smoothness to such perturbations can fail to capture properties relevant for generalization under the real-world corruptions and shifts. This suggests that the choice of augmentation strategy is critical to LMS's effectiveness, and it is non-trivial to identify such a strategy without access to *o.o.d* data.

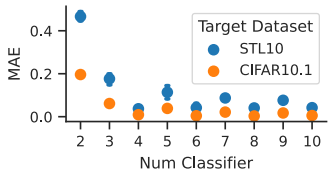

Fig. 2. Effect of ensemble size on MA performance.

Given that MA requires multiple independently trained models, we evaluate the effect of ensemble size in generating reliable scores. As shown in Fig. 2, we find that the performance of MA begins to saturate at 4 models, though further increase in the ensemble size reduces the variability. In Sec. 3.2, we will show that the diversity of the ensemble also plays an important role on GEP performance, particularly under challenging training conditions.

3.2. Impact of Training Data Fidelity on GEP

In the preceding section, we evaluated the effectiveness of different scoring functions on a large family of image-level corruptions and distribution shifts. While we found MA and Conf to be particularly effective, here, we seek to further evaluate their efficacy in the realistic, but more challenging setting where *training data* may be compromised. Specifically, we consider *label noise, measurement errors and sampling discrepancies* as sources of low-fidelity training data. We focus on these particular sources as they not only represent situations that are likely to be encountered in practical scenarios, but they also compromise the training data at different granularities, namely sample-, dataset-, and distribution-level. Given that GEPs are used as failure indicators, it is critical to evaluate the scoring functions in such settings. Additionally, we evaluate a variant of MA designed to increase ensemble diversity and further improve its performance.

Table 2. Effect of Data Fidelity Issues on GEP Performance. We evaluate the effectiveness of scoring functions when the training data contains measurement error (MN), sampling discrepancies (US) and label noise (LN). Note, we include a variant of MA , where we improve the diversity of the ensemble by synthetically injecting label corruptions to 2% of the samples.

Dataset	S(x; F)	True Accuracy			GEP Performance	
		CIFAR10	STL10	CIFAR10.1	STL10	CIFAR10.1
CIFAR10	Conf	0.9008	0.5484	0.8056	0.1488	0.0117
	LMS				0.3438	0.0961
	MA	**0.9408**	**0.5990**	**0.8665**	0.0420	**0.0054**
	$MA_{0.02}$	0.9334	0.5953	0.8555	**0.0115**	0.0179
CIFAR10 (MN)	Conf	0.8284	0.5411	0.7155	0.1010	0.0206
	LMS				0.3028	0.1464
	MA	**0.8782**	0.5987	**0.7755**	0.0528	**0.0108**
	$MA_{0.02}$	0.8716	**0.6092**	0.7670	**0.0219**	0.0123
CIFAR10 (US)	Conf	0.9015	0.5943	0.7987	0.1329	0.0207
	LMS				0.3300	0.1061
	MA	**0.9300**	**0.6439**	**0.8535**	0.0451	0.0130
	$MA_{0.02}$	0.9289	0.6250	0.8360	**0.0236**	**0.0060**
CIFAR10 (LN)	Conf	0.8497	0.4736	0.7373	0.1802	0.0253
	LMS				0.3569	0.1190
	MA	**0.9209**	**0.5394**	**0.8315**	0.0709	0.0307
	$MA_{0.02}$	0.9105	0.4611	0.8060	**0.0601**	**0.0153**

Experimental Setup. The following processes are used to create compromised data: (i) **Label Noise:** We randomly select 5% of the training set and randomly flip their labels. (ii) **Measurement Noise:** We first apply a Gaussian blur ($\sigma_1 = 0.5$) and then add standard normal noise ($\sigma_2 = 0.07$) to all training images; (iii) **Under-Sampling:** 20% of the samples are randomly dropped from the *automobile*, and *bird* classes. Given these compromised datasets, we follow the same experimental setup introduced in Sec. 3. However, models are now trained for 250 epochs, instead of 200 epochs, to achieve acceptable convergence. In the following analysis, we specifically focus on the near (CIFAR10.1) and far (STL10) distribution shift settings respectively.

MA with improved diversity. Motivated by the saturating effect of ensemble size in Fig. 2, we propose a simple variant to MA that is designed to improve the diversity of the ensemble by synthetically corrupting the data using low levels of label noise. Here, the intuition is that such label noise requires models to learn slightly different functions to effectively minimize the loss on the mis-labeled samples. Since these randomly-labeled points account for a small portion (2%) of the overall dataset, they generally do not substantially affect the ensemble accuracy (see Table 2). We denote this variant, $MA_{0.02}$, and discuss its behavior, in addition to other scoring functions below (see Table 2).

Obs. 1: MA and Conf remain considerably more effective than LMS even with compromised data. This is to be expected as the considered data infidelities are not expected to change the type of smoothness that is indicative of generalization on STL10 and CIFAR10.1.

Obs. 2: While adding measurement noise (MN) and under-

Table 3. Effect of Simplicity Bias on GEP Performance. We evaluate the effectiveness of GEPs when the training data and target data contain varying degrees of correlation between simple/complex features. MAE is reported, and we **bold** the best score per correlation level. Observe that scores suffer when the correlation is broken.

S(x; F)	Corr. = 0.95		Corr. = 0.99		Corr. = 1.0	
	Corr.	Rand.	Corr.	Rand.	Corr.	Rand.
Conf	0.0184	**0.0350**	0.0136	**0.2489**	**0.0005**	**0.7927**
LMS	0.0557	0.2438	**0.0056**	0.5571	0.0028	0.8699
MA	**0.0084**	0.1262	0.1019	0.5125	**0.0007**	0.8492

sampling (US) does minimally harm the true target accuracy, we see GEP performance of LMS and Conf improves on STL10–even better than their performance obtained by training on the clean dataset. We posit in Sec. 4 that this can be attributed to decreased reliance upon simple features that lead to over-confident, but ultimately misleading, scores.

Obs. 3: In contrast, on the near *o.o.d* setting of CIFAR10.1, we not only see that the true target accuracy decreases, but also that GEP performance decreases across all methods relative to training on the clean dataset. Given the closeness of CIFAR10.1 to CIFAR10, it is likely that low fidelity data struggles to capture the features necessary for generalization or estimation on the target distribution.

Obs. 4: Incurring only a small drop in the true target accuracy, our variant $\text{MA}_{0.02}$ improves GEP performance on STL (over 50% on CIFAR10, CIFAR10 (MN), CIFAR10 (US)), and maintains comparable GEP performance on the near *o.o.d* setting. While target accuracy does decrease noticeably on CIFAR10 (LN), we note that this is an edge case, as the underlying dataset has already been perturbed by label noise. The efficacy of $\text{MA}_{0.02}$ on CIFAR10.1 is surprising, as the perturbed distributions are certainly further from target than the clean source distribution, but this suggests diversity plays a critical role in obtaining effective scores.

4. EFFECT OF SIMPLICITY BIAS ON GEPS

In the preceding section, we found that training on compromised data may in fact decrease the GEP error. We hypothesize that this decrease can be partially attributed to the noisy datasets, mitigating *simplicity bias*, i.e., the well known propensity of deep neural networks to rely upon simple, spurious features in lieu of more complex/expressive ones [18, 19, 13, 20]. Given that simple features are not expected to generalize on *o.o.d* datasets, but DNNs remain susceptible to relying upon such simple features, we posit that GEPs will also see decreased performance on distributions where simple features are no longer indicative of the label. We test this hypothesis using a synthetic setting that controls the discriminability of simple features on target datasets, as discussed below.

Experimental Setup. We use a custom "dominoes" dataset

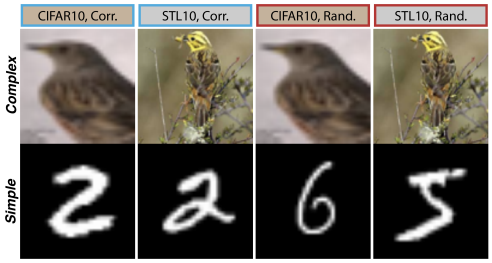

Fig. 3. Simplicity Bias Dataset, (Fig. 2 [14]). Dominoes comprised of complex (CIFAR10) and simple (MNIST) features are used to control the simplicity bias on target datasets.

[13] of complex and simple features by pairing each class from CIFAR10 (complex feature) with the corresponding digit class in MNIST (simple feature) [14]. (See Fig. 3). Three levels of correlation (95%, 99%, 100%) between the target and simples features are considered during training. When predicting generalization, we sample complex features from STL10, as well as create a variant that randomizes the spurious correlation between simple and complex features. We fine-tune a MoCo-V2 pretrained ResNet-50 [21] for 20 epochs with lr=0.001 and average results over 3 seeds.

As shown in Table. 3, we see that the prediction error often substantially increases when evaluating on the randomized target dataset, e.g., where the simple feature is no longer predictive. While we would expect that the target accuracy decreases, the decreased GEP performance is particularly troubling as such methods are intended to detect these very failures. Moreover, we note that as the correlation between the simple and complex feature increases (Corr=0.95 vs. Corr=1.0), the gap between GEP's performance on the Corr. and Rand. variants of the target dataset increases. Indeed, the Corr. MAE decreases as the training dataset correlation increases (Corr=0.95 vs. 1.0), but the Rand. MAE increases. This result further highlights the harmful role of simplicity bias on GEP performance.

5. CONCLUSION

In this work, we rigorously studied the design of scoring functions in GEPs and found that their choice is critical to produce consistently reliable predictors across different distribution shifts and noise corruptions. In fact, when the GEP construction does not involve calibration datasets or training a large family of models, even state-of-the-art scoring functions such as Conf and LMS can struggle. In comparison, we found MA to be a more reliable alternative. Furthermore, using our new MA variant, we demonstrated that improving diversity of the ensemble leads to well-calibrated GEPs, while incurring only a small drop in target accuracies. Finally, in a controlled empirical setting, we showed how reliance on simple features can adversely affect the GEP performance.

6. REFERENCES

[1] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio, "Fantastic generalization measures and where to find them," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2020.

[2] Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi, "Leveraging unlabeled data to predict out-of-distribution performance," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2022.

[3] Nathan Ng, Neha Hulkund, Kyunghyun Cho, and Marzyeh Ghassemi, "Predicting out-of-domain generalization with local manifold smoothness," *CoRR*, vol. abs/2207.02093, 2022.

[4] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt, "Predicting with confidence on unseen distributions," in *Proc. Int. Conf. on Computer Vision (ICCV)*, 2021.

[5] Mayee F. Chen, Karan Goel, Nimit Sharad Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Ré, "Mandoline: Model evaluation under distribution shift," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2021.

[6] Weijian Deng and Liang Zheng, "Are labels always necessary for classifier accuracy evaluation?," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[7] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu, "Extremely simple activation shaping for out-of-distribution detection," *CoRR*, vol. abs/2209.09858, 2022.

[8] Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J. Zico Kolter, "Assessing generalization of SGD via disagreement," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2022.

[9] Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, and Somesh Jha, "Detecting errors and estimating accuracy on unlabeled data with self-training ensembles," in *Proc. Adv. in Neural Information Processing Systems (NeurIPS)*, 2021.

[10] Preetum Nakkiran and Yamini Bansal, "Distributional generalization: A new kind of generalization," *CoRR*, vol. abs/2009.08092, 2020.

[11] Weijian Deng, Stephen Gould, and Liang Zheng, "What does rotation prediction tell us about classifier accuracy under varying testing environments?," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2021.

[12] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Advances in Neural Information Processing Systems NeurIPS*, 2017.

[13] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli, "The pitfalls of simplicity bias in neural networks," in *Proc. Adv. in Neural Information Processing Systems (NeurIPS)*, 2020.

[14] Puja Trivedi, Danai Koutra, and Jayaraman J. Thiagarajan, "A closer look at model adaptation using feature distortion and simplicity bias," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2023.

[15] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. Adv. in Neural Information Processing Systems (NeurIPS)*, 2020.

[16] Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio, "Predicting the generalization gap in deep networks with margin distributions," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2019.

[17] Dan Hendrycks and Thomas G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proc. Int. Conf. on Learning Representations, (ICLR)*, 2019.

[18] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz, "SGD learns over-parameterized networks that provably generalize on linearly separable data," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2017.

[19] Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nati Srebro, "Implicit bias of gradient descent on linear convolutional networks," in *Proc. Adv. in Neural Information Processing Systems (NeurIPS)*, 2018.

[20] 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 *Proc. Int. Conf. on Learning Representations (ICLR)*, 2019.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.