



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

A Closer Look at Scoring Functions and Generalization Prediction

P. Trivedi, D. Koutra, J. J. Thiagarajan

November 14, 2022

IEEE ICASSP 2023
Rhodes Island, Greece
June 4, 2023 through June 9, 2023

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

A CLOSER LOOK AT SCORING FUNCTIONS AND GENERALIZATION PREDICTION

Puja Trivedi*, Danai Koutra

University of Michigan

Jayaraman J. Thiagarajan

Lawrence Livermore National Laboratory

ABSTRACT

Safe model deployment requires detection and mitigation of poor performance. To this end, generalization error predictors (GEPs), which aim to predict error on previously unseen distributions using sample-level scoring functions, have become increasingly popular. Such predictors often rely upon mechanisms of disparate efficacy to convert sample-level scores to dataset-level predictions, making the benefits of using different scoring functions unclear. Thus, in this work, we rigorously study the design of GEP scoring functions, independent of such unfair mechanisms. As we show, even state-of-the-art confidence- and smoothness- based scoring functions can struggle when extensive calibration datasets are not utilized. In comparison, we find that simple ensemble-disagreement-based scores are often a more reliable alternative for predicting generalization error under distribution shifts and corruptions. We then evaluate different scoring functions in the realistic setting where the model has been trained on low fidelity data (e.g., label noise, corruptions). While ensemble-disagreement continues to perform well, our results highlight improving the ensemble diversity is important for good prediction performance. Finally, we demonstrate that simplicity bias, or the propensity of neural networks to rely upon simple but brittle features, can adversely affect a GEP’s performance, and empirically demonstrate this using a controlled setting. Overall, our work carefully studies the quality of different scoring functions to better understand GEP performance.

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

1. PROBLEM SETUP

Safe deployment of a machine learning model mandates generalize to unseen, potentially distribution-shifted data. While learning-theoretic complexity measures can partially estimate generalization under *i.i.d* assumptions [1, 2], they are as yet insufficient for estimating performance under *Out of Distribution* (*o.o.d*) data. In such situations, suitable *failure indicators* are often required to notify practitioners of potential model

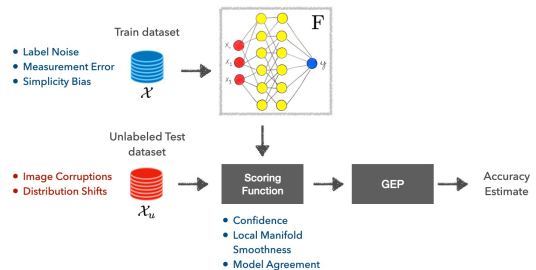


Fig. 1. We focus on the problem of generalization error prediction with classifiers and study the design of scoring functions. We consider a variety of distribution shifts/corruptions to construct the target datasets, and explore the effect of different data fidelity issues on GEP performance.

failure (e.g., the accuracy might fall below the allowed tolerance), so that such models can be pulled from deployment or updated with additional data. To this end, generalization error predictors (GEPs), which attempt to estimate a model’s performance (e.g., classification error or accuracy) on arbitrary test datasets, have become popular [3, 4, 5, 6, 7].

Formally, given a model $F : \mathbb{R}^n \rightarrow \mathbb{R}$ trained using labeled data $\mathcal{X}_l = \{(x_i, y_i)\}$, our goal is to obtain an estimate of the accuracy of F for an unlabeled, target dataset $\mathcal{X}_u = \{\bar{x}_i\}$. To design a GEP, one thus computes model features from the target data distribution, which are expected to be correlated to performance on that dataset. Here, it is common to define a sample-level *scoring function*, as follows.

Definition 1.1. For a given classifier F , a scoring function $S(x; F)$ measures the likelihood that the prediction $F(x) = y$, where y is the true label for x .

Popular scoring functions include manifold proximity [6], confidence estimates [4], local manifold smoothness [3], and agreement between independently trained models [8, 9, 10]. In this paper, we investigate the design of scoring functions when using GEPs to estimate the performance of deep neural networks under a wide-variety of distribution shifts. Specifically, given a scoring function, a GEP is defined as follows.

Definition 1.2. For an unlabeled dataset $\mathcal{X}_u$, a generalization error predictor $G(\mathcal{X}_u; S)$ returns the estimated accuracy of the pretrained classifier F , based on the scores from $S(\mathcal{X}_u; F)$.

This work began while PT was an intern at LLNL.

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.

Generally, a GEP estimates accuracy on an unseen dataset by simply aggregating thresholded sample-level scores, i.e., $\sum_i \mathbb{I}(S(\bar{x}_i; F) > \tau)$, where $\mathbb{I}$ is the indicator function and the threshold hyperparameter τ is identified by regressing G to recover the true accuracy on a pre-defined, validation dataset.

Note that, 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]. Not only is it difficult to obtain such calibration datasets in practice, using such a sophisticated strategy obfuscates the utility of a particular scoring mechanism. Hence, we focus on thresholding-based GEP construction and, for the first time, study the behavior of state-of-the-art scoring functions (confidence, local manifold smoothness, model agreement) in the following settings:

- **GEP performance under distribution shifts:** 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:** In general, scoring functions directly depend on data and model properties. Hence, we consider common data fidelity issues (label noise, measurement errors and sampling discrepancies) and study what role data quality plays on the GEP performance.
- **Effect of simplicity bias on GEP performance:** Deep neural networks are known to be susceptible to build decision rules based on simple, spurious features [13], thus affecting their generalization. We use a synthetic dataset to study the impact of this behavior on the efficacy of different scoring functions [13, 14].

2. SCORING FUNCTIONS FOR BUILDING GEP

We first revisit the scoring functions used in our study for constructing GEPs. Note that all these functions provide sample-level scores and we do not consider scores which require additional calibration datasets for implementing GEPs.

Confidence (Conf). Guillory *et al.* [4] recently showed that the difference of the average confidences between a target distribution and the original distributions can be an effective predictor of the generalization gap as it accurately reflects the uncertainty between distributions. As we require a sample-level score, we directly use the confidence of a given sample measured as the maximum softmax probability.

Local Manifold Smoothness (LMS). Based on the intuition that samples for which predictions remain smooth under input-space perturbations are more likely to be predicted correctly, Ng *et al.* [3] proposed a *local manifold smoothness* (LMS) score. Formally, for a given sample x , this can be

defined as the fraction of points sampled from the local density $q(x'|x)$, where x' is an augmented variant of x , that are assigned the same label as x :

$$S(x) = \mathbb{E}_{x' \sim q(x'|x)} [\mathbb{I}[F(x') = F(x)]].$$

Model Agreement (MA). Recently, Jiang *et al.* [8] and Chen *et al.* [9] demonstrated that the disagreement rate between two independently trained DNNs (different seeds) is an effective indicator of the “generalization gap”, i.e., difference between test accuracy on a source and target dataset, attributing this to the well-calibrated nature of DNN ensembles [12]. In this work, we follow Jiang *et al.* [8] and compute the score for a sample as the fraction of models that disagree. More specifically, given a ‘base model’, and set of other ensemble members, the disagreement between the base model and each ensemble member is first computed. Then the percentage of disagreements over the ensemble is reported as the score, and prediction error is measured with respect to the base model.

3. A CLOSER LOOK AT SCORING FUNCTIONS

As discussed in the Sec. 1, it is critical to disentangle the scoring function (Conf, LMS, MA) from the GEP implementation (thresholding vs. training with calibration data) to understand the former’s effectiveness. Using the simple, thresholding-based GEP, we evaluate the ability of different scoring functions to accurately predict generalization under distribution shifts, as well as, varying levels of corruptions.

Experimental Set-up. 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 [15] 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. This represents increasing levels of 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 for implementing GEP is determined for each scoring function using the CIFAR10 validation dataset, such that G can recover the validation accuracy. Note that all results are averaged over 10 seeds. We compute MA using a 10-member ensemble and use 10 augmentations (RandAug [16]) 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. We highlight top two methods using **bold** and underline 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-10C. 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 fails to capture smoothness that is relevant for generalization under the real-world corruptions and shifts that we considered. This suggests that the choice of augmentation strategy is critical to LMS’s effectiveness, and it is non-trivial to identify such a strategy as access to *o.o.d* cannot be assumed.

Given that MA requires multiple independently trained models, we evaluated 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 Section 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. Given that GEPs are used as failure indicators, it is critical to evaluate the scoring functions in these settings. We begin by discussing the data fidelity issues and then present our observations.

Data Infidelity. We evaluate GEPs on models which were trained on data with: label noise, measurement errors and sampling discrepancies. These particular sources of infidelity are considered as they not only represent situations that are likely to be encountered in practical scenarios, they also com-

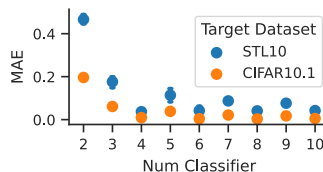


Fig. 2. Effect of ensemble size on MA performance.

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}	<u>0.9334</u>	<u>0.5953</u>	<u>0.8555</u>	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	<u>0.5987</u>	0.7755	<u>0.0528</u>	0.0108
	MA _{0.02}	<u>0.8716</u>	0.6092	<u>0.7670</u>	0.0219	<u>0.0123</u>
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	<u>0.0451</u>	<u>0.0130</u>
	MA _{0.02}	<u>0.9289</u>	<u>0.6250</u>	<u>0.8360</u>	0.0236	0.0060
CIFAR10 (LN)	Conf	0.8497	<u>0.4736</u>	0.7373	0.1802	0.0253
	LMS				0.3569	0.1190
	MA	0.9209	0.5394	0.8315	<u>0.0709</u>	0.0307
	MA _{0.02}	<u>0.9105</u>	0.4611	<u>0.8060</u>	0.0601	0.0153

promise the training data at different granularities, namely sample-, dataset-, and distribution- level. This enables a better understanding of the susceptibilities of different scoring functions. The following processes are used to create such noisy data: (i) **Label Noise:** We randomly select 5% of the training set and randomly flip their labels. The same label noise is used across different model training strategies; (ii) **Measurement Noise:** We first apply a Gaussian blur ($\sigma_1 = 0.5$) to all images and then add standard normal noise ($\sigma_2 = 0.07$) to them; (iii) **Under-Sampling:** For two of the ten classes (*automobile*, *bird*), 20% of the samples are randomly dropped.

Given these *noisy* datasets, we follow the same experimental set-up introduced in Sec. 3 and make the following observations from the results in Table 2. We note that, in this case, the models were trained for 250 epochs, instead of 200 epochs, to achieve acceptable convergence. In this 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 of 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 effect the ensemble accuracy (see Table. 2). We denote this variant, MA_{0.02}, and discuss its behavior, in addition to other scoring functions below.

Obs. 1: MA and Conf remain considerably more effective than LMS even with compromised data. This is to be expected

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. Scoring functions struggle when this 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	.0557	0.2438	0.0056	0.5571	0.0028	0.8699
MA	0.0084	0.1262	0.1019	0.5125	0.0007	0.8492

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-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. 3.3 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 the target distribution to CIFAR10, it is likely that low fidelity data does not capture the features necessary for generalization or prediction as well with the target distribution.

Obs. 4: Incurring only a small drop in the true target accuracy, our variant 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 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.

3.3. Simplicity Bias and Generalization Gap Prediction

In the preceding section, we found that training on data with fidelity issues may in fact decrease the GEP error. We hypothesize that this decrease can be 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, more expressive ones [17, 18, 13, 19]. Given that simple features are not expected to hold for *o.o.d* datasets, but the models underlying GEPs are susceptible to overly relying upon such simple features, we posit that GEPs will see decreased performance under simplicity bias. We test this hypothesis using a synthetic setting that controls the

discriminability of simple features on target datasets, as discussed below.

Experimental Set-up. We use a custom “dominoes” dataset [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 simple 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 [20] 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.



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.

4. 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. Importantly, we also showed that as low as four ensemble members are sufficient to produce effective GEPs. 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. In addition to these crucial insights, this paper also provides a rich evalu-

ation platform to effectively advance the research in generalization error prediction with deep networks.

5. 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, “Exploring the design of adaptation protocols for improved generalization and machine learning safety,” *CoRR*, vol. abs/2207.12615, 2022.
- [15] Dan Hendrycks and Thomas G. Dietterich, “Benchmarking neural network robustness to common corruptions and perturbations,” in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2019.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] 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.