On the Strong Correlation Between Model Invariance and Generalization

Weijian Deng **Stephen Gould** **Liang Zheng**
Australian National University
{firstname.lastname}@anu.edu.au

Abstract

Generalization and invariance are two essential properties of machine learning models. Generalization captures a model's ability to classify unseen data while invariance measures consistency of model predictions on transformations of the data. Existing research suggests a positive relationship: a model generalizing well should be invariant to certain visual factors. Building on this qualitative implication we make two contributions. First, we introduce effective invariance (EI), a simple and reasonable measure of model invariance which does not rely on image labels. Given predictions on a test image and its transformed version, EI measures how well the predictions agree and with what level of confidence. Second, using invariance scores computed by EI, we perform large-scale quantitative correlation studies between generalization and invariance, focusing on rotation and grayscale transformations. From a model-centric view, we observe generalization and invariance of different models exhibit *a strong linear relationship*, on both in-distribution and out-of-distribution datasets. From a dataset-centric view, we find a certain model's accuracy and invariance *linearly correlated* on different test sets. Apart from these major findings, other minor but interesting insights are also discussed.

1 Introduction

Generalization and invariance are two important model properties in machine learning. The former characterizes how well a model performs when encountering in-distribution or out-of-distribution (OOD) test data [1–5]. The latter assesses how consistent model predictions are on transformed test data [6–12]. Therefore, understanding how the two properties are related would benefit model decision analysis under dynamic environments.

The importance of model invariance on generalization has been *qualitatively* explored [9, 12–15]. For example, adding rotation invariance to the model improves its in-distribution (ID) classification accuracy [14, 16, 17]; a shift-invariant model is robust to perturbation [9]. In addition, some works provide a neuroscientific perspective to investigate the importance of invariant representations for pattern recognition [13, 15, 18–22]. Furthermore, theoretical investigations suggest that learning invariant features benefits model generalization [12, 22]. However, most existing research is limited to a few ID datasets and model architectures. As such, the relationship of interest remains unknown on many other scenarios, such as OOD and large-scale test datasets, and other types of models.

Before performing the quantitative study, it is necessary to first quantify generalization and invariance. For the former, the deep models we consider are well trained, so we simply use the accuracy on the test set, as in many previous works [3, 2, 23]. In comparison, quantifying invariance is not as straightforward. Some works use model accuracy drop when the test set undergoes transformations to indicate invariance ability [24, 7, 17]. While this strategy is useful for a single model, its effectiveness is limited when comparing invariance of multiple models. Others resort to consistency, *i.e.*, models

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

should have the same decision [9, 6], but this method neglects prediction *confidence*, which we find critical for describing invariance (see discussions in Section 3).

We make two contributions to the community. **First**, we propose a new method to measure model invariance, named effective invariance (EI), which considers both the consistency and confidence of predictions. Given a test image and its transformed counterpart, if the model predicts the same class and with high confidence, the EI value or invariance strength is high. Otherwise, if the model makes different class predictions or the confidence is low, the EI score will be low. We show this new measure solves invariance valuation in canonical cases where the commonly used metrics (*e.g.*, Jensen-Shannon divergence) may fail. **Second**, we conduct a broad correlation study to quantitatively understand the relationship between model generalization and invariance. Specifically, we use 8 test sets with various distribution types, such as the in-distribution ImageNet validation set [25], and out-of-distribution ImageNet-Rendition with style shift [4]. We evaluate 150 ImageNet models ranging from traditional convolution neural networks (*VGGs* [26]) to the very recent vision transformers (*e.g.*, BEiT [27]). Below we list two key observations and example insights.

- For *various models*, there is a strong correlation between their accuracy and invariance on both in-distribution and out-of-distribution datasets (Sections 5.1 and 5.2). This finding can be useful for unsupervised model selection, because EI does not require test ground truths.
- On *various out-of-distribution datasets*, a model's accuracy and EI scores are also strongly correlated (Section 6). This observation can be used to predict model accuracy on out-of-distribution datasets without access to ground truths.
- Compared with data augmentation, training with more data seems more effective to improve invariance and generalization (Section 5.6).

2 Related Work

Predicting generalization gap: a model-centric view. This task aims to predict the generalization gap of machine learning models on in-distribution data, *i.e.*, the difference between training and test accuracy. Most existing works focus on developing *complexity measure* of trained network parameters and training data [28–36], such as persistent topology [31] and the product of norms of the weights across layers [33]. These methods, assuming the training and test distributions are the same, *do not* consider the characteristics of the test distribution. Moreover, they only study limited types of neural networks. In comparison, we conduct a much more comprehensive study on both in-distribution and out-of-distribution test sets, using various network architectures. We show that invariance, under our definition, serves as a strong indicator of model generalization ability or accuracy on both in-distribution and out-of-distribution test data.

Closely related to our work, two recent methods [37, 24] predict ID generalization gap based on how a network performs on perturbed data points. Specifically, Kashyap *et al.* [37] use confidence drop to represent invariance, which is less effective for invariance measurement under OOD data. Schiff *et al.* [24] uses accuracy drop to measure invariance, which needs test labels, while EI does not require test labels and is more reasonable than accuracy under OOD data. In addition, drawing a response curve in [24] is computationally heavy, while our method is relatively efficient. Further, both studies are limited in their scope: they mainly study ID generalization, has few types of networks and lack large-scale test sets, while our work is much more comprehensive.

Predicting generalization gap: a dataset-centric view. The overall goal of this task is to predict the performance of a given model on various unlabeled test sets [38–43]. Many methods take into account the statistics of the test set for accuracy prediction [38, 39, 44, 40, 45], such as distribution shift [38], average Softmax score on each test set [39]. We contribute a new solution: using the model's invariance on the OOD dataset to predict its accuracy. This is supported by our new observation of strong linear correlation between a certain model's accuracy and invariance on various test sets.

Improving robustness with data augmentation. Data augmentation transforms training data to increase its diversity, which helps learn more robust models [46–52]. For example, Mixup [53, 54] and AutoAugment [48] are shown to improve model performance under distribution changes [55, 51].

Instead of using common transformations, adversarial training [56–58] augments training images with an adversarially learned noise distribution. While these works aim to improve corruption robustness with data augmentation, we instead use the latter to analyze model invariance (and generalization).

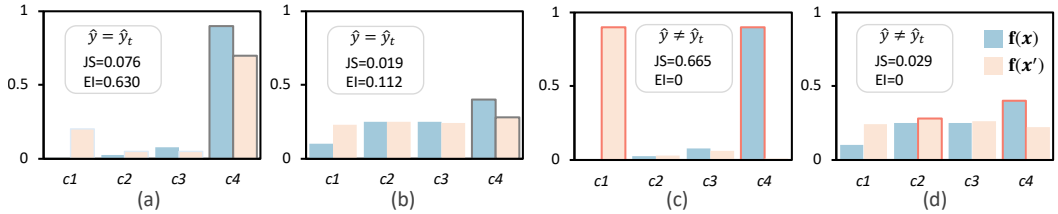

Figure 1: **An illustrative comparison of EI and Jensen-Shannon divergence (JS) as invariance measures.** Four representative cases are shown, where for each class ($c1$-$c4$) we show the Softmax output of the original image $\boldsymbol{f}(\boldsymbol{x})$ and the transformed image $\boldsymbol{f}(\boldsymbol{x'})$. On the one hand, the model exhibits higher invariance in **(a)** than **(b)**, because it makes the same class predictions ($\hat{y} = \hat{y}_t$) and has higher confidence in (a). This is correctly reflected by EI (0.630 *vs.* 0.112; higher is better), but JS incorrectly decides the opposite way (0.076 *vs.* 0.019; lower is better), because JS does not consider confidence explicitly. In cases **(c)** and **(d)**, the model makes different class predictions ($\hat{y} \neq \hat{y}_t$), so its invariance should be very low. This is again correctly captured by EI (0 value for both cases), but JS erroneously gives high invariance to (d), due to the fact that JS merely looks at the global shape of the Softmax vectors without explicitly considering class prediction consistency.

3 Proposed Effective Invariance (EI)

Notations. Considering an K-way classification task, we define input space $\mathcal{X} \in \mathbb{R}^d$ and label space $\mathcal{Y} = \{1, ..., K\}$. Given a sample $(\boldsymbol{x}, y)$ drawn from an unknown distribution π on $\mathcal{X} \times \mathcal{Y}$, a neural network classifier $\boldsymbol{f} : \mathbb{R}^d \to \Delta_K$ produces a probability distribution for $\boldsymbol{x}$ on K classes, where Δ_K denotes the $K-1$ dimensional unit simplex. Specifically, $f_i(\boldsymbol{x})$ denotes the i-th element of the Softmax output vector produced by $\boldsymbol{f}$. Then, $\hat{y} =: \arg\max_i f_i(\boldsymbol{x})$ is the predicted class, and $\hat{p} =: \max_i f_i(\boldsymbol{x})$ is the associated confidence score. Image transformation is defined as $\boldsymbol{\mathcal{T}} : \mathbb{R}^d \to \mathbb{R}^d$. Then, the transformed image is $\boldsymbol{x'} = \boldsymbol{\mathcal{T}}(\boldsymbol{x})$, and its predicted class is $\hat{y}_t =: \arg\max_i f_i(\boldsymbol{x'})$ with confidence score $\hat{p}_t =: \max_i f_i(\boldsymbol{x'})$.

Drawback of existing invariance measures. A commonly seen strategy is to directly use the distance between the Softmax vectors of two predictions as invariance measure: a lower distance means higher invariance, and vice versa. Examples of the similarity metrics are Jensen-Shannon divergence (JS) [59, 46] and ℓ_2 distance [60, 61] and Kullback–Leibler divergence [62, 63]. However, they only leverage the global similarity between two Softmax vectors without an explicitly consideration of prediction class consistency and confidence. We illustrate this drawback taking JS divergence as an example in Fig. 1. In cases (a) and (b) where the predicted classes are both consistent, JS decides classifier $\boldsymbol{f}$ in (b) has higher, which ignores the low confidence in (b). In cases (d) where the predicted classes are different, JS still gives high invariance (small JS score), indicating a clear error. Moreover, in works [46, 60, 61] cases (b) and (d) are discarded when computing the consistency loss.

Definition of EI. Unlike neuroscience works that study the invariance of an individual neuron of the network [13, 15, 18, 19], we measure the invariance at the network level. Specifically, given an image and its transformed sample, a model with high invariance should give the same predicted class, and vice versa [6, 9]. In our definition of EI, we further use the prediction confidence. Our motivation is as follows. When a model predicts the same class for the two images, if either of the two predictions is of low confidence, we should not consider it as highly invariant but give a penalty. A model should have a high invariance if and only if it is highly confident in predicting the same class. Based on these considerations, EI is defined as:

$$\mathrm{EI}(\boldsymbol{x}, \mathcal{T}(\boldsymbol{x}), \boldsymbol{f}) = \begin{cases} \sqrt{\hat{p}_t \cdot \hat{p}} & \text{if } \hat{y}_t = \hat{y}\,; \\ 0 & \text{otherwise}\,. \end{cases} \tag{1}$$

To better understand the soundness of EI, we depict four representative cases in Fig. 1. In cases (a) and (b), classifier $\boldsymbol{f}$ gives the same predicted class ($\hat{y} = \hat{y}_t$) on original and transformed images. Under EI, Classifier $\boldsymbol{f}$ has a higher invariance ability in (a) because it has high confidence scores. In case (c), the predicted class are different ($\hat{y} \neq \hat{y}_t$), and in (d), the predictions are low in confidence (and give different classes), so we define invariance in both cases to be 0.

Computation of EI in practice. Given a test image, we generate a transformed image using a certain transformation. Then, we compute the EI score based their Softmax vectors (Eq. 3). We obtain the model invariance by averaging the EI scores over all the test images. In this work, we mainly investigate the rotation transformation and grayscale transformation. For the former, to avoid interpolation that would introduce artifact, we only use three transformation angles $(90°, 180°, 270°)$. For each rotation angle, we compute an overall invariance score by comparing with the predictions of original data. By averaging the three EI scores, we obtain rotation invariance on the test set. For the grayscale transformation, we remove color information and keep only luminous intensity information. Then, we compare with the predictions of grayscale and original data and compute the overall grayscale invariance on each test set.

4 Experimental Setup

4.1 Models to Be Evaluated

We considers both very recent and classic image classification models with different architectures, including Convolutional Neural Networks (*e.g.*, standard VGGs [26], ResNets [64], and modern ConvNeXt [65]), Vision Transformers (*e.g.*, ViTs [66], Swin [67], and BEiT [27]), and all-MLP architectures [68, 69] (*i.e*, MLP-Mixer [69]).

In addition to different architectures, we also cover models with various training and regularization strategies (*e.g.*, learning rate schedule [70], label smooth [71] and data augmentation [46, 47, 49, 48]), scaling strategies in model dimension (width, depth and resolution) [72–74], and learning manners (supervised learning, semi-supervised learning [75] and knowledge distillation [76, 77]). In total, we have **150 models** provided by TIMM [78]. They are either trained or fine-tuned on the ImageNet-1k training set [25]. The selected models can be roughly divided into the following three categories:

Standard neural networks. This category includes 100 models only trained on ImageNet training set. These networks cover various architectures ranging from VGGs [26] to EfficientNet [73].

Semi-supervised learning. We include 15 models trained in a semi-supervised learning manner. They leverage a large collection of unlabelled images of YFCC100M [79] or Instagram 900M [80] to improve the performance. We use models trained based on a teacher-student paradigm (*e.g.*, SWSL-ResNet [75] and SSL-ResNet [75]). Models trained with self-training methods on unlabeled JFT-300M [81] (*e.g*, EfficientNet-L2-NS [82]) are also included.

Pretraining on more data. We use another 35 models that are pre-trained on significantly larger datasets than the standard ImageNet training set. Specifically, we consider three pre-training methods: a) weakly supervised pretraining on *IG-3.6B* (*i.e.*, RegNetY [83] and ResNeXt101-WSL [80]); b) supervised pre-training on ImageNet-21K [25] (*e.g.*, BiT [84] and Swin [67]); (c) supervised pretraining on JFT-300M [81] (*e.g.*, ViT L/16 [66]).

4.2 Test Sets

We use both in-distribution (ID) and out-of-distribution (OOD) datasets for the correlation study. Specifically, the ImageNet validation set (ImageNet-Val) is used as ID test set. For OOD test sets, we use of seven datasets, each with a different distribution from standard ImageNet. Their distribution shift can be divided into the following five types.

Dataset reproduction shift. ImageNet-V2 [23] is a recollected version of ImageNet-Val. It contains three versions resulting from different data sampling strategies : Matched-Frequency (A), Threshold-0.7 (B), Top-Images (C). Each version has $10,000$ images from $1,000$ classes.
Natural adversarial shift. ImageNet-Adv(ersarial) [85] is adversarially selected to be misclassified by ResNet-50. Its natural adversarial examples are unmodified real-world images and have been shown to be hard for other models [85, 3]. It has $7,500$ samples from 200 ImageNet classes.
Sketch shift. ImageNet-S(ketch) [86] consists of sketch-like images and matches ImageNet-Val in categories and scale. It contains $50,000$ images and shares the same $1,000$ classes as ImageNet.
Blur shift. We use ImageNet-Blur with the highest severity provided by [87]. This dataset is synthesized by blurring ImageNet-Val images with a Gaussian function.
Style shift. ImageNet-R(endition) [4] contains various abstract visual renditions (*e.g.*, art, paintings, and video game) of ImageNet classes. ImageNet-R has $30,000$ images of 200 ImageNet classes.

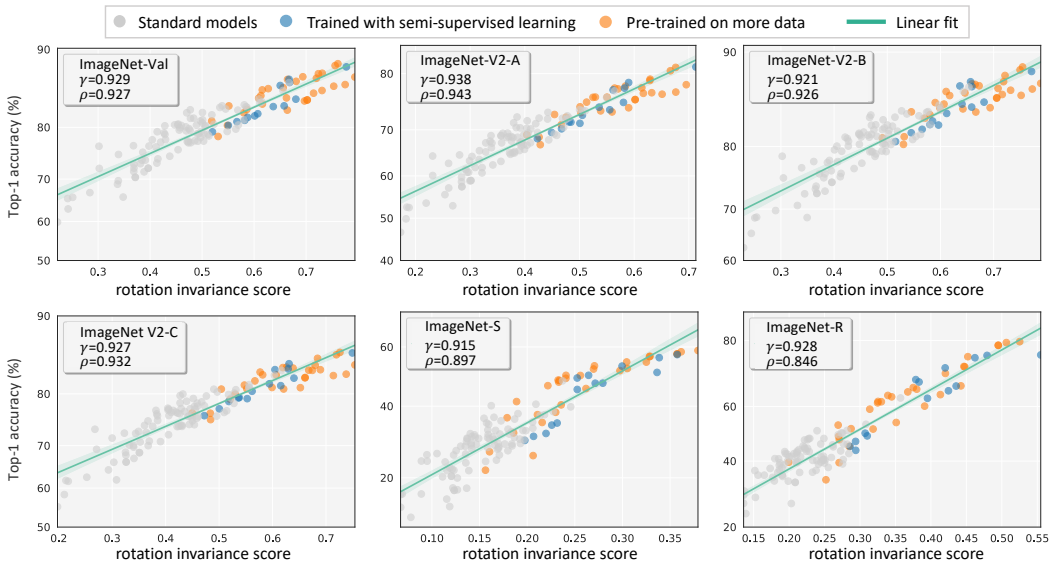

Figure 2: **Correlation between accuracy (%) and rotation invariance (EI) for** 150 **models.** Each figure is obtained from testing on a different ImageNet test set. In each figure, each dot denotes a model, and straight lines are fit by robust linear fit [90]. The shaded region in each figure is a 95% confidence region for the linear fit from $1,000$ bootstrap samples. We clearly observe a strong linear relationship (Pearson's Correlation $r > 0.915$ and Spearman's Correlation $\rho > 0.875$).

4.3 Correlation Measures

We use Pearson Correlation coefficient (r) [88] and Spearman's Rank Correlation coefficient (ρ) [89] to measure the linearity and monotonicity between invariance and generalization, respectively. Both coefficients range from $[-1, 1]$. A value closer to -1 or 1 indicates strong negative or positive correlation, respectively, and 0 implies no correlation [88]. To precisely show the correlation, we use logit axis scaling that maps the accuracy range from $[0, 1]$ to $[-\infty, +\infty]$, following [3, 2]. Unless noted otherwise, the correlation coefficients are calculated using invariance score and non-linearly scaled accuracy numbers.

5 Experimental Observations

The experiment is from a model-centric perspective, where we investigate how different models' accuracy and invariance correlate, and series of observations are made (Section 5.1 - Section 5.6).

5.1 Strong Correlation Between Model's Rotation Invariance and Accuracy

In Fig. 2, we show the correlation results of rotation invariance and generalization. We have two observations. **First**, we find that for different models, their rotation EI scores have a linear relationship with their classification accuracy. The correlation holds for both ID test and OOD test sets, various architectures and training strategies. Specifically, both correlation metrics λ and ρ are higher than 0.840. This indicates models with higher accuracy numbers are most likely to have stronger rotation invariance (measured by EI), and vice versa. To our knowledge, it is a very early observation of the quantitative relationship between generalization and invariance (to a certain factor).

Second, training with more data benefits rotation invariance and generalization. Large datasets contain images with various geometric variations. When (pre)trained with large datasets, models (blue and orange dots in Fig. 2) adapt to the rotation variations and gains stronger invariance, which, according to our study, likely means a higher generalization accuracy on ID and OOD test sets.

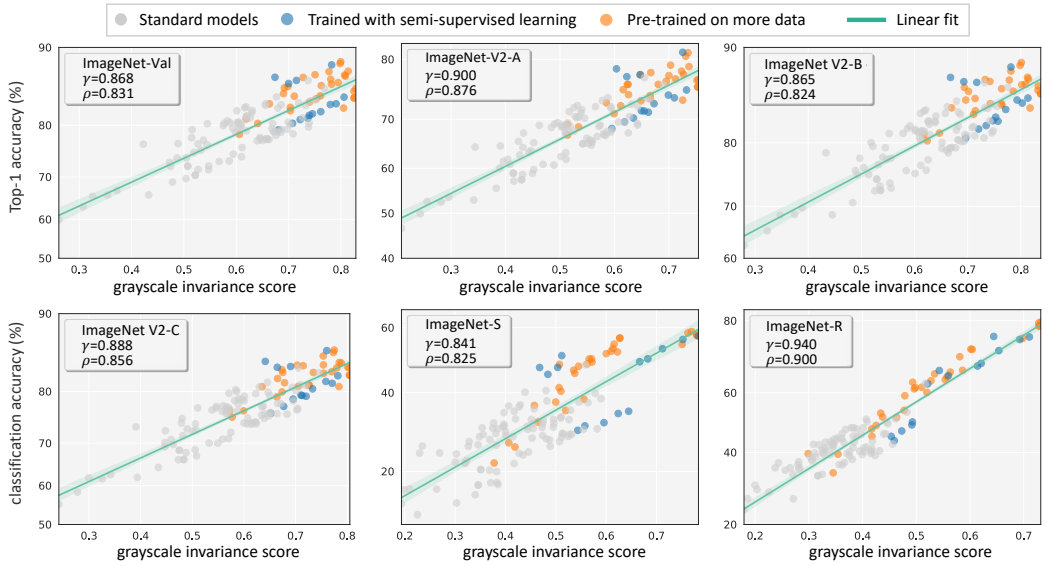

Figure 3: **Correlation between accuracy (%) and grayscale invariance (EI) of** 150 **models.** Similar to Fig. 2, each dot denotes a model, where different colors denote different training strategies (see Section 4.1). The subfigures correspond to three ImageNet test sets, respectively. We observe the correlation is also strong: Pearson's Correlation $r > 0.840$, and Spearman's Correlation $\rho > 0.820$.

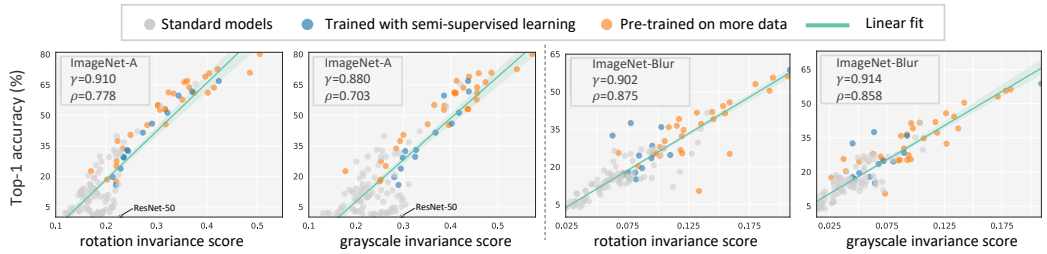

Figure 4: **Correlation study on very hard test sets.** We use the very hard ImageNet-A and ImageBlur for testing, where the accuracy of 83 models is lower than 20%. Both rotation invariance and grayscale invariance are evaluated. Overall we observe relatively solid correlation in all the four cases. Looking more closely, most standard models (gray dots) are scattered in the low-accuracy region, while models trained with more data (blue and orange ones) move away from this region and exhibit linear trends. Thus, the overall linear correlation is high ($r > 0.880$), and the overall rank correlation is slightly less consistent (ρ ranges from 0.703 to 0.858) but still has clear trends.

5.2 Strong Correlation Between Model's Grayscale Invariance and Accuracy

We now focus on grayscale invariance and report the correlation results in Fig. 3. We have the following conclusions. **First**, among the 150 models, there is strong linear correlation between accuracy and grayscale invariance measured by EI. Specifically, both correlation coefficients r and ρ are higher than 0.820 on all test sets. **Second**, we find that pretrained or semi-supervised models tend to have higher grayscale invariance and accuracy than standard models, which again indicates the usefulness of large training sets. **Third** and interestingly, the correlation is stronger on ImageNet-R than ImageNet-Val and ImageNet-V2-A (0.940 *vs.* 0.900 *vs.* 0.868). In fact, ImageNet-R is featured by style shift during its collection [4], so for this test set being invariant to color changes is an important property for a stronger generalization ability.

5.3 Correlation Exists on Very Hard Test Sets

We now study the correlation under very hard test sets and use ImageNet-A and ImageNet-GaussBlur for testing, on which accuracy of 83 models is lower than 20%. We study rotation invariance and

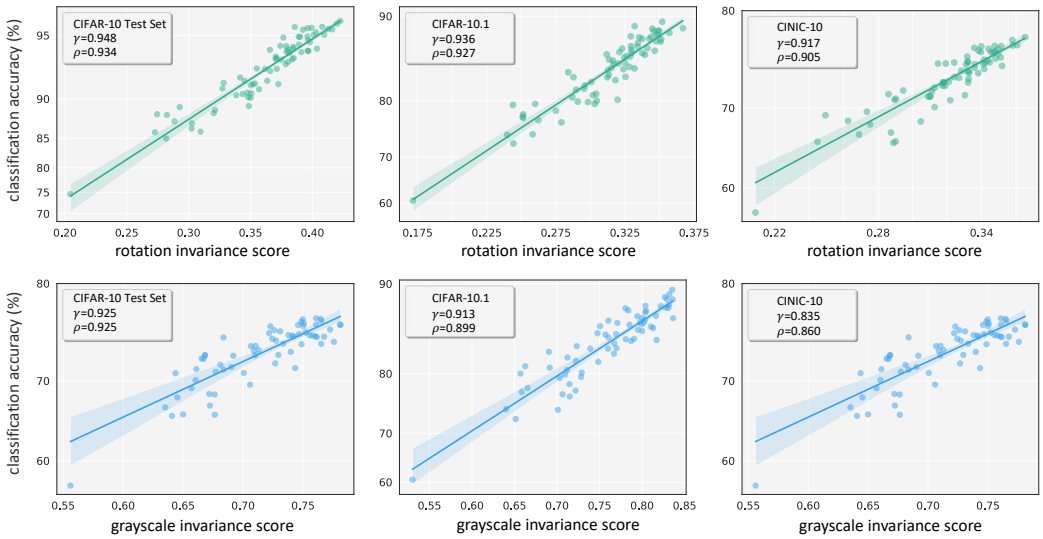

Figure 5: **Correlation study under the CIFAR-10 setup.** Each data point is a CIFAR model. We report the correlation results with rotation invariance (top) and grayscale invariance (bottom). We observe that strong correlation exists between accuracy and invariance on all the three test sets.

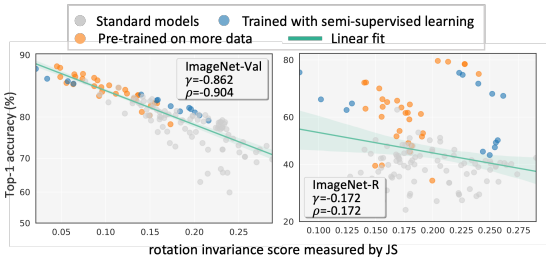

Figure 6: **Correlation between accuracy (%) and invariance (JS)**. Despite of the good correlation on ImageNet-Val, correlation on ImageNet-R, a harder test set, is weak. More results and analysis are provided in the supplementary meterials.

Test set	JS	EI
Img.-Val	-0.861 / -0.905	$+0.929$ / 0.927
Img.-R	-0.173 / -0.172	$+0.928$ / 0.846
Img.-S	$+0.131$ / $+0.148$	$+0.915$ / 0.897
Img.-A	-0.297 / -0.257	$+0.910$ / 0.778

Table 1: **Comparing EI and JS.** Under Pearson's Correlation r / Spearman's Rank Correlation ρ, JS does not show strong correlation on most datasets, while EI does on all the four datasets.

grayscale invariance. In Fig. 4, we find strong linear correlation in the four cases ($r \leq 0.88$). The rank correlation is less consistent but still indicates clear trends. Moreover, most standard models have low accuracy, while models (pre)trained with more data tend to have high accuracy and invariance. We also notice standard models are scattered differently in the low-accuracy regime of ImageNet-A and ImageNet-Blur, possibly due to their dataset bias introduced during dataset construction [91–93].

5.4 Correlation Holds Under the CIFAR-10 Setup

We conduct correlation study on CIFAR-10 setup. We collect 90 CIFAR models ranging from LeNet to EfficientNet. We use the ID CIFAR-10 test set and two OOD test sets. 1) CIFAR-10.1 [94] is a reproduction of the CIFAR-10 test set but with distribution shift arising from changes in data collection. It contains $2,000$ test images sampled from TinyImages [95]. 2) CINIC-10 test set [96] is an extended alternative for CIFAR-10. It has has $90,000$ images sampled from ImageNet.

We show the relationship between model generalization and invariance in Fig. 5. On all the three test sets, we observe the strong correlation between model accuracy numbers and rotation invariance scores, where both r and ρ are greater than 0.90. Moreover, when testing grayscale invariance, the strong correlation still exist (both r and ρ are greater than 0.83).

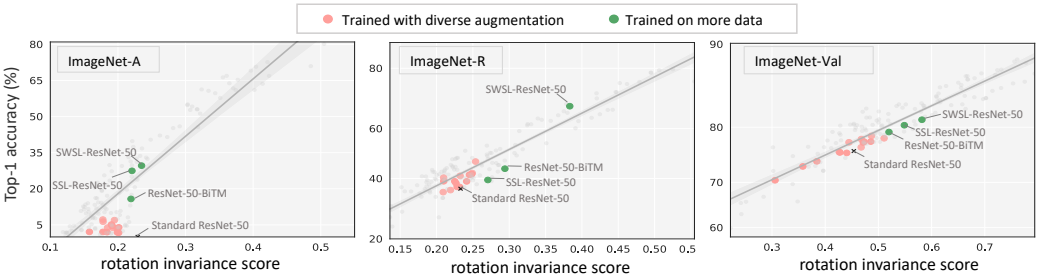

Figure 7: **Comparing data augmentation and (pre)training with more data.** We use 14 ResNet-50 models (red dots) trained with various types of augmentation methods, and 3 ResNet-50 models (green dots) trained on more data (real-world). We also mark ResNet-50 trained with the standard learning strategies. On ImageNet-Val and ImageNet-R, we observe models trained with diverse augmentation follow a linear trend. However, they deviate from the trend on ImageNet-A. In comparison, training with more data allows models to achieve relatively high accuracy and invariance on three test sets.

5.5 EI Gives Stronger Correlation Than JS

In the representative cases (Fig. 1), EI shows superiority to JS in measuring model invariance by explicitly considering Softmax consistency and confidence. Now we compare the invariance scores measured by EI and JS *w.r.t* their correlation strength with accuracy in Fig. 6 and Table 1. We find that JS provides good correlation on the ID ImageNet-Val test set, but much weaker correlation on the more difficult OOD tests than EI. As discussed in Section 3, the advantage of EI is that it not only follows the definition of invariance, but also integrates confidence to strength it, while JS only compares the overall Softmax vector. In fact, the drawback of JS is primarily reflected in hard test sets, where the Softmax vector usually has a flat shape. This explains why JS does not give strong correlation on the OOD test sets (see Section 5.3 for EI's performance on very hard test sets). We refer readers to the supplementary material for comparisons with other measures.

5.6 Comparing Two Training Manners of Their Generalization and Invariance

Existing works report that using more diverse training data artificially (*i.e.*, data augmentation) or naturally (*i.e.*, more real-world data) improves model invariance [4, 46–51, 97]. In this study we compare the two strategies of their generalization and invariance abilities. We use 14 ResNet-50 models trained with strong data augmentation such as PixMix [52] and AutoAugment [48]. For comparison, we employ another 3 ResNet-50 models that learn invariance using more training samples: SWSL-ResNet-50, SSL-ResNet-50, and ResNet-50-BiTM. In Fig. 7, we observe that models with heavy and strong augmentation exhibit linear trends on ImageNet-Val and ImageNet-R, but deviate on ImageNet-A. In comparison, models (pre)trained with more data follow the linear trend on three test sets. The latter seem more effective in improving invariance and accuracy compared with data augmentation, especially on ImageNet-A. To thoroughly understand and extend this initial observation, we will conduct more comprehensive experiment on this specific point in future.

6 Correlation of A Model's Invariance and Generalization on OOD Test Sets

From a **dataset-centric** perspective, we study the relationship between generalization and invariance of a given model on different OOD test sets. Given a model, we calculate its accuracy on every test test of ImageNet-C [87] and compute its rotation and grayscale invariance scores.

We evaluate ResNet-152, ViT-Base-16, and DenseNet-121 classifiers for the correlation study. As shown in Fig. 8, there is a very strong correlation between classifier accuracy and invariance on various datasets ($r > 0.93$ and $\rho > 0.95$). The results indicate that a classifier tends to have a high accuracy on the test set where it has a high EI score. The above analysis indicates that it is feasible to use EI to access out-of-distribution error of a model.

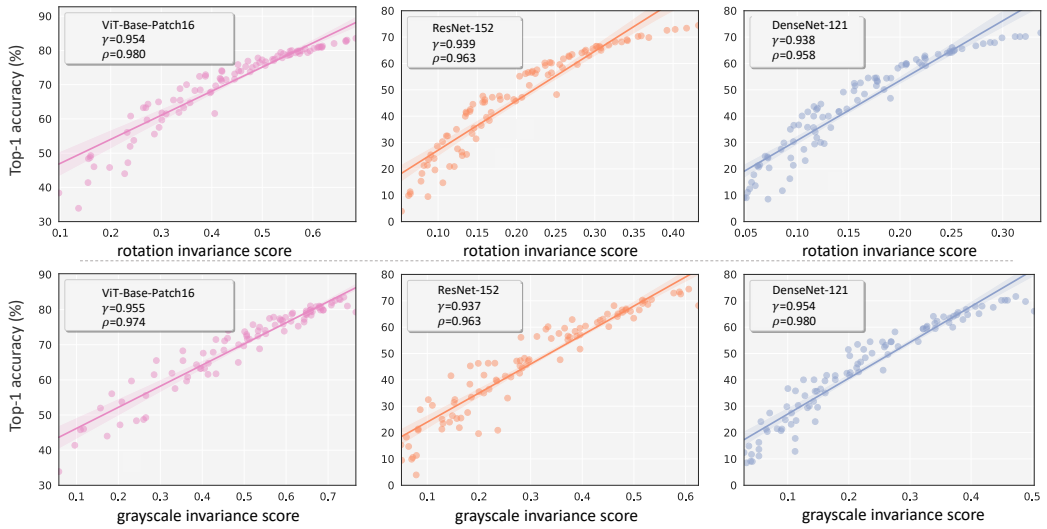

Figure 8: **Correlation between a model's invariance and accuracy on various OOD test sets.** In each figure, a data point corresponds to a test set from ImageNet-C [87]. The straight lines are fit with robust linear regression [90]. We test rotation invariance (top) and grayscale invariance (bottom). In all sugfigures, we observe a strong negative correlation (Pearson's Correlation r and Spearman's Rank Correlation ρ are greater than 0.930) between invariance and accuracy.

7 Conclusion

This work considers two critical properties of a machine learning model: invariance and generalization. To study their relationship, we first introduce effective invariance (EI) to more reasonably measure invariance and then provide in-depth and comprehensive correlation experiment. From a model-centric perspective, we report accuracy and EI of various models has a strong linear relationship on both ID and OOD datasets, which is validated on many scenarios such as large-scale test sets, CIFAR-10 and very hard test sets. From a dataset-centric perspective, we show the accuracy and EI of a model have a strong linear relationship on various OOD datasets.

Limitations and potential directions. First, some networks with specially designed modules can be highly invariant to some transformations [7–9, 16, 98], such as rotation invariance networks [14, 99, 17]. The rotation invariance scores of these models may present very different correlation from our observations. That said, their color invariance scores may still exhibit similar correlation with this work. In future works, it would be interesting to understand how these models deviate from others. **Second**, our ImageNet test sets do not include the *geographic shift* where images are captured from various locations [4]. Recent works show this shift is also a key factor influencing model accuracy [100]. We leave this question to future study. **Moreover**, we focus on classification tasks, where models are supervised by image-level annotations. In other computer vision tasks, models may be trained with different levels of supervision, such as instance-level bounding box annotations [101], pixel-level labels [102] and temporal context supervision [103]. Different types of supervision may lead to invariance ability to various factors, which may exhibit different correlation profiles. Lastly, in designing EI, we mainly study four representative cases and explain the limitations of some commonly used measures (*e.g.*, JS and L2). Considering more special cases would be beneficial.

Potential negative social impact. We study fundamental model properties with public classification datasets, which might be misused in certain applications with ethical concerns.

Acknowledgments and Disclosure of Funding

We thank all anonymous reviewers for their constructive comments in improving this paper. This work was in part supported by the ARC Discovery Early Career Researcher Award (DE200101283) and the ARC Discovery Project (DP210102801). Stephen Gould is the recipient of an ARC Future Fellowship (project number FT200100421) funded by the Australian Government. Weijian Deng is a recipient of Australian Government Research Training Program (RTP) Scholarship.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010. 1

[2] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735, 2021. 1, 5

[3] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020. 1, 4, 5

[4] 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. 2, 4, 6, 8, 9

[5] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 1

[6] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018. 1, 2, 3

[7] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pages 1802–1811, 2019. 1, 9

[8] Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: analysis and improvement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4441–4449, 2018.

[9] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334, 2019. 1, 2, 3, 9

[10] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

[11] Osman Semih Kayhan and Jan C van Gemert. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14274–14285, 2020.

[12] Sicheng Zhu, Bang An, and Furong Huang. Understanding the generalization benefit of model invariance from a data perspective. In *Advances in Neural Information Processing Systems*, volume 34, pages 4328–4341, 2021. 1

[13] Guangyu Robert Yang, Madhura R Joglekar, H Francis Song, William T Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature neuroscience*, 22 (2):297–306, 2019. 1, 3

[14] Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Oriented response networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 519–528, 2017. 1, 9

[15] Maximilian Riesenhuber and Tomaso Poggio. Just one view: Invariances in inferotemporal cell tuning. 1997. 1, 3

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, 2015. 1, 9

[17] Valentin Delchevalerie, Adrien Bibal, Benoît Frénay, and Alexandre Mayer. Achieving rotational invariance with bessel-convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2021. 1, 9

[18] Spandan Madan, Timothy Henry, Jamell Dozier, Helen Ho, Nishchal Bhandari, Tomotake Sasaki, Frédo Durand, Hanspeter Pfister, and Xavier Boix. When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations. *Nature Machine Intelligence*, 4(2):146–153, 2022. 1, 3

[19] Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. In *Advances in neural information processing systems*, 2009. 3

[20] Emanuela Bricolo, Tomaso Poggio, and Nikos K Logothetis. 3d object recognition: A model of view-tuned neurons. 1996.

[21] Tomaso A Poggio and Fabio Anselmi. *Visual cortex and deep networks: learning invariant representations*. MIT Press, 2016.

[22] Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. On invariance and selectivity in representation learning. *Information and Inference: A Journal of the IMA*, 5(2):134–158, 2016. 1

[23] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 1, 4

[24] Yair Schiff, Brian Quanz, Payel Das, and Pin-Yu Chen. Predicting deep neural network generalization with perturbation response curves. In *Advances in Neural Information Processing Systems*, 2021. 1, 2

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2, 4

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 4

[27] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2, 4

[28] Gabriel Eilertsen, Daniel Jönsson, Timo Ropinski, Jonas Unger, and Anders Ynnerman. Classifying the classifier: dissecting the weight space of neural networks. *arXiv preprint arXiv:2002.05688*, 2020. 2

[29] Thomas Unterthiner, Daniel Keysers, Sylvain Gelly, Olivier Bousquet, and Ilya Tolstikhin. Predicting neural network accuracy from weights. In *International Conference on Learning Representations*, 2020.

[30] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *arXiv preprint arXiv:1802.05296*, 2018.

[31] Ciprian A Corneanu, Sergio Escalera, and Aleix M Martinez. Computing the testing error without a testing set. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2677–2685, 2020. 2

[32] Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions. In *International Conference on Learning Representations*, 2019.

[33] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in neural information processing systems*, pages 5947–5956, 2017. 2

[34] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019.

[35] Saurabh Garg, Sivaraman Balakrishnan, Zico Kolter, and Zachary Lipton. Ratt: Leveraging unlabeled data to guarantee generalization. In *International Conference on Machine Learning*, pages 3598–3609, 2021.

[36] Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of sgd via disagreement. *arXiv preprint arXiv:2106.13799*, 2021. 2

[37] Sumukh Aithal, Dhruva Kashyap, and Natarajan Subramanyam. Robustness to augmentations as a generalization metric. *arXiv preprint arXiv:2101.06459*, 2021. 2

[38] Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15069–15078, 2021. 2

[39] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1134–1144, 2021. 2

[40] Saurabh Garg, Sivaraman Balakrishnan, Zachary C Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022. 2

[41] Christina Baek, Yiding Jiang, Aditi Raghunathan, and Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *arXiv preprint arXiv:2206.13089*, 2022.

[42] Nathan Ng, Kyunghyun Cho, Neha Hulkund, and Marzyeh Ghassemi. Predicting out-of-domain generalization with local manifold smoothness. *arXiv preprint arXiv:2207.02093*, 2022.

[43] Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. Predicting out-of-distribution error with the projection norm. In *Advances in Neural Information Processing Systems*, 2022. 2

[44] Weijian Deng, Stephen Gould, and Liang Zheng. What does rotation prediction tell us about classifier accuracy under varying testing environments? In *International conference on machine learning*, 2021. 2

[45] 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. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[46] 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. 2, 3, 4, 8

[47] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 4

[48] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 2, 4, 8

[49] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 4

[50] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[51] Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. In *Advances in Neural Information Processing Systems*, 2021. 2, 8

[52] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 8

[53] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2

[54] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Between-class learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5486–5494, 2018. 2

[55] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems*, 2019. 2

[56] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2

[57] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *Advances in Neural Information Processing Systems*, pages 3533–3545, 2020.

[58] Evgenia Rusak, Lukas Schott, Roland S Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In *European Conference on Computer Vision*, pages 53–69, 2020. 2

[59] Bent Fuglede and Flemming Topsoe. Jensen-shannon divergence and hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, page 31, 2004. 3

[60] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, pages 596–608, 2020. 3

[61] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2019. 3

[62] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 3

[63] Jiachen Sun, Akshay Mehra, Bhavya Kailkhura, Pin-Yu Chen, Dan Hendrycks, Jihun Hamm, and Z Morley Mao. Certified adversarial defenses meet out-of-distribution corruptions: Benchmarking robustness and simple baselines. *arXiv preprint arXiv:2112.00659*, 2021. 3

[64] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[65] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 4

[66] 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. 4

[67] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4

[68] Xiaohan Ding, Chunlong Xia, Xiangyu Zhang, Xiaojie Chu, Jungong Han, and Guiguang Ding. Repmlp: Re-parameterizing convolutions into fully-connected layers for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 4

[69] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems*, 2021. 4

[70] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 4

[71] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4

[72] Irwan Bello, William Fedus, Xianzhi Du, Ekin Dogus Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting resnets: Improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 34, 2021. 4

[73] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114, 2019. 4

[74] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106, 2021. 4

[75] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 4

[76] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 4

[77] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11936–11945, 2021. 4

[78] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 4

[79] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015. 4

[80] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018. 4

[81] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 4

[82] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 4

[83] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 4

[84] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pages 491–507, 2020. 4

[85] 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. 4

[86] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 4

[87] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the International Conference on Learning Representations*, 2019. 4, 8, 9

[88] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009. 5

[89] Maurice George Kendall. Rank correlation methods. 1948. 5

[90] Peter J Huber. Robust statistics. In *International encyclopedia of statistical science*, pages 1248–1251. Springer, 2011. 5, 9

[91] Kristof Meding, Luca M Schulze Buschoff, Robert Geirhos, and Felix A Wichmann. Trivial or impossible–dichotomous data difficulty masks model differences (on imagenet and beyond). In *Proceedings of the International Conference on Learning Representations*, 2022. 7

[92] Guy Hacohen, Leshem Choshen, and Daphna Weinshall. Let's agree to agree: Neural networks share classification order on real datasets. In *International Conference on Machine Learning*, pages 3950–3960, 2020.

[93] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558, 2020. 7

[94] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018. 7

[95] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008. 7

[96] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017. 7

[97] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Proceedings of the International Conference on Learning Representations*, 2021. 8

[98] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. In *Advances in Neural Information Processing Systems*, 2019. 9

[99] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018. 9

[100] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021. 9

[101] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 2015. 9

[102] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 9

[103] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 9

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1
   (b) Did you describe the limitations of your work? [Yes] Please see Section 7
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see Section 7.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We clearly illustrated the publicly available datasets and models in Section 4. We also provide details about the calculation of EI in Section 3.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] This work does not require any training process. Both datasets and models are fixed following the standard protocol.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] In our work, the trained models are provided by PyTorch Image Models (timm), the correlation study does not require retraining the model, and the datasets are also commonly used standard benchmarks. Therefore, our results do not have any randomness.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We illustrate the computational resources in Supplementary material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] In Section 4, we cite PyTorch Image Models (timm) where trained models are provided. The datasets we used are also clearly cited and introduced.
   (b) Did you mention the license of the assets? [Yes] We use publicly released datasets and timm models where the licenses are provided.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] This work does not involve any new assets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]