We thank the reviewers for their insightful comments. We address individual notes below.

Reviewer 1, Q3: We agree that the utility of a perceptual metric is much increased if it is made available for public scrutiny and further research. We are committed to making the code and model parameters available online upon publication.

Reviewer 1, Q7: We have done our best to describe the details of our experiments in the paper and the supplemental material. Due to licensing restrictions, we may not be able to make the training set public, but we will make the code and trained model parameters available to complement the description.

Reviewer 2, Q2: We thank you for your honest comments. Apparently there is some conceptual confusion here; apologies if this is due to our writing. The novelty of our approach is that we do not rely on supervised training data to solve the task (IQA), which we mention in the title, the abstract, and many points throughout the paper. Instead, we build a representation solely based on combining known physiological properties of the human visual system (scale and shift-equivariance; cell responses in cortical area V1, for instance, have been shown to have these properties) with two theories of the visual system's high-level computational goals: slowness and efficient coding. As we describe in Sec. 2.1 and 2.3, the efficient coding constraint is explicitly implemented by the compressivity of the IXYZ objective; the slowness principle is implemented by extracting the mutual information of successive frames from the training video dataset, using the objective. The fact that our learned representation is highly predictive of human quality judgements (69.4% accuracy compared to 73.9% agreement from human to human on the 2AFC task [35]) not only represents a contribution in the engineering sense (as a way to reduce costly gathering of human responses compared to supervised learning), but can also be interpreted as evidence that confirms the validity of the slowness and efficient coding principles. We have not focused on the latter point, but we will certainly improve our paper in that regard.

Reviewer 2, Q8: We are of course aware of [35], as we compare to their empirical results directly in Table 1. Although qualitatively comparing IQA metrics is difficult, we make some interesting observations about their approach vs. ours in Section 3.3. In particular, their method appears to inherit certain representational properties of a classifier, like relative equivariance to substantial amounts of fog, as it is built on pre-trained classifiers, while ours does not. Of course, humans also are sensitive to the presence of fog and many other differences that classifiers are often explicitly trained to ignore.

Reviewer 2, Q11: We did discuss the broader impact of our work in a dedicated section, as required by the submission guidelines.

Reviewer 3, Q3 & Reviewer 4, Q5: We agree that more analysis of the learned representation would be helpful, and have already begun preparations to carry out further experiments. We are also interested in evaluating whether our representation could be useful for other tasks than IQA. However, we also think that 69.4% accuracy on the 2AFC task, notably without any additional supervised regression stage, is an extremely strong result compared to 73.9% agreement from human to human [35] (which we will stress in the revised paper). In our view, this result justifies publication.

Reviewer 3, Q3/Q5: Please note that Section 3.4 is entirely devoted to assessing which of the inductive biases contribute to achieving good prediction results. Notably, a multi-scale architecture, or at least reading off activations at multiple layers, appears to be crucial; compressiveness of the objective appears to improve results substantially. Other architectural choices don't appear to matter as much. Some desirable ablation experiments are infeasible, e.g., dropping the convolutionality constraint would lead to a many times larger model, making it difficult to train in practice.

Reviewer 3, Q8: We agree that Figure 3 in particular can have a more descriptive caption, and we will improve this.

Reviewer 4, Q3: First, let us clarify that the slowness principle, as implemented in our paper, does not necessarily imply temporal persistence of the metric, it implies temporal persistence of the representation in a probabilistic sense (i.e., on average, neighboring frames will have a small perceptual distance, but individual frames may have a large distance under the metric). An intuitive depiction of the concept can be found in `http://www.scholarpedia.org/article/Slow_feature_analysis`, Fig. 2. It is correct that to train our model, we would ideally use uncompressed video, since otherwise our representation may become insensitive to compression artifacts. In practice, uncompressed video with sufficient variability is difficult to come by. We therefore applied pre-processing to our training set, including downsampling by a large randomized factor, in order to minimize compression artifacts, as described in the appendix. Spot checking the sensitivity of the metric to various corruptions in Section 3.3 and the appendix, we find that it does not have any obvious defects that may stem from compressed training data.

Reviewer 4, Q6: Thank you for this reference, which we weren't aware of. We will discuss this paper in the revision. While this work uses deep learning as well, it is an example of a supervisedly trained no-reference metric, which is targeted specifically at predicting quality of images "in the wild" without comparing to an original. Our work presents a full-reference metric and aims more at the question of which inductive biases can help unsupervised learning, in particular to make training perceptual metrics more data efficient.