We thank the reviewers for their feedback, and their positive appraisal of our work. We respond to each point below.

R1: *"I find your experiments extremely interesting with good results and insight, but unfortunately I don't think the current state of the [figure] presentation is publication-quality."* These fixes are easily made: we will use text rather than WordNet IDs in Fig. 1, increase the font sizes, and add bolding to Tables 2-4 to bring out the key results. We have updated Fig. 2, reducing the widths of the error bars and grouping results by dataset rather than by model architecture, and Fig. 3, adding a color gradient to our markers to indicate the additive accumulation of augmentations. We retained marker shapes for accessibility of colorblind readers and to ensure unambiguous mapping to the legend.

R1: *"In Figure 5, I would like to see intermediate (not after the final conv) layers included."* **R3**: *"...why not report Figure 5 for every single representations one can get..."* We initially focused on the output of the convolutional layers since this is the convolutional representation most commonly used for transfer to other computer vision tasks and most correlated with IT activity in the primate visual system. We have extended our analysis to consider each layer of AlexNet. The results recapitulate the ones reported: shape is persistently more decodable through the convolutional layers than is texture, and its decodability varies less than that of texture, which rises through them. In the FC layers, shape decodability decreases markedly whereas texture increases.

R1: *"Why did you opt for Exemplar as one of your methods in S6?"* Our choice of models was motivated by a desire for diversity of approaches that perform reasonably well on ImageNet, as well as desire to connect to the existing recent literature. The Rotation and Exemplar models were recently compared in Kolesnikov et al. 2019. Although there are differences in implementation, Exemplar, Instance Discrimination, and SimCLR are all constrastive learning methods and might be expected to behave similarly modulo augmentation. We are unaware of recent work achieving success with ordinary GANs or VAEs for unsupervised representation learning on ImageNet.

R2: *"On the broader impact section: I miss any discussion of the negative impact this work can have or rather that some of the main findings imply."* We have added the following text: "People...often have a mental model of computer vision models as similar to human vision. Our findings contribute to a body of work showing that this view is actually far from correct, especially for ImageNet, one of the datasets most commonly used to train and evaluate models. Divergences between human and machine vision of the kind we study could cause users to make significant errors in anticipating and reasoning about the behavior of computer vision systems...At the same time, we recognize the possible negative consequences of blindly constraining models' judgments to agree with people's: human visual judgments display many forms of bias that should certainly be kept out of computer models." Per **R3**'s pointer, we have removed the numbering before "Broader Impact."

R3: *"what does 'remove' mean exactly?"* In Figure 5, we presented results showing that shape information is down-weighted by (becomes less decodable through) the fully connected layers of the AlexNet classifier, whereas texture remains relatively constant. Our statement that "these models' classification layers remove shape information" refers to the fact that shape decodability is higher at the final convolutional layer (pool3) than in the fully-connected layers, and for each fully-connected layer, shape decodability is higher for its input than its output. We will add clarifying text.

R3: *"In the abstract, the paper says the differences 'arise not from differences in their internal workings, but from differences in the data that they see'...yet in the experiments, the authors demonstrate that, with more carefully designed regularizations...the model can be pushed to focus more on the shape."* Our goal with this statement was to emphasize the main finding of our study: while factors such as loss function and model architecture do indeed influence the level of texture bias in a model, they do so to a much lesser degree than do features of the data, e.g. data augmentation. However, we agree that the statement is worded too strongly. We have added a qualifier: "may not arise primarily from."

Comparison with other work. **R1**: *"I'd be interested in a comparison of your architecture findings to those of Ilyas et al."* Ilyas' main architecture analysis is Fig. 3, which investigates the transferability of adversarial examples from ResNet-50 to other architectures. While differences in the set of models studied unfortunately prevents a complete comparison, we do observe a qualitative similarity to their results: just as they find ResNet-50 adversarial examples to transfer better to DenseNet than Inception-v3, we find the shape and texture preferences of ResNet-50 to be more similar to the DenseNet variants that we study than to Inception-v3. **R3**: *"I wonder if this main arguments of [Wang et al. (2020)] also follows this understanding of shape vs. texture, why or why not."* This is an interesting question. Several studies have found that CNNs are sensitive to the Fourier statistics of training data and can use high-frequency features imperceptible to people (also Yin et al. 2019). We see this as related to texture bias, and in our data augmentation experiments, find that Gaussian blur, which removes high-frequency information, reduces texture bias. At the same time, we do not see shape and texture as entirely reducible to spatial frequency information (see e.g. Portilla and Simoncelli 2000).

Textual edits. Thank you to **R1** for noticing several errors, which we have corrected. *Future follow-up questions.* **R2** notes that an interesting future follow-up experiment would be to fine-tune the self-supervised models using the same data augmentations used in the cumulative experiment. We agree – thank you for this suggestion! We also agree with **R4** that it would be interesting to investigate texture bias in object detection or segmentation models.