We thank the reviewers for thoughtful feedback. In this paper we introduce Augerino, a method that can automatically attain the strength of group equivariant networks through learning data augmentations. Augerino is a flexible method that can be combined with any pre-existing architecture. We note that Augerino does not require one to specify invariances a priori, merely a group transformation. We show Augerino is capable of **learning** augmentations, from training data alone, which boosts accuracy when applied at test time over a wide range of tasks. We want to emphasize that Augerino is a distinctive contribution: other works on data augmentation are not geared towards learning equivariances at training time and would not even directly apply to most of the applications considered in our paper.

R1: Thank you for the constructive review, we appreciate the feedback and will incorporate the suggestions accordingly. With respect to the discussion of the mathematics, we commit to making the mathematical statements in Sections 3.1 and 3.2 more precise, we will switch from the notation of the transformation and distribution of T_ϕ,Q_θ to g, μ. Our use of uniform distribution should be made more precise, and that it could depend on the parameterization if not careful. We will make it clear that the distribution is meant to be the Haar measure on the group, and that the appropriate condition is left translation invariance $d\mu(hg) = d\mu(g)$. The transformation T_ϕ needs to act linearly so that we can pull it out of the expectation in Eqs. $9 - 10$. Unlike with learning invariances, the extension to equivariance in Section 3.1 does require the group structure and is not applicable to more general set of transformations which may lack inverses or closure, and we will add this point of discussion in the text. While this requirement is an idealization and limitation in some ways (cropping, boundary effects), we feel that this simplification is justified. Note that the invariant model only requires an invariant measure, but not necessarily the existence of inverses, and thus includes transformation semi-groups. We will mention the deep scale-spaces paper in this context. Additionally, thank you for the pointer to the Barnard and Casasent paper, this is a great early example of the utility of invariance.

R2: While we appreciate the pointers to test-time augmentation, which we are happy to reference, we want to be clear that these papers are complementary to Augerino and do not diminish our contribution. Indeed, most of these works would not even directly apply to the applications we consider. Augerino is distinctive in that it **learns** an appropriate distribution over augmentations at **train-time**. We motivate the utility of Augerino in Section 4, demonstrating both why one may want to learn a distribution over augmentations, rather than just applying them.

We also note that for Augerino to be applicable we need only the invariance to be represented by a group transformation. In Section 4 we do know the correct invariances, but these experiments serve to show that when we do know the correct distribution over invariances we can recover the ground truth, and are in no way meant to show the limiting case performance of the method. In Sections 5 and 6 we do not have a known target set of invariances yet Augerino outperforms competing methods. While in some cases the set of invariances will never be known, if we do believe that invariance is a trait we wish our model to have then group transformations are a promising path forward as they encompass a broad set of transformations that has been shown to improve performance on many well studied tasks; for examples see $[6, 7, 8, 12]$.

R3: We thank the reviewer for the constructive comments and ideas for extensions. The model used in Section 5 is an 8 layer deep convolutional neural network with a max channel width of 256. We will be sure to include the details of all architectures in the appendix of the camera ready.

We have not observed any evidence that Augerino is susceptible to loss of performance in using alternate network architectures. One of the key strengths of Augerino is its ability to be combined freely with any model. Throughout the experiments we use a range of different convolutional and feed forward neural networks, chosen only to be appropriate for the given task not for their compatibility with Augerino. In each case (the demonstrations of Section 4, and the benchmark improvements of Sections 5 – 8) Augerino performs as expected and improves accuracy over baselines.

Finally, we will incorporate larger datasets in the camera-ready submission. Preliminary results suggest Augerino will perform well on larger datasets such as CIFAR100, however we avoid presenting any incomplete results here.

R4: Thank you for your thoughtful comments. *Assumption of hyperparameters:* The parameters a and b are *learned parameters of the model* and not hyperparameters; the reparametrization trick gives us gradients with respect to these parameters. Therefore a and b can learn to accommodate an arbitrary range, and we do not find performance sensitive to their initialization. We will clarify. *Learning transformations:* as discussed in response to reviewer 2, the experiments in Section 4 highlight our ability to uncover the ground-truth augmentations. In later sections we show Augerino can outperform competing methods in which desired transformations are unknown.

Remaining points: The augmentation in Eq. 12 is done through computing the exponential map as the solution to a differential equation which allows back-propagation to flow through the affine transformations. In Section 5 the fixed augmentation indicates a standard set of transformations applied to the training data , we do expect fixed augmentation to perform worse, that is the central motivation to this work. We will be sure to include these details and clarify all notation in Section 3 in the camera-ready.