We thank the reviewers for their comments and actionable suggestions on improving the paper. Below we address the most pressing concerns. We paraphrase some of the comments for brevity.

Method	Test Accuracy	ECE
Baseline	95.12	0.023
Mixup (α=0.4)	96.16	0.019
Mixup (α=1.0)	96.04	0.1
Label Smoothing (ϵ=0.1)	95.51	0.089
ERL (kappa=0.1)	95.55	0.046

(a) CIFAR-10/ResNet-18

Method	Test Accuracy	ECE
Baseline	78.28	0.049
Mixup (alpha=0.5)	79.57	0.035
Mixup (alpha=1.0)	79.54	0.091
Label Smoothing (eps=0.1)	79.08	0.066
ERL (kappa=1.0)	78.47	0.6

(b) CIFAR-100/ResNet-18

Comment: "Justify that the baseline models are well trained, and compare with existing baselines that use ResNet-18 for CIFAR-10/100" (R2). We provide additional results on ResNet-18 for both CIFAR-10 and 100. Our baselines (w. no mixup) match the baseline accuracies reported in related work. We also provide the expected calibration error (ECE) for the best performing model as well as the mixup model that used $\alpha = 1.0$ as suggested by reviewer 2. We find that lower α gives slightly better classification and signficantly better ECE. Note that ECE can be high both due to the model being overconfident as well as under-confident, the latter being the case for $\alpha = 1.0$ since this causes the resulting training signal to have higher entropies than with smaller α.

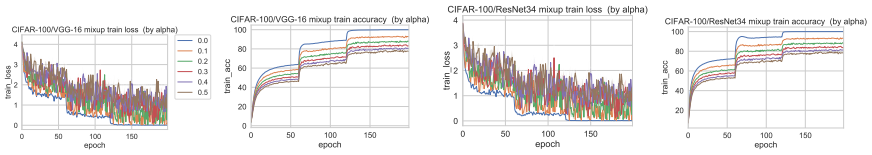

Figure 1: Train loss and accuracy for mixup for various alphas. Baseline corresponds to $\alpha = 0$

Comment: "Provide training loss curves for Figures 1 and 2. " (R2)We show training curves for some of the experiments in the paper in above Figure. Your intuition is correct: for the baseline (i.e when $\alpha = 0.$), over-fitting on the training set is indeed correlated with transitioning to overconfidence. The baseline train loss and accuracy approach 0 and 100% respectively (i.e., over-fitting), while in the mixup case (non-zero α's), the strong data augmentation prevents over-fitting and thus restricts the model from making overconfident predictions. This behavior is sustained even if one trains for much longer (see next section)

Comment: "Will the mixup models become overconfident if trained for longer?" (R2) Below we provide the ECE vs epoch for both CIFAR-10 and CIFAR-100 for the mixup models trained for 1000 epochs (original experiments only used 200 epochs). We see that the mixup model, even when trained for much longer, continues to have a low calibration error, suggesting that the mixing of data has a sustained inhibitive effect on over-fitting the training data (the training loss for mixup continues to be significantly higher than baseline even after extended training) and preventing the model from becoming overconfident.

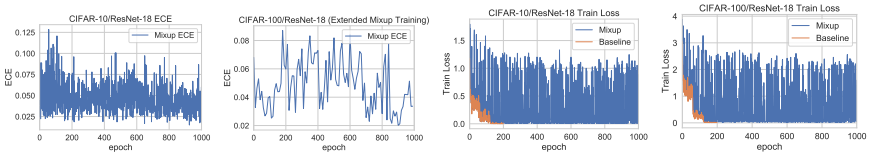

Figure 2

As for why mixup improves calibration (R2), please see discussion in Section 4:the strong data augmentation and label smoothening are both contributive factors: one can view mixup training as training with infinite data (since the model never sees the same data point twice) in which case true posteriors are learnt according to statistical learning theory, but in addition the label softening (which prevents the winning logits from becoming arbitrarily large) also prevents overconfidence. Note that mixup models can turn out to be *underconfident* if α is large. In fact, this is also related to manifold intrusion: a mixed-up sample is more likely to lie away from the original manifold and thus be affected by manifold intrusion if α is large. In our experiments, we see the resulting models are prone to under-fitting and under-confidence. We will include a discussion on ROC and AUC curves for mixup in the final version (R1). As for comparing calibration of mixup with temperature scaling (R3), this produces almost perfectly calibrated scores since it is a post-training calibration approach. We will incorporate comparisons with model ensembles and manifold mixup (R3) in the final version; we expect the latter to also produce well-calibrated scores since it is a generalization of mixup.