We thank the reviewers for their feedback. We will reflect reviewer's comments and our response in the revision.

Reviewers showed concern on the novelty and the accuracy. FixMatch represents a significant research advancement *because* of its simplicity. While discovering new techniques and tricks is important, we believe that demonstrating state-of-the-art performance through the consolidation and simplification of existing concepts should be considered just as novel and important, if not more. Prior methods have implied that it is necessary to introduce complexity (for example, ReMixMatch uses self-supervised rotation losses, mixup, soft pseudo labels with sharpening, warmup, etc.) to achieve SOTA results. We find none of these are necessary; furthermore, they result in more hyper-parameters to tune which can be impractical on realistically-sized validation sets (as argued in [2]).

Another benefit of simplicity is the extensibility – for example, we showed that Distribution Alignment of ReMixMatch can be seamlessly added to FixMatch, improving an accuracy by 10% on CIFAR-100 with 400 labeled data.

Hard versus Sharpened Soft labels (R2, R4): We show through experiments that hard pseudo-label performs just as good or better than the sharpened soft pseudo-labels, with the desirable property of using one less hyperparameter.

Barely-supervised learning (BSL) and few-shot learning (FSL) (R2): We considered to compare BSL and FSL but found that they were not directly comparable. While they are similar in that they both use very small labeled training data, FSL is a subtask of transfer learning, where the model can be trained with additional "labeled" data from normal classes that are different from few-shot classes (e.g., [6], [4]), while BSL is not.

Distribution Alignment (DA) (R4): DA is more effective when the task is more challenging. Our experiments, which were done after paper submission, showed that DA is also important when training on ImageNet with 1% labeled data. On the other hand, we find DA effective as well when the amount of labeled data is small. For example, as in Section D.1, CIFAR-10 with 40 labeled examples also benefit from DA. We will add these results in the revision.

How realistic the scenarios are (R1): We refer the reviewer to Section 4.2. The unlabeled data of STL-10 is noisy: It contains samples from different classes than the 10 classes of the labeled set. Therefore it is more realistic and challenging. Thanks to confidence-based thresholding playing a role in rejecting out-of-distribution samples, FixMatch was able to achieve SOTA results in such scenarios without modification of parameters or algorithm design.

Extension to other domains (R1): There are application domains with advanced data augmentations, such as back-translation [5] for text classification or SpecAugment [3] for speech recognition, where FixMatch is readily applicable. Developing domain-agnostic SSL methods is desirable and a worthy long term research goal. For example, we have shown some potential of domain-agnostic FixMatch in Section D.2, showing improvements over other domain-agnostic SSL methods, though the absolute performance was not as good as with domain-specific data augmentation.

Knowledge from large-labeled network training (R1): It is true that lessons from supervised learning largely inspire the success of SSL. For imagenet experiments, we adopted a few such lessons, including learning schedule, network architecture, and data preprocessing and augmentations. While this may introduce unwanted inductive biases, FixMatch showed a solid improvement over previous works (e.g., UDA) under the same setting. Moreover, we also showed SOTA on STL-10, which is not widely used for large-scale supervised learning as it comes with only 5k labeled training data, by transferring an SSL setting from CIFAR-10.

Weak and strong augmentation (R3): Weak augmentation (e.g., shift and flip) provides a reliable prediction for pseudo-label generation. Strong augmentation prevents the consistency loss from being too easily minimized and the model from overfitting. We refer Section 5.2 on their importance where we conduct ablation studies comparing different combinations of weak/strong augmentations.

Strong augmentation may refer to a more diverse set of transformations and wider range of augmentation magnitudes. Though the data augmentation becomes an essential component for deep network training, less is known on how to quantify their impact [1], and the definition of weak and strong augmentations may vary across the dataset. We think that more systematic study on data augmentation would benefit FixMatch.

References

[1] Gontijo-Lopes et al. Affinity and diversity: Quantifying mechanisms of data augmentation. *arXiv preprint arXiv:2002.08973*, 2020.

[2] A. Oliver et al. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, 2018.

[3] D. S. Park et al. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[4] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

[5] R. Sennrich et al. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015.

[6] O. Vinyals et al. Matching networks for one shot learning. In *NeurIPS*, 2016.