SI-SCORE: AN IMAGE DATASET FOR FINE-GRAINED ANALYSIS OF ROBUSTNESS TO OBJECT LOCATION, ROTATION AND SIZE

Jessica Yung, Rob Romijnders, Alexander Kolesnikov, Lucas Beyer, Josip Djolonga, Neil Houlsby, Sylvain Gelly, Mario Lucic, Xiaohua Zhai *
Google Research, Brain Team

ABSTRACT

Before deploying machine learning models it is critical to assess their robustness. In the context of deep neural networks for image understanding, changing the object location, rotation and size may affect the predictions in non-trivial ways. In this work we perform a fine-grained analysis of robustness with respect to these factors of variation using SI-SCORE, a synthetic dataset. In particular, we investigate ResNets, Vision Transformers and CLIP, and identify interesting qualitative differences between these.

1 INTRODUCTION

In practice we would like to deploy models which are robust to certain changes in their input. For some of these factors, such as weather conditions, compression artifacts, or even different object orientations, existing datasets can be readily applied to quantify models' robustness, e.g. Barbu et al. (2019); Hendrycks & Dietterich (2018). However, for other important factors such as object size or location, the effect on model performance had not yet been quantified prior to our work. This is particularly concerning because many popular image datasets suffer from *photographer's bias* (Torralba & Efros, 2011), where objects appear mostly in the center of the image.

In previous work (Djolonga et al., 2020), we open-sourced a synthetic dataset for fine-grained evaluation: SI-SCORE (Synthetic Interventions on Scenes for Robustness Evaluation). In a nutshell, we paste a large collection of objects onto uncluttered backgrounds (Figure 1), and can thus conduct controlled studies by systematically varying the object class, size, location, and orientation. We also provided extendable code for researchers to generate similar synthetic datasets and analyse the results.[1]

In this work, we take a step forward and identify interesting qualitative differences between model classes. In particular, we investigate how models based on convolutions (ResNets (He et al., 2016)) compare to models based on attention, specifically Vision Transformers (ViT) (Dosovitskiy et al., 2020). Moreover, we evaluate CLIP (Radford et al., 2021), a model trained jointly on text and images on large-scale web data and evaluated zero-shot.

Related work Creating synthetic datasets by pasting objects onto backgrounds has been used for training (Zhao et al., 2020; Dwibedi et al., 2017; Ghiasi et al., 2020) and evaluating models (Kolesnikov et al., 2020), but previous works do not systematically vary object size, location or orientation, or analyse translation and rotation robustness only at the image level (Engstrom et al., 2017). GANs have also been used to generate counterfactual images to detect bias, specifically to evaluate the effects of features such as makeup or beards on classifiers (Denton et al., 2020).

We include further related work on synthetic data generation and robustness datasets in appendix A.

*Please send correspondence to j.yung357@gmail.com.

[1]The synthetic dataset and code used to generate the dataset are available on GitHub and CVF at https://github.com/google-research/si-score.

Figure 1: Sample images from our synthetic dataset. **Left**: We paste the same foreground-background combination with the object in different sizes, locations and rotation angles. **Right**: We consider 614 foreground objects from 62 classes and 867 backgrounds, and vary the object location, rotation angle, and object size for a total of 611 608 images.

F.O.V.	DATASET CONFIGURATION	IMAGES
SIZE	Objects upright in the center, sizes from 1% to 100% of the image area in 1% increments.	92 884
LOCATION	Objects upright. Sizes are 20% of the image area. We do a grid search of locations, dividing the x-coordinate and y-coordinate dimensions into 20 equal parts each, for a total of 441 (21×21) coordinate locations.	479 184
ROTATION	Objects in the center, sizes equal to 20%, 50%, 80% or 100% of the image size. Rotation angles ranging from 1 to 341 degrees counter-clockwise in 20-degree increments.	39 540

Table 1: Synthetic dataset details. The first column shows the relevant factor of variation (F.O.V.). When there are multiple values for multiple factors of variation, we generate the full cross product.

2 SYNTHETIC DATASET DETAILS

To construct our datasets, we paste foreground images (images of objects) on uncluttered background images in a precise way according to what we wish to study. The foregrounds are extracted from OpenImages (Kuznetsova et al., 2020) using the provided segmentation masks. We include only object classes that map to IMAGENET classes. We also remove all objects that are tagged as occluded or truncated, and manually remove highly incomplete or inaccurately labeled objects. We manually filter the backgrounds to remove those with prominent objects, such as images focused on a single animal or person. This results in 614 object instances across 62 classes and 867 backgrounds.

We construct three subsets for evaluation, one corresponding to each factor of variation, as shown in Table 1. We provide further details in appendix B.

3 RESULTS

Using this dataset, we quantify fine-grained model robustness and uncover insights about models. Here we discuss three main groups of models we investigated: ResNets, Vision Transformers and CLIP. We investigate robustness to different object locations, sizes and rotation angles and include only highlights in the main paper. For full results, please see appendix C.

3.1 RESNETS

ResNets (He et al., 2016) are commonly-used architectures in computer vision. There are many decision choices involved, one of which is the normalisation method. The first widely adopted version used BatchNorm (Ioffe & Szegedy, 2015), but GroupNorm (Wu & He, 2018) has also been a popular choice since. We analyse ResNet-50 models pre-trained on IMAGENET that use BatchNorm and

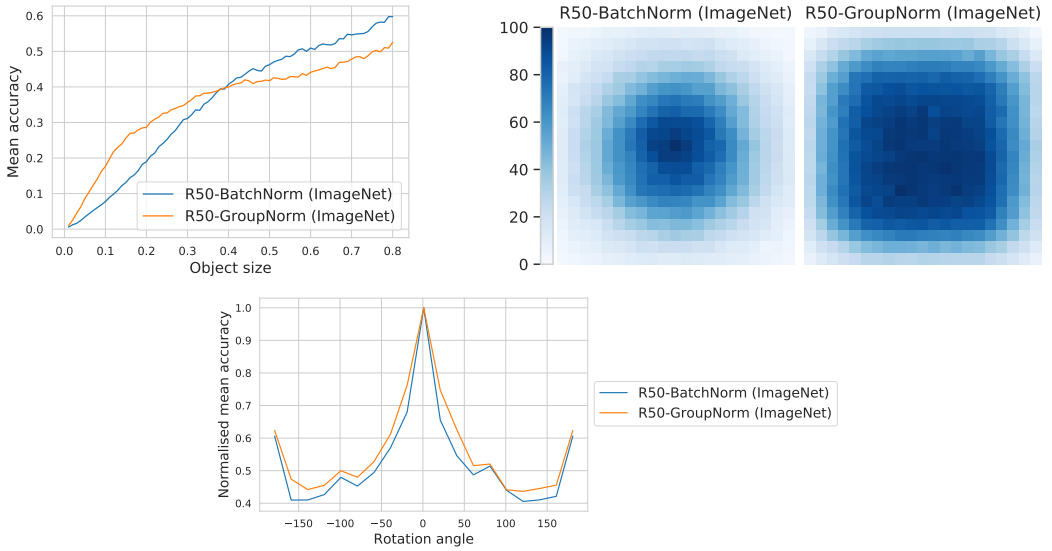

Figure 2: **Left**: We find that a ResNet-50 using GroupNorm has higher accuracy on smaller objects, whereas the one using BatchNorm has higher accuracy on large objects. **Right**: The ResNet-50 using BatchNorm is less robust to changes in location than the one using GroupNorm. Each pixel represents the average normalised top-1 accuracy of the model on images where the object is centered at that location. The accuracy is shown as a percentage of the maximum accuracy across all locations. **Bottom**: The ResNet-50 using BatchNorm seems to be slightly less robust to changes in rotation angle than the one using GroupNorm.

GroupNorm respectively, and find three qualitative differences between them. First, the model that uses GroupNorm has higher accuracy on smaller objects, whereas the model that uses BatchNorm has higher accuracy on objects that take up at least 40% of the image (fig. 2 left). Note that most objects usually take up less than 40% of the image - for example, in a self-driving scenario, each object of interest in the driver's field of view typically occupies less than 40% of it. Because of this, this tradeoff would generally be more favorable for the ResNet-50 that uses GroupNorm. Second, the model that uses BatchNorm is less robust to changes in location than the model using GroupNorm (fig. 2 right). For this experiment we use objects occupying 20% of the image. Thirdly, it seems that ResNets using GroupNorm are also slightly more robust to changes in object orientation (rotation angle) (fig. 2 bottom). Note that we measure the robustness *relative* to the model's best accuracy across locations or rotation angle respectively, so differences in absolute accuracy are accounted for. In future work, we hope to investigate whether these differences are present at scale when training on larger datasets and architectures.

3.2 Vision Transformers

Recently, Dosovitskiy et al. (2020) showed that Transformers can be effective for image classification. Since the Vision Transformer (ViT) models use image patches as input, a natural question is whether they are robust to changes in object location, and whether they exhibit a grid-like pattern in per-location accuracy. We compare Vision Transformer models with convolutional neural networks with similar ImageNet accuracy. Specifically, we use BigTransfer (BiT) (Kolesnikov et al., 2020) models, which are ResNets using GroupNorm that were pre-trained on the same datasets and fine-tuned on ImageNet with high resolution. Absolute robustness is highly correlated with ImageNet accuracy (Taori et al., 2020; Djolonga et al., 2020), therefore, we compare models with similar ImageNet accuracy to account for this confounder.

First, we find that ViT models have much higher relative accuracy when the object of interest is placed close to the edges of the image (fig. 3 rows 1 and 2). One potential explanation is that BiT models have zero padding in the convolutions, whereas ViT models do not have such padding. At the same time, in non-central, non-edge parts of the image, ViT models seem to be slightly more location-robust than BiT models in most but not all locations.

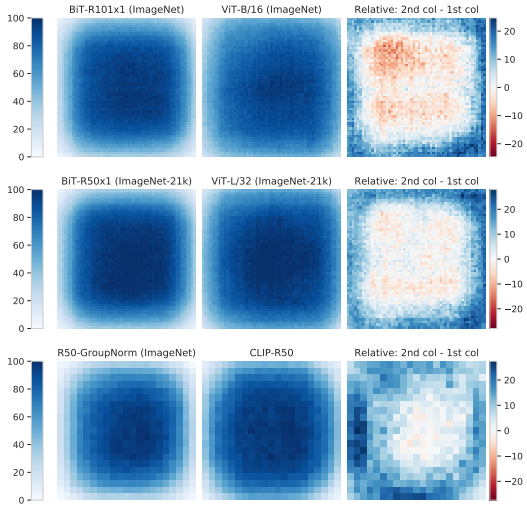

Figure 3: For each location on the grid, we compute the average accuracy on images with the object centered at that location. We show the accuracy as a percentage of the maximum accuracy across all locations. The third column indicates the difference between the second and first column. Blue indicates an improvement of the second column over the first column. This shows the difference in robustness to changes in object location between the pairs of models. Note that we compare models that have *similar* IMAGENET *accuracy*. **Rows 1 and 2**: We observe that the ViT models are more robust to location near the edges than the BiT ResNet models, as shown by the dark blue edges. We use a finer grid to investigate whether ViT models have grid-like patterns in location robustness, and do not find such patterns. **Row 3**: We observe that the CLIP model is slightly more robust to location despite having much lower IMAGENET accuracy than the vanilla ResNet-50 model.

Second, we do not find evidence for a grid-like pattern in per-location accuracy in ViT. To investigate this, we use a finer 56×56 grid compared to the 20×20 grid in the previously open-sourced dataset. The ViT-*/16 models use image patches forming a 14×14 grid, so each patch would correspond to a 4×4 patch on this grid (fig. 3 row 1). The ViT-*/32 models use a 7×7 grid, so each patch would correspond to a 8×8 patch on this grid (fig. 3 row 2). When looking at the absolute values and differences in accuracies, we do not see grid-like patterns. Thus, it seems that there are no significant grid-like patterns in per-location accuracy in ViT.

3.3 CLIP

One model that has received a great deal of attention is CLIP (Radford et al., 2021). It stands in contrast to other considered models since it was trained jointly on images and language input. Furthermore, in contrast to other models, CLIP is not fine-tuned to IMAGENET, but is evaluated in the IMAGENET label space in a zero-shot setting. As a result, its IMAGENET accuracy — at least of the small published models and without prompt ensembling — is significantly lower than that of the other models in this paper. Notably, even when we compare CLIP models to a standard ResNet-50 that has over 10% higher top-1 accuracy on IMAGENET, the CLIP model seems to be more robust to different object locations (fig. 3 row 3). This is perhaps surprising since robustness is often correlated with ImageNet accuracy, but is in line with its improved relative performance on robustness benchmarks, as reported in Radford et al. (2021).

4 DISCUSSION AND CONCLUSION

We investigated robustness of ResNets, Vision Transformers and CLIP to changes in object location and size. Additional results on robustness with respect to object rotation can be found in appendix C.

We note that there could be potential differences and confounding factors when evaluating the performance on synthetic data. We apply the following steps to mitigate the risk: Firstly, we use cut-and-pastes of real data instead of fully synthetic data. Notably, Ghiasi et al. (2020) successfully trained state-of-the-art object segmentation models on such data, which lends evidence that related artifacts may not significantly affect behaviour. Secondly, we average across over 1000 object and background combinations to minimize the effect of the choice of object or background on the results. Finally, we consider relative performance between models as opposed to absolute numbers.

We hope that the insights presented in this study will influence research on the use of synthetic data for stress-testing deep learning models.

REFERENCES

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*, 2019.

Emily Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. Image counterfactual sensitivity analysis for detecting unintended bias, 2020.

Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, Sylvain Gelly, Neil Houlsby, Xiaohua Zhai, and Mario Lucic. On robustness and transferability of convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *International Conference on Computer Vision*, 2017.

Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv: 1712.02779*, 2017.

Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv: 1807.01697*, 2018.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv: 2006.16241*, 2020.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Conference on Computer Vision and Pattern Recognition*, 2017.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. *European Conference on Computer Vision*, 2020.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv: 1811.00982*, 2020.

Yann LeCun, Fu Jie Huang, and Léon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Conference on Computer Vision and Pattern Recognition*, 2004.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Alec Radford, Ilya Sutskever, Jong Wook Kim, Gretchen Krueger, and Sandhini Agarwal. Clip: Connecting text and images, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. A systematic framework for natural perturbations from videos. *arXiv: 1906.02168*, 2019.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 2020.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Conference on Computer Vision and Pattern Recognition*, 2011.

Yuxin Wu and Kaiming He. Group normalization, 2018.

Nanxuan Zhao, Zhirong Wu, Rynson W. H. Lau, and Stephen Lin. Distilling localization for self-supervised representation learning. *arXiv: 2004.06638*, 2020.

A Further Related Work

In this section, we describe related work on synthetic data generation and datasets to measure robustness.

Other efforts on synthetic data generation include CLEVR (Johnson et al., 2017), which aims to evaluate compositional generalisation, dSprites (Matthey et al., 2017), which aims to evaluate disentanglement of latent features, or smallNorb (LeCun et al., 2004), which is also an object classification dataset, albeit with different factors of variation except rotation. These datasets use rendered shapes or models of geometric shapes or toys instead of realistic photos of ImageNet classes with photo backgrounds.

Our work focuses on synthetic data to analyse specific factors of variation. Other datasets to analyse robustness mostly include natural datasets. For example, ImageNet-R presents a dataset of alternatively rendered imagery ranging from cartoons to origami (Hendrycks et al., 2020). ImageNet-Vid (Russakovsky et al., 2015) uses frames from video sequences, and ImageNet-Vid-Robust measures whether model predictions are correct and consistent across similar frames (Shankar et al., 2019). Finally, ImageNet-C (Hendrycks & Dietterich, 2018) uses synthetic image-level perturbations on natural images to analyse robustness with respect to perturbations such as Gaussian noise, JPEG compression, variations in image brightness or motion blur. SI-SCORE focuses on object-level as opposed to image-level factors of variation.

B Synthetic dataset details

We include further details on the synthetic dataset in this section.

We construct three subsets for evaluation, one corresponding to each factor of variation we wanted to investigate (object size, location and rotation), as shown in Table 2. The table is repeated here for easy reference. For each object instance, we sample two backgrounds, and for each of these object-background combinations, we take a cross product over all the factors of variation. For the datasets with multiple values for more than one factor of variation, we take a cross product of all the values for each factor of variation in the set. For example, for the rotation angle dataset, there are four object sizes and 18 rotation angles, so we do a cross product and have 72 factor of variation combinations. For the object size and rotation datasets, we only consider images where objects are at least 95% in the image. For the location dataset, such filtering removes almost all images where objects are near the edges of the image, so we do not do such filtering. Note that since we use the central coordinates of objects as their location, at least 25% of each object is in the image even if we do not do any filtering.

F.O.V.	Dataset Configuration	Images
Size	Objects upright in the center, sizes from 1% to 100% of the image area in 1% increments.	92 884
Location	Objects upright. Sizes are 20% of the image area. We do a grid search of locations, dividing the x-coordinate dimension and y-coordinate dimensions into 20 equal parts each, for a total of 441 (21×21) coordinate locations.	479 184
Rotation	Objects in the center, sizes equal to 20%, 50%, 80% or 100% of the image size. Rotation angles ranging from 1 to 341 degrees counter-clockwise in 20-degree increments.	39 540

Table 2: Synthetic dataset details. The first column shows the relevant factor of variation (F.O.V.). When there are multiple values for multiple factors of variation, we generate the full cross product of images.

Image licenses The backgrounds are images from nature taken from *pexels.com*. The license therein allows one to reuse photos with modifications.

C FULL RESULTS FOR MODEL COMPARISONS IN THE MAIN PAPER

C.1 RESNET-50S WITH BATCHNORM AND GROUPNORM

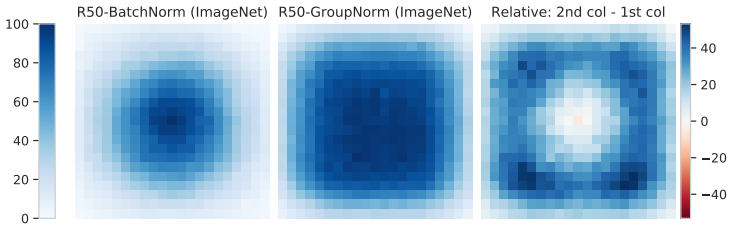

Figure 5: In the first and second columns, for each location on the grid, we compute the average accuracy of the models on images where the object is centered at that location. We show the accuracy as a percentage of the maximum accuracy across all locations. In the third column, we compute the difference between the second column and the first column. Blue indicates an improvement of the second column over the first column. This shows the difference in robustness to changes in object location between the two models. The ResNet-50 that uses BatchNorm is less robust to changes in location than the one that uses GroupNorm.

The object size and rotation plots are in Figure 2 in the main paper.

C.2 BiT (ResNet) vs ViT models

We compare BigTransfer (BiT) (Kolesnikov et al., 2020) and Vision Transformer (ViT) (Dosovitskiy et al., 2020) model pairs that have similar ImageNet accuracy. We include a ViT-L/32 model in the location study to investigate whether there is a grid-like pattern in location robustness.

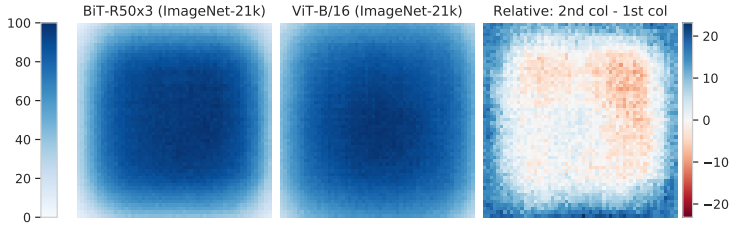

Figure 6: In the first and second columns, for each location on the grid, we compute the average accuracy of the models on images where the object is centered at that location. We show the accuracy as a percentage of the maximum accuracy across all locations. In the third column, we compute the difference between the second column and the first column. Blue indicates an improvement of the second column over the first column. This shows the difference in robustness to changes in object location between the pairs of models. We observe that the ViT models are more robust to location near the edges than the BiT ResNet models, as shown by the dark blue edges in the third column. We use a finer grid to investigate whether ViT models have grid-like patterns in location robustness, and do not find such patterns. Note that we compare models that have similar ImageNet accuracy.

The plots comparing the location robustness of BiT-R101x3 with ViT-B/16 trained on ImageNet, and BiT-R50x1 with ViT-L/32 trained on ImageNet-21k are in Figure 3 (top, middle rows) in the main paper. We compare these pairs of models that have similar ImageNet accuracy to control for differences purely due to different ImageNet accuracy.

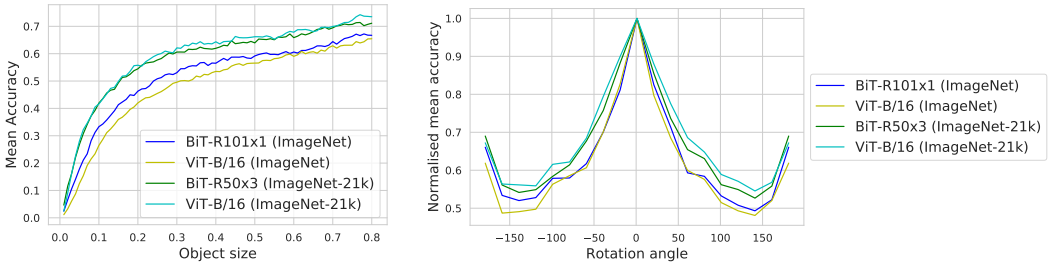

Figure 7: **Left**: We find that the BiT-R50x3 and ViT-B/16 models trained on ImageNet-21k seem to have similar robustness to changes in object size. However, the BiT-R101x1 model trained on ImageNet seems to be slightly better at classifying smaller objects than the ViT-B/16 model trained on ImageNet. **Right**: Conversely, the two BiT and ViT models trained on ImageNet seem to have similar robustness to changes in rotation angles with the BiT model perhaps being slightly better. For the two models trained on ImageNet-21k, however, the ViT model seems to be slightly more robust. The differences in both cases are quite small.

C.3 CLIP MODELS (RADFORD ET AL., 2021)

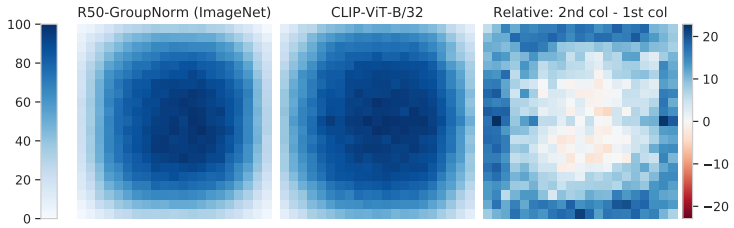

Figure 8: In the first and second columns, for each location on the grid, we compute the average accuracy of the models on images where the object is centered at that location. We show the accuracy as a percentage of the maximum accuracy across all locations. In the third column, we compute the difference between the second column and the first column. Blue indicates an improvement of the second column over the first column. This shows the difference in robustness to changes in object location between the pairs of models. We observe that the CLIP model is slightly more robust to location despite having much lower IMAGENET accuracy than the vanilla ResNet-50 model, as can be seen by the blue edges in the third column.

The plot comparing the location robustness of the ResNet-50 (GroupNorm) with the CLIP model using a ResNet-50 backbone is in Figure 3 (bottom row) in the main paper.

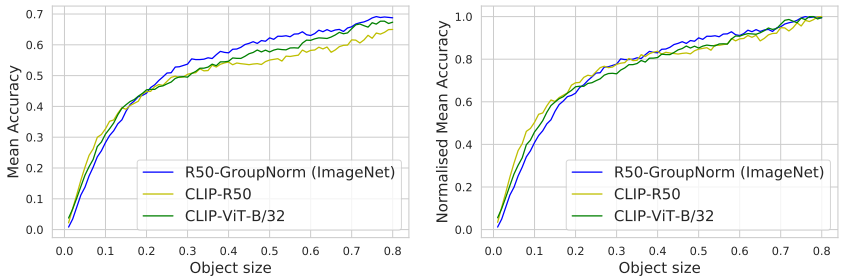

Figure 9: Note that the CLIP models have lower ImageNet accuracy than the ResNet-50 model. We used the ResNet-50 model because were not able to find standard models with lower ImageNet accuracy. Given the large difference in ImageNet accuracy between these three models, we plot the normalised accuracy (accuracy as a percentage of the highest accuracy per model across all sizes) as well (right). The plots suggest that the CLIP models may be slightly more robust than the vanilla ResNet-50 on small objects, and the vanilla ResNet-50 may be more robust on medium-sized objects.

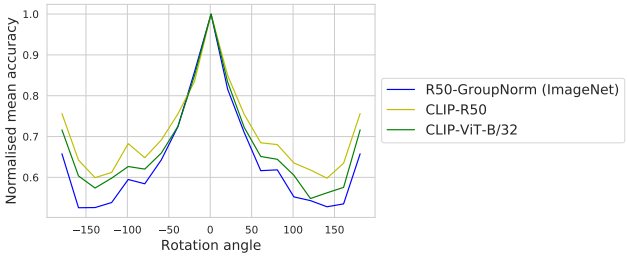

Figure 10: The CLIP models are more robust to changes in object rotation angle than the ResNet 50, with the CLIP-R50 model being more robust than the CLIP-ViT-B/32 model.