Are We Viewing the Problem of Robust Generalisation through the Appropriate Lens?

Anonymous Authors[1]

Abstract

We discuss different approaches to the challenge of robust object recognition under distribution shifts. We advocate a view of this challenge that is more closely informed by the problem of visual recognition, and which emphasizes dynamic model behaviour as opposed to centering the statistical properties of training and test distributions. We introduce an experimental setting geared towards developing models that can exhibit robust behaviour in a reliable and scalable manner. We refer to this setting as "systematic robustness", which involves excluding certain combinations of classes and image attributes are systematically during training. Unlike prior work which studies systematic generalisation in DNNs or their susceptibility to spurious correlations, we use synthetic operations and data sampling to scale such experiments up to large-scale naturalistic datasets.

1. Introduction

Automating visual perception is immensely hard – not least because the world is complex and ever-changing (Raji et al., 2020), constantly throwing up new and unlikely objects and scenes to recognise and handle. The problem of recognising objects with improbable properties – such as an unusual location (Alcorn et al., 2019) or style of appearance (Hendrycks et al., 2021) – is often subsumed under the more general problem of making predictive models robust to "distribution shifts" or to "out-of-distribution" inputs. This framing foregrounds the difference in input statistics between a model's training and deployment conditions. Variants of this general problem, such as domain generalisation and subpopulation shift (Gulrajani & Lopez-Paz, 2021; Santurkar et al., 2021; Koh et al., 2021), are then distinguished by the difference – in statistical terms – that arises after training is completed.

For example, we are faced with a domain generalisation problem when the respective marginal distributions of some property measured over training and evaluation sets have non-overlapping mass. If we consider, say, the property "depiction style", objects might be depicted using natural images or paintings during training, and during evaluation using line drawings (Peng et al., 2019). The goal is then to learn representations from the set of training "domains" or "environments" that are useful in distinct environments held-out for evaluation.

Often, *which* property changes from one environment to the other is of secondary importance; it is merely assumed that *some* property changes. Recent benchmarks reflect this abstract statistical view, collating various datasets that fit a specific problem template but which target diverse challenges in visual recognition (Gulrajani & Lopez-Paz, 2021) and even in other modalities beyond the visual (Koh et al., 2021). Methods for handling certain types of distribution shifts are expected to be agnostic to what exactly changes between environments, whether it's the viewpoint (Ghifary et al., 2015), the imaging sensor and illumination condition (Beery et al., 2018), or the time and location (David et al., 2020). Accordingly, such methods typically rely on some general objective or loss function designed to encourage robustness, e.g. via learning an invariant predictor (Arjovsky et al., 2019) or via aligning the feature statistics across environments (Li et al., 2018).

In this short paper, which describes ongoing work, we seek to question this problem framing and its underlying assumptions: that we should *primarily* view problems of robust prediction through an abstract statistical lens, and – in the case of domain generalisation specifically – that we can learn representations from arbitrary environments that will usefully carry over to others. Instead, we propose an alternate view of robust recognition and an accompanying experimental setting that we believe can help overcome the limitations of the problem-agnostic approach. We will also argue that it is not sufficient to merely design methods and benchmarks that address specific types of robustness, e.g. to image corruptions (Hendrycks & Dietterich, 2019) or adversarial perturbations (Croce et al., 2021), but that we need to rethink current training and evaluation protocols intended for such purposes to ensure that they result in a

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

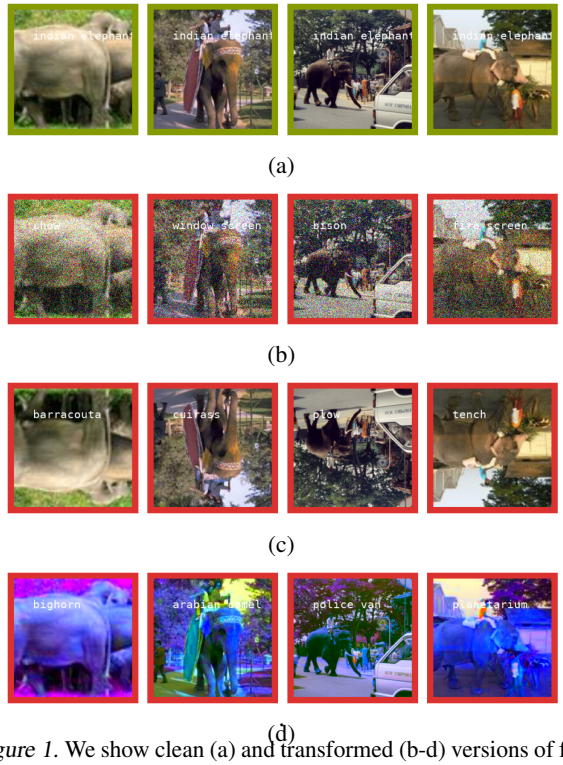


Figure 1. We show clean (a) and transformed (b-d) versions of four images from the *ImageNet* validation set belonging to the class "indian elephant", together with the corresponding predictions by a pre-trained ResNet-50. The model makes incorrect predictions for the transformed images.

reliable and scalable form of robustness.

2. What Does Robust Recognition Entail?

Before introducing our proposed setting, it is useful to consider the task of recognition in computational terms (Marr & Poggio, 1976). We can view recognition as requiring us to establish the similarity between a novel input and past experience (Ullman, 1996). Given (i) a closed set of classes C, (ii) a model trained to recognise objects belonging to any class $c \in C$, and (iii) a novel input image $\mathbf{x}$, said model needs to extract some representation of $\mathbf{x}$ as a first step. If this representation is "closest" to the stored representation(s) of some class c_k than to those of other classes, then we assign to $\mathbf{x}$ the label c.

Consider the images in Figure 1(a). These are images that a ResNet-50 model trained on *ImageNet* has no trouble recognising as depicting elephants. Rows 1(b)–1(d) contain mis-classified versions of the same images that are transformed in some manner: by applying Gaussian noise, by rotating the image, or by distorting the colours of the image. Evidently, the model extracts a representation from these inputs that is more similar to internal representations of other object classes. What then would we have to change about our model to obtain correct predictions?

We could simply treat this as a data problem. The training distribution does not contain images of purple elephants, or noisy images, and these are transformations that can be simulated, so why not modify the training distribution? Many successful approaches to robust recognition in fact rely on reshaping the training distribution, e.g by including difficult examples we might encounter during deployment through augmentation (Madry et al., 2018; Hendrycks et al., 2020; Cubuk et al., 2020), or by rebalancing the training set to emphasize under-represented examples (Sagawa et al., 2020; Idrissi et al., 2021). Is this scalable though? Can we anticipate every possible test-time image? What about transformations that cannot be easily simulated, or examples that are not represented at all in the training data? Furthermore, data augmentation schemes – whether hand-crafted or learned – are applied with equal probability to every image and thus every class. There are limitations to how much robustness this can foster to nuisance factors (Bouchacourt et al., 2021), but we also can't balance every relevant factor in the dataset, and will inevitably have factors that are spuriously correlated.

We could alternatively focus on learning representations with desirable properties, e.g. representations that are invariant across environments: in this case invariant to noise, rotations, and colour. Many approaches to domain generalisation rely on training schemes or loss terms that seek to encourage such invariances (Li et al., 2018; Arjovsky et al., 2019). Much work in self-supervised learning also aims to learn representations invariant to a pre-defined set of image transformations (Bardes et al., 2021). Is this however a sensible goal? To recognise the images in Figure 1(d) we might want to ignore the misleading colour, but do we want to discount colour cues in general? What information about shape would be preserved by a rotation-invariant representation?

Instead, we think it is beneficial to view recognition as a necessarily dynamic process that is contingent on the characteristics of any given input. For the images in Figure 1(b), removing the noise would most likely result in correct predictions. On the other hand, in an image of an old broken TV set displaying static, Gaussian noise might be useful signal. To recognise the next set of images (1(c)), we would benefit from compensating for a specific image rotation ($180°$). This could bring these images into alignment with a canonical representation of shape, but would be superfluous or even harmful for the first set of examples. For the images in Figure 1(d), we would have to correct for the unusual colours, colours that in other cases might help us distinguish objects for which colour is the salient feature.

This view of recognition that is centered around how an input image needs to be processed has a few useful implications for robustness. First of all, it suggests an approach to evaluation that focuses on desirable model behaviours:

Can a model learn to compensate reliably for things like noise, rotation, or colour? Do these behaviours generalise in a non-trivial manner? Secondly, it suggests a role for causal inference in recognition – not merely during training (Zhang et al., 2021) – but at evaluation time: We need to infer the transformations that would bring some input image into alignment with its class representation and apply these transformations accordingly.

In this work, we focus on the first point: How do we ensure that robust behaviour can generalise reliably in a non-trivial manner?

3. Systematic Robustness: Problem Setting

There is widespread interest in developing *robust* recognition models, that can recognise a variety of object classes despite the presence of unlikely object, scene, or image properties. This is reflected by the proliferation of challenging benchmarks that focus on various kinds of robust recognition behaviour, e.g. robustness to image corruptions (Hendrycks & Dietterich, 2019), to spatial transformations (Engstrom et al., 2019), to harsh weather conditions (Sakaridis et al., 2018), to imperceptible perturbations (Gilmer et al., 2018), and to abstract depictions of objects (Rusak et al., 2021). While it is hard to formulate a compact definition of robustness that covers these diverse requirements, we can nonetheless impose a useful meta-requirement that we should pursue in general: For robust behaviour to be *scalable*, it should transfer flexibly across familiar object classes, and not be separately learned for every class of interest. We refer to this problem setting as **systematic robustness**.

As argued in the previous section, in any realistic training set, certain scene factors will inevitably be correlated. Some of these correlations will be salient – e.g. elephants are typically grey, but some of them will be spurious: We typically encounter and thus photograph elephants during the day but they nontheless continue to exist at night. We need models that can learn from correlated data and ignore correlations as needed – whether salient or spurious. A self-driving car for example needs to reliably recognise unlikely obstacles, independently of their distribution in the training data.

While there is extensive work on related problems, e.g. systematic generalisation (Bahdanau et al., 2019; Ruis et al., 2020; Montero et al., 2021; Schott et al., 2021), spurious correlations (Geirhos et al., 2020), experiments are typically conducted on small datasets: Either synthetic ones where we have full control over the data generation process and can correspondingly work with tightly-controlled train/test splits, or naturalistic datasets with limited variability but which nonetheless admit some form of control. Here we aim to bridge the gap between work on controlled systematic generalisation and work on large-scale recognition models, in which models are trained on large naturalistic datasets with an "almost anything goes" approach. We also depart from related work on domain generalisation (Arjovsky et al., 2019; Gulrajani & Lopez-Paz, 2021), by acknowledging the need to learn specific robust behaviours from data: We are for example unlikely to be able to learn how to handle blurry images by only looking at clean and noisy images. These are perturbations with very different frequency characteristics, and it's not obvious that this should work without the appropriate evaluation. Our experiments also complement work on domain adaptation under label shift (Johansson et al., 2019; Zhao et al., 2019; Ben-David et al., 2010).

We formulate a variety of challenges probing systematic robustness w.r.t. (1) photometric transforms, (2) geometric transforms, (3) rendition styles. Our proposed benchmarks are a compromise between a tightly-controlled but overly simplistic setups and large-scale datasets without meaningful experimental controls, and they build on top of existing data (Russakovsky et al., 2015; Hendrycks & Dietterich, 2019; Peng et al., 2019). See Figure 2 for an example.

We now introduce the general problem setting of systematic robustness, describe a simple metric for measuring it, and then describe different instantiations of the problem that we examine in this paper.

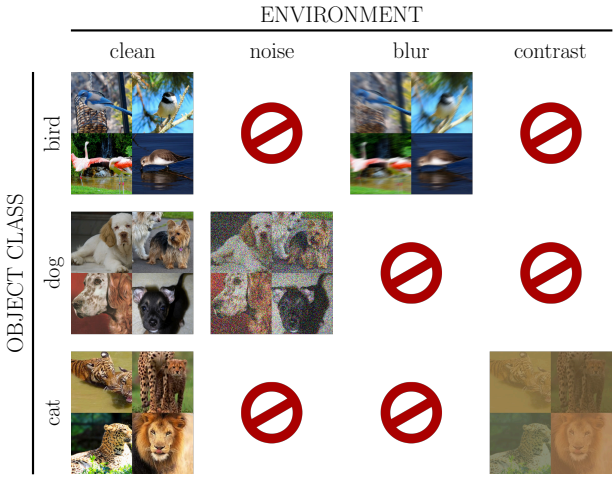


Figure 2. **Systematic Robustness:** We study the task of image classification in the presence of unlikely image properties, e.g. image corruptions, but under strict experimental controls: We systematically exclude certain combinations of class and property from the training data while ensuring that each is individually represented. At test time, we consider all possible combinations, seen and unseen. The goal is to encourage flexible robust behaviour that transfers in a non-trivial manner.

3.1. Preliminaries

In this paper, we focus on the task of image classification. Our goal is to train a classifier to recognise object

classes from some closed set $\mathrm{C} = \{\mathrm{c}_i\}_{i=1}^{N}$ given a training set of images annotated with class labels. The training data is drawn from a set of "environments" or "domains" $\mathcal{E} = \{\mathrm{E}_0, \ldots, \mathrm{E}_M\}$. Each environment $\mathrm{E}_k = \mathrm{T}_k \times \mathrm{C}_k$ can be characterised by a set of classes C_k and some property T_k that applies only to images from that environment, e.g. "contain Gaussian noise", " are rotated by 90° ", and "objects are depicted with line drawings". E_0 is the default environment in which all classes C are present. The remaining training environments are restricted to disjoint subsets of classes $\{\mathrm{C}_1, \ldots, \mathrm{C}_M\}$, s.t.: (1) $\bigcup_{k=1}^{M} \mathrm{C}_k \subseteq \mathrm{C}$, and (2) $\forall (i,j) \in \{1, \ldots, M\} : i \neq j \Rightarrow \mathrm{C}_i \cap \mathrm{C}_j = \emptyset$.

At test time, we consider all possible environments: $\{\mathrm{T}_0 \times \mathrm{C}\} \cup \{\mathrm{T}_i \times \mathrm{C}_j \mid i,j \in \{1, \ldots, M\}\}$. Thus for any valid i, j, if $i \neq j$ then $\mathrm{T}_i \times \mathrm{C}_j$ has not been seen during training. This allows us to contrast performance in seen and unseen environments, and determine how well robust behaviour learned for a subset of available classes transfers to the rest. See Figure 2 for an illustration.

3.2. Measuring Systematic Robustness

To measure systematic robustness of a model f w.r.t. some scene factor T_k, we first need to measure classification accuracy $\mathrm{acc}^f(\cdot)$ in two environments: (1) the control environment $\mathrm{T}_k \times \mathrm{C}_k$, i.e. the set of classes C_k for which robustness to T_k was learned during training, and (2) the experimental environment $\mathrm{T}_k \times \mathrm{C}_{\overline{k}}$, i.e. the environment in which the property T_k holds for classes $\mathrm{C}_{\overline{k}} = \mathrm{C} \setminus \mathrm{C}_k$. We can then simply normalise the accuracy in the experimental environment $\mathrm{acc}^f(\mathrm{T}_k \times \mathrm{C}_{\overline{k}})$ by the accuracy in the control environment $\mathrm{acc}^f(\mathrm{T}_k \times \mathrm{C}_k)$. This represents the degree to which robust behaviour learned in a seen environment transfers to an unseen one. Sometimes it will make sense to compensate for some baseline robustness, e.g. if we are fine-tuning a pretrained model f_0 that already shows some degree of robust accuracy. We simply subtract average robust accuracy of the baseline model $\mathrm{acc}^{f_0}(\mathrm{T}_k \times \mathrm{C})$ from the accuracy measured in both environments. We refer to the resulting quantity as ρ^k, i.e. systematic robustness w.r.t. $\mathrm{T}_{scenefac}$, and compute it as follows while limiting the range to $[0, 1]$:

$$\rho^k = \min(1,\ \frac{\max(0,\ \mathrm{acc}^f(\mathrm{T}_k \times \mathrm{C}_{\overline{k}}) - \mathrm{acc}^{f_0}(\mathrm{T}_k \times \mathrm{C}))}{\min(1,\ \mathrm{acc}^f(\mathrm{T}_k \times \mathrm{C}_k) - \mathrm{acc}^{f_0}(\mathrm{T}_k \times \mathrm{C}))}) \tag{1}$$

It should be noted that a value of 1 is trivial to achieve if f is not much more robust than the reference classifier f_0. Thus on its own, this measure of robustness is insufficient and has to be paired with the other metrics that measure absolute robust accuracy. A related metric is that of *effective robustness* (Taori et al., 2020).

3.3. Problem Instantiations

We opt to study three instantiations of this problem: systematic robustness w.r.t. (1) image corruptions (e.g. noise, blur) (Hendrycks & Dietterich, 2019), (2) in-plane image rotations (Engstrom et al., 2019), and (3) rendition styles (e.g. natural images, line drawings) (Hendrycks et al., 2021; Rusak et al., 2021). These represent basic challenges in recognition that robust models should arguably handle in a flexible manner. They also differ from one another qualitatively: The first are photometric perturbations that can drastically affect image statistics, whereas recognising objects from different viewpoints can require reasoning of a more geometric nature. While the way an object is depicted can also heavily affect the statistics of the corresponding image, successfully managing this challenge requires non-trivial abstract reasoning about the essence of an object's texture and shape. Most importantly, we can study this problems in the large-scale setting building on prior efforts to collect and generate the relevant data. We can study systematic robustness w.r.t. corruptions and rotations by applying selective data augmentation to *ImageNet* (Russakovsky et al., 2015; Hendrycks & Dietterich, 2019), and for rendition styles, we can sample images from *DomainNet* (Peng et al., 2019) as required. For the image corruptions we select four out of the 19 suggested for the *ImageNet-C* benchmark, each representing one of four categories: noise ($\mathcal{N}$), blur ($\mathcal{B}$), digital ($\mathcal{D}$), and ($\mathcal{W}$). More details can be found in the appendix.

4. Summary of Ongoing Experiments

Our findings so far can be summarised as follows (see Appendix A for selected results): We train a standard CNN (He et al., 2016) from scratch on *ImageNet* with the default "clean" data as well as data that is selectively transformed to exclude certain combinations of class and image corruption – or alternatively class and in-plane image rotation. We find that the network picks up on image corruptions as a spurious feature resulting in a huge gap between seen and unseen combinations (Tab. 1). The difference when it comes to image rotations is much less dramatic (Figure 4) suggesting that a unified approach to both will have limitations. We find that the learning rate is a critical hyperparameter for avoiding reliance on spurious correlations, which is why fine-tuning a network pretrained on clean data results in more – albeit unstable – systematic robustness (Figure 3). We find that the class composition of the different environments matters for systematic robustness (Tab. 2), i.e. whether these contain groups of fine-grained classes or not. We also find that popular domain generalisation methods result in improved systematic robustness for corruptions but in limited improvements for rotations (Appendix A.3) and non-existent improvements for alternate renditions (not included). Aligning a subset of the available environments only results in robustness to those environments Tab. 3.

References

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W., and Nguyen, A. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4845–4854. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00498. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Alcorn_Strike_With_a_Pose_Neural_Networks_Are_Easily_Fooled_by_CVPR_2019_paper.html.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019. URL http://arxiv.org/abs/1907.02893.

Bahdanau, D., Murty, S., Noukhovitch, M., Nguyen, T. H., de Vries, H., and Courville, A. C. Systematic generalization: What is required and can it be learned? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HkezXnA9YX.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *CoRR*, abs/2105.04906, 2021. URL https://arxiv.org/abs/2105.04906.

Beery, S., Horn, G. V., and Perona, P. Recognition in terra incognita. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*, pp. 472–489. Springer, 2018. doi: 10.1007/978-3-030-01270-0_28. URL https://doi.org/10.1007/978-3-030-01270-0_28.

Ben-David, S., Lu, T., Luu, T., and Pál, D. Impossibility theorems for domain adaptation. In Teh, Y. W. and Titterington, D. M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pp. 129–136. JMLR.org, 2010. URL http://proceedings.mlr.press/v9/david10a.html.

Bouchacourt, D., Ibrahim, M., and Morcos, A. S. Grounding inductive biases in natural images: invariance stems from variations in data. *CoRR*, abs/2106.05121, 2021. URL https://arxiv.org/abs/2106.05121.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/a3c65c2974270fd093ee8a9bf8ae7d0b-Abstract-round2.html.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. Randaugment: Practical automated data augmentation with a reduced search space. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html.

David, E., Madec, S., Sadeghi-Tehran, P., Aasen, H., Zheng, B., Liu, S., Kirchgeßner, N., Ishikawa, G., Nagasawa, K., Badhon, M. A., Pozniak, C., de Solan, B., Hund, A., Chapman, S. C., Baret, F., Stavness, I., and Guo, W. Global wheat head detection (GWHD) dataset: a large and diverse dataset of high resolution RGB labelled images to develop and benchmark wheat head detection methods. *CoRR*, abs/2005.02162, 2020. URL https://arxiv.org/abs/2005.02162.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., and Madry, A. Exploring the landscape of spatial robustness. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1802–1811. PMLR, 2019. URL http://proceedings.mlr.press/v97/engstrom19a.html.

Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R. S.,

Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Ghifary, M., Kleijn, W. B., Zhang, M., and Balduzzi, D. Domain generalization for object recognition with multi-task autoencoders. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2551–2559. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.293. URL https://doi.org/10.1109/ICCV.2015.293.

Gilmer, J., Adams, R. P., Goodfellow, I. J., Andersen, D. G., and Dahl, G. E. Motivating the rules of the game for adversarial example research. *CoRR*, abs/1807.06732, 2018. URL http://arxiv.org/abs/1807.06732.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=S1gmrxHFvB.

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. The many faces of robustness: A critical analysis of out-of-distribution generalization. pp. 8320–8329, 2021. doi: 10.1109/ICCV48922.2021.00823. URL https://doi.org/10.1109/ICCV48922.2021.00823.

Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. Simple data balancing achieves competitive worst-group-accuracy. *CoRR*, abs/2110.14503, 2021. URL https://arxiv.org/abs/2110.14503.

Johansson, F. D., Sontag, D. A., and Ranganath, R. Support and invertibility in domain-invariant representations. In Chaudhuri, K. and Sugiyama, M. (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 527–536. PMLR, 2019. URL http://proceedings.mlr.press/v89/johansson19a.html.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL http://proceedings.mlr.press/v139/koh21a.html.

Li, H., Pan, S. J., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 5400–5409. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00566. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Li_Domain_Generalization_With_CVPR_2018_paper.html.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Marr, D. and Poggio, T. From understanding computation to understanding neural circuitry. *AI Memo*, 1976. URL http://mit.dspace.org/handle/1721.1/5782.

Montero, M. L., Ludwig, C. J., Costa, R. P., Malhotra, G., and Bowers, J. The role of disentanglement in generalisation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=qbH974jKUVy.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain

adaptation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 1406–1415. IEEE, 2019. doi: 10.1109/ICCV.2019.00149. URL https://doi.org/10.1109/ICCV.2019.00149.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. Ai and the everything in the whole wide world benchmark. In *ML-Retrospectives, Surveys & Meta-Analyses @ NeurIPS 2020 Workshop*, 2020. URL https://ml-retrospectives.github.io/neurips2020/camera_ready/18.pdf.

Ruis, L., Andreas, J., Baroni, M., Bouchacourt, D., and Lake, B. M. A benchmark for systematic generalization in grounded language understanding. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e5a90182cc81e12ab5e72d66e0b46fe3-Abstract.html.

Rusak, E., Schott, L., Zimmermann, R. S., Bitterwolf, J., Bringmann, O., Bethge, M., and Brendel, W. A simple way to make neural networks robust against diverse image corruptions. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J. (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pp. 53–69. Springer, 2020. doi: 10.1007/978-3-030-58580-8_4. URL https://doi.org/10.1007/978-3-030-58580-8_4.

Rusak, E., Schneider, S., Gehler, P. V., Bringmann, O., Brendel, W., and Bethge, M. Adapting imagenet-scale models to complex distribution shifts with self-learning. *CoRR*, abs/2104.12928, 2021. URL https://arxiv.org/abs/2104.12928.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Sakaridis, C., Dai, D., and Gool, L. V. Semantic foggy scene understanding with synthetic data. *Int. J. Comput. Vis.*, 126(9):973–992, 2018. doi: 10.1007/s11263-018-1072-8. URL https://doi.org/10.1007/s11263-018-1072-8.

Santurkar, S., Tsipras, D., and Madry, A. BREEDS: benchmarks for subpopulation shift. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=mQPBmvyAuk.

Schott, L., von Kügelgen, J., Träuble, F., Gehler, P. V., Russell, C., Bethge, M., Schölkopf, B., Locatello, F., and Brendel, W. Visual representation learning does not generalize strongly within the same domain. *CoRR*, abs/2107.08221, 2021. URL https://arxiv.org/abs/2107.08221.

Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., and Schmidt, L. Evaluating machine accuracy on imagenet. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8634–8644. PMLR, 2020. URL http://proceedings.mlr.press/v119/shankar20c.html.

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. How to train your vit? data, augmentation, and regularization in vision transformers. *CoRR*, abs/2106.10270, 2021. URL https://arxiv.org/abs/2106.10270.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d8330f857a17c53d217014ee776bfd50-Abstract.html.

Ullman, S. *High-Level Vision: Object Recognition and Visual Cognition*. MIT Press, 1996. ISBN 978-0-262-71007-7.

Zhang, Y., Gong, M., Liu, T., Niu, G., Tian, X., Han, B., Schölkopf, B., and Zhang, K. Adversarial robustness through the lens of causality. *CoRR*, abs/2106.06196, 2021. URL https://arxiv.org/abs/2106.06196.

Zhao, H., des Combes, R. T., Zhang, K., and Gordon, G. J. On learning invariant representations for domain adaptation. In Chaudhuri, K. and Salakhutdinov, R. (eds.),

Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of *Proceedings of Machine Learning Research*, pp. 7523–7532. PMLR, 2019. URL http://proceedings.mlr.press/v97/zhao19a.html.

Training Environment				Top-1 Accuracy																	
				$(\mathrm{T}_0 \times \cdot)$					$(\mathrm{T}_\mathcal{N} \times \cdot)$			$(\mathrm{T}_\mathcal{B} \times \cdot)$			$(\mathrm{T}_\mathcal{W} \times \cdot)$			$(\mathrm{T}_\mathcal{D} \times \cdot)$			$(\mathrm{T}_\mathcal{A} \times \cdot)$
$\mathrm{E}_\mathcal{N}$	$\mathrm{E}_\mathcal{B}$	$\mathrm{E}_\mathcal{D}$	$\mathrm{E}_\mathcal{W}$	C	$\mathrm{C}_\mathcal{N}$	$\mathrm{C}_\mathcal{B}$	$\mathrm{C}_\mathcal{W}$	$\mathrm{C}_\mathcal{D}$	C	$\mathrm{C}_\mathcal{N}$	$\mathrm{C}_{\overline{\mathcal{N}}}$	C	$\mathrm{C}_\mathcal{B}$	$\mathrm{C}_\mathcal{B}$	C	$\mathrm{C}_\mathcal{W}$	$\mathrm{C}_{\overline{\mathcal{W}}}$	C	$\mathrm{C}_\mathcal{D}$	$\mathrm{C}_{\overline{\mathcal{D}}}$	C
-	-	-	-	76	76	78	76	75	34	-	-	41	-	-	36	-	-	47	-	-	38
✓	-	-	-	77	74	78	77	76	38	79	25	41	-	-	33	-	-	47	-	-	40
✓	✓	-	-	76	75	76	78	77	39	79	25	36	81	21	34	-	-	47	-	-	39
✓	✓	✓	✓	76	76	78	76	75	38	80	24	37	82	22	41	82	27	50	85	38	41

Table 1. Top-1 accuracy for four ResNet-50 models trained from scratch on *ImageNet*-1K with various data augmentation settings. We split the 1000 classes into four disjoint sets of 250 classes each, that are separately augmented as indicated. We observe that average robustness to corruptions (indicated by $\mathrm{T}_\mathcal{A} \times \mathrm{C}$) tends to increase relative to the baseline, but that the image corruptions are used as a spurious feature. Average clean performance ($\mathrm{T}_0 \times \mathrm{C}$) remains stable.

A. Appendix

A.1. Motivating Experiment

As a first exploration of this setting we conduct the following experiment: We define five environments $\{\mathrm{E}_0, \mathrm{E}_\mathcal{N}, \mathrm{E}_\mathcal{B}, \mathrm{E}_\mathcal{D}, \mathrm{E}_\mathcal{W}\}$. The default environment $\mathrm{E}_0 = \mathrm{T}_0 \times \mathrm{C}$ includes all 1000 *ImageNet* classes and consists of the original, untransformed images. All classes are equally distributed among the remaining environments resulting in class sets $\{\mathrm{C}_\mathcal{N}, \mathrm{C}_\mathcal{B}, \mathrm{C}_\mathcal{W}, \mathrm{C}_\mathcal{D}\}$ and corresponding image corruptions $\{\mathrm{T}_\mathcal{N}, \mathrm{T}_\mathcal{B}, \mathrm{T}_\mathcal{W}, \mathrm{T}_\mathcal{D}\}$. Images from all environments additionally undergo simple geometric transformations during training including resizing, cropping, and horizontal flipping

We train ResNet-50 models with cross-entropy loss for 90 epochs with four different configurations: a baseline with data only from the default environment, and three other models with data from either one, two, or four additional environment(s). In all cases, we sample data from the default environment with a probability of $p = 0.75$, and otherwise sample uniformly from the rest.

This is a simple baseline, often referred to in the domain generalisation literature as *ERM* (empirical risk minimisation), where risk is taken to mean the classification loss. Data from all environments is effectively treated as a single unified training set. This baseline has been shown to be as resilient to domain shifts as most specialised methods (Gulrajani & Lopez-Paz, 2021; Koh et al., 2021) and thus is a good starting point for our analysis.

We report top-1 accuracy for these different configurations in Tab. 1. We observe the following: Relative to the baseline, corruption robustness tends to improve on average. However, the large improvement on seen environments (e.g. $\mathrm{T}_\mathcal{N} \times \mathrm{C}_\mathcal{N}$) comes at the cost of reduced accuracy on unseen environments (e.g. $\mathrm{T}_\mathcal{N} \times \mathrm{C}_{\overline{\mathcal{N}}}$).

Does starting from a pre-trained model help? We start from a ResNet-50 pre-trained on data from the default environment $\mathrm{E}_0 = \mathrm{T}_0 \times \mathrm{C}$. We fine-tune this model with cross-entropy loss for 25 epochs using data from two environments: $\{\mathrm{E}_0, \mathrm{E}_\mathcal{N}\}$. Since the model has already been trained on uncorrupted data, we sample images from the default environment with a probability of $p = 0.5$. $\mathrm{E}_\mathcal{N}$ contains data for 25% of the classes and with a probability of $p = 0.5$ we apply Gaussian noise at one of five severity levels (Hendrycks & Dietterich, 2019). This is very similar to the *Gaussian Noise Training* baseline of (Rusak et al., 2020). During our initial search over the space of hyperparameters, we observed that the learning rate plays a key role above all others. We report the results of a learning rate sweep in Figure 3 together with two reference results: [I] the *ImageNet*-pretrained model, [II] a pre-trained ResNet-50 fine-tuned using unrestricted data augmentation with all available corruptions. We focus on accuracy on clean and noisy data and observe the following: Robustness to seen environments is at odds with performance on unseen environments. There is also a trade-off between achieving a small systematic robustness gap and good robust accuracy that doesn't exist when we apply noise augmentation to all classes. We conducted an identical learning rate sweep for the latter case, and found that performance is much less sensitive to the learning rate and monotonically increases with it for the entirety of the range we examined.

Corrupting an image introduces spurious but salient visual features that can easily be picked up by the model to distinguish classes from different environments. What happens we consider a transformation that does not affect the image statistics as dramatically? To this end, we fine-tune pre-trained ResNet-50 models with different learning rates on data from two environments: $\{\mathrm{E}_0, \mathrm{E}_\mathcal{R}\}$. Images from the second environment are rotated with a probability of $p = 0.5$ by an angle

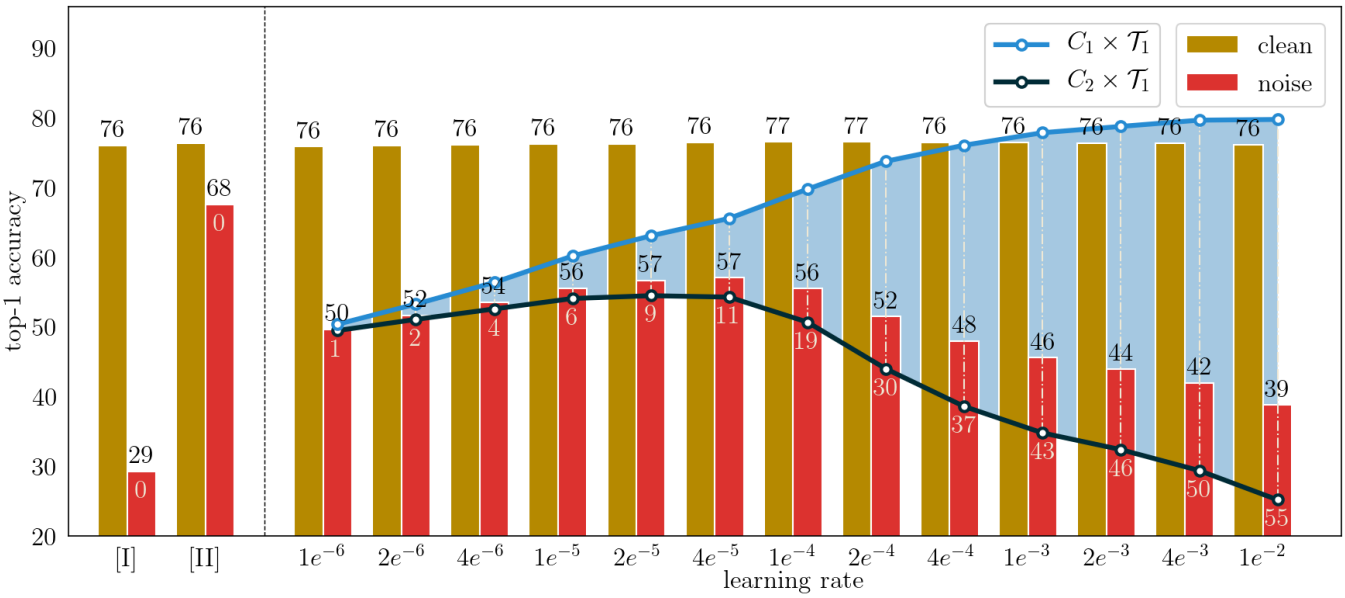


Figure 3. Clean and noise-robust accuracy of pre-trained ResNet-50 models after additional fine-tuning with selective Gaussian noise augmentation. Gaussian noise is applied with p=0.5 to 25% of the available classes. [I] is a reference pre-trained model, and [II] is a model fine-tuned with noise augmentation applied to every class. This illustrates the trade-off between in-domain and out-of-domain robustness.

sampled from the set $\{18°, 36°, 54°72°, 90°\}$. Since rotating an image by a right angle does not introduce any artifacts from interpolation, we measure robustness rotation only for $90°$. We plot the validation accuracy throughout training for three such models together with three models trained with selective noise augmentation. The results can be seen in Figure 4. We see a significant difference in behaviour between models trained for noise-robustness vs. models trained for rotation robustness.

A.2. Comparing across Class Sets and Models

So far we have examined a single model, and split classes randomly among different environments. (1) What happens when we consider different class splits, especially in light of the numerous fine-grained classes in *ImageNet*? (2) Do we observe different behaviour when considering a different class of models?

The size and diversity of *ImageNet* classes allows us to analyse the behaviour of the ERM baseline under various conditions, which we summarise in Tab. 2. We both increase and decrease the size of C_1 from 25% to 50% and 10% respectively. We also consider the smaller subsets `200adv` and `150det`. Finally, we can leverage the class structure of *ImageNet*, specifically the availability of a large number of fine-grained classes. This echoes the recommendation of Shankar et al. (2020), who recommend distinguishing between animate and inanimate objects for evaluation. We go further and examine the case where the training conditions for (mostly) animate objects depart from other classes. We find that for example when all the fine-grained classes are in the corruption set, that the robustness gap increases, also w.r.t. to the opposite case: when no super-classes are in the corruption set. For more details on how we construct these splits please consult the supplementary material.

Next we compare different models for both systematic robustness to corruptions and to rotations. Besides a ResNet-50, we compare two vision transformer models (Dosovitskiy et al., 2021) of varying sizes (ViT-S and ViT-B), as well as a CNN-Transformer hybrid (R26+ViT-S/32). Since attention-based models tend to require significant amounts of training data, we use models pre-trained on *ImageNet*-21k but without any augmentations that could confound our results (Steiner et al., 2021). For each type of model, we train two instances: one for corruptions and one for rotations. Results are summarised in Figure 5. As we've seen in the previous section, the learning rate is a critical parameter. We found that the optimal learning rate not only varies by model class, but also by scene factor. In terms of systematic robustness, there is no salient difference between model types. Flexible rotation robustness, however, appears to be harder to achieve than to corruptions.

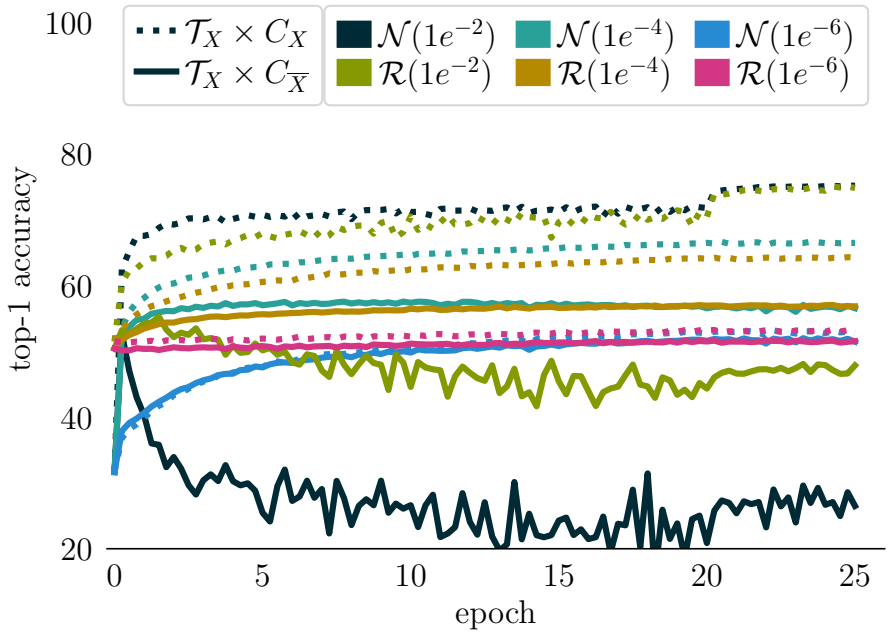


Figure 4. Validation accuracy for two sets of models trained either with selective noise ($\mathcal{N}$) or selective ($\mathcal{R}$) augmentation at different learning rates. We observe that with the models trained for noise robustness, a quick adaptation occurs in the early iterations, then noise is used as a spurious feature depending on the learning rate. This is different from the behaviour for the rotation-robust models, where performance on the unseen environment is largely unaffected by performance on seen ones.

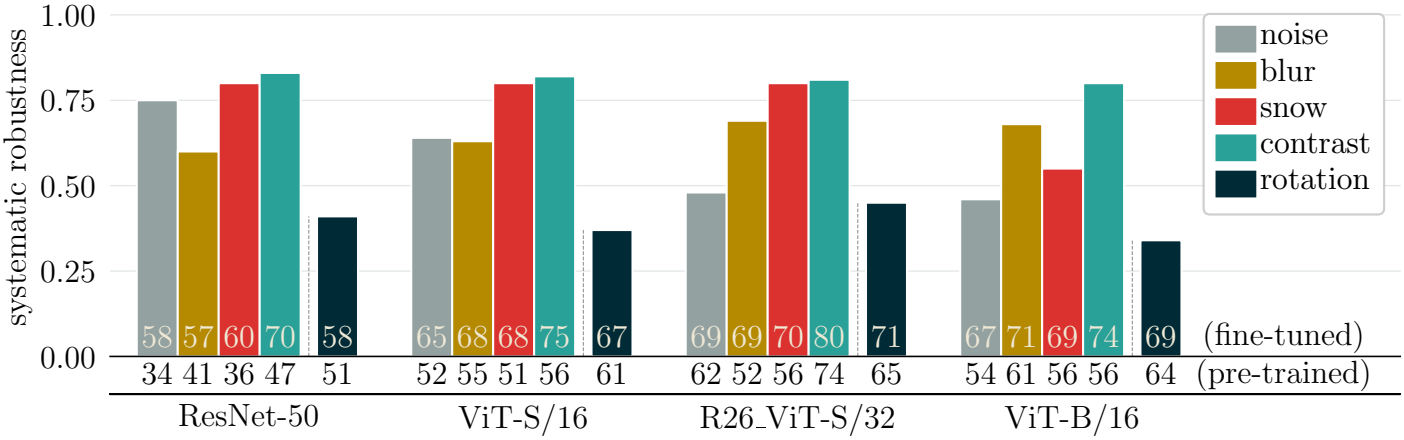
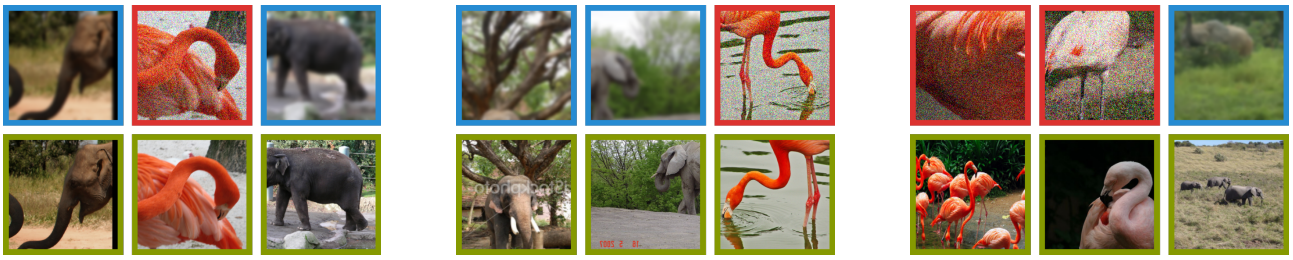


Figure 5. Systematic robustness of various models to corruptions and rotations. We report accuracy both before and after fine-tuning as well as systematic robustness to each scene factor separately. For each model class, we train one model for corruptions and another for rotations.



(a) pairs (same view) (b) pairs (same instance) (c) pairs (same class)

Figure 6. We experiment with three different schemes for constructing image pairs for feature alignment objectives. This shows a simple dataset with three environments: clean (green), noisy (red), and blurry (blue). In the first setting, we pair copies of the same image from different environments, in the second we sample a different view of the same instance, and in the third we pair images from the same class.

A.3. Feature Alignment: Corruptions vs. Rotations

Training Settings		**Performance**				
Dataset	$\lvert\mathbf{C}_\mathcal{N}\rvert/\lvert\mathbf{C}\rvert$	$(\mathrm{T}_0 \times \cdot\,)$ C	$(\mathrm{T}_\mathcal{N} \times \cdot\,)$ C	$\mathrm{C}_\mathcal{N}$	$\mathrm{C}_{\overline{\mathcal{N}}}$	$\rho^\mathcal{N}$
150det	Ø	80	29	29	30	-
	0.5	78	61	70	53	0.59
	1.0	79	71	71	71	-
200adv	0.0	88	40	40	40	-
	0.5	88	73	79	66	0.66
	1.0	88	81	81	81	-
1000cls	0.0	76	32	32	32	-
	0.1	77	51	69	49	0.46
	0.25	76	57	66	54	0.69
	0.5	76	61	64	57	0.78
	0.5 (all superc.)	76	60	68	52	0.56
	0.5 (no superc.)	76	60	57	63	0.81
	1.0	76	68	67	68	-

Table 2. Top-1 accuracy measured on the *ImageNet* and *ImageNet-C* validation sets, for different class splits together with combinatorial robustness measure ρ as applicable. Interestingly, the class split appears to have a significant effect on transfer ability.

We experimented with several domain adaptation and generalisation methods from the literature, but found that one class of methods was able to improve over the simple ERM baseline. These methods are based on aligning features across environments or instances by means of an appropriate loss term. These are also relevant in "self-supervised" learning, where the goal is to extract representations that are invariant to a predefined set transformations that normally preserve label semantics, such as changes in scale or illumination. We experiment with three different schemes to construct image pairs depicted in Appendix A.3: (1) pair_{view} results in pairs that differ only in a single property T_k e.g. clean vs. noisy, (2) $\text{pair}_{instance}$ samples two different crops from the same image before applying T_k to one of them, and (3) pair_{class} samples two images from the same class. This give us a more fine-grained picture of any improvements to systematic robustness. In addition to a simple classification loss (**CLS**) we consider three different loss terms: (1) **COS**, a simple cosine similarity term applied to the final layer embeddings of an image pair, (2) **VICReg** (Bardes et al., 2021), which is a recent loss term used for self-supervised learning that adds variance and covariance terms to an invariance loss similar to **COS**, and (3) **MMD** (Li et al., 2018) which is a domain adaptation method for aligning the statistics across different environments. Unlike the first two, the latter does not require aligned image pairs. The results can be seen in Appendix A.3. What we can observe is that the different objective functions do not improve significantly over the baseline trained with just classification loss. We should note that we reduced the learning rate for **CLS** in keeping with our findings from earlier sections. What also sticks out is that the alignment losses only improve over **CLS** in the pair_{view} setting, i.e. when the images are perfectly aligned, suggesting that the network perhaps is learning to compensate for specific corruptions. To follow up on this, we trained ResNet-18 models on a subset of *ImageNet*, applying the cosine similarity loss to certain environments (see Tab. 3). Perhaps surprisingly, even when we align three out of four environments to the default one, there is no benefit to the remaining one.

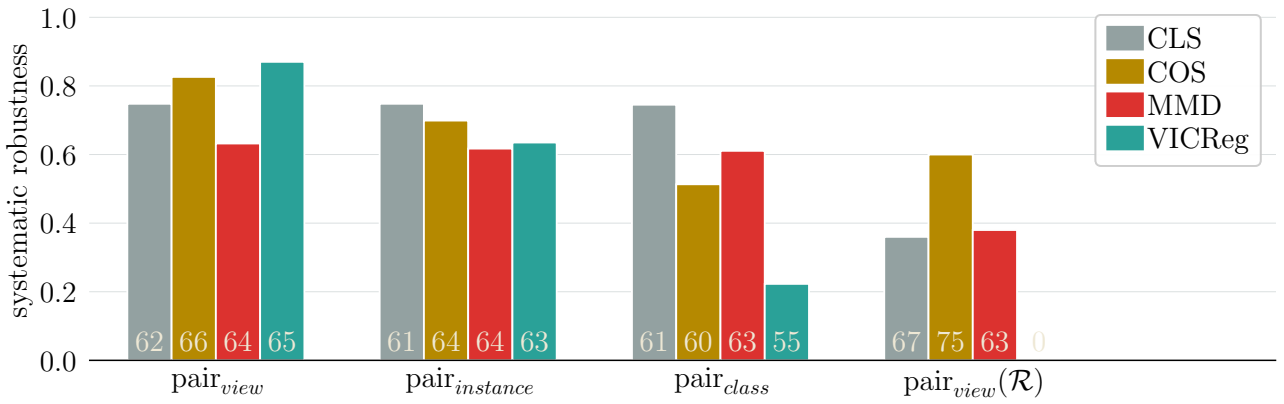


Figure 7. Various feature alignment objectives applied to both models trained or corruption robustness (three groups on the left) and rotation robustness. The alignment objectives improve over the vanilla classification loss only when we have the same view of the same instance.

Objective	**Cosine Similarity**											
Aligned to $\mathcal{C}$	Noise ($\mathcal{N}$)			Blur ($\mathcal{B}$)			Digital ($\mathcal{D}$)			Weather ($\mathcal{W}$)		
$\mathcal{N}$ $\mathcal{B}$ $\mathcal{W}$ $\mathcal{D}$	C_1	C_2	$\rho^{\mathcal{N}}$	C_1	C_2	$\rho^{\mathcal{B}}$	C_1	C_2	$\rho^{\mathcal{D}}$	C_1	C_2	$\rho^{\mathcal{W}}$
(pretrained)	29	26	-	43	44	-	32	25	-	32	31	-
- - - -	72	42	0.34	68	45	0.05	76	32	0.10	75	48	0.37
✓ ✓ ✓ ✓	67	60	**0.81**	64	57	**0.65**	72	52	**0.56**	71	64	**0.81**
✓ - - -	70	61	**0.80**	63	39	0.00	72	24	0.00	69	29	0.00
- ✓ - -	70	37	0.23	66	56	**0.54**	74	26	0.00	72	35	0.10
- - ✓ -	68	29	0.04	65	39	0.00	76	54	**0.56**	72	35	0.09
- - - ✓	70	33	0.13	66	37	0.00	73	27	0.00	76	70	**0.87**
✓ ✓ ✓ -	68	60	**0.81**	64	56	**0.61**	73	22	0.00	72	64	**0.82**
✓ ✓ - ✓	68	59	**0.79**	63	55	**0.61**	73	52	**0.54**	70	31	0.00
✓ - ✓ ✓	68	60	**0.80**	64	37	0.00	73	54	**0.58**	72	65	**0.83**
- ✓ ✓ ✓	68	25	0.00	64	56	**0.59**	74	52	**0.54**	72	65	**0.82**

Table 3. We apply an objective function for aligning features across environments selectively to different subsets of environments. What we see is that any improvements to robustness are limited to the the aligned environments. This suggests that we are learning specific invariances – or rather learning to compensate for specific corruptions – rather than learning domain-invariant features.