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

Supplementary Material

1. Labeling task on Amazon Mechanical Turk

For reliable benchmarks, we use Amazon Mechanical Turk (MTurk) [2, 4, 5] to evaluate the labeling quality of ImageNet-D.

1.1. Labeling task design

Labeling instructions. Since ImageNet-D includes images with diverse object and nuisance pairs that may be rare in the real world, we take both the appearance and functionality of the main object as the labeling criteria. Specifically, we ask the workers from MTurk to answer the following two questions:

`Question 1`: Can you recognize the desired object ([`ground truth category`]) in the image? It may have rare backgrounds, textures, materials, or styles.

`Question 2`: Can the object in the image be used as the desired object ([`ground truth category`])?

Labeling pipeline. To ensure that the workers understand these two criteria, we ask the workers to label two example images for practice, which provides the correct answer for the above two questions. After the practice session, the workers are required to label up to 20 images in one task, and answer both two questions for each image. The worker selects 'yes' or 'no' for each question.

Labelling UI. The labeling page is designed as in Figure 1. The workers can proceed to the next image only if they finish both questions on the current page.

1.2. Quality control of human labelling

We use sentinels to ensure high-quality annotations. For each labeling task with multiple images, we design three types of sentinels as follows.

Positive sentinel: Image that belongs to the desired category and is correctly classified by multiple models. If the workers do not select 'yes' for this image, they may not understand the concept well and their annotations will be removed.

Negative sentinel: Image that does not belong to the desired category. For example, if the desired category is a chair, the negative sentinel may be a ladle. If the workers select 'yes' for the ladle image, they may not answer the questions seriously and their annotations will be removed.

Consistent sentinel. We assume that the workers should select the same answer for the same image if it appears multiple times. Consistent sentinels are images that appear twice in a random order. If the workers answer differently for the same image, their annotations are not consistent and will be removed.

For each labeling task with up to 20 images, we include one positive sentinel, one negative sentinel and two consistent sentinels. We discard the responses if the workers do not pass all the sentinel checks.

1.3. Results

For each image, we collect independent annotations from 10 workers and filter out responses from the workers that do not pass the quality check. A total of 679 qualified workers submitted 1540 labeling tasks, resulting an agreement of 91.09% on sampled image from ImageNet-D.

2. Experimental results on ImageNet-D

More results for Section 4. We compare the model accuracy of Image-D with existing test sets, including ImageNet [6], ObjectNet [1]s, ImageNet-9 [7] and Stylized-ImageNet [3]. All the accuracy numbers are reported in Table 1, which also includes the numbers of Figure 8 in the main manuscript.

Training setups for Table 6. We introduce experimental details of Table 6 in the main manuscript. We finetune a pre-trained ResNet18 model on various training sets. To examine the effect of incorporating synthetic images into the fine-tuning training set, we sample ImageNet and Synthetic-easy for same data distributions, where Synthetic-easy includes diffusion-generated images correctly classified by surrogate models. Each set contains 111098 images, and both sets have same number of images per category. All models are finetuned on on a pre-trained ResNet18 at epoch 90 for 10 epochs further, using a SGD optimizer with a learning rate of 0.0001. Apart from sampled ImageNet and Synthetic-easy, we include original ImageNet-1K as training data for smooth training.

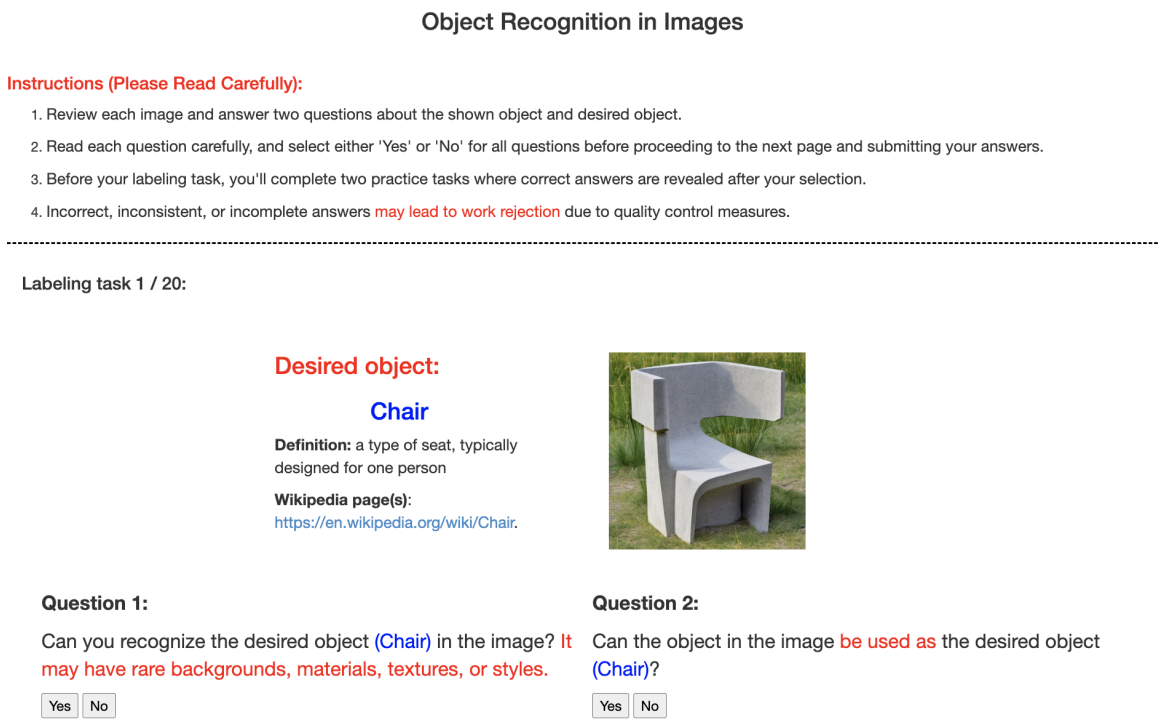

Figure 1. User interface for MTurk studies. The workers can proceed to the next image only if they finish both questions on the current page.

Table 1. 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	ImageNet-D			ImageNet-D
						Background	Texture	Material	Total
Vision model (CNN)	VGG11	56.85	21.85	68.59	13.12	6.46(-50.39)	9.64(-47.21)	11.87(-44.98)	7.43(-49.42)
	VGG13	58.42	23.23	68.96	13.59	7.39(-51.03)	8.63(-49.79)	9.6(-48.82)	7.78(-50.64)
	VGG16	60.86	25.96	73.28	13.83	9.94(-50.92)	10.84(-50.02)	13.79(-47.07)	10.49(-50.37)
	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)
	ResNet18	57.15	22.62	71.65	21.17	7.41(-49.74)	10.64(-46.51)	12.22(-44.93)	8.31(-48.84)
	ResNet34	61.81	26.15	75.31	21.33	8.87(-52.94)	12.25(-49.56)	12.74(-49.07)	9.68(-52.13)
	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)
	ResNet152	69.18	34.41	83.41	23.05	13.79(-55.39)	13.86(-55.32)	18.85(-50.33)	14.4(-54.78)
	Densenet121	63.1	28.74	82.05	19.92	9.99(-53.11)	12.65(-50.45)	15.36(-47.74)	10.9(-52.20)
	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)
	Densenet169	64.13	30.13	83.8	22.97	10.73(-53.40)	12.25(-51.88)	12.39(-51.74)	11.09(-53.04)
	Densenet201	66.58	31.36	83.43	21.8	9.88(-56.70)	10.84(-55.74)	15.71(-50.87)	10.67(-55.91)
	Wideresnet50	69.06	32.65	81.41	19.45	8.69(-60.37)	11.24(-57.82)	11.17(-57.89)	9.25(-59.81)
	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	Vicuna 13B	86.83	79.97	91.47	62.61	75.91(-10.92)	64.56(-22.27)	60.39(-26.44)	72.9(-13.93)
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)

References

[1] 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] 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. 1

[3] 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

[4] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1

[5] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 1

[6] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1

[7] 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