We would like to thank all the reviewers for their time and for giving our paper this encouraging feedback.

Reviewer #1 mentioned the issue of transferability of the student features when trained without any real data, which is also something we consider interesting. Since submission we tested our student on the ImageNet component from the CINIC-10 dataset (https://arxiv.org/abs/1810.03505), corresponding to a subset of ImageNet images chosen and resized to match CIFAR. We observe that the student generalizes equally well to its teacher, i.e. that the difference in performance between teacher and student remains the same under such dataset shift. This is not as much of a shift as transferring to a segmentation task (in general classification features generalize poorly to segmentation tasks even in normal circumstances). But in general we find that the student's representation shows similar transferability to that of the teacher, because of the student's tendency to directly match the decision boundaries of the teacher. However we cannot make any claim of increased transferability at this stage.

Vaccination is also an interesting thought. In practice the nature of the problem forces the student to match the teacher on the entire input space regardless of how it was trained, but having a teacher agnostic to out of distribution data could reduce the task-relevant information that leaks out to the student. On the other hand, depending on the vaccination approach it could be that the agnosticism of the teacher helps identifying samples close to the training data more easily, because they would have a specific entropy signature (high confidence). We have not pushed further on vaccination at this stage because it seems to be a use-case that is harder to justify in practice: someone releasing a trained model for people to use typically wouldn't take measures against people compressing it.

We agree with the reviewers that additional large scale datasets would be complementary to this work. However we also note that many of the competing papers we cited are limited to MNIST experiments due to the difficulty of the problem, so our tests are more extensive than many. In practice, a limitation to our current model is computational cost, since our distillation trains two networks instead of one, and several gradient steps per batch are taken for the student. This amounts to a handful times more computational cost than training a vanilla network, and our limited computational resources have made ImageNet scale experiments a challenge. This is a focus of future work; however those working in this field need to know of methods that are effective on the smaller problems so we can focus our efforts on scale in the right direction. Hence we think it is important to not wait for huge-scale experiments before releasing this work.

In particular in future work we plan to investigate more closely the relationship between the size of the teacher's training set and the difficulty of zero-shot distillation. Extra challenges relating to large scale datasets may include the larger number of classes, some of which may not be sampled enough by the generator.

We thank **Reviewer #2** for his comments on layout, and have modified our paper in light of that for the camera ready version.

Reviewer #4 correctly and pertinently noticed that in Figure 4 we took adversarial steps on the students rather than the teacher. Attacking the students rather than the teacher allows us to show 4 curves per dataset instead of 3, and as such it conveys a bit more information. For instance it shows that it is easier to cross decision boundaries on our zero-shot student than on the baseline student when stepping on each network respectively (orange vs red curve per row). Another possible motivation comes from considering the case that the teacher and student are sets of linear decision boundaries. If we assume that the students is a smaller set (less capacity) mostly contained within the teacher set, then crossing decision boundaries of the student should correspond to crossing decision boundaries of the teacher, to a greater extent than the other way around.

Lastly we note that the batchnorm layers of the WRNs in our paper were initially not trained (gamma and beta fixed). This was a mistake because a trainable batchnorm is the most common practice for WRNs in the literature, and it facilitates the training procedure. We re-ran all the experiments in the paper and obtained a 1 to 2% boost to most of our accuracies, and reduced the Transition Error (TE) on CIFAR-10 by a third. For the sake of reproducibility we will ensure the camera ready version has these updated values, so that all values match our latest published code.

Again we thank the reviewers for their time and hope that we have answered their questions.