Thank you for the thorough reviews.

1. *New insights about detectors (all reviewers) and which detectors are better (reviewer 1).* At least six new insights:
i) Existing object detectors largely fail to generalize well not because of a generic problem with performance but because of specific phenomena (lack of invariance to background, viewpoint and rotation) that they are brittle to in a way that humans are not. When we selected a subset of images in ObjectNet which had the same biases as those in ImageNet, detector performance on ObjectNet reached that of humans, just as it is in ImageNet. This indicates that focusing on structural methods to handle these phenomena in detectors can have significant payoff.
ii) We demonstrate that there is a large performance gap with respect to humans even over object classes that are part of ImageNet. Human performance over ObjectNet is around 95% on a free-form text answer question, which is considerably higher than networks' performance. This is in stark contrast to other work which indicates that object detectors are approaching human-level performance.
iii) Our results indicate that data alone will not fix these problems. We systematically fine-tuned on ObjectNet using a ResNet-152 trained on ImageNet. Fine-tuning was with 1-64 per class images in steps of powers of 2. Performance increased by $\approx$2% when doubling the data. This shows that even when networks learn biases that lurk in ObjectNet they remain brittle with respect to the controls used.
iv) ObjectNet helps explain why a large gap in performance is seen between developers and users of object detectors in a way that no other modern dataset does. On other datasets fine-tuning recovers the vast majority of the performance very quickly, which is not the case with ObjectNet, and which is not observed, for example, in robotics applications.
v) Large-scale data can be collected in a way that is less biased, not just for object detection but for machine learning in general. Bias is rarely addressed despite affecting all modern machine learning methods.
vi) In a sense, detectors are roughly the same. The gap between ObjectNet and ImageNet is not closing with newer models. 1% gained on ImageNet is also 1% gained on ObjectNet. The community as a whole is making steady measured progress and a breakthrough awaits.

2. *Occlusion and clutter, introduced biases due to people holding objects. (reviewer 1)* Without a doubt, ObjectNet has its own biases which include the fact that certain object shapes are likely to be held when imaged from certain angles. Had we removed more biases, model performance would have likely further decreased. Not all biases are the same though: some are easy to learn and others are difficult. Our fine-tuning results show that learning about how specific objects appear in ObjectNet does not help performance much — far less than the large performance jump seen in other datasets. What we do show is the importance of three particular biases that humans are known to be extremely resilient to from an early age, but networks fail to understand. These overwhelm other types of biases that might help performance. The reviewer astutely points out that we do not consider occlusion or clutter. In the next version of the dataset we intend to do so.

3. *Real-world biases are not necessarily bad, as it can reflect real-world distributions. These biases exist in the world, and leveraging them is useful for task performance in many cases. (all reviewers)* Is it true that real-world biases can be useful and they do increase average-case performance. At the same time, relying on them can be very dangerous. Applications of computer vision to robotics, autonomous cars, surveillance, and other domains require high performance in unusual situations. Indeed, that's where robustness is most needed because those are likely to be situations that were not encountered during the development phase. Additionally, many applications of computer vision can have adversarial opponents — humans that attempt to fool systems by defacing lane markings or hiding dangerous objects from detectors. Most datasets that exist today, ObjectNet being an exception to this, focus on the usual case and do not shed light on the performance one will encounter in unusual cases.

4. *Limitations (all reviewers), simulation and AR (reviewers 2 and 3).* ObjectNet like all methods has many limitations: it consists of indoor objects, the objects have to be available to many people, objects must be mobile, must not be too large or small or fragile or dangerous. We cannot ask users to damage, paint, cut, or otherwise permanently change objects.

In simulation one can create datasets that minimize bias although current methods produce images which are on the whole easy for object detectors even when bias is removed. Foreground/background separation is easier in rendered images and there are fewer sources of noise. In addition, there is a kind of variety and bias that is not easy to control in simulation: shape. There are many possible shapes for a chair but in simulation one will have access to only a small number of variations of static models. Methods exist to synthesize objects of different shapes for the same class but these generally require specifying class boundaries or generative shape models — itself an open and difficult problem.

AR can be useful for collecting data and will play a role as we ask users to pose more complex scenes, like those which include occlusion. We are excited to bring an awareness of more systematic experimental design methods, and show how these can identify the most important phenomena for different machine learning tasks, can help characterize the state of models and compare models in more fine-grained ways, and ultimately help develop new types of models.