On the surprising similarities between supervised and self-supervised models

Anonymous Author(s)
Affiliation
Address
email

Abstract

How do humans learn to acquire a powerful, flexible and robust representation of objects? While much of this process remains unknown, it is clear that humans do not require millions of object labels. Excitingly, recent algorithmic advancements in self-supervised learning now enable convolutional neural networks (CNNs) to learn useful visual object representations without supervised labels, too. In the light of this recent breakthrough, we here compare self-supervised networks to supervised models and human behaviour.

We tested models on 15 generalisation datasets for which large-scale human behavioural data is available (130K highly controlled psychophysical trials). Surprisingly, current self-supervised CNNs share four key characteristics of their supervised counterparts: (1.) relatively poor noise robustness (with the notable exception of SimCLR), (2.) non-human category-level error patterns, (3.) non-human image-level error patterns (yet high similarity to supervised model errors) and (4.) a bias towards texture. Taken together, these results suggest that the strategies learned through today's supervised and self-supervised training objectives end up being surprisingly similar, but distant from human-like behaviour. That being said, we are clearly just at the beginning of what could be called a self-supervised revolution of machine vision, and we are hopeful that future self-supervised models behave differently from supervised ones, and—perhaps—more similar to robust human object recognition.

1 Introduction

"If intelligence is a cake, the bulk of the cake is unsupervised learning, the icing on the cake is supervised learning and the cherry on the cake is reinforcement learning", Yann LeCun famously said [1]. Four years later, the entire cake is finally on the table—the representations learned via self-supervised learning now compete with supervised methods on ImageNet [2] and outperform supervised pre-training for object detection [3]. But given this fundamentally different learning mechanism, *how do recent self-supervised models differ from their supervised counterparts in terms of their behaviour?*

We here attempt to shed light on this question by comparing eight flavours of "cake" (`PIRL`, `MoCo`, `MoCoV2`, `InfoMin`, `InsDis`, `SimCLR-x1`, `SimCLR-x2`, `SimCLR-x4`) with 24 common variants of "icing" (from the `AlexNet`, `VGG`, `Squeezenet`, `DenseNet`, `Inception`, `ResNet`, `ShuffleNet`, `MobileNet`, `ResNeXt`, `WideResNet` and `MNASNet` star cuisines). Specifically, our culinary test buffet aims to investigate:

1. Are self-supervised models more robust towards distortions?
2. Do self-supervised models make similar errors as either humans or supervised models?
3. Do self-supervised models recognise objects by texture or shape?

Submitted to 2nd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM).

For all of these questions, we compare supervised and self-supervised[1] models against a comprehensive set of openly available human psychophysical data totalling over 130,000 trials [4, 5]. This is motivated on one hand by the fact that humans, too, rapidly learn to recognise new objects without requiring hundreds of labels per instance; and on the other hand by a number of fascinating studies reporting increased similarities between self-supervised models and human perception. For instance, Lotter et al. [6] train a model for self-supervised next frame prediction on videos, which leads to phenomena known from human vision, including perceptual illusions. Konkle and Alvarez [7] and Zhuang et al. [8] report an improved match with neural data; Zhuang et al. [8] also find more human-like error patterns for semi-supervised models and Storrs and Fleming [9] observe that a self-supervised network accounts for behavioural patterns of human gloss perception. While methods and models differ substantially across these studies, they jointly provide evidence for the intriguing hypothesis that self-supervised machine learning models may better approximate human vision.

2 Methods

Models. `InsDis` [10], `MoCo` [11], `MoCoV2` [12], `PIRL` [3] and `InfoMin` [13] were obtained as pre-trained models from the PyContrast model zoo. We trained one linear classifier per model on top of the self-supervised representation. A PyTorch [14] implementation of `SimCLR` [2] was obtained via simclr-converter. All self-supervised models use a `ResNet-50` architecture and a different training approach within the framework of contrastive learning [e.g. 15]. For supervised models, we used all 24 available pre-trained models from the PyTorch model zoo version 1.4.0 (`VGG`: with batch norm).
Datasets. Models were tested on 12 different image degradations from [4], as well as on texture-vs-shape datasets from [5]. Plotting conventions follow these papers (unless indicated otherwise).

3 Results

3.1 With the exception of SimCLR, supervised and self-supervised models show similar (non-human) out-of-distribution generalisation

Motivation. Given sufficient quantities of labelled training data, CNNs can learn to identify objects when the input images are noisy. However, supervised CNNs typically generalise poorly to novel distortion types not seen during training, so-called out-of-distribution images [4]. In contrast, human perception is remarkably robust when dealing with previously unseen types of noise. Given that recent self-supervised networks are trained to identify objects under a variety of transformations (like scaling, cropping and colour shifts), have they learned a more robust, human-like representation of objects, where high-level semantic content is unimpaired by low-level noise?

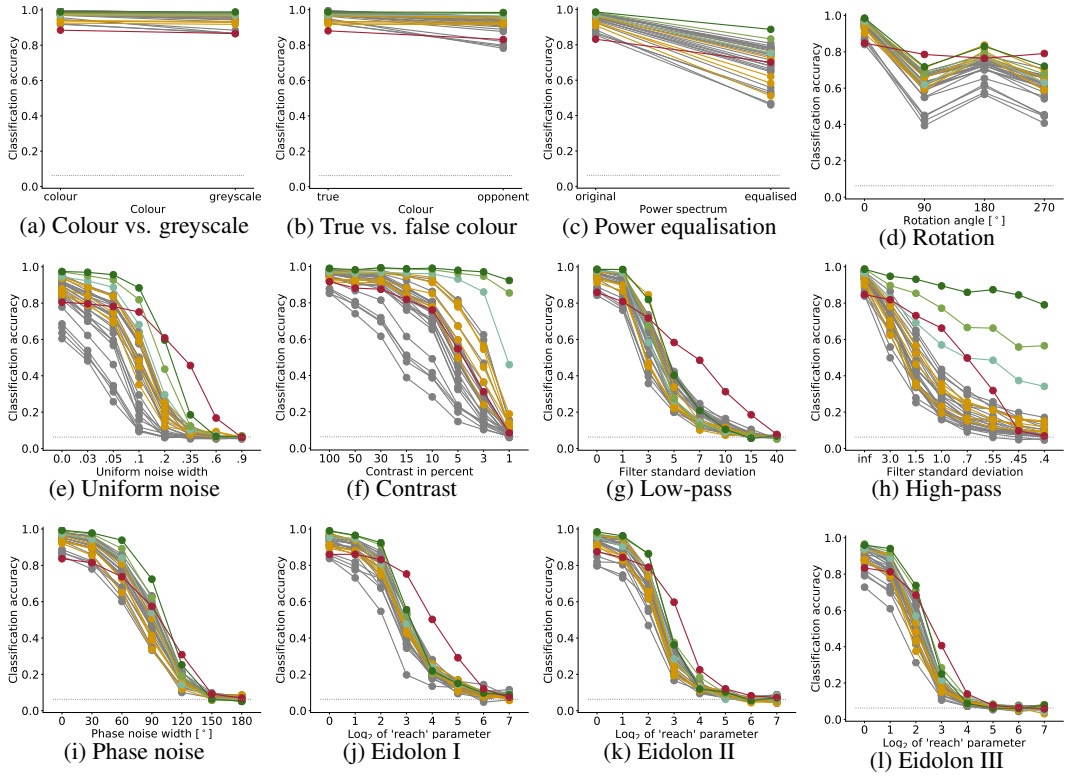
Results. In Figure 1, we compare self-supervised and supervised networks on twelve different types of image distortions. Human observers were tested on the exact same distortions by [4]. Across distortion types, self-supervised networks are well within the range of their poorly generalising supervised counterparts. However, there is one exception: `SimCLR` shows strong generalisation improvements on uniform noise, low contrast and high-pass images—quite remarkable given that the network was trained using other augmentations (random crop with flip and resize, colour distortion, and Gaussian blur). Apart from `SimCLR`, however, we do not find benefits of self-supervised training for distortion robustness. These results for ImageNet models contrast with [16] who observed some robustness improvements for a self-supervised model trained on the CIFAR-10 dataset.

3.2 Self-supervised models make non-human category-level errors

Motivation. On clean images, CNNs now recognise objects as well as humans. But do they also confuse similar categories with each other (which can be investigated using confusion matrices)?

Results. In Figure 3, we compare category-level errors of humans against a standard supervised CNN (`ResNet-50`) and three self-supervised CNNs. We chose uniform noise for this comparison since this is one of the noise types where `SimCLR` shows strong improvements, nearing human-level accuracies. Looking at the confusion matrices, both humans and CNNs start with a dominant diagonal indicating

[1] "Unsupervised learning" and "self-supervised learning" are sometimes used interchangeably. We use the term "self-supervised learning"' since the methods use (label-free) supervision.



(a) Colour vs. greyscale (b) True vs. false colour (c) Power equalisation (d) Rotation

(e) Uniform noise (f) Contrast (g) Low-pass (h) High-pass

(i) Phase noise (j) Eidolon I (k) Eidolon II (l) Eidolon III

Figure 1: Noise generalisation results for humans (red) vs. supervised models (grey) vs. self-supervised models (orange). Self-supervised SimCLR variants: green. Best viewed on screen.

correct categorisation. With increasing noise level, however, all CNNs develop a strong tendency to predict one category and only one: "knife" for `ResNet-50`, "clock" for `InsDis`, `MoCoV2` and `SimCLR-x4`. Human observers, on the other hand, more or less evenly distribute their errors across classes. For supervised networks, this pattern of errors was already observed by [4]; we here find that this peculiar idiosyncrasy is shared by self-supervised networks, indicating that discriminative supervised training is not the underlying reason for this non-human behaviour.

3.3 Self-supervised models make non-human image-level errors (error consistency)

Motivation. Achieving human-level accuracies on a dataset does not necessarily imply using a human-like strategy: different strategies can lead to similar accuracies. Therefore, it is essential to investigate *on which stimuli* errors occur. If two decision makers—for instance, a human observer and a CNN—use a similar strategy, we can expect them to consistently make errors on the same individual images. This intuition is captured by *error consistency* (κ), a metric to measure the degree to which decision makers make the same image-level errors [17]. $\kappa > 0$ means that two decision makers systematically make errors on the same images; $\kappa = 0$ indicates no more error overlap than what could be expected by chance alone.

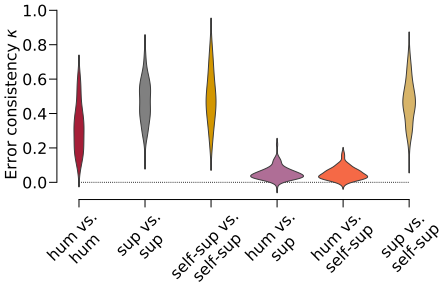


Figure 2: Self-supervised models make errors on similar images as supervised models (high κ = consistent errors).

Results. Figure 2 plots the consistency of errors (measured by κ). Humans make highly similar errors as other humans (mean $\kappa = 0.32$), but neither supervised nor self-supervised models make human-like errors. Instead, error consistency *between* model groups (self-supervised vs. supervised) is just as high as consistency *within* model groups: self-supervised models make errors on the same images as supervised models, an indicator for highly similar strategies.

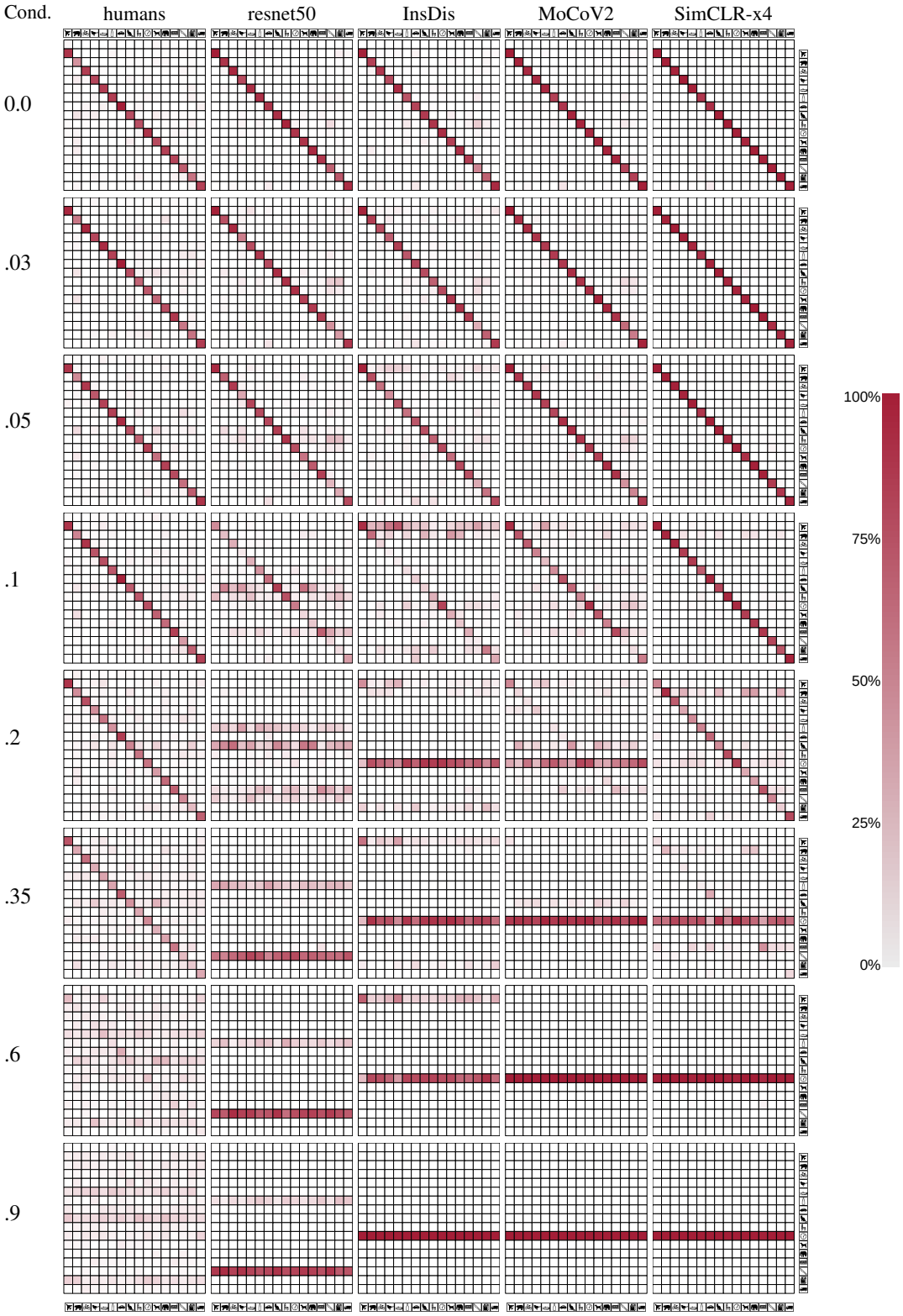


Figure 3: Confusion matrices for different conditions ("Cond.") of the uniform noise experiment. Columns show ground truth object categories, rows indicate predicted categories. Supervised `ResNet-50` and self-supervised networks `InsDis`, `MoCoV2` & `SimCLR-x4` all preferentially select a single category with increasing noise level.

3.4 Self-supervised models are biased towards texture

Motivation. Standard supervised networks recognise objects by relying on local texture statistics, largely ignoring global object shape [5, 18, 19]. This striking difference to human visual perception has been attributed to the fact that texture is a shortcut sufficient to discriminate among objects [20]—but is texture also sufficient to solve self-supervised training objectives?

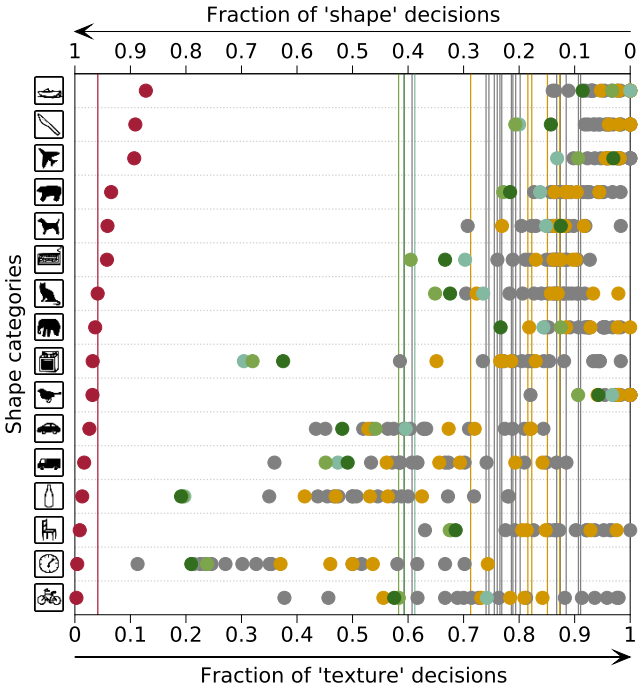


Figure 4: Self-supervised models are biased towards texture.

Results. We tested a broad range of CNNs on the texture-shape cue conflict dataset from [5]. This dataset consists of images where the shape belongs to one category (e.g. cat) and the texture belongs to a different category (e.g. elephant). When plotting whether CNNs prefer texture or shape (Figure 4), we observe that most self-supervised models have a strong texture bias known from traditional supervised models. This texture bias is less prominent for `SimCLR` (58.3–61.2% texture decisions), which is still on par with supervised model `Inception-V3` (60.7% texture decisions). These findings are in line with [21], who observed that the influence of training data augmentations on shape bias is stronger than the role of architecture or training objective. Neither supervised nor self-supervised models have the strong shape bias that is so characteristic for human observers, indicating fundamentally different decision making processes between humans and CNNs.

4 Discussion

Comparing self-supervised networks to supervised models and human observers, we here investigated four key behavioural characteristics: out-of-distribution generalisation, category-level error patterns, image-level error patterns, and texture bias. Overall, we find that self-supervised models resemble their supervised counterparts much more closely than what could have been expected given fundamentally different training objectives. While standard models are notoriously non-robust [4, 20, 22–24], `SimCLR` represents a notable exception in some of our experiments as it is less biased towards texture and much more robust towards some types of distortions. It is an open question whether these benefits arise from the training objective or whether they are instead induced by the specific set of data augmentations used during `SimCLR` model training.

Perhaps surprisingly, error consistency analysis suggests that the images on which supervised and self-supervised models make errors overlap strongly, much more than what could have been expected by chance alone. This provides evidence for similar processing mechanisms: It seems that switching label-based supervision for a contrastive learning scheme does not have a strong effect on the inductive bias of the resulting model—at least for the currently used contrastive approaches. Furthermore, we find little evidence for human-like behaviour in the investigated self-supervised models.

We are clearly just witnessing the beginning of what could be called a self-supervised revolution of machine vision, and we expect future self-supervised models to behave significantly differently from supervised ones. What we are showing is that the current self-supervised CNNs are not yet more human-like in their strategies and internal representations than plain-vanilla supervised CNNs. We hope, however, that analyses like ours may facilitate the tracking of emerging similarities and differences, whether between different types of models or between models and human perception.

References

[1] Yann LeCun. Predictive learning, 2016. URL `https://www.youtube.com/watch?v=0unt2Y4qxQo`.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[3] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[4] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*, 2018.

[5] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[6] William Lotter, Gabriel Kreiman, and David Cox. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence*, 2(4):210–219, 2020.

[7] Talia Konkle and George A Alvarez. Instance-level contrastive learning yields human brain-like representation without category-supervision. *bioRxiv*, 2020.

[8] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael Frank, James DiCarlo, and Daniel Yamins. Unsupervised neural network models of the ventral visual stream. *bioRxiv*, 2020.

[9] Katherine R Storrs and Roland W Fleming. Unsupervised learning predicts human perception and misperception of specular surface reflectance. *bioRxiv*, 2020.

[10] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[12] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[13] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019.

[15] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

[16] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pages 15663–15674, 2019.

[17] Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *arXiv preprint arXiv:2006.16736*, 2020.

[18] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, 14(12):e1006613, 2018.

[19] Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In *International Conference on Learning Representations*, 2019.

[20] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

[21] Katherine L Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *arXiv preprint arXiv:1911.09071v2*, 2020.

[22] Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. *International Conference on Quality of Multimedia Experience*, 2016.

[23] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[24] Arturo Deza and Talia Konkle. Emergent properties of foveated perceptual systems. *arXiv preprint arXiv:2006.07991*, 2020.

Appendix

Method details: Linear classifier training procedure for PyContrast models

The PyContrast repository by Yonglong Tian contains a Pytorch implementation of unsupervised representation learning methods, including pre-trained representation weights. The repository provides training and evaluation pipelines, but it supports only multi-node distributed training and does not (currently) provide weights for the classifier. We have used the repository's linear classifier evaluation pipeline to train classifiers for InsDis [10], MoCo [11], MoCoV2 [12], PIRL [3] and InfoMin [13] on ImageNet. Pre-trained weights of the model representations (without classifier) were taken from the provided Dropbox link and we then ran the training pipeline on a NVIDIA TESLA P100 using the default parameters configured in the pipeline. Detailed documentation about running the pipeline and parameters can be found in the PyContrast repository (commit #3541b82).

Additional results

Figure 5 shows error consistency for two additional experiments, and Figure 6 shows additional confusion matrices. Both confirm the results reported in the main text.

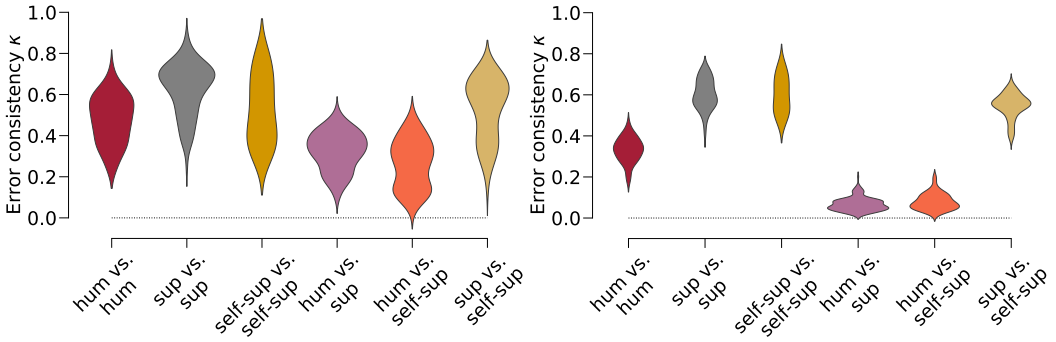


Figure 5: Error consistency (high κ = consistent errors) between all combinations of the following three groups: humans (hum), supervised networks (sup) and self-supervised networks (self-sup). In Section 3.3, "edge" images were shown; here are two additional experiments: "silhouettes" on the left and "cue conflict" images on the right. Original stimuli: [5]. For all three experiments, consistency between networks is much higher than between networks and humans: CNNs make errors on the same images as other CNNs, whether these are supervised or self-supervised.

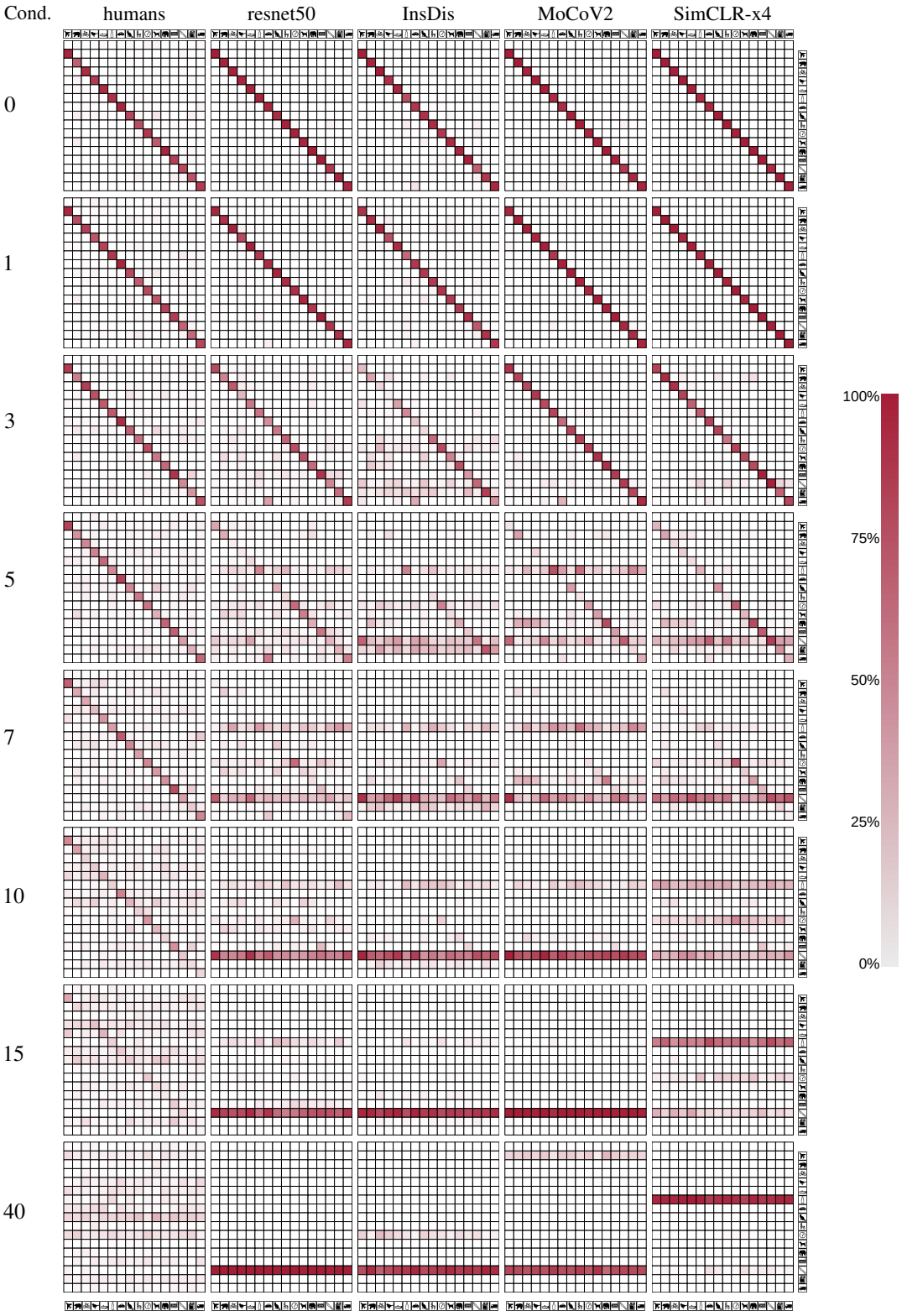


Figure 6: Confusion matrices for additional experiment (low-pass filtering). Again, CNNs develop a preference for a certain category as the distortion strength increases.