A. Implementation Detail

A.1. Hyperparameters

As mentioned in section 4, our hyperparameters are almost identical to that of MixMatch [2] and FixMatch [26]. We use the same network architecture and similar hyperparameters as FixMatch for CIFAR-10, SVHN, and CIFAR-100 (WRN 28-8). We conducted ablation study on WRN 28-2 with hyperparameters similar to that of MixMatch for simplicity. We also evaluated SimPLE on Mini-ImageNet with WRN 28-2 and ResNet 18. We use the same α (beta distribution parameter for mix-up [35]) and T (temperature for sharpening) across all experiments. Notice that only MixMatch and MixMatch Enhanced use mix-up.

The full detail of our hyperparameters choices can be found in table 7 and 8. Our transfer experiment configurations are in table 9.

	CIFAR-10	SVHN	CIFAR-100
τ_c	0.95		
τ_s	0.9		
$\lambda_{\mathcal{U}}$	75	250	150
$\lambda_{\mathcal{P}}$	75	250	150
lr	0.03		
K	7		4
T	0.5		
α	0.75		
weight decay	0.0005		0.001
batch size	64		
EMA decay	0.999		
backbone	WRN 28-2		WRN 28-8
optimizer	SGD		
Nesterov	True		
momentum	0.9		
lr scheduler	cosine decay		
lr decay rate	7π / 16		

Table 7: Hyperparameters for CIFAR-10, SVHN, and CIFAR-100 (with WRN 28-8).

	CIFAR-100	Mini-ImageNet	
τ_c	0.95		
τ_s	0.9		
$\lambda_{\mathcal{U}}$	150	300	
$\lambda_{\mathcal{P}}$	150	300	
lr	0.002		
K	2	7	
T	0.5		
α	0.75		
weight decay	0.04	0.02	
batch size	64	16	
EMA decay	0.999		
backbone	WRN 28-2	WRN 28-2	ResNet 18
optimizer	AdamW		

Table 8: Hyperparameters for CIFAR-100 (WRN 28-2) and Mini-ImageNet.

	DN-R to M-IN	IN-1K to DN-R
τ_c	0.95	
τ_s	0.9	
$\lambda_{\mathcal{U}}$	300	
$\lambda_{\mathcal{P}}$	300	
feature lr	0.0002	0.00002
classifier lr	0.002	
K	2	
T	0.5	
α	0.75	
weight decay	0.02	
batch size	16	
EMA decay	0.999	
backbone	WRN 28-2	ResNet 50
optimizer	AdamW	

Table 9: Hyperparameters for Transfer: DomainNet-Real to Mini-ImageNet (DN-R to M-IN) and Transfer: ImageNet-1K to DomainNet-Real (IN-1K to DN-R) experiments.

A.2. Optimization

For CIFAR-10, SVHN, and CIFAR-100 (WRN 28-8), we use SGD with Nesterov momentum set to 0.9. We also use cosine learning rate decay [18] with a decay rate of $\frac{7\pi}{16}$ following FixMatch. For CIFAR-100 (WRN 28-2), Mini-ImageNet, and transfer experiments, we use AdamW [20] without learning rate scheduling follows that of MixMatch. Details are available in table 7, 8 and 9.

A.3. Augmentations

Our augmentations are implemented on GPU with Kornia [24]. In table 10, we list the transformations used by the fixed augmentations of table 4 and 5. For RandAugment [6], we follows the exact same settings as FixMatch [26]. Note that we only reported the changed augmentation parameters while the omitted values are the same as the default parameters in Kornia [24].

B. Further Analysis on Pair Loss

B.1. Analysis on Confidence Threshold

Theorem 1 $\forall p, q \in \Delta^N$, *if* $\varphi_{\tau_c}(\max(p)) \cdot \varphi_{\tau_s}(f_{\text{sim}}(p, q)) > 0$, *then* $\max(q) > \cos(\cos^{-1}(\sqrt{\tau_c}) + \cos^{-1}(\tau_s))^2$.

Transformation	Description	Parameter
Random Horizontal Flip	Horizontally flip an image randomly with a given probability p	$p = 0.5$
Random Resized Crop	Random crop on given size and resizing the cropped patch to another	scale $= (0.8, 1)$, ratio $= (1, 1)$
Random 2D GaussianBlur	Creates an Gaussian filter for image blurring. The blurring is randomly applied with probability p	$p = 0.5$, kernel size $= (3, 3)$, sigma $= (1.5, 1.5)$
Color Jitter	Randomly change the brightness, contrast, saturation, and hue of given images	contrast $= (0.75, 1.5)$
Random Erasing	Erases a randomly selected rectangle for each image in the batch, putting the value to zero	$p = 0.1$
Random Affine	Random affine transformation of the image keeping center invariant	degrees $= (-25, 25)$, translate $= (0.2, 0.2)$, scale $= (0.8, 1.2)$, shear $= (-8, 8)$

Table 10: Augmentation details. Applied in order. Descriptions are from [24].

Since $\varphi_{\tau_c}(\max(p)) \cdot \varphi_{\tau_s}(f_{\text{sim}}(p,q)) > 0$, we have:

$$\begin{cases} \max(p) > \tau_c \\ f_{\text{sim}}(p,q) > \tau_s \end{cases}$$

Denote $j = \arg\max_i p_i$, i.e., the confidence of p is attained at the j-th coordinate, $p_j = \max(p)$.

Denote $e_j \in \Delta^n$ as the elementary vector with the j-th element to be 1 and all other elements to be 0.

In the square root probability space, we have:

$$\begin{cases} \sqrt{e_j}^\top \sqrt{p} = \max\left(\sqrt{p}\right) \quad > \sqrt{\tau_c} \\ \sqrt{p}^\top \sqrt{q} > \tau_s \end{cases}$$

Notice, because $\|p\|_1 = \|q\|_1 = \|e_j\|_1 = 1$, we have $\left\|\sqrt{p}\right\|_2 = \left\|\sqrt{q}\right\|_2 = \left\|\sqrt{e_j}\right\|_2 = 1$. Therefore, $\sqrt{p}$, $\sqrt{q}$, and $\sqrt{e_j}$ are on the unit n-sphere S_n. Denote the geodesic distance between any two points $x, y \in S_n$ as $d_{S_n}(x,y) = \cos^{-1}(\frac{x^\top y}{\|x\|_2 \cdot \|y\|_2}) = \cos^{-1}(x^\top y)$.

$$\begin{cases} d_{S_n}\left(\sqrt{p}, \sqrt{e_j}\right) > \cos^{-1}(\sqrt{\tau_c}) \\ d_{S_n}\left(\sqrt{p}, \sqrt{q}\right) > \cos^{-1}(\tau_s) \end{cases}$$

As the geodesic distance preserves triangular inequality:

$$\begin{aligned} d_{S_n}\left(\sqrt{q}, \sqrt{e_j}\right) &\geq d_{S_n}\left(\sqrt{q}, \sqrt{p}\right) + d_{S_n}\left(\sqrt{p}, \sqrt{e_j}\right) \\ &> \cos^{-1}(\sqrt{\tau_c}) + \cos^{-1}(\tau_s) \end{aligned}$$

$$\sqrt{q_j} = \sqrt{q}^\top \sqrt{e_j} > \cos(\cos^{-1}(\sqrt{\tau_c}) + \cos^{-1}(\tau_s))$$

$$\max(q) \geq q_j > \cos(\cos^{-1}(\sqrt{\tau_c}) + \cos^{-1}(\tau_s))^2$$

B.2. More on Pair Loss

In this section, we provide additional information to two existing ablation studies in table 6 on CIFAR-100, to demonstrate the effectiveness of Pair Loss in encouraging more unlabeled samples to have accurate and high confidence predictions. Specifically, we compare the performance of the SimPLE algorithm with and without the Pair Loss enabled in the following measurements: 1) the percentage of unlabeled samples with high confidence pseudo labels; 2) the percentage of unlabeled sample pairs that pass both confidence and similarity thresholds; 3) the percentage of false-positive unlabeled sample pairs that pass both confidence and similarity thresholds but are in different categories.

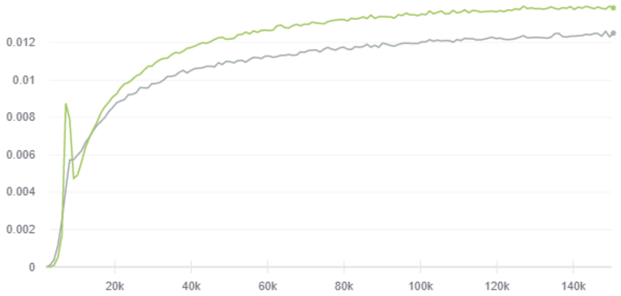

Figure 4: Ratio of pairs pass both confidence and similarity thresholds. The green line is SimPLE and the grey line is SimPLE without Pair Loss

From figure 4, the ratio of pairs that pass both the confidence threshold and similarity threshold is increased by 16.67%, with a consistently nearly 0% false positive rate, which indicates that Pair Loss encourages the model to make more consistent and similar predictions for unlabeled samples from the same class.

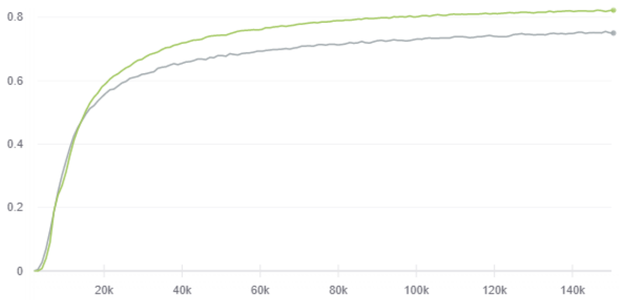

Figure 5: Ratio of high confidence prediction. The green line is SimPLE and the grey line is SimPLE without Pair Loss

As shown in figure 5, with Pair Loss, the percentage of unlabeled sample with high confidence labels is increased by 7.5%, and the prediction accuracy is increased by 2% as shown in table 6. These two results indicate that Pair Loss encourages the model to make high confidence and accurate predictions on more unlabeled samples, which follows our expectation that Pair Loss aligns samples with lower confidence pseudo labels to their similar high confidence counterparts during the training and improves the prediction accuracy.