CNS-BENCH: BENCHMARKING MODEL ROBUSTNESS UNDER CONTINUOUS NUISANCE SHIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

One important challenge in evaluating the robustness of vision models is to control individual nuisance factors independently. While some simple synthetic corruptions are commonly applied to existing models, they do not fully capture all realistic distribution shifts of real-world images. Moreover, existing generative robustness benchmarks only perform manipulations on individual nuisance shifts in one step. We demonstrate the importance of gradual and continuous nuisance shifts, as they allow evaluating the sensitivity and failure points of vision models. In particular, we introduce **CNS-Bench**, a **C**ontinuous **N**uisance **S**hift **Bench**mark for image classifier robustness. CNS-Bench allows generating a wide range of individual nuisance shifts in continuous severities by applying LoRA adapters to diffusion models. After accounting for unrealistic generated images through an improved filtering mechanism for such samples, we perform a comprehensive large-scale study to evaluate the robustness of classifiers under various nuisance shifts. Through carefully-designed comparisons and analyses, we find that model rankings can change for varying shifts and shift scales, which is not captured when averaging the performance over all severities. Additionally, evaluating the model performance on a continuous scale allows the identification of model failure points, providing a more nuanced understanding of model robustness. Overall, our work demonstrated the advantage of using generative models for benchmarking robustness across diverse and continuous real-world nuisance shifts in a controlled and scalable manner.

1 INTRODUCTION

Machine learning models are typically validated and tested on fixed datasets under the assumption of independent and identically distributed samples. This, however, does not fully cover the true capabilities and potential vulnerabilities of models when deployed in dynamic real-world environments. The robustness in out-of-distribution (OOD) scenarios is important and decision-makers might need to know how models perform under various distribution shifts and severity levels in safety-critical scenarios. Therefore, it is crucial to continue building richer and more systematic benchmarks.

In the past few years, various benchmarks have been proposed to evaluate the robustness of computer vision models. One line of benchmarks manually collects data with nuisance shifts (Zhao et al., 2022; Hendrycks et al., 2021a; Wang et al., 2019; Geirhos et al., 2022; Barbu et al., 2019; Idrissi et al., 2022; Hendrycks et al., 2021b; Recht et al., 2019). Yet, such approaches are not scalable and often include only a small variety of nuisance shifts.

On the other hand, synthetic datasets offer opportunities to evaluate deep neural networks since various instances of an object class with specified context and nuisance shifts can be generated. While rendering pipelines allow precise control of several variables and are applied for benchmarking (Bordes et al., 2024; Shu et al., 2020; Kar et al., 2022; Li et al., 2023c), some nuisance shifts such as weather variations (*e.g.*, snow) are very hard to perform using traditional pipelines. While Hendrycks & Dietterich (2018) report accuracy drops for various types and levels of synthetic corruptions, they lack relevant real-world nuisance shifts.

Recent developments in diffusion models have enabled the application of generative models for training (He et al., 2022b; Fan et al., 2024) and benchmarking vision models (Mofayezi & Medghalchi, 2023; Metzen et al., 2023; Vendrow et al., 2023; Zhang et al., 2024). However, all

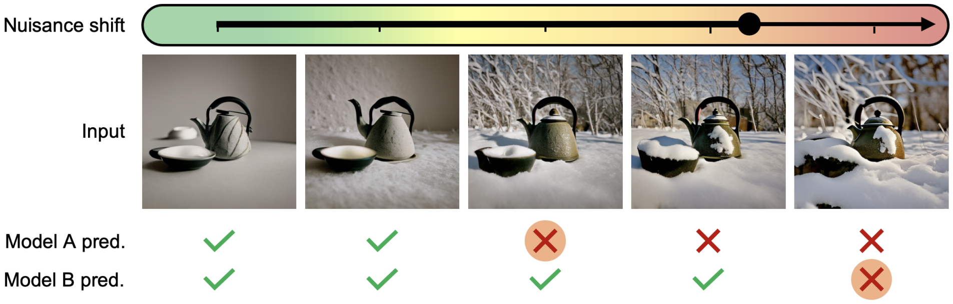


Figure 1: **Benchmarking under continuous nuisance shifts.** We evaluate the robustness of different models under gradually increasing nuisance shifts. This allows identifying the *failure point* (highlighted in red) of a model.

previous approaches define *categorical* or *binary* nuisance shifts by considering the existence or absence of a shift, which contradicts their continuous realization in real-world scenarios. For example, as shown in Fig. 1, the snow level in an environment can range from light snowfall to objects fully covered with snow. While one model might fail at all snow levels, a different model might only fail when the object is heavily occluded. In most real-world applications, it is important to know the expected performance at specific nuisance shift levels, rather than just a global accuracy drop. For instance, an autonomous driving company may need to determine the fog density at which system performance falls below a critical threshold. Evaluating such failure points to probe the sensitivity of models requires realizing continuous shifts.

To overcome this shortcoming, we establish a **C**ontinuous **N**uisance **S**hift **Bench**mark for model robustness, dubbed as **CNS-Bench**. Specifically, we apply LoRA (Hu et al., 2021) adapters to diffusion models to perform a continuous variation of specified nuisance shifts, and use them to benchmark a variety of classifiers along the following axes: (i) architecture, (ii) number of parameters, (iii) pre-training paradigm and data. In contrast to previous works conducting analysis on *binary* or *categorical* shifts, our study advocates multiple scales of shifts. We caveat that model rankings can change when considering several scales. It is also essential to consider failure points, *i.e.*, the shift severity at which a model fails. Thus, measuring robustness as a spectrum instead of aggregating it into a single average metric allows a more comprehensive understanding of OOD robustness (Drenkow et al., 2021; Hendrycks et al., 2021a). With our benchmark, we evaluate more than 40 classifiers and demonstrate that a rigorously-designed generative benchmark allows systematically studying the robustness behaviors of vision models in a controlled and scalable manner.

One essential requirement when using synthetic images for benchmarking is to ensure that the considered images correspond to the class distribution. Manually checking the quality of images to find those not aligned with the desired condition is still a common practice (Zhang et al., 2024). However, it has difficulty in scaling up the analysis (Hastie et al., 2009; Angelopoulos et al., 2023). Some approaches have been proposed for automatic filtering, but no standard datasets are available to evaluate filtering strategies. With this in mind, we also provide a dataset with manually annotated out-of-class (OOC) images. We show that our proposed filtering mechanism outperforms previous strategies in removing such problematic samples.

In summary, our work makes the following contributions: **1)** We propose CNS-Bench to benchmark vision models under continuous nuisance shifts. We publish a dataset with 14 diverse and realistic nuisance shifts that represent various style and weather variations at five severity levels. In addition, we also provide trained LoRA sliders for all shifts that can be used to compute shift levels in a fully continuous manner. **2)** We collect an annotated dataset to benchmark OOC filtering strategies and propose a novel filtering mechanism that achieves higher filter accuracies than previous methods. **3)** We evaluate the robustness of more than 40 classifiers along different axes and reveal multiple valuable findings, underlining the importance of considering continuous shift severities of real-world nuisance shifts.

2 RELATED WORK

Robustness. When referring to robustness, we consider the relative accuracy drop of a classifier *w.r.t.* interventions that alter images from a base distribution, building upon the formalism introduced in Drenkow et al. (2021). While the averaged accuracy drops provide an aggregated measure of the robustness, we consider the robustness *w.r.t.* specific nuisance shifts that can be modeled as causal interventions on the environment, the appearance, the object, or the renderer. We define such continuous interventions on metric scale.

Benchmarking robustness. Early approaches for benchmarking the performance and generalizability of models use fixed datasets, assuming independent and identically distributed samples (Deng, 2012; Deng et al., 2009; Lin et al., 2014). However, this lacks scalability and fails to capture the performance in real-world applications facing OOD scenarios. To tackle this challenge, a line of research involves manually collecting data with nuisance shifts (Zhao et al., 2022; Hendrycks et al., 2021a; Wang et al., 2019; Geirhos et al., 2022; Barbu et al., 2019; Idrissi et al., 2022; Hendrycks et al., 2021b; Recht et al., 2019). However, these methods are often time-consuming and labor-intensive since they require data crawling and human annotations. Moreover, they usually capture only a subset of nuisance shifts that models may encounter in the real world and it is challenging to ensure the disentanglement of these annotated nuisances.

Another line of research uses synthetic data for benchmarking, which offers the ability to generate a large and diverse range of nuisance shifts with precise control (Hendrycks & Dietterich, 2018; Bordes et al., 2024; Shu et al., 2020; Kar et al., 2022). However, these works are limited to nuisances that can be easily modelled (*e.g.*, lighting, fog, occlusions) or restricted to what can be expressed in rendering pipelines. Recent developments in diffusion models shed light on creating realistic and diverse synthetic benchmark datasets (Mofayezi & Medghalchi, 2023; Metzen et al., 2023; Vendrow et al., 2023; Zhang et al., 2024) with realistic data and more possibilities to control nuisances (*e.g.*, text-guided corruptions, counterfactual). In our work, we propose a framework to benchmark vision models *w.r.t.* nuisance shifts under multiple severity levels. To address the need to remove OOC images from generative models, which are essential for benchmarking applications, we additionally propose a novel strategy to remove such samples from the dataset.

3 CONTINUOUS NUISANCE SHIFT BENCHMARK

In this section, we present how CNS-Bench is created. We first discuss the strategy to replicate the in-domain distribution in Section 3.1. We then present our methodology to perform continuous shifts to evaluate the model's sensitivity to various nuisance factors in Section 3.2. Finally, we detail our filtering dataset and the selected filtering strategy in Section 3.3.

3.1 REPLICATING THE IMAGENET DISTRIBUTION

We aim to evaluate a model's robustness to specific nuisance shifts that alter the base ImageNet (Deng et al., 2009) distribution $p(X_{\text{IN}}|c)$, which is conditioned on an ImageNet class c. For a more accurate estimate of the robustness concerning a single considered shift, we desire a model accuracy comparable to the in-domain (ImageNet) distribution. As pointed out by Vendrow et al. (2023), the distribution of Stable Diffusion (SD) (Rombach et al., 2022) generated images $p(X_{\text{SD}}|c)$ differs from the ImageNet distribution, resulting in lower classification accuracies of ImageNet-trained classifiers. Therefore, we use the text embeddings provided by Vendrow et al. (2023) after training them via textual inversion (Gal et al., 2023) on the ImageNet training dataset. We call this distribution IN*: $p(X|c) = p(X_{\text{IN*}}|c)$.

3.2 CONTINUOUS NUISANCE SHIFTS FOR BENCHMARKING

To evaluate the robustness of vision models *w.r.t.* continuous nuisance shifts, the following characteristics are desirable: (i) The severity of the considered shift can be controlled, allowing the estimation of the shift scale where a considered model fails. (ii) Realizing a nuisance shift should not come along with factors of variations that might alter the class identity. (iii) The variations should be subtle, allowing a fine-grained analysis also for specific images.

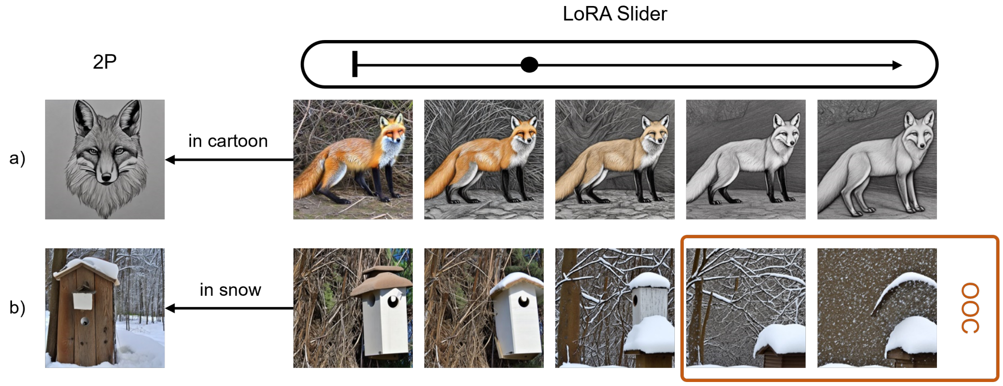


Figure 2: **Qualitative examples for prompt-based and LoRA-based shifts with out-of-class samples.** On the left, we present two images in a different a) style and b) weather condition generated from a text prompt a) "fox in cartoon style" and b) "birdhouse in heavy snow", respectively. On the right, we show the gradual variation performed by our LoRA sliders. a) Unlike the prompt-based shift, our LoRA sliders successfully generated images showing a gradual shift. b) Our LoRA sliders sometimes result in out-of-class (OOC) samples for higher scales, as depicted with the orange box.

Realizing continuous nuisance shifts. A natural way to perform synthetic nuisance shifts are methods based on text prompts (Metzen et al., 2023; Liu et al., 2023; Vendrow et al., 2023). They follow the two prompt (2P) templates: "`A picture of a <class>`" and "`A picture of a <class> in <shift>`". However, this approach does not allow the gradual increase of a nuisance for a given image. In addition, the generated shifts largely vary for different seeds and classes when applying the prompt addition "`in <shift>`"—for some seeds, the generated shift is more prominent, while for others, it is barely visible. Additionally, the semantic structure of the generated image can be significantly changed.

We leverage LoRA (Hu et al., 2021) adapters that represent low-rank matrices added to the original weight matrices to perform continuous shifts. Such adapters are trained to characterize the effect of a considered nuisance shift. Gandikota et al. (2023) propose a strategy to learn concept sliders using LoRA adapters that allow a continuous modulation of the considered concept, which is achieved by learning low-rank matrices that increase the expression of a specific attribute when applied to a class concept c. The low-rank parameters θ_{LoRA} that modify the original model parameters θ to $\theta^* = \theta + s \cdot \theta_{\text{LoRA}}$ with scale s are trained to capture a concept of interest c_+: $P_{\theta^*}(X|c) \leftarrow P_\theta(X|c) \cdot P_\theta(X|c_+)^\eta$, where η refers to weighting factor that is fixed during training. Following Gandikota et al. (2023), we optimize with the MSE objective (Sohl-Dickstein et al., 2015) using the Tweedie's formula (Efron, 2011) and the reparametrization trick (Ho et al., 2020) by formulating the scores as a denoising prediction $\epsilon(X, c, t)$ with the diffusion timestep t: $\text{MSE}(\epsilon_{\theta^*}(X, c, t); \epsilon_\theta(X, c, t) + \epsilon_\theta(X, c_+, t))$. We model the class concept c and the nuisance concept c_+ by two text embeddings "`<class>`" and "`<class> in <shift>`". Different to (Gandikota et al., 2023), we specifically use class concepts c that are acquired from the IN* distribution. After training, the learned LoRA adapters capture the direction between the two language concepts, i.e., they characterize attributes of the concept of interest c_+. Weighting their effect using the scale s modulates the effect of the applied shift. Gandikota et al. (2023) stated that the LoRA adapters generalize to other concepts and images. We found that learning class-specific LoRA sliders produces higher-quality shifts. This choice also allows capturing the class-specific characteristics and confounders of the considered shifts that occur in the real world. Hence, we train separate LoRA adapters for each ImageNet class and shift. As qualitatively shown in Fig. 2, applying these learned directions enables gradual nuisance shifts. We show examples of more shifts in Fig. 33 and Fig. 34.

Following (Mokady et al., 2023; Gandikota et al., 2023), we evaluate the shift severity based on the CLIP similarity of the generated image to the text prompt describing the shift, *i.e.*, "`A picture in <shift>`". Similarly, we also compute the CLIP (Radford et al., 2021) similarity to the class prompt "`A picture of a <class>`". To measure the performed shift, we compute the CLIP shift difference by $\Delta\text{CLIP}_{\text{shift}}(I_k, I_0) = \cos(\text{CLIP}_{\text{img}}(I_k), \text{CLIP}_{\text{text}}(\text{“in }\{\text{shift}\}\text{”})) - \cos(\text{CLIP}_{\text{img}}(I_0), \text{CLIP}_{\text{text}}(\text{“in }\{\text{shift}\}\text{”}))$ for the generated image with scale 0 and scale k, and similarly for the class similarity.

In contrast to simply applying a second text prompt (2P) to perform a *binary* shift, our LoRA adapters allow performing a variety of shift scales, as measured by the CLIP shift difference (Section 3.2). This allows gradual shifts, as also illustrated in Fig. 2.

Activating the LoRA adapter at different time steps throughout the diffusion process will modulate the effect of the adapter on the generation process. (Meng et al., 2021; Gandikota et al., 2023) If the LoRA adapter is active for all noise steps, it will significantly influence the semantic structure and the appearance of the generated image, while deactivating the adapter for earlier time steps will keep the semantic structure. Since we aim to perform more fine-grained edits that do not heavily change the semantic structure, we deactivate the LoRA adapter for early steps. This allows realizing edits as, *e.g.*, visualized in Fig. 2 a), where the semantic structure is kept but only the appearance changes.

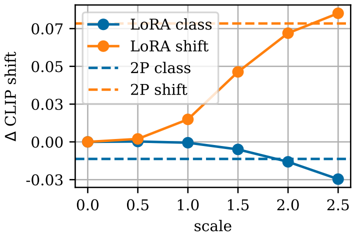


Figure 3: **Average delta CLIP evaluation for various scales of the snow shift.** Our sliders perform a gradual shift, while a naive application (2P) only allows *binary* shifts.

Failure point concept. We define a failure point $s = \min\{S \in \mathbb{R} | f(X(S)) \neq c\}$ as the smallest shift scale where a classifier $f(X(s))$ fails to correctly classify an image $X(s)$ with a class c and a scale s of a considered shift. The failure point distribution captures the ratio of failed samples for the considered scales. We estimate this distribution in our work with a histogram, where the number of elements in one bin I_k is computed by $H(I_k) = \sum_{n=1}^{N} \mathbb{1}_{I_k}(s_n)$ with the indicator function $\mathbb{1}(\cdot)$ and the scale of the n-th element of the set of images with N images. We compute and report the ratio of failure points for the scales $s \in \{0.5, 1, 1.5, 2, 2.5\}$, dividing $H(I_k)$ by the number of considered images N.

3.3 FILTERING DATASET AND STRATEGY

To evaluate filtering strategies for removing out-of-class (OOC) samples, we collect a manually labeled dataset. This section presents this dataset and the selected filtering strategy.

Filtering of OOC samples. Current diffusion models allow the generation of diverse and realistic images $x \sim p(X|\mathbf{z})$ that are conditioned on $\mathbf{z} = [c, s_i]$, which involves the considered ImageNet class $c \in \mathbb{N} \mid 1 \leq c \leq 1000$ and the variable $s_i \in \mathbb{R}$ corresponding to the severity of a considered nuisance shift i. However, due to their probabilistic formulation, the generated sample might deviate from the condition $\mathbf{z}$. For benchmarking applications, we are particularly concerned about generated samples deviating from the original class c, *i.e.*, the considered class cannot be characterized anymore (c.f., Fig. 2). We call such samples "OOC" samples (Metzen et al., 2023). Evaluating the sensitivity to specific nuisance shifts requires removing the OOC samples generated by the shift's application. Therefore, we collect a dataset of generated images to evaluate the sliding process and strategies to automatically remove OOC samples.

Dataset for evaluating OOC filtering strategies. To evaluate various OOC filtering strategies, we manually label a dataset consisting of 18k generated images with two shifts, five scales, and 100 random ImageNet classes. We select *snow* as one weather variation and *cartoon* as one style shift to represent two rather different nuisance shifts. Before manually labeling the dataset, we remove easy samples that have a high CLIP text alignment and are classified correctly by multiple classifiers. Then, all hard images are labeled by two human annotators, where each annotator can choose from the following labels: "class", "partial class properties", and "not class". More details on the labeling strategy and the dataset statistics are provided in Appendix A.6.

OOC filtering strategy. A filter serves its purpose if it removes all OOC samples, corresponding to a high true positive rate (TPR), while not removing too many in-class samples, corresponding to a low false positive rate (FPR). Instead of simply applying a CLIP threshold as in Vendrow et al. (2023), we consider a combinatorial selection approach, which requires two out of four filters to be active. For the first and the second filter, we consider text alignment to "`A picture of a <class>`" and "`A picture of a <class> in <shift>`", respectively, computed via CLIP. For the third and fourth filter, we measure the cosine similarity to the starting images using the CLIP image encoder and the class tokens of DINOv2 (Oquab et al., 2023), respectively. We select the filtering

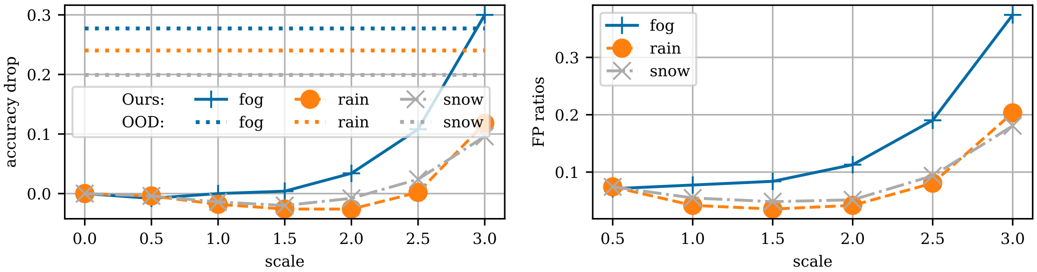


Figure 4: **Accuracies and failure point ratios of a ResNet-50 classifier on OOD-CV and our benchmark.** *Left*: Accuracies on OOD-CV and various scales of our benchmark. Horizontal lines show the average score for each weather nuisance of the OOD-CV test dataset , while our benchmark allows identfying the performance drop at various shift scales. *Right*: Distribution of failure points. While the OOD-CV dataset only provides the accuracy drop, our continuous nuisance shifts allow identifying the shift scales that result in a failure. Note that models fail earlier for fog, potentially due to heavier occlusions than snow and rain.

threshold for each filter such that 90% of the labeled OOC samples are removed. Note that none of these filters are trained on ImageNet data.

4 EXPERIMENTS

In this section, we discuss our benchmark results. First, we compare our bechmarking strategy with the OOD-CV benchmark. Then, we perform a large-scale analysis by evaluating more than 40 ImageNet classifiers on CNS-Bench.

4.1 COMPARING CONTINUOUS SHIFTS WITH OOD-CV DATASET

Zhao et al. (2022; 2024) introduce OOD-CV to measure out-of-distribution (OOD) robustness, a benchmark dataset that includes OOD examples of ten object categories for five different individual nuisance factors (*e.g.*, weather) on real data. OOD-CV is the only real-world dataset that provides accurate labels of various individual weather shifts. This allows comparing our generated images with real-world weather realizations of the considered shifts. We use our trained LoRA adapters to create a benchmark for the OOD-CV classes and scales up to 3.0 to directly compare with the original manually labeled dataset. We refer to the supplementary for exemplary images of both benchmarks and CLIP alignments to the considered shifts.

First, we train a ResNet-50 classifier on the training set of the OOD-CV benchmark. Then, we evaluate the performance on our data and the OOD-CV benchmark. Fig. 4 presents the results for each nuisance independently. The accuracies remain more or less constant with an accuracy around 95% up to a nuisance scale of 1.5. From a nuisance scale of 2.0, the accuracy starts dropping, with the nuisance of *fog* having the biggest impact. This could be explained by the fact that fog can lead to severe occlusion, while rain and snow can be considered as corruption factors. We hypothesize that the partially bigger drop for the OOD-CV benchmark is due to a major limitation of its dataset: The nuisances are not completely disentangled, and part of the accuracy drop originates from various other factors (*e.g.*, image quality, image size, and noise). In contrast, our benchmark allows for fine-grained control of nuisances with multiple shift levels, leading to a more complete and scalable analysis of the model's performance.

4.2 EVALUATED MODELS AND EXPERIMENTAL SETUP

We use our large-scale benchmark to evaluate the models along the following axes:
(i) *Architecture.* To compare architectures with a comparable number of parameters, we consider both CNN and ViT architectures with different training recipes: ResNet-152 (He et al., 2016), ViT-B/16 (Dosovitskiy et al., 2020), DeiT-B/16 (Touvron et al., 2021), DeiT-3-B/16 (Touvron et al., 2022), and ConvNeXt-B (Liu et al., 2022). All models are trained in a supervised manner.
(ii) *Model size.* For ViT, we consider the small, medium, base, large, and huge variants of DeiT-3.

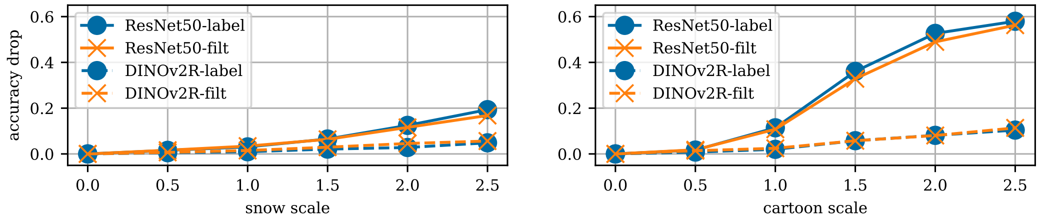


Figure 5: **Classification accuracies on the labeled and the filtered datasets.** The accuracy curves of ResNet-50 and DINOv2 classifiers on the filtered and the labeled dataset of snow and cartoon shift are comparable, demonstrating the effectiveness of our automatic filtering strategy. We provide results for more classifiers in Fig. 8.

For CNN, we consider the ResNet variants: 18, 34, 50, 101, and 152.
(iii) *Pre-training paradigm and data.* We evaluate a set of models with the same backbone but different pre-training strategies. The following models are pre-trained on IN1k with a self-supervised objective: MAE (He et al., 2022a), DINOv1 (Caron et al., 2021), and MoCov3 (Chen et al., 2021). To study the impact of more data during training, we compare their performance to a supervised model that is trained only on ImageNet-1k and a supervised model that is pre-trained on ImageNet-21k. All Transformer-based models use ViT-B/16 as the backbone. Furthermore, we evaluate an ImageNet-trained diffusion classifier (Li et al., 2023b) on a smaller subset due to its heavy computational cost.

Metrics. We typically report the average accuracy drops averaged over the images of one shift or all shifts. In Table 1, we report the mean relative corruption error (rCE) as introduced by Hendrycks & Dietterich (2018). It is defined by the average over all relative corruption errors $\text{CE}_{\text{shift}} = \left(\sum_s E^f_{\text{shift},s} - E^f_{\text{shift},0}\right) / \left(\sum_s E^{\text{alex}}_{\text{shift},s} - E^{\text{alex}}_{\text{shift},0}\right)$ with the average error E for scale s, model f, and a specific shift.

Slider details. As pointed out in Section 3.1, we use textual inversions to replicate the ImageNet distribution. To evaluate the relevance of this approach, we generate 200 images of 100 randomly selected ImageNet classes using standard SD2.0 and SD2.0 with the textual inversions of IN*. To illustrate the distribution gap, we compute the accuracies for ResNet-50 (DeiT). They achieve an accuracy of 68.2% (71.6%) for the SD distribution and 74.1% (79.1%) for the IN* distribution, which equals accuracy drops of 6% (8%) for both classifiers. This is significantly closer to the performance on the original ImageNet distribution. We perform all the following experiments using the IN* distribution. We use SD2.0 and we activate the LoRA adapters with the selected scale for the last 75% of the noise steps.

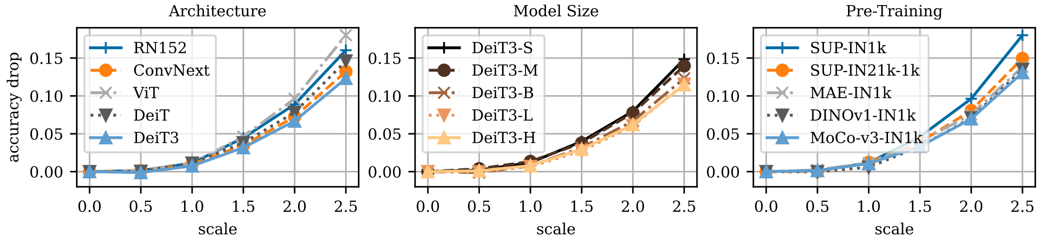
Due to the computational complexity, we perform sliding for 100 classes. To get an estimate of the robustness on the full scale of ImageNet, we classify based on 1000 classes using off-the-shelf classifiers without applying classifier masking, as done by Hendrycks et al. (2021a). We ablate how the number of classes influences the robustness evaluations in Appendix A.5.2.

The selection of the shifts is mainly inspired by ImageNet-R Hendrycks et al. (2021a) (8 shifts) and the OOD-CV dataset Zhao et al. (2022) (6 shifts) to consider a diverse set of nuisance shifts that modulate the appearance and style or the background and occlusion. Specifically, we consider the following 14 shifts: cartoon style, plush toy style, pencil sketch style, painting style, design of sculpture, graffiti style, video game renditions style, style of a tattoo, heavy snow, heavy rain, heavy fog, heavy smog, heavy dust, and heavy sandstorm.

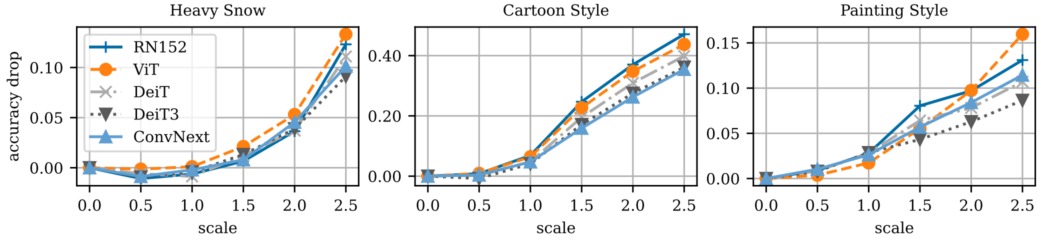
Filtering details. Our OOC filtering mechanism reaches a TPR of 87.9% and an FPR of 12.0% with an accuracy of 88.0%, while the naive CLIP-based thresholding reaches a TPR of 89.9% and an FPR of 35.7% with an accuracy of 65.1%. We plot the classification accuracy of DINOv2-R and ResNet-50 for the labeled and the filtered versions in Fig. 5. We observe comparable accuracy drops on both the manually-labeled and the filtered datasets. To further support the realism of our generated images, we fine-tune ResNet-50 with our data and show more than 10% gains on ImageNet-R (see Appendix A.3).

Table 1: **rCE along the model axes.** We choose the average relative corruption error Hendrycks & Dietterich (2018) as a single metric to measure the performance of a model on our benchmark (lower is better). We provide results for all models in Table 2.

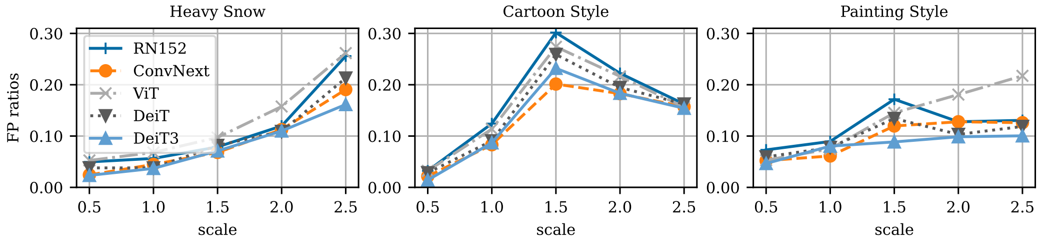
Architecture		Size		Pre-Training	
ConvNext	0.686	DeiT3-S	0.747	DINOv1-IN1k	0.636
DeiT3	0.610	DeiT3-M	0.758	MAE-IN1k	0.732
DeiT	0.746	DeiT3-B	0.610	MoCov3-IN1k	0.669
RN152	0.790	DeiT3-L	0.574	SUP-IN1k	0.926
ViT	0.926	DeiT3-H	0.583	SUP-IN21k-1k	0.722



(a) Accuracy drops averaged over the whole benchmark. Architecture (*left*): We show models with the same training data and similar parameter counts. The selection of the architecture influences the accuracy drop. Model size (*center*): We show DeiT3 with various numbers of parameters. Increasing the model capacity results in lower accuracy drops. Pre-training paradigm and data (*right*): We show different pre-training paradigms: supervised, self-supervised (MAE, DINO, MoCo), and more data (IN21k), all using ViT-B/16. We present results for all shifts in Fig. 9.



(b) Accuracy drops for three selected shifts. Models exhibit varying performance changes depending on the considered shifts. For snow and painting shifts, the ranking of the models changes. In contrast, the cartoon style shift results in a consistent model ranking. However, the OOD performance on cartoon-shifted images is drastically worse than the other shifts.



(c) Ratio of failure points per scale for various models and shifts. The distribution allows inferring at which scales various models fail most often. Different models fail at varying stages depending on the considered shifts. While the number of failure points gradually increases for the snow shift, most failure points occur around scale 1.5 for the cartoon style shift. We present results for all shifts in Fig. 10.

Figure 6: **Evaluation of accuracy drops and failure points.** We plot the averaged accuracy drops and failure points of selected models and provide the results of all evaluated models in Appendix A.2.

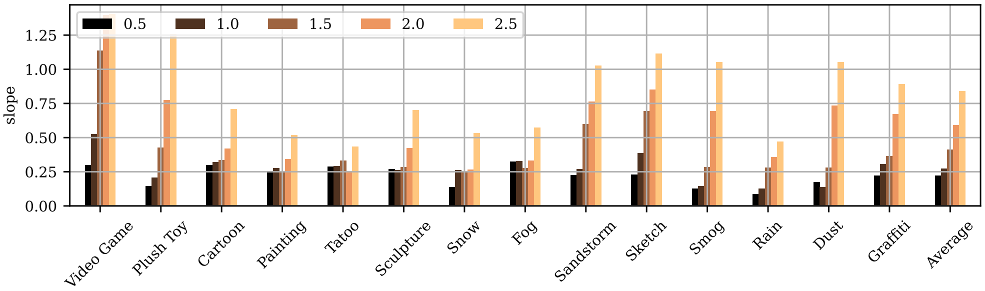


Figure 7: **Relation between ID and OOD accuracy.** We report the slope of the linear fit between ID and OOD accuracy using 16 supervised ImageNet-trained models for all evaluated shifts. The relation varies for different shifts and scales between 0.5 and 2.5.

4.3 ANALYSIS AND FINDINGS

In this subsection, we discuss the main findings on our benchmark. Following Hendrycks et al. (2021a), we report the accuracy drops for 5 scales averaged over 14 diverse shifts as a measure of robustness in Fig. 6a. Table 1 compares models using the average relative corruption errors as proposed by Hendrycks & Dietterich (2018). We also provide results for three exemplary shifts in Fig. 6b. In addition, we report the distribution of failure points in Fig. 6c. We provide more evaluations in Appendix A.2.

Considering multiple scales of a shift allows a more nuanced analysis of OOD robustness. We present the accuracy drops for multiple scales and classifiers along the architecture axis in Fig. 6b. The results indicate that the model rankings measured by the accuracy drop change for different scales and shifts. For example, while the rankings remain consistent for the cartoon style (*right*) for all scales, the model rankings change significantly for the painting style shift: Here, ViT outperforms the other models on a lower scale but performs worse on large shift scales. Varying rankings also occur for other shifts (see Fig. 9 in the supplementary). To validate the observation of changed model rankings, we also evaluate multiple corruption levels of an examplary ImageNet-C corruption and show the results in Fig. 23 in the supplementary.

We conclude from this observation that the average accuracy drop and the accuracy drops at specific nuisance scales do not always indicate the same model behavior, which provides experimental evidence for the need for a multi-scale robustness benchmarking dataset and adequate metrics.

Model failure points differ across different types of shifts. A failure point captures at which scale a model fails for the first time. Comparing the failure point distribution of various models largely differs for different shift types, as shown in Fig. 6c. We provide more results in Fig. 10 in the supplementary. Weather shifts, such as snow, typically correspond to slight appearance changes and mainly add a disturbance factor or occlusions to the image. Therefore, the failure rate increases gradually compared to some style shifts, for which models tend to fail more abruptly at a specific scale, as, *e.g.*, for the cartoon style at scale $s = 1.5$. An exemplary explanation for the abrupt shift for the cartoon shift might be the wrong classification of a class as the ImageNet class *comic book*.

The relation between ID and OOD accuracy depends on the considered nuisance factor and its scale. Miller et al. (2021) formalize the positive correlation between ID and OOD accuracy—classifiers tend to have a better OOD accuracy if they perform better on the training data ("*Accuracy-on-the-line*" phenomenon). To analyze the linear relation between ID and OOD accuracy for our benchmark, we compute the slope of the linear fit between ID and OOD accuracies of 16 ImageNet-trained models. Miller et al. (2021) have already shown that the slope varies for different datasets. In Fig. 7, we further observe that not only the considered shift but also its severity influence the slope of the linear fit. Refer to Appendix A.2.3 for the test statistics. We believe using our benchmark to investigate this relation more extensively is an interesting direction for future work.

Transformers with modern training recipes outperform modern CNNs across all shift severities. We present the average accuracy drops of various models with the same training data and a comparable number of parameters in Fig. 6a (*left*). DeiT3 consistently achieves the highest robustness on our benchmark, increasing the gap towards DeiT and ViT for stronger shifts. Interestingly, ResNet-152 is more robust than the standard ViT variant, but ConvNeXt outperforms the ResNet-152 architecture. A modern CNN (ConvNext) outperforms vision transformers (ViT,DeiT) of the same size but it is less robust than a transformer with modern training recipes (DeiT3), despite having a higher ID accuracy. This observation is in line with the performance on ImageNet-R. However, our benchmark shows that the gap between ConvNext and DeiT3 does not increase for stronger shifts. We can observe that this behavior is not consistent for all shifts. Consider, *e.g.*, the failure point distribution in Fig. 6c (*Painting Style*), where DeiT3 has a gradually increasing failure point rate, while ConvNext depicts a sharp increase for scale $s = 1.5$.

Self-supervised pre-training improves the OOD robustness. To study the impact of the pre-training paradigm, we compare different learning objectives with the same ViT-B backbone and the same training data in Fig. 6a (*right*). We consider both the supervised and self-supervised (MAE, DINOv1, and MoCov3) paradigms. Using a self-supervised objective for pre-training followed by a fine-tuning protocol results in a better robustness for the same training data and model size. Considering the rCE metric in Table 1, the fine-tuned DINOv1 model achieves the best performance.

Diffusion classifiers are less robust than discriminative models. In addition, we also compare the robustness of an ImageNet-trained diffusion classifier (Li et al., 2023b) on our benchmark. Due to the heavy computational cost, we evaluate the accuracy drop of the DiT-based diffusion classifier for 1k images on a subset of our dataset (around 12k images) for the snow and the cartoon style shift. We apply the L1 loss computation strategy as proposed by Li et al. (2023b) since it results in the best performance. We compute the average accuracy drops as 0.106 / 0.07 / 0.05 for DiT / supervised ViT / MAE. Comparing on the smaller dataset with discriminative models, the diffusion classifier demonstrates a lower robustness on the evaluated shifts than the compared discriminative models despite having substantially more parameters. The gap is increasing for larger severity levels. We present more results in Fig. 21 in the supplementary.

More training data improves the robustness. In Fig. 6a (*right*), we observe that more training data benefits OOD robustness for all scales. For example, compared with the supervised model trained on IN1k, pre-training on IN21k has a positive impact on the OOD robustness for small and large scales. This might be explained by the fact that the tested distribution is less OOD for the model (Miller et al., 2021).

In summary, we show that benchmarking with generative continuous shifts allows systematically studying the model robustness via easily scalable synthetic data. Our study underscores that considering multiple-scale nuisance shifts provides a more nuanced view of the model robustness, as the performance drops can vary across different nuisance shifts and scales. Besides, the relation between ID and OOD accuracy not only depends on the considered nuisance factor but also on its severity. Therefore, instead of aggregating the robustness evaluation into a single metric, we motivate the community to report the accuracy with different shift scales and the failure points for a more comprehensive understanding of model robustness.

5 CONCLUSION

The key advantage of using generative models for benchmarking is the ability to perform diverse nuisance shifts in a controlled and scalable way. This work filled a gap in generative benchmarking by introducing CNS-Bench, an evaluation method that performs diverse, realistic, fine-grained, and continuous nuisance shifts at multiple scales. We further added a new dimension for benchmarking robustness by introducing the concept of failure points. Our systematic evaluation of classifiers revealed new insights along three axes (architecture, number of parameters, pre-training paradigm and data) and demonstrated the importance of continuous shifts in assessing the model robustness. Furthermore, we studied the necessity of removing out-of-class samples when benchmarking with diffusion-generated images. We hope this benchmark can encourage the community to adopt generated images for evaluating the robustness of vision models.

6 REPRODUCIBILITY STATEMENT

All steps of our benchmarking pipeline are reproducible: We provide our datasets and implementation as part of the supplementary material, which includes code to reproduce training of LoRA adapters, generation of images, filtering, and evaluation of all classifiers. We also include all evaluated classification results for all images of the dataset in the shared code. All classifiers are evaluated in a standardized way using the *easyrobust* (Mao et al., 2022) framework.

The supplementary material contains more details about the implementation, the computation of metrics, the labeling, and the filtering strategies.

We also refer to our datasheet in Appendix B.

REFERENCES

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Teodor Zrnic. Prediction-powered inference. *Science*, 382:669–674, Nov 2023.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in t2i models by identifying semantic directions, 2024. URL `http://arxiv.org/abs/2403.17064`.

Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari Morcos. Pug: Photorealistic and semantically controllable synthetic data for representation learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9620–9629, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

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. In *ICLR*, 2020.

Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. Robustness in deep learning for computer vision: Mind the gap? *CoRR*, abs/2112.00639, 2021. URL `https://arxiv.org/abs/2112.00639`.

A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA). http://www.robots.ox.ac.uk/ vgg/software/via/, 2016. Version: X.Y.Z, Accessed: 2024-05-12.

Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6889-6/19/10. doi: 10.1145/3343031.3350535. URL `https://doi.org/10.1145/3343031.3350535`.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338, 2010.

Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. The scaling law of synthetic images for model training, for now. In *CVPR*, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: LoRA adaptors for precise control in diffusion models, 2023. URL `http://arxiv.org/abs/2311.12092`.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. Datasheets for datasets, 2021.

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, 2022. URL `http://arxiv.org/abs/1811.12231`.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York, NY, 2 edition, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022a.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022b.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

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 *ICCV*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. Imagenet-x: Understanding model mistakes with factor of variation annotations. *arXiv preprint arXiv:2211.01866*, 2022.

Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *CVPR*, pp. 18963–18974, 2022.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier, 2023a.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2206–2217, 2023b.

Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, and Hui Xue. Imagenet-e: Benchmarking neural network robustness via attribute editing. In *CVPR*, pp. 20371–20381, 2023c.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL `http://arxiv.org/abs/1405.0312`.

Jiang Liu, Chen Wei, Yuxiang Guo, Heng Yu, Alan Yuille, Soheil Feizi, Chun Pong Lau, and Rama Chellappa. Instruct2attack: Language-guided semantic adversarial attacks. *arXiv preprint arXiv:2311.15551*, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022.

Xiaofeng Mao, Yuefeng Chen, Xiaodan Li, Gege Qi, Ranjie Duan, Rong Zhang, and Hui Xue. Easyrobust: A comprehensive and easy-to-use toolkit for robust computer vision. `https://github.com/alibaba/easyrobust`, 2022.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2021.

Jan Hendrik Metzen, Robin Hutmacher, N Grace Hua, Valentyn Boreiko, and Dan Zhang. Identification of systematic errors of image classifiers on rare subgroups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5064–5073, 2023.

John 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*, 2021.

Mohammadreza Mofayezi and Yasamin Medghalchi. Benchmarking robustness to text-guided corruptions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 779–786, 2023. doi: 10.1109/CVPRW59228.2023.00085.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pp. 6038–6047, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Michelle Shu, Chenxi Liu, Weichao Qiu, and Alan Yuille. Identifying model weakness with adversarial examiner. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11998–12006, 2020.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, pp. 2256–2265. JMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *ECCV*, 2022.

Joshua Vendrow, Saachi Jain, Logan Engstrom, and Aleksander Madry. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. *arXiv preprint arXiv:2302.07865*, 2023.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

Chenshuang Zhang, Fei Pan, Junmo Kim, In So Kweon, and Chengzhi Mao. Imagenet-d: Benchmarking neural network robustness on diffusion synthetic object. In *CVPR*, pp. 21752–21762, 2024.

Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *ECCV*, 2022.

Bingchen Zhao, Jiahao Wang, Wufei Ma, Artur Jesslen, Siwei Yang, Shaozuo Yu, Oliver Zendel, Christian Theobalt, Alan Yuille, and Adam Kortylewski. Ood-cv-v2: An extended benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *NeurIPS*, 2023.