Assessing Visually-Continuous Corruption Robustness of Neural Networks Relative to Human Performance

Huakun Shen
University of Toronto
huakunshen@cs.toronto.edu

Boyue Caroline Hu
University of Toronto
boyue@cs.toronto.edu

Krzysztof Czarnecki
University of Waterloo
kczarnec@gsd.uwaterloo.ca

Lina Marsso
University of Toronto
lina.marsso@utoronto.ca

Marsha Chechik
University of Toronto
chechik@cs.toronto.edu

Abstract

Neural Networks (NNs) have surpassed human accuracy in image classification on ImageNet, yet they often lack robustness against image corruption, i.e., corruption robustness, with such robustness being seemingly effortless for human perception. In this paper, we propose visually-continuous corruption robustness *(VCR) – an extension of corruption robustness to allow assessing it over the wide and continuous range of changes that correspond to the human perceptive quality (i.e., from the original image to the full distortion of all perceived visual information), along with two novel human-aware metrics for NN evaluation. To compare VCR of NNs with human perception, we conducted extensive experiments on 14 commonly used image corruptions with 7,718 human participants and state-of-the-art robust NN models with different training objectives (e.g., standard, adversarial, corruption robustness), different architectures (e.g., convolution NNs, vision transformers), and different amounts of training data augmentation. Our study showed that: 1) assessing robustness against continuous corruption can reveal insufficient robustness undetected by existing benchmarks; as a result, 2) the gap between NN and human robustness is larger than previously known; and finally, 3) some image corruptions have a similar impact on human perception, offering opportunities for more cost-effective robustness assessments.*

1. Introduction

For Neural Networks (NNs) used in safety-critical domains, ensuring robustness against potential corruptions is crucial [15]. As NNs in these domains automate tasks usually performed by humans, comparing their robustness to human performance is essential.

Human VS NN robustness. Corruption robustness measures the average-case performance of an NN or humans on a set of image corruption functions [15]. Existing studies, including out-of-distribution anomalies [16], benchmarking [15, 18], and comparison with humans [10, 20], generally evaluate robustness against a pre-selected, fixed set of transformation parameter values that represent varying degrees of image corruption. However, these parameters may not capture how different levels of corruption affect human perception. For instance, the same parameter can impact visual perception differently depending on the image's brightness [20]. Humans can perceive a continuous spectrum of visual corruptions, from subtle to extreme [9, 41], so relying on fixed parameters may lead to incomplete coverage of the full range of visual corruptions and biased evaluations of NN robustness compared with humans.

Contributions and Outlook. To address these issues, we introduce *visually-continuous corruption robustness* (VCR), focusing on NN robustness across a continuous range of image corruption levels. We also present two novel human-aware metrics (HMRI and MRSI) for comparing NN performance with human perception. Our extensive experiments, involving 7,718 Mechanical Turk participants and 14 common image transformations from three sources [1], reveal a significant robustness gap between NNs and humans. No NN fully matches human performance across the entire continuous range of corruption levels in terms of both accuracy and prediction consistency, and only a few exceed human performance by a small margin in specific levels of corruption. Our experiments yield insightful findings about the robustness of human and state-of-the-art (SoTA) NNs concerning accuracy, degrees of visual corruption, and consistency of classification, which can contribute towards the development of NNs that match or surpass human perception. We also discovered classes of corruption transformations for which humans showed similar robustness (e.g., different types of noise),

[1] The number is comparable to 15 corruptions included in IMAGENET-C.

while NNs reacted differently. Recognizing these classes can contribute to reducing the cost of measuring human robustness and elucidating the differences between humans and computational models. Our validation set with 14 image corruptions, human robustness data, and the evaluation code is provided as a toolbox and a benchmark[2].

2. Related Work

We briefly review related work on the comparison of human and NN robustness, adversarial robustness, robustness benchmarks and improving robustness.

Human VS NN Robustness. Prior studies have used human performance to study the existing differences between humans and neural networks [6, 55], to study invariant transformations [23], to compare recognition accuracy [19, 44], to compare robustness against image transformations [9, 10], or to specify expected model behaviour [20]. The main difference between our study and existing work, specifically, the most recent study by [10], is three-fold: 1) we are the first to quantify robustness across the full continuous visual corruption range, thus revealing previous undetected robustness gap; 2) our experiments for obtaining human performance are designed to include more participants for measuring the *average* human robustness, resulting in more generalizable results and reduced influence of outliers; 3) we identified visually similar transformations for humans but not NNs, potentially reducing experiment costs.

Robustness Benchmarks. Hendrycks et al. built the IMAGENET-C and -P benchmarks for checking NN model classification robustness against common corruptions and perturbations on IMAGENET images [15]. They have inspired other benchmarks for different corruption functions, datasets, and tasks [2, 21, 22, 32, 33, 46, 51]. However, these benchmarks generate images by applying corruption functions with only five pre-selected values per parameter. IMAGENET-CCC [36] is the only prior work targeting a more continuous range of corruptions, by using 20 pre-selected values per parameter. It does not check the coverage in terms of the visual effects on the images, which we do with an Image Quality Assessment (IQA) metric Visual Information Fedility (VIF) [41]. Further, their work focuses on continuous changes over time for benchmarking test-time adaptation, which is different from a general robustness benchmark, and the dataset has not been released as the time of writing. In contrast to all these previous works, our method randomly and uniformly samples parameter values to cover the full range of visual change that a corruption function can achieve, which is modeled and assessed for coverage using an IQA metric. Our work also compares robustness of NNs with humans.

Adversarial Robustness. Adversarial robustness measures the worst-case performance on images with added 'small' distortions or perturbations tailored to confuse a classifier [15]. However, changes that can be encountered in the real-world situations are often of a much bigger range [22]. Thus, in this paper, we focus on *average-case performance* over a *realistic* range of changes.

Improving Robustness. Numerous methods for improving model robustness have been proposed, e.g., data augmentation with corrupted data [8, 30, 31, 38], texture changes [11, 14], image compositions [53, 54] and corruption functions [17, 52]. All of these have different abilities to generalize to unseen data [22]. While not our primary focus, we demonstrate that NN robustness compared to humans can be improved through data augmentation and fine-tuning with our generated images for VCR.

3. Visually-Continuous Corruption Robustness

To study robustness against a wide and continuous spectrum of visual changes, we define *visually-continuous corruption robustness* (VCR) and describe our method for generating test sets. To study VCR of NNs in relation to humans, we also present the human-aware metrics.

3.1. VCR Definition

A key difference between corruption robustness and VCR is that the latter is defined relative to the *visual impact* of image corruption on human perception, rather than the transformation parameter domain. To quantify visual corruption, VCR uses the Image Quality Assessment (IQA) metric Visual Information Fidelity (*VIF*) [28, 41]. VIF measures the perceived quality of a corrupted image x' compared to its original form x by measuring the visual information unaffected by the corruption. Thus, we define the *change* in the perceived quality caused by the corruption as $\Delta_v(x, x') = max(0, 1 - VIF(x, x'))$. See App. C for more detail on Δ_v. With Δ_v, whose value ranges from 0 and 1, we can consider VCR against the wide, finite, and continuous spectrum of visual corruptions ranging from no degradation to visual quality (i.e., the original image) ($\Delta_v = 0$) to the full distortion of all visual information ($\Delta_v = 1$).

Limitation: VCR is limited to image corruption applicable to the chosen IQA metric, thus by using VIF, VCR is limited to only pixel-level corruption. Metrics suitable for other types of corruption (e.g., geometric) need further research.

For VCR, we consider a classifier NN $f : X \to Y$ trained on samples of a distribution of input images P_X, a ground-truth labeling function f^*, and a parameterized image corruption function T_X with a parameter domain C. We are interested in robustness of f against images with all degrees of visual corruption *uniformly* ranging from $\Delta_v = 0$ to $\Delta_v = 1$.[3] Given a value $v \in [0, 1]$, we define $P(x, x'|v)$

[2] https://github.com/HuakunShen/VCR

[3] Note that distributions other than uniform can be used based on the application. For example, one may wish to favour robustness against heavy snow conditions for NNs deployed in arctic areas.

as the *joint distribution* of original images (x) and corresponding corrupted images ($x' = T_X(x, c)$, $c \in C$) with $\Delta_v(x, x') = v$. We define VCR in the presence of a robustness property γ that f should satisfy given T_X:

$$\mathcal{R}_\gamma = \mathbb{E}_{v\sim Uniform(0,1)}(P_{x,x'\sim P(x,x'|v)}(\gamma)) \quad (1)$$

In this paper, we instantiate VCR with two existing robustness properties (see Fig. 1). The first one is *accuracy* (a_v), requiring that the prediction on corrupted images should be correct, i.e., $f(x') = f^*(x)$. It is also used in the existing definition of corruption robustness [15]. Thus,

$$\mathcal{R}_a = \mathbb{E}_{v\sim Uniform(0,1)}(P_{x,x'\sim P(x,x'|v)}(f(x') = f^*(x))) \quad (2)$$

The second property is *prediction consistency* (p_v), requiring consistent predictions before and after corruptions, i.e., $f(x') = f(x)$ [20]. It is applicable when ground truth is not available, which is common during deployment. Thus,

$$\mathcal{R}_p = \mathbb{E}_{v\sim Uniform(0,1)}(P_{x,x'\sim P(x,x'|v)}(f(x') = f(x))) \quad (3)$$

Summary of VCR Definitions. Fig. 1 gives a visual summary of the VCR metrics, starting with the general definition $\mathcal{R}_\gamma$ at the top, and instantiating it for accuracy as $\mathcal{R}_a$ and consistency as $\mathcal{R}_p$. Each of them is the average accuracy or prediction consistency, respectively, over the full and continuous range of visual change $\Delta_v \in [0, 1]$.

3.2. Testing VCR

VCR of a subject (a human or an NN) is measured by first generating a test set through sampling and then estimating it using the sampled data. The test set is generated by sampling images and applying corruption to obtain $P(x, x'|v)$ for different Δ_v values v. We sample $x \sim P_X$ and $c \sim$ *Uniform*(C), and obtain $x' = T_X(x, c)$ and $v = \Delta_v(x, x')$, resulting in samples (x, x', c, v). Then, we divide them into groups of (x, x', c), each with the same v value. Next, by dropping c, we obtain groups of (x, x') with the same v, which are samples from $P(x, x'|v)$. Note that we consider each group separately, thus this procedure requires only sufficient data in each group but not uniformity, i.e., $v \sim$ *Uniform*$(0, 1)$ is not required. The varying size of each group, i.e., the non-uniformity of v distribution, will not distort VCR estimates, but only impact the estimate uncertainty at a given v. Further, interpolation in the next step helps address any missing points.

With the test set, we estimate the performance w.r.t. the property γ for each v. For each v in the test data, we compute the *rate* of accurate predictions $f(x') = f^*(x)$ to estimate accuracy, i.e., $a_v = P_{x,x'\sim P(x,x'|v)}(f(x') = f^*(x))$ [resp. consistent predictions $f(x') = f(x)$ to estimate consistency, i.e., $p_v = P_{x,x'\sim P(x,x'|v)}(f(x') = f(x))$]. Then by plotting (v, a_v) and (v, p_v) and applying monotonic smoothing splines [25] to reduce randomness and outliers, we obtain smoothed spline curves s_a and s_p, respectively. The curves s_γ (namely, s_a and s_p) describe how the performance w.r.t. the robustness property γ (namely, a and p) decreases as the visual corruption in images increases. Finally, we estimate $\mathcal{R}_a = \mathbb{E}_{v\sim Uniform(0,1)}(a_v)$ [resp. $\mathcal{R}_p = \mathbb{E}_{v\sim Uniform(0,1)}(p_v)$] as the area under the spline curve, i.e., $\hat{\mathcal{R}}_a = A_a = \int_0^1 s_a(v)dv$ [resp. $\hat{\mathcal{R}}_p = A_p = \int_0^1 s_p(v)dv$]. See Alg. 1 for the pseudo-code of VCR estimation.

3.3. Human-Aware Metrics for VCR

A commonly used metric for measuring corruption robustness is the *Corruption Error (CE)* [15]—the top-1 classification error rate on the corrupted images, normalized by the error rate of a baseline model. CE can be used to compare an NN with humans if the baseline model is set to be humans. However, CE is not able to determine whether an NN can exceed humans, and NN models could potentially have super-human accuracy for particular types of perturbations or in some Δ_v ranges. Inspired by CE, we propose two new human-aware metrics, *Human-Relative Model Robustness Index* (*HMRI*) that measures NN VCR relative to human VCR; and *Model Robustness Superiority Index* (*MRSI*) that measures how much an NN exceeds human VCR.

Auxiliary VCR metrics to compute *HMRI* and *MRSI*. *HMRI* and *MRSI* take as inputs the estimated spline curves for humans (s^h_γ) and for NN (s^m_γ). We denote areas under these curves as A^h_γ and A^m_γ, respectively (see Fig. 2a). To compare NN model and human performance, VCR w.r.t. prediction consistency or accuracy is estimated using Alg. 1 using both model and human performance data, as illustrated by the yellow (A^h_γ) and blue (A^m_γ) areas in Fig. 2a, respectively. Both the blue and yellow areas also include the green area representing their overlap. Additionally, the VCR lead of humans over a model $A^{h>m}_\gamma$, the girded area in Fig. 2a, and the VCR lead of a model over humans $A^{m>h}_\gamma$, the striped area in Fig. 2a, are estimated. The definitions of these four auxiliary metrics are summarized in Tab. 2b, and they are used to define *HMRI* and *MRSI*.

Definition 1 (***HMRI***). *Given s^h_γ and s^m_γ, let $A^{h>m}_\gamma = \int_0^1 (s^h_\gamma(v) - s^m_\gamma(v))^+ dv$ denote the average (accuracy or preservation) performance lead of humans over a model across the visual change range, where the performance lead is defined as the positive part of performance difference, i.e., $(s^h_\gamma(v) - s^m_\gamma(v))^+ = max(0, s^h_\gamma(v) - s^m_\gamma(v))$. Human-Relative Model Robustness Index (HMRI), which quantifies the extent to which a DNN can replicate human performance, is defined as $\frac{A^h_\gamma - A^{h>m}_\gamma}{A^h_\gamma} = 1 - \frac{A^{h>m}_\gamma}{A^h_\gamma}$.*

The *HMRI* value ranges from $[0, 1]$; a higher *HMRI* indicates a NN model closer to human VCR, and *HMRI* $= 1$ signifies that s^m_γ is the same as or completely above s^h_γ in the entire Δ_v domain, i.e., the NN is at least as robust as an average human (see Fig. 2a).

Definition 2 (***MRSI***). *Given s^h_γ and s^m_γ, let $A^{m>h}_\gamma = \int_0^1 (s^m_\gamma(v) - s^h_\gamma(v))^+ dv$ denote the average performance lead of a model over a human across the visual change*

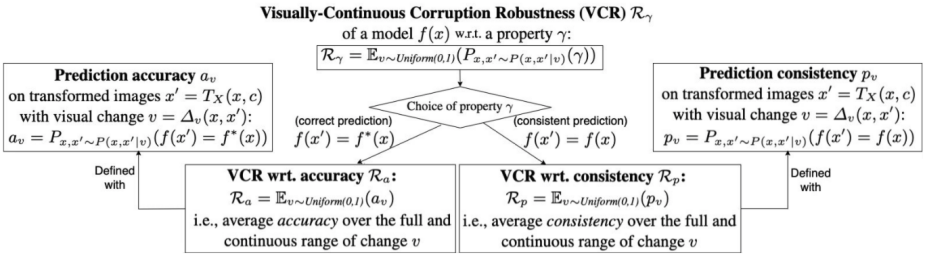


Figure 1. Summary of VCR definitions with respect to accuracy and consistency.

range. Model Robustness Superiority Index (MRSI), which quantifies the extent to which a DNN model can surpass human performance, is defined as $\frac{A_\gamma^{m>h}}{A_\gamma^m}$.

The *MSRI* value ranges from $[0, 1)$, with the higher value indicating better performance than humans. $MSRI = 0$ means that the given NN model performs worse than or equal to humans in the entire Δ_v domain. A positive *MSRI* value indicates that the given NN model performs better than humans at least in some ranges of Δ_v (see Fig. 2a). Comparing humans and NNs with *HMRI* and *MRSI* yields three possible scenarios: (1) humans' performance fully exceeds NN's, i.e., $0 < HMRI < 1$ and $MRSI = 0$; (2) NN's performance fully exceeds humans', i.e., $HMRI = 1$ and $MRSI > 0$; and (3) humans' performance is better than NN's in some Δ_v intervals and worse in others, i.e., $HMRI < 1$ and $MRSI > 0$.

4. Experiments

In this section, we describe experiments that check the VCR of NN models against human performance.

NN models. Tab. 1 summarizes models included in our study. We have selected a wide range of architectures (CNN and transformer architectures) and training methods (supervised, adversarial, semi-weakly, and self-supervised), including dinov2_giant [34], which is on the top of the IMAGENET-C leaderboard as of time of writing. In total, we studied 11 standard supervised models, 4 adversarial learning models, 2 SWSL models, 1 CLIP (clip-vit-base-patch32) model and 3 ViT models. For CLIP, we used the prompt "a picture of (ImageNet class)" while tokenizing the labels.

Image Corruptions. As shown in Fig. 3, we focus on studying VCR of NNs in relation to humans regarding 14 commonly used image corruptions from three different sources: Shot Noise, Impulse Noise, Gaussian Noise, Glass Blur, Gaussian Blur, Defocus Blur, Motion Blur, Brightness and Frost from IMAGENET-C [15]; Blur, Median Blur, Hue Saturation Value and Color Jitter from Albumentations [1]; and Uniform Noise from [9].

Crowdsourcing. Since VCR focuses on the average-case performance, we used crowdsourcing to measure human performance, as it allows for a larger participant pool and more accurate estimation. The experiment is designed following [20] and [9]. The experiment procedure is a *forced-choice image categorization task*: humans were shown one image at a time for 200 ms to limit recurrent processing and asked to choose the correct category from 16 entry-level labels [9]. For NN models, the 1,000-class decision vector was mapped to the same 16 classes using the WordNet hierarchy [9]. The time to classify each image was set to ensure fairness in the comparison between humans and machines [6]. Between images, we showed a noise mask to minimize feedback influence in the brain [9]. Qualification tests and sanity checks were used to filter out misunderstandings and instances of spam [35], resulting in 7,718 participants with 70,000 predictions on corrupted images and 50,000 on original images. The same image was never shown to a participant more than once.

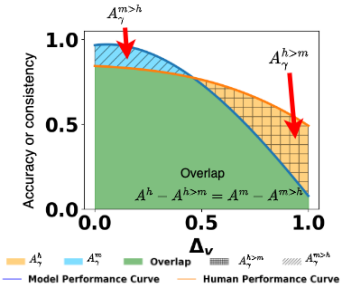
4.1. Testing Robustness against Visual Corruption

IMAGENET-C is a SoTA benchmark for corruption robustness, using 5 pre-selected parameter values for each corruption type on IMAGENET validation images [15]. This section compares robustness results from IMAGENET-C with those from VCR across all 9 IMAGENET-C corruption functions in our study. Full results are available in the Appendix due to page limitations.

Visual Corruption in Test Sets. For each corruption, our generated tests contain $50,000$ images, mirroring the size of the IMAGENET [39] validation set, while IMAGENET-C includes $5 \times 50,000$ images. Due to differences in test set generation, the corruption distributions differ in coverage and peak at different values (e.g., Fig. 4).

To assess the coverage of Δ_v in the test sets, Tab. 2 shows the percentage of the full Δ_v covered. The distribution is divided into 40 equal-width bins, with coverage defined as having 20 or more images per bin. IMAGENET-C exhibits low Δ_v coverage, particularly for Gaussian blur at 56.4%, focusing mainly on the center and missing the low and high Δ_v values, leading to biased evaluation (Fig. 4 and Tab. 2. In contrast, our test sets cover nearly the entire domain, with 97.4% coverage. Our test sets have consistently higher coverage than IMAGENET-Cfor other corruptions as well. Full details are in the Appendix.

Among all corruptions studied, Shot Noise and Impulse Noise have relatively low coverage, because the level of noise these functions add is exponential to their parameters. As a result, uniform sampling of the parameter range C fails to cover small Δ_v values. When using uniform sampling



(a) Visualization of auxiliary metrics for model vs. human performance.

Auxiliary metric (cf. Fig. 2a)
VCR of humans w.r.t. a property γ, estimated as an area under performance curve A_γ^h: $\hat{\mathcal{R}}_\gamma^h = A_\gamma^h = \int_0^1 s_\gamma^h(v)dv$
VCR of a model $f(x)$ w.r.t. a property γ, estimated as an area under performance curve A_γ^m: $\hat{\mathcal{R}}_\gamma^m = A_\gamma^m = \int_0^1 s_\gamma^m(v)dv$
VCR lead of humans over a model $f(x)$ w.r.t. a property γ, estimated as a difference area $A_\gamma^{h>m}$: $\hat{\mathcal{R}}_\gamma^{h>m} = A_\gamma^{h>m} = \int_0^1 \max(0, s_\gamma^h(v) - s_\gamma^m(v))dv$
VCR lead of a model $f(x)$ **over humans** w.r.t. a property γ, estimated as a difference area $A_\gamma^{m>h}$: $\hat{\mathcal{R}}_\gamma^{m>h} = A_\gamma^{m>h} = \int_0^1 \max(0, s_\gamma^m(v) - s_\gamma^h(v))dv$

(b) Summary of auxiliary metrics for defining *HMRI* and *MRSI*.

Figure 2. Auxiliary VCR metrics to compute *HMRI* and *MSRI*.

Model	Architecture	Training Method	Model	Architecture	Training Method
NoisyMix [5]	ResNet-50	Supervised	NoisyMix_new [5]	ResNet-50	Supervised
SIN [11]	ResNet-50	Supervised	SIN_IN [11]	ResNet-50	Supervised
SIN_IN_IN [11]	ResNet-50	Supervised	HMany [14]	ResNet-50	Supervised
HaugMix [17]	ResNet-50	Supervised	Standard_R50 [12]	ResNet-50	Supervised
AlexNet [26]	AlexNet	Supervised	Tian_DeiT-S [47]	DeiT Small	Supervised ViT
Tian_DeiT-B [47]	DeiT Base	Supervised ViT	Do_50_2_Linf [40]	WideResNet-50-2	Adversarial
Liu_Swin-L [29]	Swin-L	Adversarial	Liu_ConvNeXt-L [43]	ConvNeXt-L	Adversarial
Singh_ConvNeXt-L-ConvStem [43]	ConvNeXt-L + ConvStem	Adversarial	swsl_resnet18 [49]	ResNet-18	Semi-weakly sup.
swsl_resnext101_32x16d [49]	ResNext-101	Semi-weakly sup.	CLIP [37]	Clip	Supervised CLIP
dinov2_giant [34]	ViT	Self-supervised ViT			

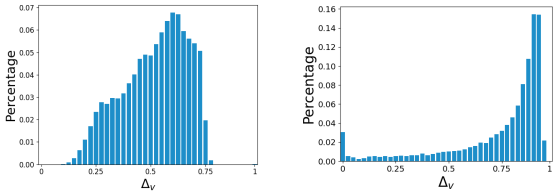
Table 1. Summary of the models included in our study.

Figure 3. Image corruption functions.

over C, reaching the full coverage of Δ_v would require a large amount of data. Note, however, Alg. 1 still computes VCR over the full Δ_v range of $[0..1]$, and the lack of samples for low values of Δ_v has a limited impact on the VCR estimate. This is because we fit a monotonic spline that is anchored with a known initial performance for $\Delta_v = 0$, as discussed in App. D.

Remark: The reported accuracy of ImageNet-C can be directly impacted both by a lack of coverage and by non-uniformity, as it is computed as the average accuracy of all corrupted images. In contrast, the shape of the Δ_v distribution in the test images does not impact VCR once sufficient coverage is achieved to estimate the spline curves s_γ.

Robustness Evaluation Results. Next, we compare robustness evaluation results obtained with ImageNet-C and VCR test sets. Consider results for Gaussian Noise in Fig. 5. NoisyMix and NoisyMix_new have almost the same robust accuracy on ImageNet-C, but NoisyMix_new has higher $\hat{\mathcal{R}}_a$; similarly, SIN has higher ImageNet-C robust



(a) ImageNet-C. (b) VCR Test Set.

Figure 4. Histograms showing Δ_v distribution between ImageNet-C and our VCR test sets for Gaussian Blur.

Corruption	Coverage	
	ImageNet-C	VCR Test Set
Brightness	0.590	1.000
Gaussian Blur	0.564	0.974
Defocus Blur	0.538	0.923
Shot Noise	0.462	0.590
Frost	0.436	1.000
Gaussian Noise	0.436	0.872
Impulse Noise	0.385	0.641
Motion Blur	0.333	0.974
Glass Blur	0.333	0.949

Table 2. Δ_v Coverage Comparison with ImageNet-C.

accuracy but lower $\hat{\mathcal{R}}_a$ than SIN_IN_IN due to the almost complete lack of coverage for $\Delta_v < 0.5$ for Gaussian Noise in IMAGENET-C (see Tab. 2), which can lead to biased evaluation results (i.e., biased towards $\Delta_v \geq 0.5$). Checking VCR allows us to detect such biases.

In addition to accuracy, VCR can also check whether the NN can preserve its predictions after corruption, i.e., the prediction consistency property p_v, giving additional information about NN robustness. Fig. 5b and Fig. 5c show that the model TIAN_DEIT_B has a higher $\hat{\mathcal{R}}_a$ than SINGH-CONVNEXT-L-CONVSTEM but a lower $\hat{\mathcal{R}}_p$. This suggests that even though TIAN_DEIT_B has better accuracy for corrupted images, it labels the same image with different labels before and after the corruption. Since ground truth can be hard to obtain during deployment, having low prediction consistency indicates issues with model stability and could raise concerns about when to trust the model prediction. Results for the remaining corruptions are in App. E.

Summary: Robustness must be tested before deploying NNs into an environment with a wide and continuous range of visual corruptions. Our results confirmed that **testing robustness in this range using a fixed and pre-selected number of parameter values can lead to undetected robustness issues**, which can be avoided by checking VCR. Also, **accuracy cannot accurately represent model stability when facing corruptions**, which can be addressed by testing $\mathcal{R}_p$.

4.2. VCR of DNNs Compared with Humans

In this experiment, we use *HMRI* and *MRSI* metrics and the data from the human experiment data to compare VCR of the studied models against human performance.

For Gaussian Noise, Fig. 6 shows our measured *HMRI* and *MRSI* values for $\mathcal{R}_a$ and $\mathcal{R}_p$, where higher values indicate better robustness. Fig. 6a reveals that no NN achieves 1.0 for $HMRI_a$, and in Fig. 6d, only 3 out of 21 NNs DINOV2_GIANT, TIAN_DEIT-B and SINGH-CONVNEXT-L-CONVSTEM reached 1.0 for $HMRI_p$, indicating that there are still unclosed gaps between human and NN robustness. These three top-performing models have also the highest *HMRI* values for both $\mathcal{R}_a$ and $\mathcal{R}_p$, making them closest to human robustness. In Fig. 6b, these three models have $MRSI_a$ values above 0.0, indicating that they surpass human accuracy in certain ranges of visual corruption. This can be visualized by checking the estimated curves s_a as shown in Fig. 6c. The top-three models exceed human accuracy (the red curve) when $\Delta_v > 0.85$. For prediction consistency, Fig. 6e shows that all NNs have the $MRSI_p$ value above 0.0 and this is because, as shown in Fig. 6f, all NN curves are above the human curve when the Δ_v value is small.

Similarly, for Uniform Noise, as shown in Fig. 7a and Fig. 7d, no models reached 1.0 for $HMRI_a$ and the top-three models, reached 1.0 for $HMRI_p$. Together with Fig. 7b and Fig. 7e, we can see that for both $\mathcal{R}_a$ and $\mathcal{R}_p$, TIAN_DEIT-B has higher *HMRI* values but TIAN_DEIT-S has higher *MRSI* values. This suggests that while TIAN_DEIT-B is closer to human performance, TIAN_DEIT-S exceeds human performance more. This counter-intuitive result can be explained with the curves s_a and s_p representing how the performance w.r.t. the robustness properties a and p decreases as Δ_v increases, as shown in Fig. 7c and Fig. 7f. Based on s_a and s_p, TIAN_DEIT-B performs better than TIAN_DEIT-S when $\Delta_v < 0.8$, resulting in a higher *HMRI*. However it performs worse and drops more rapidly when $\Delta_v > 0.8$, leading to a lower *MRSI*. This suggests that both *HMRI* and *MRSI* are useful for comparing NN robustness, and our curves s_a and s_p can provide further information on NN robustness with different degrees of visual corruption.

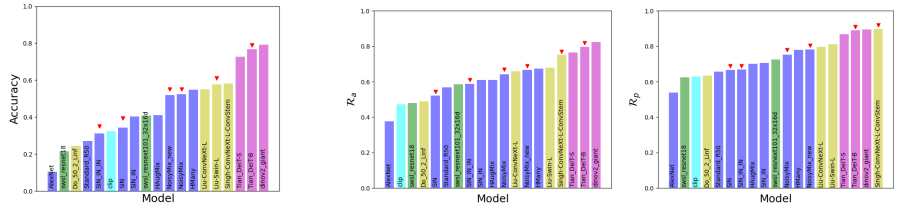
Summary: When considering the full range of visually-continuous corruption, **no NNs can match human accuracy, especially for blur corruptions, though some can match human prediction consistency. Few NNs can outperform humans in specific degrees of corruption.** This highlights a more substantial gap between human and NN robustness than previously identified by [10]. By evaluating VCR using our human-centric metrics, we can better understand the robustness gap and work towards models closer to human performance.

4.3. Training with Data Augmentation

VCR considers a different distribution of corruptions in the images (i.e., continuous) than existing benchmarks (i.e., selected parameter values), so model performance is expected to improve once the model is fine-tuned on the new distribution. We show a small retraining example to demonstrate the usefulness of our benchmark in improving VCR.

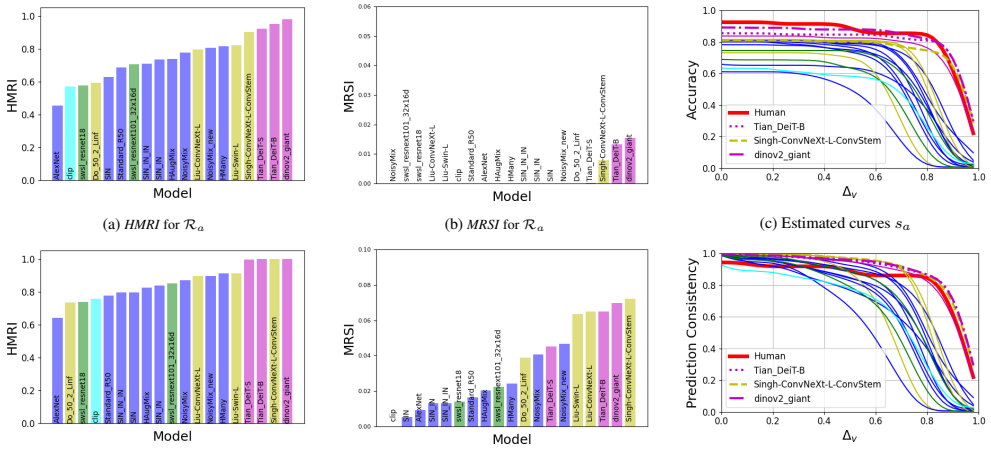
The image classification model was fine-tuned with parameters optimized using stochastic gradient descent (learning rate=0.001, momentum=0.9) and Cross-Entropy Loss. The training set was generated from a subset sampled from the IMAGENET [39] training set with a size of around 12,000, and typically five epochs were sufficient to see progress. While state-of-the-art NNs are optimized for the corruption functions in IMAGENET-C, for certain corruption functions, such as Motion Blur, Frost and Glass Blur, IMAGENET-C images do not cover a wide range of visual changes, leaving room for robustness improvement, as detailed in Tab. 2. Results for SIN [11] and Standard_R50 [3] are shown in Tab. 3; additional details can be found in the codebase[2].

Summary: Our results indicate that retraining with VCR-generated images improves all metrics of NN model performance compared to human performance, even for models optimized for IMAGENET-C's corruption functions. VCR introduces a new distribution of corruptions that models weren't previously exposed to, highlighting that **the gap between human and NN robustness is larger than benchmarks like IMAGENET-C detect. VCR not only identifies this gap but also helps to bridge it.**



(a) IMAGENET-C Gaussian Noise Accuracy. (b) Gaussian Noise $\hat{\mathcal{R}}_a$. (c) Gaussian Noise $\hat{\mathcal{R}}_p$.

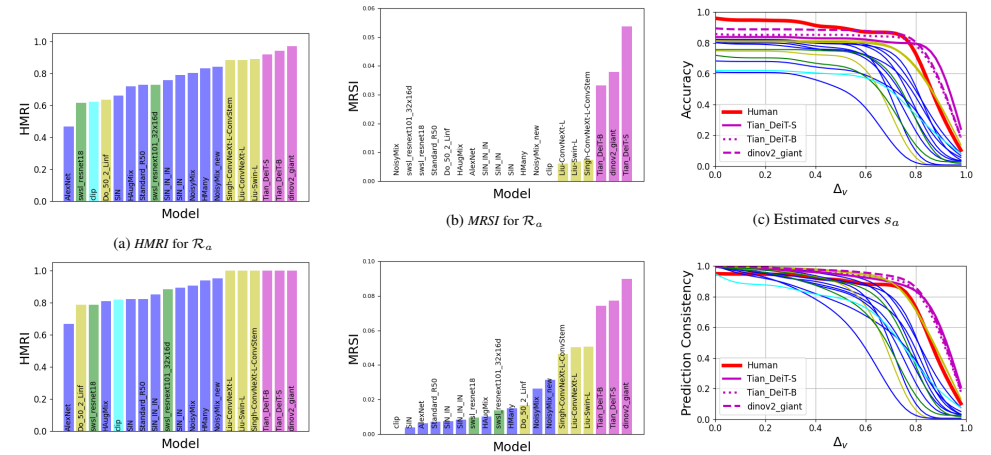
Figure 5. Comparison between IMAGENET-C and VCR with Gaussian Noise. Models discussed in the text are marked by a red triangle.



(a) *HMRI* for $\mathcal{R}_a$ (b) *MRSI* for $\mathcal{R}_a$ (c) Estimated curves s_a

(d) *HMRI* for $\mathcal{R}_p$ (e) *MRSI* for $\mathcal{R}_p$ (f) Estimated curves s_p

Figure 6. VCR evaluation results for Gaussian Noise. Results include, for each NN, the estimated curves s_a and s_p (representing how the performance w.r.t. the robustness properties a and p decreases as Δ_v increases); and the corresponding *HMRI* and *MRSI* values. Results are colored based on their category: Human, Vision Transformer, Supervised Learning, SWSL, Adversarial Training, CLIP.



(a) *HMRI* for $\mathcal{R}_a$ (b) *MRSI* for $\mathcal{R}_a$ (c) Estimated curves s_a

(d) *HMRI* for $\mathcal{R}_p$ (e) *MRSI* for $\mathcal{R}_p$ (f) Estimated curves s_p

Figure 7. VCR evaluation results for Uniform Noise.

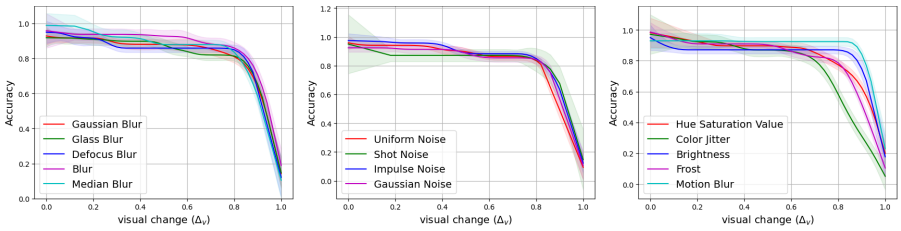
corruption function	Results for Standard_R50 [3]												Results for SIN [11]											
	Before Retraining						After Retraining						Before Retraining						After Retraining					
	Accuracy			Prediction similarity			Accuracy			Prediction similarity			Accuracy			Prediction similarity			Accuracy			Prediction similarity		
	$\hat{\mathcal{R}}_a$	*HMRI*	*MRSI*	$\hat{\mathcal{R}}_p$	*HMRI*	*MRSI*	$\hat{\mathcal{R}}_a$	*HMRI*	*MRSI*	$\hat{\mathcal{R}}_p$	*HMRI*	*MRSI*	$\hat{\mathcal{R}}_a$	*HMRI*	*MRSI*	$\hat{\mathcal{R}}_p$	*HMRI*	*MRSI*	$\hat{\mathcal{R}}_a$	*HMRI*	*MRSI*	$\hat{\mathcal{R}}_p$	*HMRI*	*MRSI*
Median Blur	0.532	0.635	0.000	0.573	0.673	0.000	**0.694**	**0.828**	**0.003**	**0.728**	**0.854**	**0.001**	0.522	0.624	0.00	0.605	0.710	0.00	**0.650**	**0.774**	**0.004**	**0.729**	**0.852**	**0.004**
Frost	0.429	0.521	0.011	0.473	0.572	0.012	**0.575**	**0.690**	**0.025**	**0.678**	**0.804**	**0.031**	0.423	0.512	0.015	0.513	0.618	0.016	**0.517**	**0.625**	**0.016**	**0.647**	**0.768**	**0.031**
Glass Blur	0.468	0.569	0.003	0.502	0.603	0.003	**0.647**	**0.770**	**0.024**	**0.744**	**0.866**	**0.034**	0.334	0.407	0.000	0.397	0.478	0.000	**0.572**	**0.687**	**0.016**	**0.684**	**0.809**	**0.018**

Note: all numbers are rounded.

Table 3. VCR comparison before and after retraining. Red indicates improvement.

4.4. Visually Similar Corruption Functions

Our experiments revealed the existence of *visually similar* corruption functions, which can potentially reduce experiment costs and enhance our understanding of differences between humans and NNs. Different corruptions affect aspects like color, contrast, and noise, influencing human perception in varied ways [9]. For example, Gaussian and Impulse noise may have similar visual effects, making them hard for an average human to distinguish. We call such functions *visually similar*. We postulate that since visually similar functions,



(a) Blur class s_a^h (b) Noise class s_a^h (c) Dissimilar s_a^h

Figure 8. Comparing human performance spline curves s_a^h for similar and dissimilar corruption functions. For each curve, the coloured region around the curve is the 83% confidence interval used for comparison of similarity. See s_p^h in App. E.

by definition, affect human perception similarly, they would affect human robustness similarly as well. This suggests that human data for one function can be reused for other similar functions, potentially lowering experiment costs.

Since VCR is estimated with the spline curves s_a^h and s_p^h, if the difference among the curves of a set of functions is statistically insignificant, human data (i.e., the spline curves) can be reused among the functions in this set. In Fig. 8, we plot the smoothed spline curves s_a^h and s_p^h obtained for all 14 corruption functions included in our experiments. We can observe that, for all corruption functions shown, human performance decreases slowly for small values of visual degrade (Δ_v), but once Δ_v reaches a turning point, human performance starts decreasing more rapidly. Then, we observe that spline curves obtained for certain blur and noise transformations have similar shapes, while those for dissimilar transformations start decreasing at different turning points with different slopes. More specifically, the differences between two spline curves are statistically insignificant if their 83% confidence intervals overlap [25].

Summary: By checking statistical significance with 83% confidence interval for each corruption function, we empirically observed two classes of visually similar corruptions in our experiments with humans: (1) the noise class: Shot Noise, Impulse Noise, Gaussian Noise, and Uniform Noise; and (2) the blur class: Blur, Median Blur, Gaussian Blur, Glass blur, Defocus Blur. The remaining corruptions are dissimilar (see Fig. 8).

NN Robustness for Visually Similar Corruption Functions. Due to fundamental differences between humans and NNs, such as computational power, NNs may respond differently to visually similar corruptions. VCR allows us to empirically analyze these differences. For instance, during deployment, NNs may encounter noise with unknown distributions (e.g., Uniform, Gaussian, Poisson) that do not affect humans as shown in Fig. 8, but could pose safety concerns if NNs are particularly sensitive to specific distributions.

For example, Gaussian Noise and Uniform Noise (visually similar) both add noise to images but from different distributions. Our results in Fig. 6 and Fig. 7 suggest that the NNs detect the distribution difference. Models generally exhibit higher *HMRI* and *MRSI* values for Uniform Noise compared to Gaussian Noise. While performance differences are not statistically significant with low corruption ($\Delta_v < 0.8$), models perform better with Uniform Noise than Gaussian Noise at higher corruption levels (Δ_v between $[0.8..1.0]$). Studying VCR helps analyze how different noise distributions impact NN performance, an impractical task with human data for all possible distributions. Identifying visually similar corruption functions and reusing human data can significantly reduce experimental costs.

Identifying Visually Similar Transformations. We propose a simple method for identifying classes of visually similar corruptions by determining if humans can distinguish between them. Participants view corrupted images and indicate if they believe the corruptions are the same or different. By measuring accuracy across multiple trials, we use a binomial test to assess statistical significance. Our method can detect visually similar transformations quickly, reducing experiment time from about 5.55 hours with 2,000 images and five participants to just 5 minutes.

Limitation: Our method's results depend on participants having normal eyesight and basic knowledge of image corruptions, and may not always accurately identify visually similar transformations. For instance, transformations with different visual effects but similar impacts on human robustness might not be detected. Despite these limitations, we hope our approach encourages further research into how NNs and humans respond differently to corruptions.

5. Conclusion

In this paper, we introduce *visually-continuous corruption robustness* (VCR) and two novel human-aware metrics for evaluating NNs. Our findings reveal **a larger robustness gap between humans and NNs than previously detected**, particularly for blur corruptions. We emphasize that using a comprehensive range of visual changes is crucial for accurate robustness estimation, as **insufficient coverage can lead to biased results**. We also identify classes of image corruptions that similarly affect human perception, which can reduce the cost of measuring human robustness and assessing gaps with computational models. While our study focused on object recognition, human and machine vision comparisons could extend to other aspects like neural data [27, 50], contrasting Gestalt effects [24], object similarity judgments [13], or mid-level properties [45]. We hope our results inspire further robustness research and offer our benchmark datasets and code as open source.

References

[1] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2), 2020. Licensed with MIT License. To view a copy of this license see https://github.com/albumentations-team/albumentations/blob/master/LICENSE. 4

[2] Prithvijit Chattopadhyay, Judy Hoffman, Roozbeh Mottaghi, and Aniruddha Kembhavi. RobustNav: Towards Benchmarking Robustness in Embodied Navigation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 15671–15680. IEEE, 2021. 2

[3] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: A Standardized Adversarial Robustness Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. Licensed with MIT license. To view a copy of this license see https://github.com/RobustBench/robustbench/blob/master/LICENSE. 6, 7

[4] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. 12

[5] N. Benjamin Erichson, Soon Hoe Lim, Winnie Xu, Francisco Utrera, Ziang Cao, and Michael W. Mahoney. NoisyMix: Boosting Model Robustness to Common Corruptions, 2022. 5

[6] Chaz Firestone. Performance vs. Competence in Human–Machine Comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, 2020. 2, 4

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision Meets Robotics: The KITTI Dataset. *Int. J. of Robotics Research (IJRR)*, 2013. 13

[8] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, 2020. 2

[9] R Geirhos, CR Medina Temme, J Rauber, HH Schütt, M Bethge, and FA Wichmann. Generalisation in Humans and Deep Neural Networks. In *NeurIPS 2018*, pages 7549–7561. Curran, 2019. 1, 2, 4, 5, 7

[10] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 23885–23899, 2021. 1, 2, 6

[11] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. 2, 5, 6, 7

[12] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[13] Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185, 2020. 8

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8320–8329. IEEE, 2021. 2, 5

[15] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. Licensed with Apache-2.0 license. To view a copy of this license see https://github.com/hendrycks/robustness/blob/master/LICENSE. 1, 2, 3, 4

[16] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 1

[17] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020. 2, 5

[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15262–15271. Computer Vision Foundation / IEEE, 2021. 1

[19] T. Ho-Phuoc. CIFAR10 to Compare Visual Recognition Performance between Deep Neural Networks and Humans. *ArXiv*, abs/1811.07270, 2018. 2

[20] Boyue Caroline Hu, Lina Marsso, Krzysztof Czarnecki, Rick Salay, Huakun Shen, and Marsha Chechik. If a Human Can See It, So Should Your System: Reliability Requirements for Machine Vision Components. In *Proceedings of the 44th International Conference on Software Engineering (ICSE'2022), Pittsburgh, USA*. ACM, 2022. 1, 2, 3, 4, 12

[21] Christoph Kamann and Carsten Rother. Benchmarking the Robustness of Semantic Segmentation Models with Respect

to Common Corruptions. *Int. J. Comput. Vis.*, 129(2):462–483, 2021. 2

[22] Oguzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3D Common Corruptions and Data Augmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18941–18952. IEEE, 2022. 2

[23] Saeed Reza Kheradpisheh, Masoud Ghodrati, Mohammad Ganjtabesh, and Timothée Masquelier. Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. *Scientific reports*, 6(1):1–24, 2016. 2

[24] B Kim, E Reif, M Wattenberg, S Bengio, and MC Mozer. Neural networks trained on natural scenes exhibit gestalt closure. arxiv. *arXiv preprint arXiv:1903.01069*, 2019. 8

[25] Roger Koenker, Pin Ng, and Stephen Portnoy. Quantile Smoothing Splines. *Biometrika*, 81(4):673–680, 1994. 3, 8, 15

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012. 5

[27] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019. 8

[28] Abhinau Kumar. Python 3 implementation of the visual information fidelity (vif) image quality assessment (iqa) metric. https://github.com/abhinaukumar/vif, 2020. Licensed with MIT license. To view a copy of this license see https://github.com/abhinaukumar/vif/blob/main/LICENSE. 2

[29] Chang Liu, Yinpeng Dong, Wenzhao Xiang, Xiao Yang, Hang Su, Jun Zhu, Yuefeng Chen, Yuan He, Hui Xue, and Shibao Zheng. A comprehensive study on robustness of image classification models: Benchmarking and rethinking, 2023. 5

[30] Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D. Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *CoRR*, abs/1906.02611, 2019. 2

[31] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 2

[32] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. *arXiv preprint arXiv:1907.07484*, 2019. 2

[33] Eric Mintun, Alexander Kirillov, and Saining Xie. On Interaction Between Augmentations and Corruptions in Natural Corruption Robustness. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 3571–3583, 2021. 2

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 4, 5

[35] Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari. Training Object Class Detectors with Click Supervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 180–189. IEEE Computer Society, 2017. 4

[36] Ori Press, Steffen Schneider, Matthias Kümmerer, and Matthias Bethge. Rdumb: A simple approach that questions our progress in continual test-time adaptation, 2023. 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 5

[38] E. Rusak, Lukas Schott, Roland S. Zimmermann, Julian Bitterwolf, O. Bringmann, M. Bethge, and Wieland Brendel. Increasing the Robustness of DNNs Against Image Corruptions by Playing the Game of Noise. *ArXiv*, abs/2001.06057, 2020. 2

[39] Olga Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 4, 6, 14

[40] 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 5

[41] H. R. Sheikh and A. C. Bovik. Image Information and Visual Quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006. 1, 2, 12

[42] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006. 12

[43] Naman D Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models, 2023. 5

[44] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. Computer: Benchmarking Machine Learning Algorithms for Traffic Sign Recognition. *Neural Networks*, 32:323 – 332, 2012. Selected Papers from IJCNN 2011. 2

[45] Katherine R Storrs, Barton L Anderson, and Roland W Fleming. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, 5(10):1402–1417, 2021. 8

[46] Jiachen Sun, Qingzhao Zhang, Bhavya Kailkhura, Zhiding Yu, Chaowei Xiao, and Z. Morley Mao. Benchmarking Robustness of 3D Point Cloud Recognition Against Common Corruptions. *CoRR*, abs/2201.12296, 2022. 2

[47] Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, and Yu-Gang Jiang. Deeper Insights into the Robustness of ViTs towards Common Corruptions, 2022. 5

[48] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. on Image Processing*, 13(4):600–612, 2004. 12

[49] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *CoRR*, abs/1905.00546, 2019. 5

[50] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of Sciences*, 111(23):8619–8624, 2014. 8

[51] Chenyu Yi, Siyuan Yang, Haoliang Li, Yap-Peng Tan, and Alex C. Kot. Benchmarking the Robustness of Spatial-Temporal Models Against Corruptions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. 2

[52] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A Fourier Perspective on Model Robustness in Computer Vision. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13255–13265, 2019. 2

[53] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6022–6031. IEEE, 2019. 2

[54] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 2

[55] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[56] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 12