We thank the reviewers for their thorough comments, which we address below:

Reviewer 1. We ran additional experiments based on your detailed suggestions. Their results have been added to our manuscript and are summarized below.

1. Loss function design. We neglected to mention this fact in the paper: we linearly warm-up the lambda from 0 to its final value as previously done e.g. in [25, 34, 43]. Below we present the results with and without warmup:

Method	CIFAR-10, 250 labels	CIFAR-10, 4000 labels
Without Warmup	87.60	93.73
With Warmup	**88.62**	**93.77**

2. Accuracy & Averaging. The reason we report the average over the last checkpoints is to be more realistic and not use large validation sets (as discussed in [6, 34, 38]). If we simply return the model accuracy of the last trained checkpoint our results are consistent but have higher variance.

Dataset	Labels	Method	Median of Last 20	Accuracy of Last 5 Models				
CIFAR-10	250	Mean Teacher	46.34	46.01	46.83	45.61	45.58	45.55
CIFAR-10	250	MixMatch	85.60	85.55	85.65	85.62	85.75	85.68
CIFAR-10	4000	Mean Teacher	88.47	88.43	88.55	88.60	88.61	88.54
CIFAR-10	4000	MixMatch	93.16	93.07	93.12	93.05	93.14	93.14

3. 13-Layer ConvNet. As suggested we run experiments based on the TensorFlow Implementation of the 13-layer ConvNet from the Mean Teacher paper. MixMatch has a larger advantage when using this 13-layer network than when using the ResNet in our paper.

Method	CIFAR-10		SVHN	
	250	4000	250	1000
Mean Teacher	46.34	88.57	94.00	96.00
MixMatch	**85.69**	**93.16**	**96.41**	**96.61**

4. Table 1 Comparison with Mean Teacher. The purpose of Table 1 is to show the strongest reported results in prior papers along with our strongest MixMatch results. Unfortunately Mean Teacher does not report on CIFAR-100.

5. CIFAR-100 Comparisons. When new CIFAR-100 experiments finish we plan to include them in the final paper.

6. Ablation Studies. We believe we have included most of the ablations requested in Table 4 where we run MixMatch without MixUp (row 5), without sharpening (row 3), and without EMA (row 4). The best way to evaluate MixMatch without distribution averaging in our view is to set $K = 1$ augmentations (row 2). We address other values of K below. We hope this clarifies the ablation studies we included; if you have other suggestions we would like to add them.

7. ImageNet. We also are excited for the possibility of MixMatch on ImageNet and hope to study it in future work.

8. K=3, 4 Augmentations. Thank you for the suggestion, also made by **R3**; we find that in practice $K = 2$ gives the best results for the least performance penalty. We included additional experiments in our revised manuscript which we also list below. Using $K > 1$ augmentations is necessary; further augmentations do not give as much of a gain.

Dataset	$K = 1$	$K = 2$	$K = 3$	$K = 4$
CIFAR-10, 250 labels	84.02	**88.62**	88.45	87.55
CIFAR-10, 4000 labels	92.00	93.73	93.77	**94.12**

Reviewer 2. To address your question, Equation (3) uses the standard cross-entropy loss. The reason we must move to a L2 loss in Equation (4) is to help stabilize the training process as previously identified in [25, 34].

Reviewer 3. As outlined in item 8 of our response to R1, we added a discussion of $K = 3$ or $K = 4$ augmentations in our revised manuscript. Thank you for bringing up the clarification for weight decay: we always multiply the weight decay value by the learning rate (2e-3 in all experiments) but neglected to mention this in the paper. Finally, we tried other regularization strategies (such as cutout) but found it gives inferior results. Cutmix may work even better and we hope that future work explores additional regularizations (e.g., cutmix, manifold mixup).