This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ImageNet-D: Benchmarking Neural Network Robustness on Diffusion Synthetic Object

Chenshuang Zhang[1] Fei Pan[2] Junmo Kim[1*] In So Kweon[1] Chengzhi Mao[3,4*]
KAIST[1], University of Michigan, Ann Arbor[2], McGill University[3], MILA[4]

Abstract

We establish rigorous benchmarks for visual perception robustness. Synthetic images such as ImageNet-C, ImageNet-9, and Stylized ImageNet provide specific type of evaluation over synthetic corruptions, backgrounds, and textures, yet those robustness benchmarks are restricted in specified variations and have low synthetic quality. In this work, we introduce generative model as a data source for synthesizing hard images that benchmark deep models' robustness. Leveraging diffusion models, we are able to generate images with more diversified backgrounds, textures, and materials than any prior work, where we term this benchmark as ImageNet-D. Experimental results show that ImageNet-D results in a significant accuracy drop to a range of vision models, from the standard ResNet visual classifier to the latest foundation models like CLIP and MiniGPT-4, significantly reducing their accuracy by up to 60%. Our work suggests that diffusion models can be an effective source to test vision models. The code and dataset are available at `https://github.com/chenshuang-zhang/imagenet_d`.

1. Introduction

Neural networks have achieved remarkable performance in tasks ranging from image classification [33, 34, 55] to visual question answering [10, 28, 31, 65]. These advances have inspired the application of neural networks in various fields, including security and safety-critical systems such as self-driving cars [27, 32, 42], malware detection [9, 43, 60] and robots [6, 7, 25]. Due to their wide adaptation, it is becoming increasingly important to identify the robustness of neural networks [30, 41] for safety reasons.

To evaluate the robustness of neural networks, ObjectNet [3] collects real-world object images on controlled factors like background with human workers, which is time-consuming and labor-intensive. To scale up data collection, synthetic images are proposed as test images [15, 19, 58]. For example, ImageNet-C [19] introduces a set of low-level common visual corruptions, such as gaussian noise and blur, to test models' robustness. ImageNet-9 [58] uses simple cutting and paste technique to create robustness benchmark on object background, yet the images are not realistic. Stylized-ImageNet [15] generates new images by altering the textures of ImageNet images, which cannot control the global factors like background.

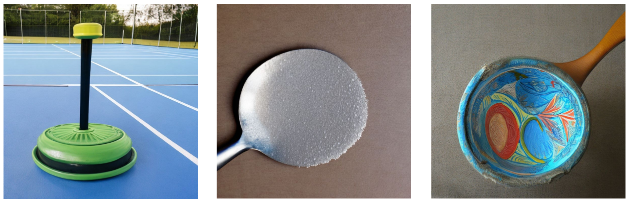


Figure 1. Top predictions from CLIP (ViT-L/14) on ImageNet-D. We synthesize the images by changing their background, texture and material. The groundtruth for the images are *plunger*, *spatula*, and *ladle* in order, together with the background (badminton court), texture (freckled), and material (painted).

In this work, we introduce ImageNet-D, a synthetic test set generated by diffusion models for object recognition task. Capitalizing on the capability of pioneering Stable Diffusion models[48], we show that we can steer diffusion models with language to create realistic test images that cause vision models fail. Figure 1 shows three failure cases of CLIP model on our synthetic ImageNet-D dataset. Since we rely on language to create images, we can vary the high-level factors in the images in contrast to the local corruptions and texture in prior work, introducing addition factors that one can evaluate robustness on.

To enhance sample difficulty of our dataset, we selectively retain images that cause failures in multiple chosen vision models. Our results show that images triggering errors in chosen models can reliably transfer their challenging nature to other, previously untested models. This leads to a notable decrease in accuracy, even in state-of-the-art foundation models like MiniGPT-4 [65] and LLaVa [31], suggesting our dataset reveals common failures in vision models.

Visualizations demonstrate that Imagenet-D signifi-

---

[*] Corresponding author. Junmo Kim <junmo.kim@kaist.ac.kr>, Chengzhi Mao <chengzhi.mao@mila.quebec>.

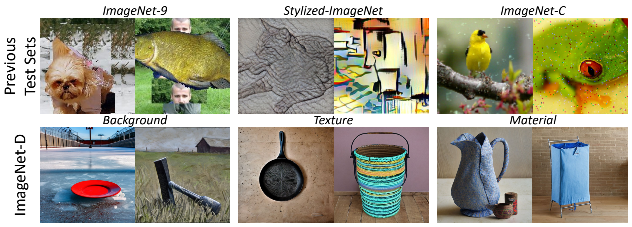


Figure 2. Examples from ImageNet-9 [58], Stylized-ImageNet [15] and ImageNet-C [19] and our ImageNet-D. For the second row, we show images from ImageNet-D with different backgrounds, textures and materials orderly. Take the background for example (the two columns on the left), ImageNet-9 [58] generates new images by simply cutting and paste foreground and background from different images, leading to object deformation and dislocation. By contrast, ImageNet-D includes images with diverse backgrounds by diffusion generation, achieving superior visual fidelity.

cantly enhances image quality compared to previous synthetic robustness benchmarks, as evidenced in Figure 2. Imagenet-D serves as an effective tool for reducing the performance and assessing model robustness, including ResNet 101 (reducing 55.02%), ViT-L/16 (reducing 59.40%), CLIP (reducing 46.05%), and transfer well to unforeseen large vision language models like LLaVa [31] (reducing 29.67%), and MiniGPT-4 [65] (reducing 16.81%). Our approach of utilizing generative models to evaluate model robustness is general, and shows significant potential for even greater effectiveness with future advancements in generative models.

2. Related work

Robustness of neural networks. Neural networks have evolved from CNN [16, 24], ViT [33, 55], to large foundation models [5, 12, 54]. Previous work has investigated neural networks robustness from multiple aspects, such as adversarial examples [35, 36, 38, 63, 64] and out-of-domain samples [17, 20, 22, 37]. Foundation models have shown greater robustness on out-of-distribution samples [45]. Robust explanation has also been investigated [31, 39, 65]. To systematically evaluate the robustness of deep models, test sets that cover different factors are urgently needed.

Dataset for benchmarking robustness. To evaluate neural network robustness, a branch of studies source images online, including ImageNet-A [23], Imagenet-R [22] and ImageNet-Sketch [57]. However, they are limited to images that exist on the web. ObjectNet [3] manually collects images with the help of 5982 workers, which is time-consuming and resource-intensive.

To overcome the limitations of web images and reduce the cost of manual collection, synthetic images are proposed for robustness evaluation [15, 19, 58]. ImageNet-C [19] benchmarks model robustness on low-level corruptions. ImageNet-9 [58] generates new images by combining foreground and background from different images, however, limited by poor image fidelity. Stylized-ImageNet [15] alters the textures of ImageNet images by AdaIN style transfer [26] or introducing texture-shape cue conflict, which cannot control other factors like backgrounds. In this work, we introduce a new test set ImageNet-D, which is generated by controlling diffusion models and includes novel images with diverse backgrounds, textures, and materials.

Image generation. Diffusion models have achieved great success in various tasks including image generation [46, 49, 51, 62]. As a milestone work, Stable diffusion [48] enables high-fidelity image synthesis controlled by language. InstructPix2Pix [8] provides a more sophisticated control by editing a given image according to human instructions. In this paper, we build our pipeline with the standard Stable Diffusion model, yet our algorithm is compatible with other generative models that can be steered by language.

Enhancing perception with diffusion images. Diffusion-generated images have been used for vision perception tasks. A branch of studies [1, 2, 53, 59] improves classification accuracy by using synthetic images as training data augmentation. DREAM-OOD [14] finds the outliers by decoding sampled latent embeddings to images. However, their method lacks specific control over image space, which is crucial for benchmarks like ImageNet-D. [40] identifies under-represented attribute pairs, while our study focuses on hard images with a single attribute. Unlike [29, 44, 56] that modify existing datasets, our work generates new images and mines the most challenging ones as the test set, achieving greater accuracy drop than [29, 44, 56].

3. ImageNet-D

We first present how ImageNet-D is created in Section 3.1, followed by an overview of its statistics in Section 3.2.

3.1. Dataset design

While neural networks excel in various applications, their robustness needs rigorous evaluation for safety. Traditional evaluations use existing test sets, including either natural [3, 23] or synthetic images [15, 19, 58]. Compared to manual image collection, collecting a synthetic test set is more efficient [15, 58]. However, the diversity of current synthetic test sets is limited due to their reliance on existing images for attribute extraction. These synthetic images are also not realistic, as shown in Figure 2. ImageNet-D is introduced to evaluate model robustness across various object and nuisance combinations, addressing these limitations.

Image generation by diffusion models. To construct ImageNet-D, diffusion models are used to create a vast image pool by combining all possible object and nuisances,

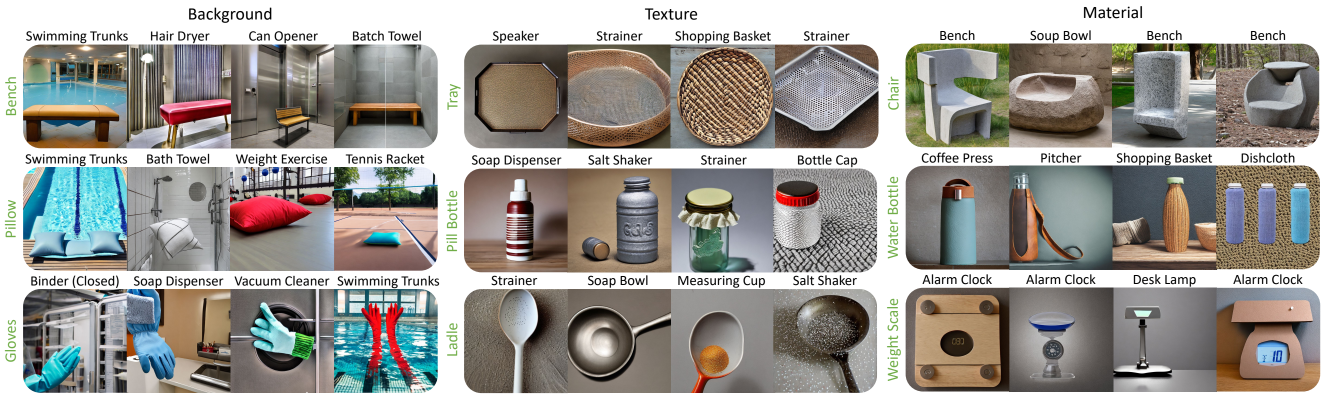


Figure 3. The ImageNet-D test set. Each group of images is generated with the same object and nuisance, such as background, texture, and material. For each group of images, the ground truth label is color green, while the predicted categories by CLIP (ViT-L/14) on each image are in black. Leveraging diffusion models for image generation, we can create a test set with diverse combinations of objects and nuisances. For example, the top left corner shows a bench in the swimming pool background. Interestingly, CLIP (ViT-L/14) recognizes the bench in this image as swimming trunks.

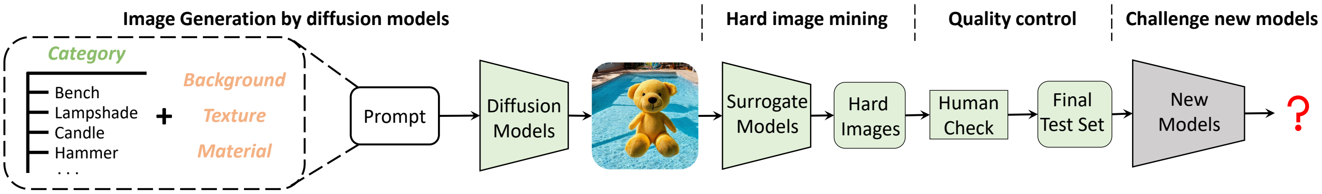


Figure 4. ImageNet-D creation framework. ImageNet-D is created by first combining various object categories and nuisances, including background, texture, and material. To make the test set challenging, we only keep the hard images from the large pool that commonly make multiple surrogate models fail to predict the correct object label. The test set is then refined through human verification to ensure the images are valid, single-class, and high-quality, making ImageNet-D suitable for evaluating the robustness of different neural networks.

enabling high-fidelity image generation based on user-defined text inputs. We use Stable Diffusion model[48] for image generation, while our approach is compatible with other generative models that can be steered by language. The image generation process is formulated as follows:

$$\text{Image}(C, N) = \text{Stable Diffusion}(\text{Prompt}(C, N)), \quad (1)$$

where C and N refer to the object category and nuisance, respectively. The nuisance N includes background, material, and texture in this work. Table 1 presents an overview of nuisances and prompts to diffusion models. Using the backpack category as an example, we first generate images of backpacks with various backgrounds, materials, and textures (e.g., a backpack in a wheat field), offering a broader range of combinations than existing test sets. Each image is labeled with its prompt category C as ground truth for classification. An image is viewed misclassified if the model's predicted label does not match the ground truth C.

After creating a large image pool with all object category and nuisance pairs, we evaluate CLIP (ViT-L/14) model on these images in Table 2. Experimental details are reported in Section 4.1. Table 2 shows that CLIP achieves high accuracy on all the test sets, with an accuracy of around 94% on synthetic image pool. To create a challenging test set for robustness evaluation, we propose an efficient strategy to find the hard test samples from all generated images as follows.

Hard image mining with shared perception failures. Before introducing how to identify hard samples from synthetic image pool, we first define the concept of shared perception failure as follows.

`Shared failure:` An image is a shared failure if it leads multiple models to predict object's label incorrectly.

An ideal hard test set should include images that all tested models fail to identify. However, this is impractical due to the inaccessibility of future test models, termed target model. Instead, we construct the test set using shared failures of known surrogate models. If the failures of surrogate models lead to low accuracy in unknown models, the test set is deemed challenging. This is defined as transferable failure as follows:

`Transferable failure:` Shared failures of known surrogate models are transferable if they also result in low accuracy for unknown target models.

To test if shared failures of surrogate models are trans-

Table 1. Overview of nuisances and input prompts to diffusion models. During ImageNet-D construction, 468 backgrounds, 47 textures, and 32 materials from the Broden dataset [4] are used as nuisances. Images are generated by pairing each object with all nuisances in diffusion model prompts. This approach allows for efficient scaling of ImageNet-D with additional categories and nuisances.

Nuisance	Nuisance number	Prompt to diffusion models	Prompt example
Background	468	A [`category`] in the [`background`]	A `backpack` in the `wheat field`
Texture	47	A [`texture`] [`category`]	A `knitted backpack`
Material	32	A [`category`] made of [`material`]	A `backpack` made of `leather`

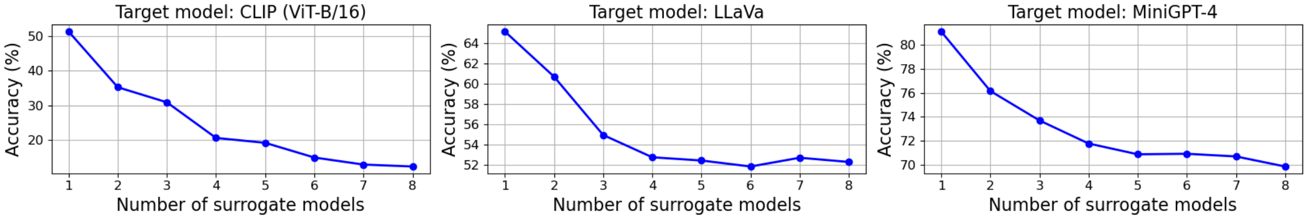


Figure 5. Test accuracy of target models on shared failures of surrogate models. We adopt known surrogate models to identify their shared failure images as the test set, then evaluate a new target model on this test set. We show that the shared failures of surrogate models can also deceive a new target model, leading to low test accuracy. Increasing the number of surrogate models lowers the target models' test accuracy, suggesting a more difficult test set.

Table 2. Test accuracy of CLIP (ViT-L/14) on the synthetic image pool by **exhausting all** the object category and nuisance combinations. We show that CLIP achieves high accuracy on the synthetic image pool. To create a challenging test set for robustness evaluation, we further mine the hard samples as the final test set.

Test Set	ImageNet	ObjectNet	Synthetic image pool		
			Background	Texture	Material
Acc (%)	74.64	66.91	95.79	94.02	93.75

ferable for diffusion-generated images, we evaluate test sets created with shared failures from 1 to 8 surrogate models in Figure 5. We evaluate the accuracy of three target models that are not used during test set construction, including CLIP(ViT-B/16), LLaVa, and MiniGPT-4. Figure 5 shows that target model accuracy decreases as more surrogate models are used. The test sets are created with diverse backgrounds, while experiments for texture and material show the same trend. This trend demonstrates that failure images of multiple surrogate models can form a challenging test set for unseen new models. Notably, the accuracy decrease slows when the number of surrogate models exceeds four.

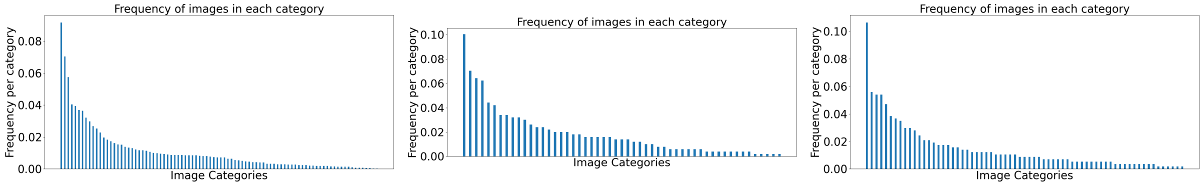
Quality control by human-in-the-loop. The above process allows us to automatically find a challenging test set to unseen models. However, generative models can produce incorrect images not matching the prompt category. We resort to human annotation to ensure the ImageNet-D images are simultaneously valid, single-class, and high-quality. After first-round annotation by graduate students, we use Amazon Mechanical Turk [11, 22, 47] to evaluate labeling quality. We ask the workers to select the images that they can either recognize the main object or the main object can be used functionally as the ground truth category. Moreover, we design sentinels to ensure high-quality responses, including positive, negative and consistent sentinels. We report details of the labeling task in the appendix. A total of 679 qualified workers participated in 1540 labeling tasks, achieving an agreement of 91.09%. Figure 3 displays images from ImageNet-D, demonstrating high fidelity and diversity in object and nuisance pairs. We summarize the framework of creating ImageNet-D in Figure 4.

3.2. Dataset statistics

ImageNet-D includes 113 overlapping categories between ImageNet and ObjectNet, and 547 nuisances candidates from the Broden dataset [4](see Table 1), resulting in 4835 hard images featuring diverse backgrounds (3,764), textures (498), and materials (573). Our pipeline to create ImageNet-D is general and efficient, allowing easy addition of new categories and nuisances. ImageNet-D's category distribution exhibits a natural long-tail pattern, as shown in Figure 6. The sparse and non-uniform category-attribute distribution in Figure 7 shows the necessity of exhausting all category and nuisance pairs in test set creation.

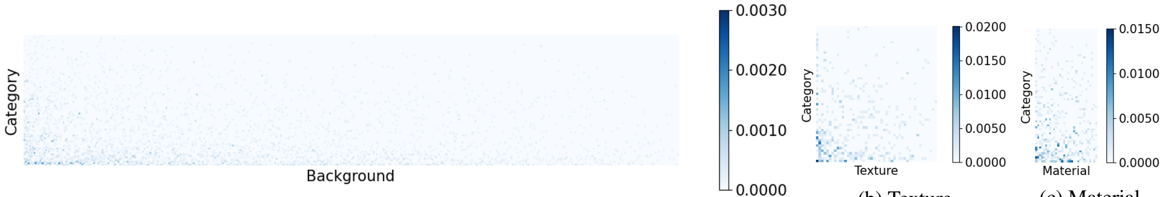
4. Experiments

We evaluate various models on ImageNet-D benchmark. We find that ImageNet-D significantly decreases all models' accuracy by up to 60%. We then show whether prior



(a) Background (b) Texture (c) Material

Figure 6. Histogram of the image frequency per category in our test set, following a natural long-tail distribution.



(a) Background (b) Texture (c) Material

Figure 7. Frequency of object category and nuisance pairs. Each dot indicates a pair, while the x-axis and y-axis indicate the nuisance and category, respectively. As shown in Table 1, we adopt 468 backgrounds, 47 textures, and 32 materials from the Broden dataset [4], leading to different widths of Figure (a) (b) (c). A darker color indicates more samples. The sparse and non-uniform distribution shows the necessity of exhausting all category and nuisance pairs in ImageNet-D creation.

advancements could improve ImageNet-D robustness, such as data augmentation. Lastly, we discuss ImageNet-D from various aspects, such as nearest neighbor retrieval.

4.1. Experimental setups

Test set construction setups. We use Stable Diffusion [48] to create ImageNet-D, and adopt the pretrained weight of version stable-diffusion-2-1 from Hugging Face. To find the hard images, we finalize ImageNet-D with shared failures of 4 surrogate models, including CLIP [45] (ViT-L/14, ViT-L/14-336px and ResNet50), and vision model (ResNet50 [16]). The candidate set of surrogate models in Figure 5 also includes CLIP (Resnet101,ViT-B/32) and vision model (ViT-L/16 [13] and VGG16 [52]).

Evaluation of classification models. Robustness on ImageNet-D is measured by top-1 accuracy in object recognition, the ratio of correctly classified images to total images. We evaluate classification models with the open-source pretrained weights. For CLIP [45], we follow the original paper [45] to adopt *A photo of a* [*category*] as the text template. The zero-shot accuracy of CLIP is reported.

Evaluation of visual question answering (VQA) models. We evaluate the accuracy of the state-of-the-art open-source VQA models on ImageNet-D, including LLaVa [31], and MiniGPT-4 [65]. Given an input image, VQA models output answers based on input text prompt. However, the textual output of VQA models is not limited to a certain template, thus may not include the category name in pre-defined category list of object recognition tasks. This makes it hard to assess the accuracy based on diverse answers.

A common prompt that asks VQA models to recognize the object is: `What is the main object in this image?` To make VQA models choose from pre-defined category list, we ask VQA models as follows: `What is the main object in this image? Choose from the following list: [GT category], [failure category].` GT category refers to the image's ground truth category. As for the failure category, we adopt the category that achieves the highest CLIP (ViT-L/14) confidence among all wrong categories. With this prompt, we find that both MiniGPT-4 and LLaVa can choose from provided category list in their output. If the model chooses ground truth category, this image is viewed to be correctly recognized. Therefore, we can compute the accuracy of VQA models.

4.2. Robustness evaluation

Quantitative results. We evaluate ImageNet-D on 25 models, and plot test accuracy trend in Figure 8. The horizontal axis and vertical axis indicate the test accuracy on ImageNet and ImageNet-D, respectively. Figure 8 shows that as ImageNet accuracy increases, ImageNet-D accuracy also gets higher. ImageNet-D accuracy is much lower than ImageNet accuracy for all models, indicated by the lower distribution below the $y = x$ reference line. We report the accuracy of 14 models on different test sets in Table 3, and all models' accuracy in appendix. Table 3 shows that ImageNet-D achieves the lowest test accuracy for all models, except for the comparable result on Stylized-ImageNet for VQA models. Note that ImageNet-D achieves higher image fi-

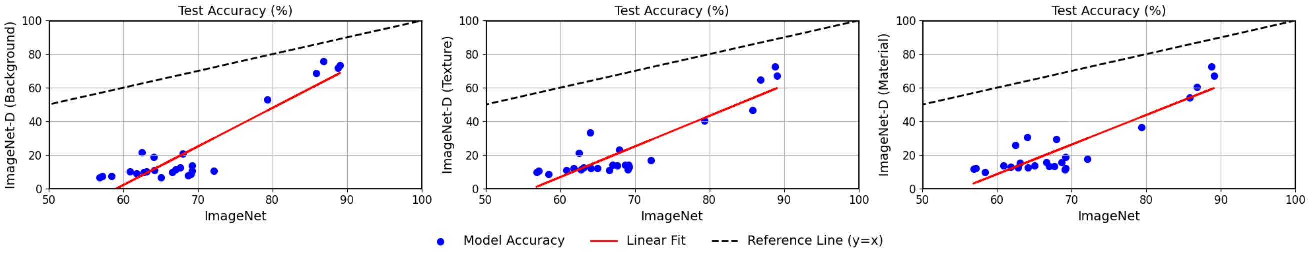


Figure 8. Model accuracy on ImageNet vs. ImageNet-D. Each data point corresponds to one tested model. The plots reveal that there is a significant accuracy drop from ImageNet to our new test set ImageNet-D. As the model's accuracy on ImageNet increases, the accuracy on ImageNet-D is also higher. These results show the effectiveness of ImageNet-D to evaluate the robustness of neural networks. We report the exact number of 14 models of this figure in Table 3, and the results for all models can be found in the appendix.

Table 3. Test accuracy of vision models and large foundation models (%). We show the test accuracy for the vision models and large foundation models (rows) on different test sets (columns). The numbers in green refer to the accuracy drop of ImageNet-D compared to ImageNet. For MiniGPT-4 and LLaVa, ImageNet-D reduces the accuracy by 16.81% and 29.67% compared to the ImageNet, respectively. Our results show that ImageNet-D is effective to evaluate the robustness of neural networks.

Model	Architecture	ImageNet	ObjectNet	ImageNet-9	Stylized	ImageNet-D Background	ImageNet-D Texture	ImageNet-D Material	ImageNet-D Total
Vision model (CNN)	VGG19	62.77	27.19	74.84	16.25	9.8(-52.97)	11.45(-51.32)	12.39(-50.38)	10.28(-52.49)
	ResNet101	67.66	32.34	81.85	22.66	12.38(-55.28)	13.65(-54.01)	13.44(-54.22)	12.64(-55.02)
	Densenet161	66.99	31.86	84.91	22.5	11.34(-55.65)	14.06(-52.93)	13.26(-53.73)	11.85(-55.14)
	Wideresnet101	69.2	34.37	82.17	21.48	10.55(-58.65)	13.05(-56.15)	12.04(-57.16)	10.98(-58.22)
Vision model (ViT)	ViT-B/32	65.02	27.59	77.51	42.34	6.64(-58.38)	12.25(-52.77)	13.79(-51.23)	8.07(-56.95)
	ViT-B/16	72.14	34.79	82.49	31.02	10.49(-61.65)	16.87(-55.27)	17.63(-54.51)	12.0(-60.14)
	ViT-L/16	68.67	32.7	78.91	29.38	7.68(-60.99)	14.06(-54.61)	15.53(-53.14)	9.27(-59.40)
CLIP	RN101	62.48	42.89	83.09	22.58	21.47(-41.01)	21.29(-41.19)	25.83(-36.65)	21.96(-40.52)
	ViT-B/32	64.06	43.67	79.56	44.22	18.73(-45.33)	33.33(-30.73)	30.37(-33.69)	21.61(-42.45)
	ViT-B/16	67.95	54.87	85.16	40.62	20.64(-47.31)	22.89(-45.06)	29.32(-38.63)	21.9(-46.05)
MiniGPT-4	Vicuna 13B	88.77	77.57	89.46	69.88	71.81(-16.96)	72.48(-16.29)	72.5(-16.27)	71.96(-16.81)
LLaVa	Vicuna 13B	79.32	76.02	90.84	61.94	52.89(-26.43)	40.53(-38.79)	36.28(-43.04)	49.65(-29.67)
LLaVa-1.5	Vicuna 13B	89.08	78.66	93.88	64.14	73.31(-15.77)	67.27(-21.81)	67.08(-22.00)	71.95(-17.13)
LLaVa-NeXT	Hermes-Yi-34B	85.83	77.54	90.52	57.98	68.77(-17.06)	46.67(-39.16)	54.11(-31.72)	64.76(-21.07)

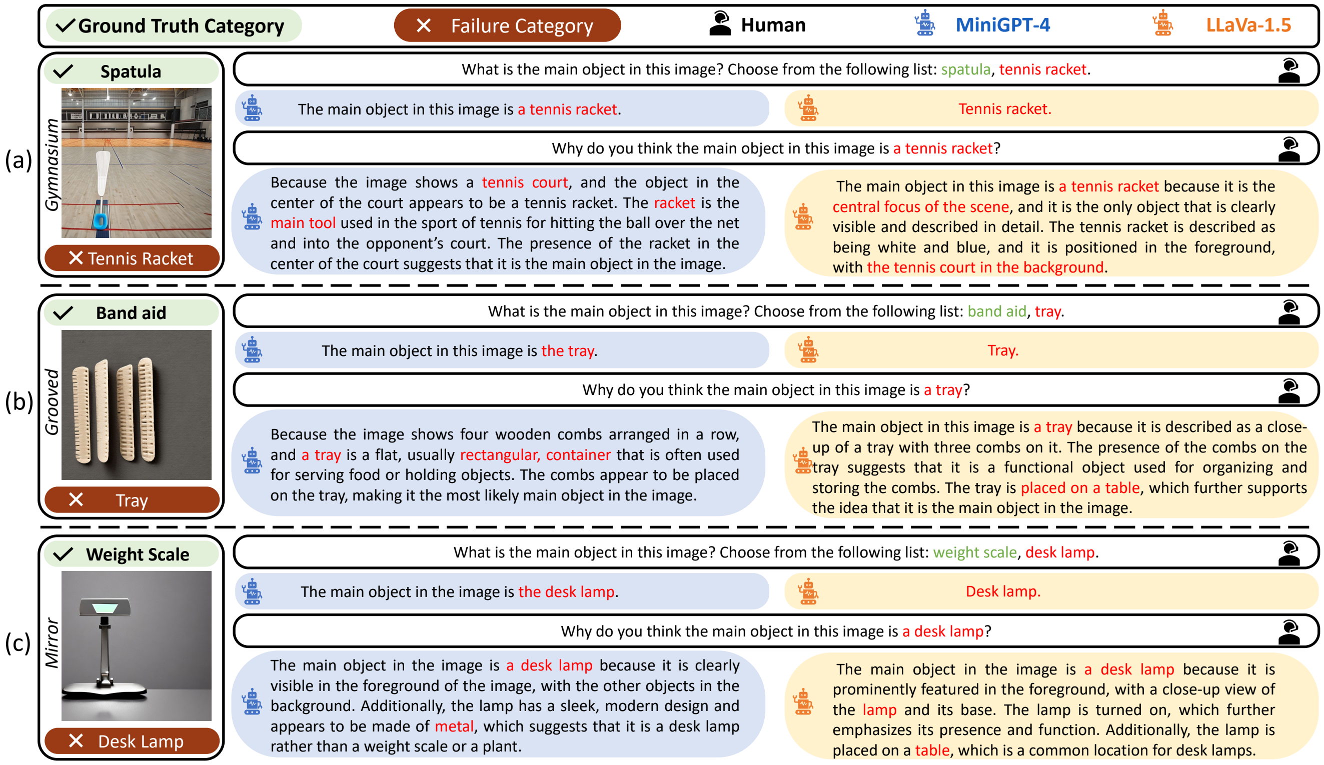
delity than Stylized-ImageNet as shown in Figure 2. Although ObjectNet changes multiple attributes for each image, it still results in higher accuracy than ImageNet-D that specifies only one attribute per image. Compared to ImageNet, ImageNet-D yields a test accuracy drop of more than 16% for all models, including LLaVa (reducing 29.67%) and MiniGPT-4 (reducing 16.81%). Our ImageNet-D can also cause significant accuracy drop of the latest LLaVa-1.5 and LLaVa-NeXT. Although LLaVa-NeXT outperforms LLaVa-1.5 on benchmarks like MMMU [61], it achieves lower accuracy on ImageNet-D, demonstrating the uniqueness of ImageNet-D. For vision models, the accuracy drop is even close to 50% to 60%. The results in Figure 8 and Table 3 show the effectiveness of ImageNet-D in evaluating the robustness of neural networks.

Visualization results. Figure 3 displays image examples from ImageNet-D, demonstrating high quality. Although humans can easily recognize the main object, CLIP (ViT-L/14) mistakenly classifies these images into a wrong category. Figure 9 shows conversations with MiniGPT-4 and LLaVa-1.5 on ImageNet-D images, indicating that MiniGPT-4 and LLaVa-1.5 can also fail to recognize the main object from ImageNet-D.

4.3. Robustness improvement

Data augmentation. Prior studies reveal that data augmentation is effective to improve model robustness, such as on ImageNet-C [19]. We evaluate data augmentation methods on ImageNet-D, including SIN [15], AugMix [21], ANT [50] and DeepAugment [22]. Table 4 shows the results with ResNet50 backbone for all methods. We report test accuracy for ImageNet and ImageNet-D, mean Corruption Error (mCE) for ImageNet-C following [19, 21, 22, 50]. Although these methods improve ImageNet-C robustness (lower mCE) than baseline model, Table 4 shows that their ImageNet-D robustness is comparable or even worse than baseline. This suggests that current benchmarks like ImageNet-C do not accurately represent real robustness of neural networks in our setting, making ImageNet-D a necessary benchmark for robustness evaluation.

Model architecture. We compare ImageNet-D robustness of different model architectures in Figure 11. When



What is the main object in this image? Choose from the following list: spatula, tennis racket.

The main object in this image is a tennis racket.

Tennis racket.

Why do you think the main object in this image is a tennis racket?

Because the image shows a tennis court, and the object in the center of the court appears to be a tennis racket. The racket is the main tool used in the sport of tennis for hitting the ball over the net and into the opponent's court. The presence of the racket in the center of the court suggests that it is the main object in the image.

The main object in this image is a tennis racket because it is the central focus of the scene, and it is the only object that is clearly visible and described in detail. The tennis racket is described as being white and blue, and it is positioned in the foreground, with the tennis court in the background.

What is the main object in this image? Choose from the following list: band aid, tray.

The main object in this image is the tray.

Tray.

Why do you think the main object in this image is a tray?

Because the image shows four wooden combs arranged in a row, and a tray is a flat, usually rectangular, container that is often used for serving food or holding objects. The combs appear to be placed on the tray, making it the most likely main object in the image.

The main object in this image is a tray because it is described as a close-up of a tray with three combs on it. The presence of the combs on the tray suggests that it is a functional object used for organizing and storing the combs. The tray is placed on a table, which further supports the idea that it is the main object in the image.

What is the main object in this image? Choose from the following list: weight scale, desk lamp.

The main object in the image is the desk lamp.

Desk lamp.

Why do you think the main object in this image is a desk lamp?

The main object in the image is a desk lamp because it is clearly visible in the foreground of the image, with the other objects in the background. Additionally, the lamp has a sleek, modern design and appears to be made of metal, which suggests that it is a desk lamp rather than a weight scale or a plant.

The main object in the image is a desk lamp because it is prominently featured in the foreground, with a close-up view of the lamp and its base. The lamp is turned on, which further emphasizes its presence and function. Additionally, the lamp is placed on a table, which is a common location for desk lamps.

Figure 9. Failure cases of MiniGPT-4 and LLaVa-1.5 from ImageNet-D.We ask MiniGPT-4 and LLaVa-1.5 for the category of the input image and the reason for their predictions. Figure (a) to (c) are images with different background, texture and material, respectively. Our results show that images from ImageNet-D can also make the state-of-the-art foundation models fail.

Table 4. Robustness of different augmentation methods. Despite superior ImageNet-C robustness, these methods fail to improve ImageNet-D robustness, demonstrating the uniqueness of ImageNet-D from ImageNet-C.

Model	ImageNet (↑)	ImageNet-C(mCE)(↓)	ImageNet-D(↑)
Baseline	65.82	81.74	10.22
SIN	63.42(-2.40)	78.45(-3.29)	8.81(-1.41)
Augmix	66.88(+1.06)	74.7(-7.04)	8.75(-1.47)
ANT	65.95(+0.13)	76.74(-5.00)	10.09(-0.13)
DeepAugment	66.54(+0.72)	70.31(-11.43)	9.37(-0.85)

we change the model from ViT to Swin Transformer [33] and ConvNeXt [34], the test accuracy on both ImageNet-D (Background) and ImageNet improve. However, the robustness on Texture and Material test set even decreases slightly. These results show the difficulty of improving ImageNet-D robustness by model architecture.

Pretraining with more data. Pretraining on a large data set is effective to improves model performance, such as ImageNet accuracy [18]. Figure 11 compares ConvNext, that is trained directly on ImageNet-1K, with ConvNext (Pretrained) which is first pretrained on ImageNet-22K. We find that ConvNext (Pretrained) achieves higher robustness than ConvNext on all three sets of ImageNet-D, especially for the Background set. These results show that pretraining on a large data set helps improve robustness on ImageNet-D.

4.4. Further discussions

Can CLIP find the correct neighbors of ImageNet-D images? CLIP model [45] shows potential in nearest neighbor search tasks. With ImageNet-D images as query, we retrieve the most similar images from ImageNet to investigate whether CLIP can find correct neighbors, as shown in Figure 10. Take Background for example, retrieved images may either have a similar background to the query image (Figure 10(a)) or include the object that is related to the query image's background(Figure 10(b)).Our results show that ImageNet-D can find the failure cases of neural networks in nearest neighbor retrieval.

Can ImageNet-D match natural test sets in failure transferability? Section 3.1 defines `transferable failure` and finalize ImageNet-D with shared failures of surrogate models. We conduct the same experiment on ImageNet, introducing ImageNet (Failure) with the shared failure images of surrogate models. Table 5 show that ImageNet-D achieves similar accuracy to ImageNet (Failure), indicating that that synthetic images can achieve similar failure transferability as natural images. In contrast to

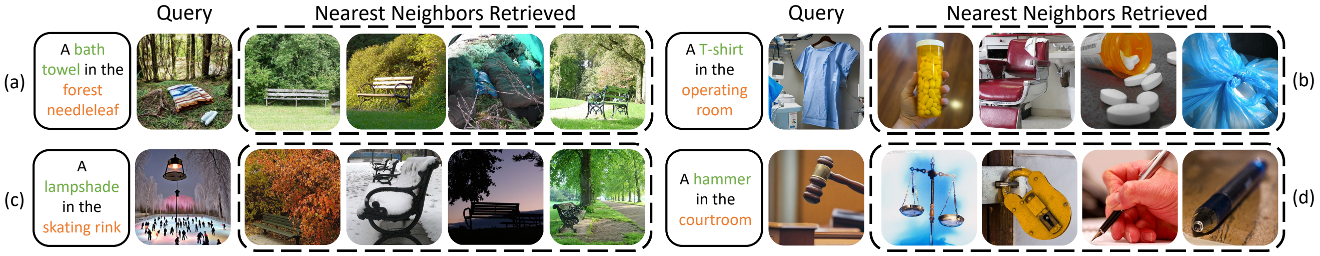

Query
Nearest Neighbors Retrieved
Query
Nearest Neighbors Retrieved
(a)
A bath towel in the forest needleleaf
(b)
A T-shirt in the operating room
(c)
A lampshade in the skating rink
(d)
A hammer in the courtroom

Figure 10. Visualizations of nearest neighbor images. We visualize the nearest neighbor images from ImageNet with ImageNet-D image as the query image. Instead of following the same object category as the query image, the nearest neighbor images either follow a similar background or follow another object category that is highly correlated with the background of query image. Our results show that ImageNet-D can find the failure cases of neural networks in nearest neighbor retrieval.

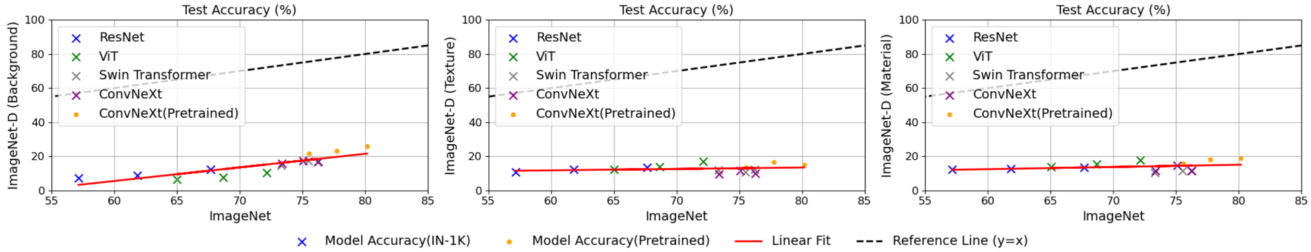


Figure 11. Test accuracy of different architectures and training data. Each data point corresponds to one tested model. Apart from the ConvNeXt (Pretrained), all other models are trained on ImageNet-1K. The plots show that pretraining achieves higher robustness on all three subsets, while changing model architectures only improves on ImageNet-D (Background).

Table 5. Results of failure transferability. We create ImageNet (Failure) with shared failures of surrogate models from original ImageNet, achieving comparable accuracy to ImageNet-D. These results show that that our synthetic images achieve similar ability to natural images in finding the failures of new models.

Model	Architecture	ImageNet	ImageNet (Failure)	ImageNet-D
CLIP	ViT-B/16	67.95	11.09	21.9
LLaVa	Vicuna 13B	79.32	41.43	49.65
MiniGPT-4	Vicuna 13B	88.77	65.22	71.96

Table 6. Test accuracy of models finetuned on synthetic data. We finetune a pretrained ResNet18 model on ImageNet-1K together with different extra training data. Training on synthetic images achieves highest robustness on both ImageNet-D and ObjectNet.

Model	Extra training data	ImageNet	ObjectNet	ImageNet-D
A	/	55.68	21.51	8.6
B	ImageNet	59.78(+4.10)	24.78(+3.27)	10.4(+1.80)
C	Synthetic-easy	56.56(+0.88)	26.12(+4.61)	27.86(+19.26)

natural datasets like ImageNet, ImageNet-D enjoys a lower cost in data collection and can be scaled efficiently.

Training on diffusion-generated images. By contrast to shared failure images in ImageNet-D, we term generated images correctly classified by surrogate models as Synthetic-easy, and investigate their influence as training data. We finetune a pre-trained ResNet18 model on different training sets in Table 6. Table 6 shows that training on Synthetic-easy significantly improves ImageNet-D robustness by 19.26%. Remarkably, model C outperforms model B in ObjectNet accuracy by 1.34%, indicating model C's superior generalization. These results imply that diffusion-generated images with diverse object and nuisance pairs could enhance model robustness as training samples.

5. Conclusion

In this paper, we introduce a test set ImageNet-D and establish a rigourous benchmark for visual perception robustness. Capitalizing the image generation ability of diffusion models, ImageNet-D includes images with diverse factors including background, texture and material. Experimental results show that ImageNet-D significantly decreases the accuracy of various models, including CLIP (reducing 46.05%), LLaVa [31] (reducing 29.67%), and MiniGPT-4 [65] (reducing 16.81%), demonstrating the effectiveness in model evaluation. Our work makes a step forward in improving synthetic test sets, and will create more diverse and challenging test images as generative models improve.

Acknowledgments: This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00951, Development of Uncertainty-Aware Agents Learning by Asking Questions).

References

[1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *Transactions on Machine Learning Research*, 2023. 2

[2] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. In *Deployable Generative AI Workshop at ICML 2023*, 2023. 2

[3] 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*, pages 9448–9458, 2019. 1, 2

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 4, 5

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 1

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 1

[8] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2

[9] Li Chen, Carter Yagemann, and Evan Downing. To believe or not to believe: Validating explanation fidelity for dynamic malware analysis. In *CVPR Workshops*, pages 48–52, 2019. 1

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. *arXiv preprint arXiv:2010.11929*, 2010. 5

[14] Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023. 2

[15] 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. *arXiv preprint arXiv:1811.12231*, 2018. 1, 2, 6

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 5

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 7

[19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1, 2, 6

[20] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 2

[21] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 6

[22] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 2, 4, 6

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 2

[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[25] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 1

[26] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2

[27] Markus Kängsepp and Meelis Kull. Calibrated perception uncertainty across objects and regions in bird's-eye-view. *arXiv preprint arXiv:2211.04340*, 2022. 1

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1

[29] Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, and Hui Xue. Imagenet-e: Benchmarking neural network robustness via attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20371–20381, 2023. 2

[30] Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, and Hao Su. Distilling large vision-language model with out-of-distribution generalizability. *arXiv preprint arXiv:2307.03135*, 2023. 1

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 5, 8

[32] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *International Conference on Machine Learning*, pages 22352–22369. PMLR, 2023. 1

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1, 2, 7

[34] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 1, 7

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

[36] Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7838–7847, 2021. 2

[37] Chengzhi Mao, Lu Jiang, Mostafa Dehghani, Carl Vondrick, Rahul Sukthankar, and Irfan Essa. Discrete representations strengthen vision transformer robustness. *arXiv preprint arXiv:2111.10493*, 2021. 2

[38] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022. 2

[39] Chengzhi Mao, Revant Teotia, Amrutha Sundar, Sachit Menon, Junfeng Yang, Xin Wang, and Carl Vondrick. Doubly right object recognition: A why prompt for visual rationales. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2722–2732, 2023. 2

[40] Jan Hendrik Metzen, Robin Hutmacher, N Grace Hua, Valentyn Boreiko, and Dan Zhang. Identification of systematic errors of image classifiers on rare subgroups. *ICCV*, 2023. 2

[41] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems*, 35:35087–35102, 2022. 1

[42] Tommaso Nesti, Santhosh Boddana, and Burhaneddin Yaman. Ultra-sonic sensor based object detection for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 210–218, 2023. 1

[43] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*, pages 1–18, 2017. 1

[44] Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. Lance: Stress-testing visual models by generating language-guided counterfactual images. *Advances in Neural Information Processing Systems*, 36, 2023. 2

[45] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5, 7

[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[47] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 4

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 5

[49] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[50] Evgenia Rusak, Lukas Schott, Roland S Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 53–69. Springer, 2020. 6

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour,

Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[53] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2023. 2

[54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2

[56] Joshua Vendrow, Saachi Jain, Logan Engstrom, and Aleksander Madry. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. *arXiv preprint arXiv:2302.07865*, 2023. 2

[57] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[58] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020. 1, 2

[59] Jianhao Yuan, Francesco Pinto, Adam Davies, and Philip Torr. Not just pretty pictures: Toward interventional data augmentation using text-to-image generators. 2023. 2

[60] Zhenlong Yuan, Yongqiang Lu, Zhaoguo Wang, and Yibo Xue. Droid-sec: deep learning in android malware detection. In *Proceedings of the 2014 ACM conference on SIGCOMM*, pages 371–372, 2014. 1

[61] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 6

[62] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 2

[63] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019. 2

[64] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *arXiv preprint arXiv:2305.16934*, 2023. 2

[65] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2, 5, 8