Progress and Limitations of Deep Networks to Recognize Objects in Unusual Poses

Amro Abbas[1], **Stéphane Deny**[2]

[1]The African Institute For Mathematical Sciences
[2]Aalto University
afagiri@aimsammi.org, stephane.deny@aalto.fi

Abstract

Deep networks should be robust to rare events if they are to be successfully deployed in high-stakes real-world applications. Here we study the capability of deep networks to recognize objects in unusual poses. We create a synthetic dataset of images of objects in unusual orientations, and evaluate the robustness of a collection of 38 recent and competitive deep networks for image classification. We show that classifying these images is still a challenge for all networks tested, with an average accuracy drop of 29.5% compared to when the objects are presented upright. This brittleness is largely unaffected by various design choices, such as training losses, architectures, dataset modalities, and data-augmentation schemes. However, networks trained on very large datasets substantially outperform others, with the best network tested—Noisy Student trained on JFT-300M—showing a relatively small accuracy drop of only 14.5% on unusual poses. Nevertheless, a visual inspection of the failures of Noisy Student reveals a remaining gap in robustness with humans. Furthermore, combining multiple object transformations—3D-rotations and scaling—further degrades the performance of all networks. Our results provide another measurement of the robustness of deep networks to consider when using them in the real world. Code and datasets are available at https://github.com/amro-kamal/ObjectPose.

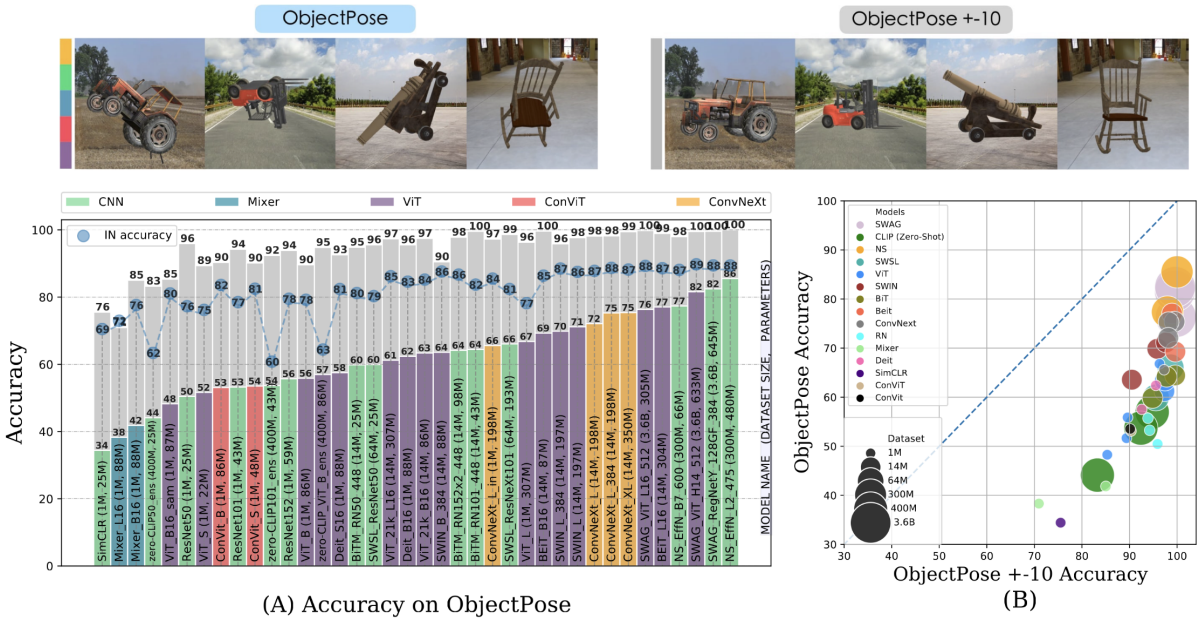
1 Introduction

In real-world applications, deep networks are often deployed and used to detect harder examples than those seen in the development test set. This has led researchers to investigate the performance of networks using more challenging test examples, in the so-called out-of-distribution (OOD) regime [Hendrycks et al. 2021b, Hendrycks and Dietterich 2019, Hendrycks et al. 2021a, Wang et al. 2019, Dodge and Karam 2017, Recht et al. 2019a, Shankar et al. 2021, Geirhos et al. 2018, 2021]. Many of the previous studies on out-of-distribution generalization have focused on measuring the generalization capabilities of networks to distorted images. Notably, [Geirhos et al. 2021] show that the newest generation of very large deep networks is closing the human-machine robustness gap on 17 different out-of-distribution image distortion types. However, most of the transformations used to generate these datasets are local distortions that only affect the texture of objects, but do not change the global structure of the image. [Taori et al. 2020] shows that robustness to these kinds of distortions does not transfer entirely to natural shifts and does not represent a comprehensive measure of the network's robustness.

Pose transformations (e.g., a bus seen upside-down) represent an interesting case-study for networks' robustness, as (1) unlike simple distortions, these transformations affect the global structure of the image, and (2) it would be technically challenging to augment an image dataset at scale with this type of 3D transformations. In a rare study of its kind, [Alcorn et al. 2019] show that two deep networks, Inception-V3 [Szegedy et al. 2016] and ResNet50 [He et al. 2016], drastically fail to recognize objects in unusual poses, incorrectly classifying most of the poses explored. Later, [Madan et al. 2021] also reveal a brittleness of ResNet18 and CLIP [Radford et al. 2021] to small changes in pose, in an adversarial setting.

In this study, we revisit the question of networks' robustness to unusual poses using a diverse set of the latest and best publicly available networks for image classification. We test 38 networks with a variety of different architectures, sizes, training datasets, and training objectives on a custom dataset of images of objects in unusual poses. Our contributions are:

- We observe that, on unusual poses, the networks of our collection suffer from a 14.5% to 45.5% accuracy drop compared to usual poses. A visual inspection of networks' failures reveal that, even for the best models, a robustness gap remains with the human visual system.
- We provide a detailed study of the effect of in-(image)-plane vs. out-of-plane object rotations, background-foreground congruency, and rotation angle on performance. We show that the best networks rely on a different strategy than weaker networks when it comes to incorporating information from the background.
- By combining multiple unusual transformations—such as object rotations and scaling—we show that such combinations lead to further performance degradation for all networks, as predicted by a combinatorial model of error.
- In an effort to go beyond synthetic datasets, we test the networks on images from the *Common Objects in 3D Dataset* (CO3D), a dataset of objects seen in various poses,



(A) Accuracy on ObjectPose (B)

Figure 1: Performance of all networks on ObjectPose. (A) Rotating the objects in unusual poses (ObjectPose, colored bars) induces a top-1 accuracy drop of 14.5%-45.5% compared to when objects are presented upright (ObjectPose +-10, grey bars). Bar colors indicate different architectures. Blue dots: Top-1 accuracy on ImageNet (as reported in the papers). (B) Accuracy on usual vs. unusual poses. Networks trained on ImageNet1k cluster together with low accuracy on ObjectPose. Networks trained on ImageNet21k perform better, and networks trained on extremely large datasets—Noisy Student models (300M images) and SWAG (3.6B images)—perform the best, with the exception of CLIP models (400M images), which were not fine-tuned on ImageNet categories.

and show a generalization gap of 5.2% on average across networks compared to a benchmark of objects presented in their usual pose, ImageNetV2.

2 Dataset and Networks

ObjectPose Dataset We generated a synthetic dataset of objects in unusual poses, ObjectPose. The dataset contains 27,540 images of 17 high-quality 3D objects rendered in a range of different orientations and over different background images, following the pipeline of [Alcorn et al. 2019]. Briefly, the object is first placed in an initial upright position. We then choose one of the *YAW*, *ROLL*, or *PITCH* axes to rotate the object along it and render it on top of a background image. We used three different background images with each object. Two of the backgrounds are images chosen manually from the internet to match the object's usual context and not to contain any other ImageNet object. We chose the third background to be grey with all its RGB pixel values equal to (0.485, 0.456, 0,406), corresponding to the average pixel color of ImageNet images. In order to focus on robustness to unusual poses, each object is chosen carefully so that the resulting images for that object are correctly classified with $> 90\%$ accuracy by a ResNet-50 when the object's orientation is less than 10° apart from its upright pose. We gather these images where the object is rotated by -10° to 10° only in the ObjectPose +-10 dataset (1,683 images in total), and exclude them from ObjectPose, such that ObjectPose only contains unusual poses (11° to 349° from the upright pose). Each of the 17 objects belonging to one of 1000 ImageNet [Russakovsky et al. 2015] classes.

Deep Networks We tested a collection of 38 networks on ObjectPose. We chose a diverse set of networks with different architectures, training datasets sizes, number of parameters, and training objectives. The networks have varying number of parameters (from 22M up to 645M), and different architectures including convolutional neural networks (CNNs) [He et al. 2016, Xie et al. 2020, Kolesnikov et al. 2020, Chen et al. 2020b, Liu et al. 2022], Vision Transformers (ViTs) [Dosovitskiy et al. 2020, Chen, Hsieh, and Gong 2021, Liu et al. 2021] [Bao, Dong, and Wei 2021, Touvron et al. 2021], and MLP-Mixers [Tolstikhin et al. 2021]. We also include ConViT [d'Ascoli et al. 2021], a hybrid CNN-ViT architecture that adds a convolutional inductive bias to the Vision Transformer.

We chose networks trained under different objectives, including (1) Supervised learning, including convolutional architectures, such as [He et al. 2016, Xie et al. 2020, Liu et al. 2022, Kolesnikov et al. 2020], Vision Transformers, such as [Dosovitskiy et al. 2020, Chen, Hsieh, and Gong 2021, Bao, Dong, and Wei 2021, Touvron et al. 2021, Liu et al. 2021, 2022], and MLP-mixer [Tolstikhin et al. 2021], (2) Self-supervised learning, such as SimCLR [Chen et al. 2020a], and BEiT [Bao, Dong, and Wei 2021], (3) Semi-weakly supervised learning, such as SWSL-ResNet50 and SWSL-ResNeXt101

Figure 2: Presentation of the 17 3D objects composing ObjectPose. These objects were carefully selected from https://sketchfab.com/.

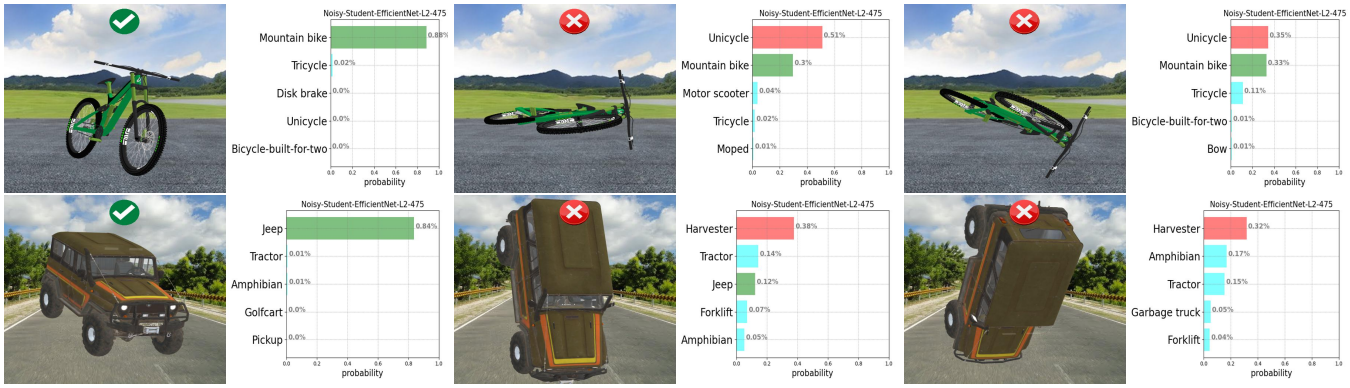


Figure 3: Selected failures of Noisy Student EfficientNet-L2—the best-performing network of our collection—on ObjectPose. Left column: Objects presented upright and top-5 predictions from the network (as measured by the softmax layer activations). Other columns: Objects presented in incorrectly classified poses and top-5 predictions from the network. Some of these errors reveal a brittleness compared to the human visual system (e.g., a tank at 90° is confused with a shield).

[Chen et al. 2020b], and weakly supervised learning, such as SWAG [Singh et al. 2022], (4) Text supervision, such as CLIP [Radford et al. 2021].

The networks were trained on datasets with different sizes ranging from 1M to 3.6B images. Among the networks we use, Noisy Student EfficientNet [Xie et al. 2020] (300M images) and SWAG [Singh et al. 2022] (3.6B images) are the only networks pretrained on extremely large datasets and fine-tuned on ImageNet. Although CLIP [Radford et al. 2021] was pretrained on a very large dataset (400M image-caption examples), it was not fine-tuned on ImageNet. The best network of our collection according to its performance on ImageNet is the SWAG-RegNetY-128GF-384 model [Singh et al. 2022], pretrained with a weakly-supervised learning approach on 3.6B Instagram images (IG). It achieves 88.55% ImageNet top-1 accuracy.

3 Results

All networks exhibit a performance drop on unusual poses compared to usual poses (Fig. 1). We measure networks' robustness to unusual object poses by testing their accuracy on our synthetic dataset, ObjectPose. Our collection of networks show a top-1 accuracy drop in a range of 14.5%-45.5% on unusual poses compared to usual poses (ObjectPose +-10). Examples of network failures are shown in Fig. 3 for the best model tested, Noisy Student EfficientNet-L2.

Scaling both training dataset size and network capacity is helpful on ObjectPose. Networks trained on larger datasets show a narrower performance gap than networks trained on smaller datasets (Fig. 1B), with the exception of CLIP models which were not fine-tuned on ImageNet. It is also important for the network to have a sufficient size (i.e. number of parameters) to benefit from the large dataset. This is deduced by the performance of the EfficientNet-L2 model (480M parameters) compared to EfficientNet-B7 (66M parameters), both trained using the Noisy Student method on the JFT-300M dataset. While EfficientNet-L2 outperforms the rest of the networks on ObjectPose, EfficientNet-B7 performs on par with many networks pretrained on smaller datasets (such as ImageNet21k).
Although SWAG models are pretrained on more images (3.6B) and have more parameters, they do not outperform Noisy Student EfficientNet-L2. This also shows that scaling the dataset is not the only factor that matters, here the training procedure and architecture also play an important role. In particular, Noisy student was trained using the RangAugment data augmentation method [Cubuk et al. 2020] which applies

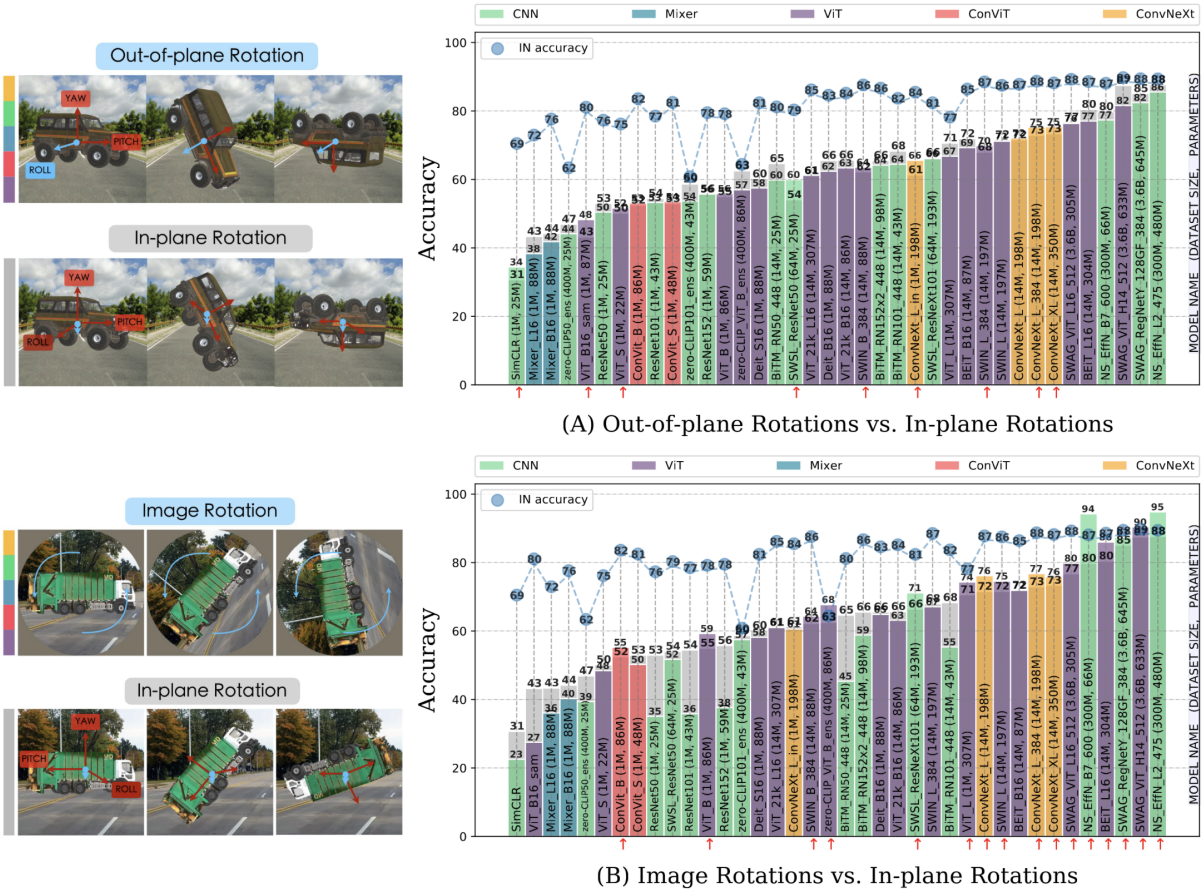


Figure 4: (A) In-plane rotation accuracy. Most networks are slightly more robust to in-plane rotations (colored bars) than out-of-plane rotations (grey bars). Exceptions are marked by red arrows. (B) Image Rotation. The best networks benefit from the background image being rotated with the object (colored bars), unlike most weaker networks which prefer an upright background (in-plane condition, grey bars), revealing a difference in strategy between these two groups of networks.

intensive data augmentation including rotations and shears. In contrast, SWAG only used a limited augmentation strategy of cropping and flipping and not including rotations and shears.

A visual inspection of networks' failures reveals room for improvement even for the best networks tested (Fig. 3). By visually inspecting the errors made by the networks (Fig. 3), we find that even the best model tested, Noisy Student EfficientNet-L2 (NS), makes errors that a human observer would not.

Which types of rotation are most problematic for deep networks (Fig. 4)? We compare the effect of object rotations in the plane of the image, vs. out-of-plane rotations seen in ObjectPose (Fig. 4A). We find that both conditions are problematic for all networks, with the out-of-plane condition only slightly worse than the in-plane condition for most networks. We then compare the in-plane condition with simple image rotations, where the background is rotated with the foreground object (Fig. 4B). By comparing these two conditions, we find that the best networks on ObjectPose rely on a different strategy than the weaker networks when it comes to incorporating background information. Indeed, weak networks perform better on in-plane rotations than on image rotations, in contrast to the best networks (e.g., Noisy Student EfficientNets and SWAG) which perform better on image rotations. Our interpretation is that weaker networks benefit from seeing features of the background upright, whereas the best networks suffer from the incongruency between the upright background and the rotated foreground object. The interpretation that the best networks suffer from the incongruency of the background is reinforced by the observation that they are the only ones to see an increase in accuracy on ObjectPose when the natural backgrounds are removed altogether. The interpretation that the weaker networks benefit from seeing the background features upright is confirmed by the observation that their accuracy drops when the background alone is rotated.

We next study the relation between network accuracy and rotation angle for different rotation types (Fig. 5). We find that all networks are most fragile when the object is rotated by 90° in the out-of-plane condition (ObjectPose). We

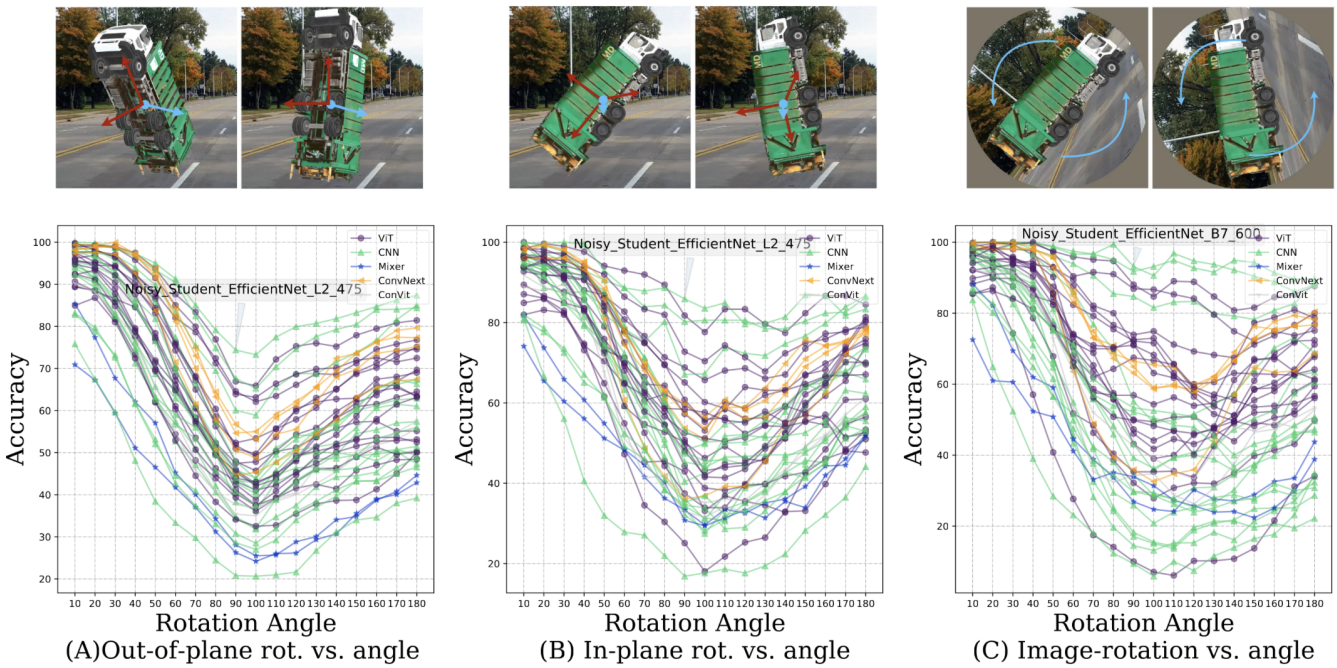


Figure 5: Accuracy decreases as we increase rotation angle, with a low point at 90°. Networks are less robust to (A) out-of-plane rotations than (B) in-plane rotations and (C) image rotations. Noisy Student EfficientNet models are very robust to image rotations, perhaps for the reason that they were trained with this type of augmentation (RandAugment). Yet they are not completely robust to the two other types of rotations.

also find that the best networks are more robust across the full range of rotation angles in the in-plane and image-rotation conditions than in the out-of-plane condition. Noisy Student EfficientNets are especially robust to image rotations, which might be explained by the fact that they are trained with image-rotation augmentations. Yet, we see that this type of data augmentation is not enough to guarantee full robustness to in-plane and out-of-plane rotations.

Combining more than one transformation degrades performance of all networks further, as predicted by a combinatorial model of error (Fig. 6). We next sought to study how the combination of multiple transformations would affect the performance of the networks. For this set of experiments, we use a grey background for all images in order to avoid complex interferences between foreground and background. First, we try combining rotations along the three axes of rotations, YAW, PITCH, and ROLL together. We find that this combination leads to a degradation of performance for all networks (Fig. 6A) in a range 6%-16.5% compared to the condition where only one axis is rotated at a time. Next we investigate the effect of combining three-axes rotations with scaling (Fig. 6C). We find that this combination of transformations further degrades the performance of all networks, with an accuracy drop in the range 24.5%-78.2% compared to usual poses, larger than the accuracy drop seen for ObjectPose in the range 14.5%-45%. We find that a simple combinatorial model of errors (grey bars in Fig. 6C) recapitulates the accuracy drop well for most of the networks: this model assumes that the probability of the network being correct in the scaled-rotated condition (panel C) is simply the product of the probabilities of being correct in the scaled-only (panel B) and rotated-only (panel A) conditions respectively (see Discussion for implications).

How well do our findings on synthetics datasets transfer to real-world datasets (Fig. 7)? In an attempt to go beyond synthetic datasets, we explore the robustness of our collection of networks to a dataset of real objects filmed from various points of view, the *Common Objects in 3D Dataset* (CO3D) [Reizenstein et al. 2021]. Originally designed for 3D reconstruction and new-view synthesis tasks, this dataset was collected by workers turning around and filming common objects from their environment. We sample 1000 images from each of 10 categories common to CO3D and ImageNet, to get a total of about 10,000 images. We then estimate the performance of our collection of networks on these images (Fig. 7). As a point of comparison, we measure the accuracy of networks on images of the same object categories taken from ImageNetV2, a datatset where objects are mostly presented in their usual canonical view. ImageNetV2 can be seen as a fairer comparison benchmark to CO3D than ImageNet, as networks were not overfitted to the the exact statistics of ImageNetV2 [Recht et al. 2019b]. We observe an average accuracy drop across networks of 10.4% on ImageNetV2 over ImageNet, and an average accuracy drop of 5.2% on CO3D over ImageNetV2 (Fig. 7A). However, four networks trained on very large datasets perform nearly as well on both datasets: CLIP-RN-101, CLIP-ViT-B/16, Noisy Student EfficientNet-B7, and SWAG-ViT (Fig. 7B). In summary, the variety of

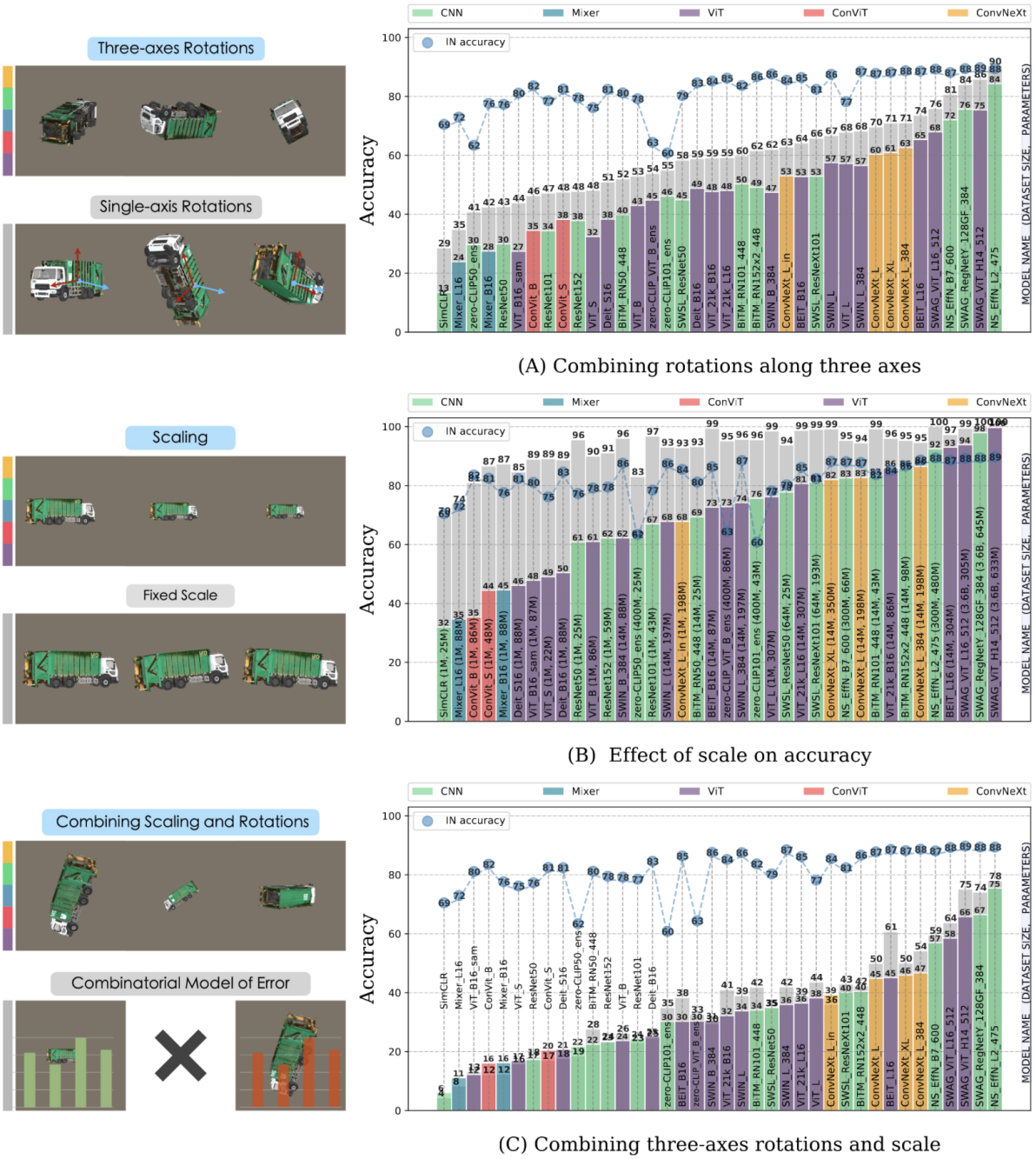


Figure 6: Effect of combining multiple transformations on performance. **(A)** Combining rotations along the three axes ROLL, PITCH and YAW decreases the accuracy of all networks (colored bars) compared to the single-axis rotation condition (grey bars). **(B)** Scaling the object size in the image decreases the accuracy severely for most networks (colored bars) compared to the fixed-scaled upright condition (grey bars), but only slightly for the best networks. **(C)** Combining three-axes rotations and scaling strongly affects the accuracy of all networks (colored bars), with the best network (Noisy Student) at 75% accuracy only in this condition. The grey bars represent a combinatorial model of error which predicts performance degradation well (see text).

object views seen in CO3D seems less problematic for our collection of networks than the views from of our synthetic dataset ObjectPose. This discrepancy could be due to the fact that although the workers turn around the objects in CO3D, they do not necessarily explore all the unusual views that we can explore in our synthetic dataset.

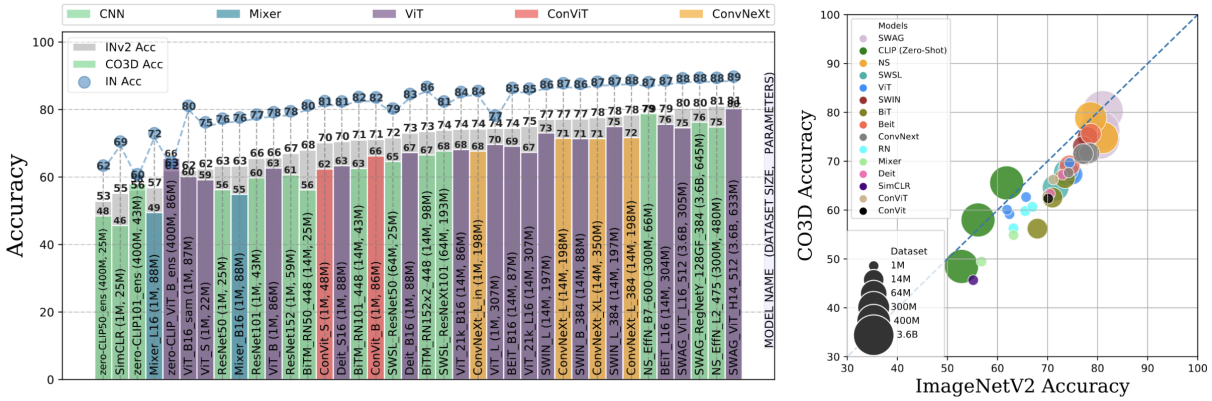


Figure 7: Going beyond synthetic datasets. Most networks perform worse on CO3D (colored bars), a dataset exploring various object views, than on ImageNetV2 (grey bars), which mostly presents objects in their usual canonical views (average accuracy drop of 5.2% between the two datasets). ImageNetV2 is a fairer comparison benchmark to CO3D than ImageNet as there was no overfitting it.

4 Discussion

Probing the robustness of deep networks to objects in unusual poses is interesting for multiple reasons. First, this type of transformation changes the global structure of the image, which might represent a bigger challenge for deep networks than local image distortions. Second, data-augmentation strategies do not provide an easy fix to the problem of generalization to unusual poses. Indeed, a classical mitigation strategy in deep learning consists in augmenting the dataset with the out-of-distribution case that one would like to become robust to, effectively making that case in-distribution. For most image distortions, this augmentation strategy is easy to implement. But for objects in unusual poses, one would need to collect a very large number of images of objects in unusual poses, which is practically challenging, especially at the scale needed to compete with the best current networks (e.g., Noisy Student is trained on JFT-300M comprising 300 million images). Alternatively, one could synthetically generate such a dataset with 3D models of objects, but this would require a very large database of high-quality 3D models which is currently lacking.

A striking result of our study is that very large networks trained on very large datasets (e.g., Noisy Student trained on JFT-300M, SWAG trained on 3.6B Instagram images) are quite robust to unusual poses. However, a careful visual inspection of the errors made by these networks reveal that they do not yet meet the robustness of the human visual system. A thorough comparison to human would be useful to estimate the fraction of the errors that are due to the images themselves not containing enough information about the class (e.g., a jeep seen from the bottom could be any sort of car) vs. out-of-distribution generalization errors that a human would not fall into (e.g., a cannon seen at a 90° angle is confused by the network with a harp). In future work, we plan to establish this human benchmark on our custom dataset ObjectPose in order to precisely quantify the gap in robustness between humans and networks.

When combining object rotations and scaling, we find that a combinatorial model of error—which assumes that the probability of the network being correct on the combination of transformations is equal to the product of the probabilities of it being correct on each respective transformation—accounts well for the degradation of performance of most networks. It would be interesting to study whether networks' performance degrade according to this combinatorial model when combining even more transformations, such as translations, texture-removals etc. Indeed, an implication of this combinatorial model of error is that networks should be brittle in the face of a large combination of transformations, as each factor of variation adds its own source for potential errors.

It is an open question how the human visual system builds robustness to unusual poses. When performing mental rotation, a task consisting in comparing two 3D shapes in different orientations, human subjects take a time to respond that scales linearly with the angle of rotation between the two shapes [Shepard and Metzler 1971]. A similar phenomenon happens when recognizing every-day-life objects in unusual orientations [Jolicoeur 1985]: the recognition time is again linear with the angle of the object with respect to its upright pose. These observations are striking as they suggest that different mechanisms take place in the brain compared to feed-forward deep networks, where the processing time is fixed and does not depend on the complexity of the input. We hypothesize that recurrent mechanisms may play a key role in recognizing objects in unusual poses in the brain, as there is mounting evidence that such mechanisms are critical for recognizing images that are challenging to deep networks [Kar et al. 2019, Bonnen, Yamins, and Wagner 2021].

5 Acknowledgments

We thank [Alcorn et al. 2019] and [Geirhos et al. 2021] for making their codebases publicly available. We thank Robert Geirhos, Michael Alcorn, Anh Nguyen, and the OOD group at Meta AI (led by Pascal Vincent) for their kind feedback and suggestions to improve this work. We also thank Google for providing GCP credits for the project.

References

Alcorn, M. A.; Li, Q.; Gong, Z.; Wang, C.; Mai, L.; Ku, W.-S.; and Nguyen, A. 2019. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4845–4854.

Bao, H.; Dong, L.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Bonnen, T.; Yamins, D. L. K.; and Wagner, A. D. 2021. When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Neuron*, 109(17): 2755–2766.e6.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. E. 2020b. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33: 22243–22255.

Chen, X.; Hsieh, C.-J.; and Gong, B. 2021. When vision transformers outperform ResNets without pre-training or strong data augmentations. *arXiv preprint arXiv:2106.01548*.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.

d'Ascoli, S.; Touvron, H.; Leavitt, M.; Morcos, A.; Biroli, G.; and Sagun, L. 2021. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. *arXiv preprint arXiv:2103.10697*.

Dodge, S.; and Karam, L. 2017. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, 1–7. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Geirhos, R.; Narayanappa, K.; Mitzkus, B.; Thieringer, T.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2021. Partial success in closing the gap between human and machine vision. *arXiv preprint arXiv:2106.07411*.

Geirhos, R.; Temme, C. R.; Rauber, J.; Schütt, H. H.; Bethge, M.; and Wichmann, F. A. 2018. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; Song, D.; Steinhardt, J.; and Gilmer, J. 2021a. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. *ICCV*.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15262–15271.

Jolicoeur, P. 1985. The time to name disoriented natural objects. *Memory & Cognition*, 13(4): 289–303. Place: US Publisher: Psychonomic Society.

Kar, K.; Kubilius, J.; Schmidt, K.; Issa, E. B.; and DiCarlo, J. J. 2019. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6): 974–983. Number: 6 Publisher: Nature Publishing Group.

Kolesnikov, A.; Beyer, L.; Zhai, X.; Puigcerver, J.; Yung, J.; Gelly, S.; and Houlsby, N. 2020. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, 491–507. Springer.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A ConvNet for the 2020s. *arXiv preprint arXiv:2201.03545*.

Madan, S.; Sasaki, T.; Li, T.-M.; Boix, X.; and Pfister, H. 2021. Small in-distribution changes in 3D perspective and lighting fool both CNNs and Transformers. *arXiv*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019a. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 5389–5400. PMLR.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019b. Do ImageNet Classifiers Generalize to ImageNet? In *International Conference on Machine Learning*.

Reizenstein, J.; Shapovalov, R.; Henzler, P.; Sbordone, L.; Labatut, P.; and Novotny, D. 2021. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10901–10911.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Shankar, V.; Dave, A.; Roelofs, R.; Ramanan, D.; Recht, B.; and Schmidt, L. 2021. Do image classifiers generalize across time? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9661–9669.

Shepard, R. N.; and Metzler, J. 1971. Mental rotation of three-dimensional objects. *Science*, 171(3972): 701–703. Place: US Publisher: American Assn for the Advancement of Science.

Singh, M.; Gustafson, L.; Adcock, A.; de Freitas Reis, V.; Gedik, B.; Kosaraju, R. P.; Mahajan, D.; Girshick, R.; Dollár, P.; and van der Maaten, L. 2022. Revisiting Weakly Supervised Pre-Training of Visual Perception Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 804–814.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Taori, R.; Dave, A.; Shankar, V.; Carlini, N.; Recht, B.; and Schmidt, L. 2020. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33: 18583–18599.

Tolstikhin, I. O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357. PMLR.

Wang, H.; Ge, S.; Lipton, Z.; and Xing, E. P. 2019. Learning Robust Global Representations by Penalizing Local Predictive Power. In *Advances in Neural Information Processing Systems*, 10506–10518.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10687–10698.