A. Contrastive Regularization for Semi-Supervised Learning

In this section, we introduce the detailed derivation of the gradients of contrastive regularization. For a model parameterized by θ, the label prediction of an unlabeled sample $u \in \mathcal{U}$ is $\hat{p}(y|u) = \text{softmax}[W^\top h_\theta(u)]$, where a mini-batch of unlabeled sample $\mathcal{U}$, K-class weight matrix $W = [w_1, w_2, ..., w_K] \in \mathbb{R}^{H\times K}$ and the penultimate features of u $h_\theta(u) \in \mathbb{R}^H$. We assume that the model parameter θ comprises the class weight matrix W. For a stochastic strong augmentation α, we define the set of m strongly augmented samples in an unlabeled mini-batch as $\mathcal{A}_m(\mathcal{U}) = \{u'_i | u \in \mathcal{U}, u'_i = \alpha(u), 1 \leq i \leq m\}$. $\hat{q}_u$ is the pseudo-label of u' and defined as $\hat{q}_u = \arg\max q_u$, where $q_u = \text{sg}[\hat{p}(y|u)]$ and sg is the stop gradient. We define the set of *pseudo*-positive pairs of u' as $\hat{P}(u') = \{p' | p' \in \mathcal{A}_m(\mathcal{U})/u', \hat{q}_p = \hat{q}_u\}$, where $\hat{q}_p$ and $\hat{q}_u$ are the pseudo-label of p' and u', respectively. Note that pseudo-labels of strongly augmented samples are defined by the label predictions on the samples *before* the strong data augmentation to improve the reliability of pseudo-labeling. Then, for an unlabeled sample u in an unlabeled mini-batch $\mathcal{U}$, contrastive regularization, $\mathcal{R}_{CR}$, is defined as follows:

$$\mathcal{R}_{CR}(\mathcal{U}) = \frac{1}{|\mathcal{A}_m(\mathcal{U})|} \sum_{u' \in \mathcal{A}_m(\mathcal{U})} \mathbb{1}[\max q_u > \delta'] r(u'), \tag{1}$$

$$r(u') = \frac{-1}{|\hat{P}(u')|} \sum_{p' \in \hat{P}(u')} \log \frac{\exp(\langle z_{u'}, z_{p'} \rangle / \tau)}{\sum_{v' \in \mathcal{A}_m(\mathcal{U})/u'} \exp(\langle z_{u'}, z_{v'} \rangle / \tau)}, \tag{2}$$

where a confidence threshold δ', a temperature scaling parameterτ, and a normalized vector of the projection head output $z_{u'}$.

Assuming that an augmented sample $u' = \alpha A(u)$ has a confident pseudo-label, we first show that the contrastive regularization moves the features of u' toward the centroid of the feature cluster having the same pseudo-label, while pushing away the features in different clusters. Without the loss of generalizability, we assume $z = h$ and $\tau = 1$. In addition, we omit θ for the notation brevity. The first-order derivative of $\mathcal{R}_{CR}$ with respect to a feature vector of u' is as follows:

$$\begin{aligned} -\frac{\partial r(u')}{\partial h(u')} &= \frac{-1}{|\hat{P}(u')|} \sum_{p' \in \hat{P}(u')} (-h(p') + \sum_{v' \in \mathcal{A}(\mathcal{U})/u'} s[u', v'] h(v')) && (3) \\ &= \frac{-1}{|\hat{P}(u')|} \sum_{p' \in \hat{P}(u')} [-h(p') + \sum_{v'_p \in \hat{P}(u')} s[u', v'_p] h(v'_p) + \sum_{v'_n \in N(u')} s[u', v'_n] h(v'_n)] && (4) \\ &= \sum_{p' \in \hat{P}(u')} (\frac{1}{|\hat{P}(u')|} - s[u', p']) h(p') - \sum_{v'_n \in N(u')} s[u', v'_n] h(v'_n) && (5) \end{aligned}$$

where $s[u', p']$ is the softmax score of the pair of $h(u')$ and $h(p')$, $N(u') = \{n' | n' \in \mathcal{A}_m(\mathcal{U}), \hat{q}_n \neq \hat{q}_u\}$, n is the original sample of strongly augmented n'. Since the sum of minus log-sum-exp terms is the convex function, the first-order optimality condition holds when the $s^*[u', v'] = 1/|\hat{P}(u')|$ if $v' \in \hat{P}(u')$ and $s^*[u', v'] = 0$ otherwise. Thus, the contrastive regularization on an unlabeled sample pushes the features of different pseudo-labels and pulls those of the same pseudo-label. Assuming that the softmax scores of the negative pairs are small enough, Eq. (5) is summarized as follows:

$$-\frac{\partial r(u')}{\partial h(u')} = \sum_{p' \in \hat{P}(u')} (\frac{1}{|\hat{P}(u')|} - s[u', p']) h(p') + R(u'), \tag{6}$$

where $R(u')$ is a remainder term. The feature vector of u' is updated toward the centroid, which is the weighted sum of positive features regardless of the confidence of the pseudo-labels.

For another unlabeled sample $v \in \mathcal{U}$ and $v' = \mathcal{A}(v)$, the contrastive regularization on features with confident pseudo-labels pushes the features of v' if v' has a different pseudo-label, and pulls them otherwise. By the same process of the derivation of above $-\partial r(u')/\partial h(u')$, we can derive $-\partial r(u')/\partial h(v')$ as follows:

$$-\frac{\partial r(u')}{\partial h_\theta(v')} = \begin{cases} (\frac{1}{|\hat{P}(u')|} - s[u', v']) h_\theta(u'), & \text{if } v' \in \hat{P}(u') \\ \quad - s[u', v'] h_\theta(u'), & \text{if } v' \notin \hat{P}(u') \end{cases}. \tag{7}$$

Table 1. The details of hyper-parameters on training datasets. LR and WD describe initial learning rate and weight decay, respectively.

Dataset	Model	B	μ	Epochs	λ_{CS}	λ_{CR}	LR	WD	δ	δ'
SVHN	WRN-28-2	16, 64	7	6500	1.0	1.0	0.03	0.0005	0.95	0.95
CIFAR-10	WRN-28-2	32, 64	7	6500	1.0	1.0	0.03	0.0005	0.95	0.95
CIFAR-100	WRN-28-8	64	7	2500	1.0	10.0	0.03	0.001	0.95	0.95
STL-10	WRN-37-2	64	7	5000	1.0	10.0	0.03	0.0005	0.95	0.95
ImageNet	ResNet-50	1024	5	300	10.0	1.0	0.05	0.003	0.7	0.7

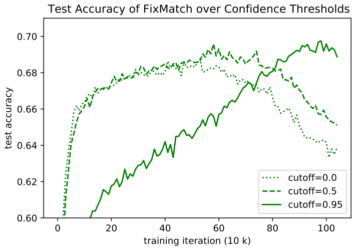


Figure 1. Test accuracy of FixMatch with different confident thresholds views over training.

B. Implementation Details

For the implementation, we follow the setting of the original FixMatch paper [1] except the hyper-parameters related to the contrastive regularization. We use Pytorch 1.6.0 to reproduce FixMatch and implement the contrastive regularization on the same codebase. We conduct all experiments using four Tesla V100 32GB GPUs, except the ImageNet dataset. For ImageNet, we use 32 V100 GPUs for FixMatch and 64 GPUs for FixMatch with the contrastive regularization. For all datasets, we use the stochastic gradient descent (SGD) optimizer with Nesterov momentum $\beta = 0.9$, and temperature parameter $\tau = 0.01$. We use an exponential moving average (EMA) of model parameters with 0.999 momentum and cosine learning scheduling used in [1]. The batch size of labeled data (B) is 64 for SVHN, CIFAR-10, CIFAR-100, and STL-10 except SVHN with 20 and 40 labels, and CIFAR-10 with 20 labels. Considering the small number of labeled samples, we use 16 labeled samples for training SVHN with 20 and 40 labels, and 32 labeled samples for CIFAR-10 with 20 labels per training iteration of WRN-28-2. For WRN-28-8, we use 16 labeled samples for training SVHN with 20 and 40 labels, and 32 labeled samples for CIFAR-10 with 20 and 40 labels.

In the ablation study for different hyper-parameters, we use WRN-28-4 for CIFAR-100 with 2500 labels, because WRN-28-8 requires over two times more training time than WRN-28-4 but the accuracy gain is marginal (+0.83% in Table **??**). We use random horizontal flipping and the random crop for both weak and strong augmentations of training datasets. For the SVHN dataset, we do not use horizontal flipping, considering that a classifier can be easily confused to discriminate some classes such as eight and three. For strong augmentations, we use RandAugment [1] following the original paper of FixMatch. Please refer to the original paper of FixMatch and our source codes to check the details of augmentation policies according to the datasets. The other training details are available in Table 1.

C. Additional Experimental Results

In this study, we have claimed that the previous consistency regularization suffers from the training inefficiency by the exclusion of unconfident pseudo-labels to ensure the reliability of pseudo-labeling. Figure 1 shows the relationship between the confident threshold δ and the convergence speed of SSL training in FixMatch. When low confidence thresholds are used such as 0.0 or 0.7, FixMatch can leverage more unlabeled samples than FixMatch with $\delta = 0.95$. Thus, SSL training with low confidence thresholds can achieve the best performance much faster and increase the training efficiency with respect to the training iterations and time. However, the low δs lower the best test accuracies, because the unreliable pseudo-labeling propagates wrong labeling information to other unlabeled samples. Meanwhile, when a high confidence threshold is used, the accuracy increases slower especially in the early stage of training, because it cannot leverage many unlabeled samples. Despite the low training efficiency, FixMatch with $\delta = 0.95$ can achieve high performance, while leveraging only its confident labeling information. The results imply that the training inefficiency of the previous consistency regularization comes from the trade-off between the reliability of pseudo-labeling and the number of used unlabeled samples in training.

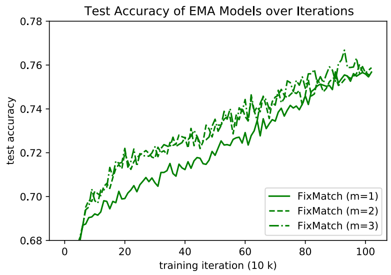
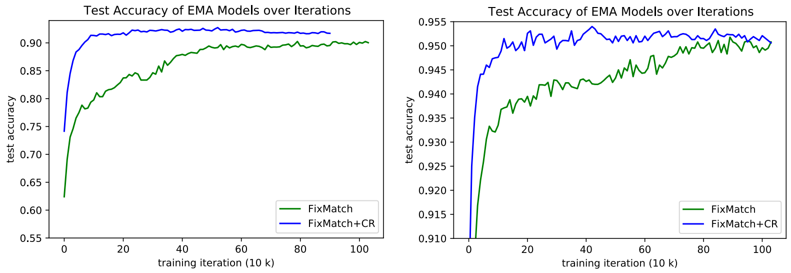


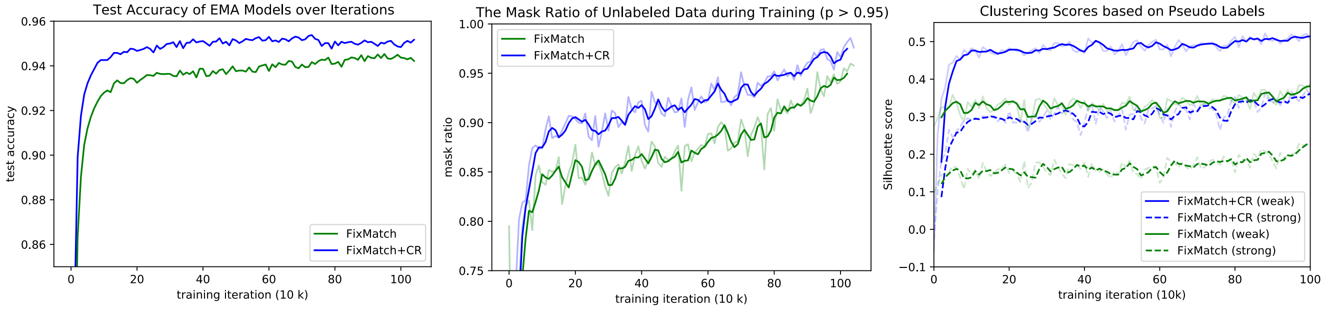
Figure 2. Test accuracy of FixMatch with different views over training.



(a) Test Accuracy of the EMA Model (STL-10, 1000 labels)

(b) Test Accuracy of the EMA Model (STL-10, 5000 labels)

Figure 3. Test accuracy of FixMatch and FixMatch+CR of WRN-37-2 trained on STL-10 with (a) 1000 and (b) 5000 labels.



(a) Test Accuracy of the EMA Model

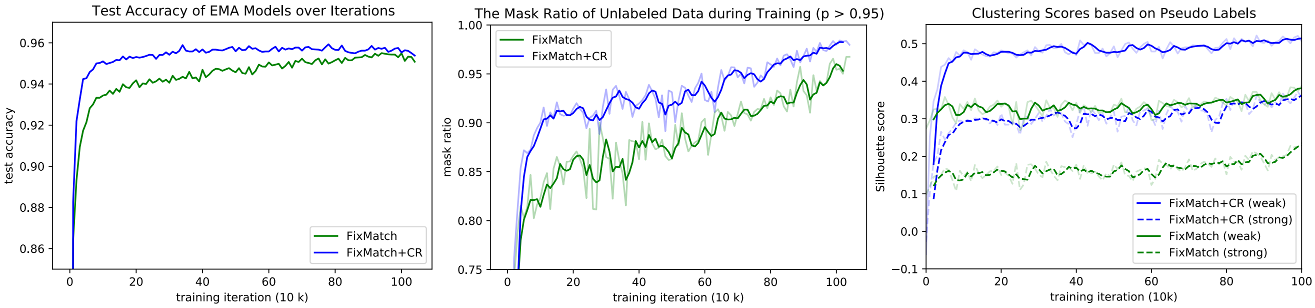
(b) The Ratio of Selection Mask in Training

(c) Clustering Score

Figure 4. Empirical Results of FixMatch and FixMatch+CR with WRN-28-2 trained on the CIFAR-10 with 250 labels. (a) Test accuracy of EMA models, (b) the ratio of selection mask, and (c) Silhouette score of penultimate features of unlabeled samples based on pseudo-labels.

In the ablation study, we have shown that the increased views (m) of unlabeled samples also increase the accuracy of FixMatch. However, we show that more numbers of views cannot improve the training efficiency, and the accuracy over training still increases gradually in the training. The results imply that the effect of our contrastive regularization does not result from the increased views of unlabeled samples, but well-clustered representations for SSL.

We show the additional results to show the more efficient training of contrastive regularization than consistency regularization. For better visualization, we try to train FixMatch+CR in the same epochs of FixMatch, 10,500 epochs, until FixMatch+CR starts to be overfitted. First, we show that the results described in Section 4.3 are also consistent with other datsets such as CIFAR-10 (Figure 6) and STL-10 (Figure 3). Figure 4 and 5 show the test accuracy of EMA models, the average ratio of selection mask in a training mini-batch, and clustering scores of features in training on the CIFAR-10 dataset. The results imply that our contrastive regularization learns well-clustered features for SSL and requires fewer iterations for high performance. The accuracy of FixMatch gradually increases over the entire training iterations, but the accuracy of FixMatch+CR increases much faster than FixMatch, especially in the early stage of training. Moreover, we find that about 10% of iterations for FixMatch+CR are enough to achieve the performance of FixMatch.

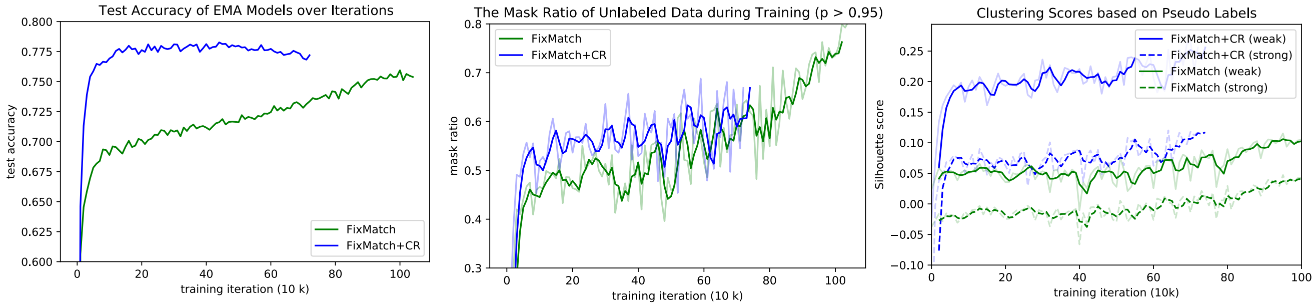


(a) Test Accuracy of the EMA Model

(b) The Ratio of Selection Mask in Training

(c) Clustering Score

Figure 5. Empirical Results of FixMatch and FixMatch+CR with WRN-28-2 trained on the CIFAR-10 with 4000 labels. (a) Test accuracy of EMA models, (b) the ratio of selection mask, and (c) Silhouette score of penultimate features of unlabeled samples based on pseudo-labels.



(a) Test Accuracy of the EMA Model

(b) The Ratio of Selection Mask in Training

(c) Clustering Score

Figure 6. Empirical Results of FixMatch and FixMatch+CR with WRN-28-4 trained on the CIFAR-100 with 1000 labels. (a) Test accuracy of EMA models, (b) the ratio of selection mask, and (c) Silhouette score of penultimate features of unlabeled samples based on pseudo-labels.

References

[1] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33, 2020. 2