We thank the Reviewers for their thoughtful assessment of our work and valuable comments. Below we address the main questions raised in the reviews.

Reviewer 1

- *About the employed adversarial attacks:* Adversarial attacks are usually meant to be subtle, making unnoticeable changes in the original image (and keeping the original label unchanged). Thus, "swapping items in a scene" or converting a 7 via truncation into a 1, as mentioned in the review, are usually not considered to be adversarial attacks. Indeed, translation attacks (introduced by Azulay and Weiss, 2018) are much less effective than the most popular gradient-based attacks (e.g., the fast gradient sign method or PGD), however, the effects of the latter are often noticeable on the images (while keeping the labels intact) and change the image distribution quite a lot. Our original intention was to use the gradient-based attacks, however, we could not deal properly with the resulting distribution shift (i.e., calculate the corresponding Radon-Nikodym derivative). Translation attacks were selected because there we could address this issue, and—perhaps surprisingly—they work reasonably effectively (on CIFAR-10, with a standard test set error rate of about 2-8%, the adversarial translations were successful for an additional 5-8% of the test set, while on ImageNet, with the standard test set error rates in the range of 20-30%, adversarial attacks were successful for an additional 7-13% of the test images). We also think that translation attacks are quite realistic in the sense that they capture the real-life phenomenon that photos of the same subjects can be framed differently, for example when trying out different compositions or taking several photos in a sequence. From this point of view, image translations are actually more realistic than adding model-dependent noise, which is essentially what gradient-based attacks do. Figure 1 on page 2 shows an attack example for ImageNet; we will include more examples in the appendix, also for CIFAR-10.
- We will work on improving the writing for the final version, as suggested. We will aim to make the text more concise (without making it harder to follow) in the camera-ready version, separate out the mathematical formalism and include more adversarial examples in the supplementary material. We will add a more complete justification for translational attacks, either in the main text or as an appendix.
- Just for clarity, we would like to point out that our method was also applicable to ImageNet, where we found overfitting to the training set but not to the test set.

Reviewer 2

The test can naturally be applied at any point of the training process to see if overfitting has happened. However, our independence test itself uses the test data, so using the results of this test already leaks information from the test set to the training process, hence induces some degree of overfitting. Also, using the test multiple times increases the risk of a false positive, which one has to protect against by using, e.g., the Bonferroni correction (i.e., applying a union bound over the Type I error probabilities of the multiple tests).

Reviewer 3

- We will provide more details about the experimental settings and the training methods (including the selection of hyperparameters) in the appendix. In all the training procedures, the number of epochs and the corresponding learning rate schedules were fixed in advance, following the recommendations of previous work in the literature. We used different random seeds for each training process.
- Indeed, hyperparameter selection is one of the potential sources of overfitting. When averaging over multiple i.i.d. training runs (as we do in our strongest tests), the only possible causes of overfitting are tweaking either (i) the hyperparameters or (ii) the model architecture in order to minimize the test set error rate. The suggestion of tuning the hyperparameters to CIFAR-10.1 could help to distinguish between the two: if choosing the hyperparameters of a CIFAR-10 model (trained on a CIFAR-10 training set) to minimize the CIFAR-10.1 test set error rate were to lead to a model overfitted to the latter but not to the CIFAR-10 test set, it would suggest that (i) is a more important source of overfitting than (ii).