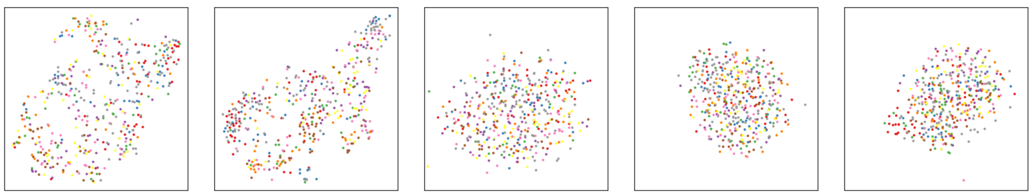
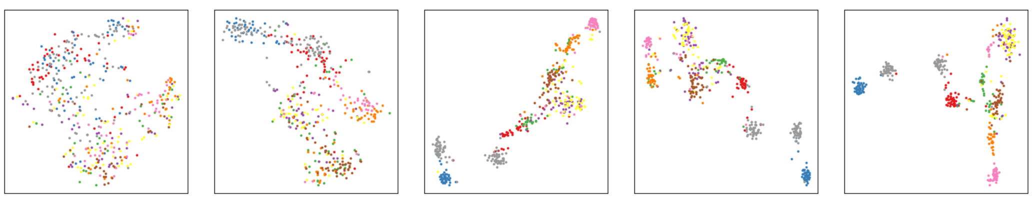
A T-SNE visualization

To visually validate the ability of separating different classes in FixMatch and our method, we observe T-SNE results of their feature representation during the training process. The results are shown in Figure 1 2. Due to the lack of labeled samples, FixMatch is difficult to distinguish samples from different categories during the training process (*i.e.*, 1-st to 200-th epoch). We notice that, in FixMatch, only imposing the consistent constraint causes the samples gradually being closer together as shown in Figure 1. On the contrary, by mining the super-class relation between samples, our method can escape from this dilemma with more informative representations in Figure 2. In the early stage of training, although only a small part of the categories could be distinguished, with the refinement of the super-class, more categories of samples will be gradually distinguished in the later training process.



(a) 1-st epoch (b) 10-th epochs (c) 50-th epochs (d) 100-th epochs (e) 200-th epochs

Figure 1: Feature visualization of FixMatch in the training process (CIFAR-10 with 10 labels)



(a) 1-st epoch (b) 10-th epochs (c) 50-th epochs (d) 100-th epochs (e) 200-th epochs

Figure 2: Feature visualization of our method in the training process (CIFAR-10 with 10 labels)

B Why the progressive form of the super-class is important

In our method, at the beginning of training process, when the number of super-classes is relative small, the learning of discriminative information is safe and reliable. However, with the training process, the performance will be largely limited by this small number of super-classes. Here, we provide proofs from an perspective of information amount.

Theorem 1. *Given the number K of super-class and dataset $\mathcal{D}$, the upper bound of information amount produced by dividing the samples of $\mathcal{D}$ into super-classes is $|\mathcal{D}| \log_2 K$.*

Proof 1. *Assuming that each sample $u_i \in \mathcal{D}$ is a signal source, when the information amount becomes its theoretical largest value, the entropy is also up to its largest value (i.e., u_i belongs to each super-class with equal probability). Then the upper bound of information amount produced from u_i is:*

$$\sup H(u_i) = -\sum_{i=1}^{K} \frac{1}{K} \log_2 \frac{1}{K} = \log_2 K,$$

where K is the number of super-class. Then the upper bound of information amount produced from $\mathcal{D}$ is:

$$\sup H(\mathcal{D}) = \sum_{i=1}^{|\mathcal{D}|} H(u_i) = |\mathcal{D}| \log_2 K$$

C Hyperparameter setting

We show the detailed hyperparameters setting for each dataset in table 1.

Table 1: The detailed hyperparameter setting in our method

	CIFAR-10	CIFAR-100	STL-10
Learning Rate		0.03	
SGD Momentum		0.9	
EMA Momentum		0.99	
Batch Size		64	
τ_1		0.95	
τ_2		0.8	
λ_{con}		1	
λ_{dis}		1	
Net	WRN-28-2	WRN-28-8	ResNet-18
Weight Decay	5e-4	1e-3	5e-4
Set of K	{3, 5, 10}	{5, 10, 20}	{3, 5, 10}
α	0.3	0.5	0.3