Supplementary Material

1. Overview

In section 2, we describe our Manifold DivideMix algorithm with more details, then hyperparameter settings and experimental results are explained in section 3. Finlay, in Section 4 more ablation studies are discussed.

2. Manifold DivideMix Algorithm

Algorithm1 delineates the full steps of our proposed method. After training the backbone model in a self-supervised way, and before warm-up training, we first apply KNN on the embedding space to identify a small and constant percentage of the OOD samples (see hyperparameters settings in Table1). Then, we add the classification head on top of the backbone and perform warm-up training on the classification layer, followed by fine-tuning of the backbone model with a smaller learning rate. Subsequently, we initiate the Semi-supervised learning step by applying sample selection mechanisms to identify both ID and OOD labeled noise samples, using GMM and KNN every 10 epochs.

2.1. Loss Functions

The Semi-SL loss function is computed based on the strongly augmented copies as well as the MixUp augmentations of the input and embedding spaces of samples as follows:

$$\mathcal{L}^{semi} = \mathcal{L}^{sup} + \lambda_u \mathcal{L}^{unsup} + \lambda_c \mathcal{L}^{self}, \tag{1}$$

where λ_u and λ_c are unsupervised loss coefficient and contrastive loss coefficient which set to 1 and 100, respectively to balance the scale of each loss term in proposed loss function(1).

The supervised loss function ($\mathcal{L}^{sup}$) consists of two symmetric cross-entropy loss terms on the mixed-up augmented input ($\mathcal{X}'$) and the mixed-up augmented embedding spaces ($\mathcal{H}'_x$) of the labeled set. The symmetric cross-entropy proposed in [12] as $\ell_{sce} = \ell_{ce} + \ell_{rce}$ where ℓ_{ce} is the cross-entropy and ℓsce is the reverse cross-entropy defined as follows:

$$\ell_{ce} = -\sum_{c=1}^{C} q(c|x) \log p(c|x), \tag{2}$$

$$\ell_{rce} = -\sum_{c=1}^{C} p(c|x) \log q(c|x), \tag{3}$$

where $q = q(c|x)$ represents the true distribution of class labels for sample x and $p(c|x)$ represents the predicted distribution of class labels based on a model consisting of a backbone network f_1 and a classifier f_2. Both distributions are conditional on the input sample x.

The unsupervised loss function ($\mathcal{L}^{unsup}$) consists of two Mean Squared loss terms. One of them defined on the mixed-up augmented input ($\mathcal{U}'$) denoted as follows:

$$\mathcal{L}_{\mathcal{U}'} = \frac{1}{|\mathcal{U}'|} \sum_{u',q \in \mathcal{U}'} \|q - \mathrm{p}(u'; \theta, \phi)\|_2^2, \tag{4}$$

where $\mathrm{p}(u'; \theta, \phi)$ is the probability output of our model with parameter θ, ϕ for backbone and classification head, respectively. The other Mean Squared loss defined on the mixed-up augmented embedding spaces ($\mathcal{H}'_u$) of the unlabeled set denoted as follows:

$$\mathcal{L}_{\mathcal{H}'} = \frac{1}{|\mathcal{H}'|} \sum_{h',q \in \mathcal{H}'} \|q - \mathrm{p}(h'; \phi)\|_2^2, \tag{5}$$

The contrastive loss function ($\mathcal{L}^{self}$), defined in Equation (1) in the submitted manuscript, is applied to the projection head of the embedding spaces of $\mathcal{H}_x$ and $\mathcal{H}_u$.

2.2. ID and OOD Labeled Noise Detection

In Figure 1, the Area Under a Curve (AUC) is presented, which shows the performance of clean/ID labeled noise detection using GMM on the training data of CIFAR-100 corrupted with ImageNet32 at various levels of ID and OOD noise. The results indicate that warm-up training on top of the pre-trained SSL model is highly effective in separating clean and ID labeled noise samples, even in the first epoch of the Semi-SL step when there are high levels of noisy samples. This demonstrates the efficacy of the proposed method in improving the performance of noise detection, which is crucial for training robust and generalized model in the presence of label noise.

The results presented in Figure 2 illustrate the effectiveness of our OOD labeled noise detection mechanism in various noise settings. Our method consistently improves the accuracy of OOD detection regardless of the level of noise present. When high rates of OOD labeled noise are present, the KNN algorithm often struggles to distinguish between clean, ID, and OOD labeled noise samples. However, our approach of separating the OOD labeled noise samples proves to be effective even in such challenging scenarios. As the accuracy improves, the network learns more discriminative features from labeled data and achieves better generalization to unlabeled data by iteratively detecting ID labeled noise and removing OOD labeled noise samples.

3. Training Details

Our proposed method uses consistent hyperparameter settings where the majority of parameters remain constant across different datasets, demonstrating its versatility. The list of all hyperparameters settings of different steps of our method described in Table 1. Additionally, we incorporate the Cosine Annealing method as a learning rate scheduler.

Algorithm 1 Manifold DivideMix

Input: θ, ψ, ϕ ▷ Backbone, Projection head and classification head parameters
Input: $(\mathcal{X}, \mathcal{Y})$ ▷ Training Data
Input: λ_u, λ_c ▷ Unsupervised and contrastive loss weights
Input: $k, initRemoval, p$ ▷ Parameter of KNN, constant ratio for OOD removal, percentage of removal
Input: τ_2, α ▷ GMM threshold, Parameter of Beta distribution for Mixup

$\theta, \psi \leftarrow$ SSLModel$(\mathcal{X}, \theta, \psi)$

OODScore $\leftarrow$ KNN$(\mathcal{X}, \theta, \phi, k)$ ▷ Compute OOD score based on the average distance of k-nearest neighbors
RemovalRate $\leftarrow initRemoval$ ▷ Initial sample removal rate
OODMask $\leftarrow$ TopScore$(OODScore, RemovalRate)$ ▷ Select samples with highest OOD score to discard them

$(\theta, \phi) \leftarrow$ Warmup$(\mathcal{X}, \mathcal{Y}, \theta, \phi,$ OODMask$)$ ▷ Warmup training of f_2 and fine-tuning of f_1

while $e \leq$ MaxEpoch / 10 **do**

  $OODScore \leftarrow$ KNN$(\mathcal{X}, \theta, \phi, k)$ ▷ Compute OOD score based on the average distance of k-nearest neighbors
  $RemovalRate \leftarrow RemovalRate + p$ ▷ Compute percentage of sample removal
  OODMask $\leftarrow TopScore(OODScore, RemovalRate)$ ▷ Select samples with highest OOD score to discard them
  $\mathcal{W} \leftarrow$ GMM$(\mathcal{X}, \mathcal{Y}, \theta, \phi,$ OODMask$)$ ▷ Model per-sample loss distribution to obtain clean probability

  for $iter \leftarrow 1$ **to** 10 **do**

    $\mathcal{X} \leftarrow \{(\mathbf{x}_i, \mathbf{y}_i, w_i) : w_i = p\left(x_i = \text{clean} | \ell^{sup}, \gamma\right) \geq \tau_2, \forall (x_i, y_i, w_i) \in (\mathcal{X}, \mathcal{Y}, \mathcal{W})\}$ ▷ Construct labeled set
    $\mathcal{U} \leftarrow \{(\mathbf{x}_i, \mathbf{y}_i, w_i) : w_i = p\left(x_i = \text{clean} | \ell^{sup}, \gamma\right) < \tau_2, \forall (x_i, w_i) \in (\mathcal{X}, \mathcal{W})\}$ ▷ Construct unlabeled set

    $\mathcal{L}^{semi} \leftarrow \mathcal{L}^{sup} + \lambda_u \mathcal{L}^{unsup} + \lambda_c \mathcal{L}^{self}$ ▷ Compute total loss
    $\theta, \phi = SGD(\mathcal{L}^{semi}, \theta, \phi)$ ▷ Update model parameters

4. Ablation Studies

In this section, we analyze the performance of proposed method under different scenarios.

First, we compare our proposed pipeline with the most previous works on Webvision dataset that used Inception-ResNetV2 model [10] which has about $65M$ parameters, while we train PreActResNet-50 which has about $25.6M$ parameters. Compared to the state-of-the-art, our algorithm performs on par with those methods that used two or an ensemble of $2\times$ larger model. We believe that by using the self-supervised pre-training step and considering contrastive loss during the semi-supervised step, the model learns more generalizable features, which reduces the risk of overfitting to noisy samples as well as overconfident prediction on the semantically different class samples in the noisy real-world dataset even with a smaller model. Also, while PropMix [3] and SNCF [1] utilize contrastive learning based on the Inception-ResNetV2 model, Manifold DivideMix has comparable performance using a smaller model with the idea of mixing up the input and embedding spaces. As our method only engages a single network, we highlight the methods that utilise an ensemble model with"$\star$" in the Table 2.

Second, to show the robustness of our method to deal with label noise, we compare the training accuracy of our method with the standard training method. In Figure 3, we observe that with standard training (SSL+LC), the accuracy gradually improves during training over different epochs, indicating that the network is memorizing the label noise. In contrast, our proposed method quickly saturates the training accuracy, meaning that it prevents the network from memorizing the incorrect labels at a later stage of training. As a result, with the help of semi-supervised learning and ID/OOD labeled noise detection, the performance of the model on the clean test set increases(Figure 3b).

Third, Figure 4 shows the impact of varying k values on the KNN model's performance. The test-time performance of the proposed model is found to be relatively consistent across different k values. For training on the CIFAR-100 dataset corrupted by ImageNet32 with 20% ID and 20% OOD labeled noise, we choose 10, 100, and 300 as potential k values. Since these values result close and similar perfor-

Steps	Hyperparameters	CIFAR10/100	WebVision
All	Optimizer	SGD	SGD
SSL	Initial Learning Rate	0.5	0.5
	Total Epochs	1000	1000
	Temperature (τ)	0.1	0.1
Warmup and Semi-Supervised	Initial Learning Rate (Classifier f_2)	0.2	0.02
	Initial Learning Rate (Backbone f_1)	0.002	0.0002
	Momentum	0.9	0.9
	Weight Decay	$5e^{-4}$	$5e^{-4}$
	Mini-batch Size	64	32
	Total Epochs	300	100
	Warmup Epochs	20	20
	Percentage of removal (p)	0.05	0.05
	Initial OOD removal	0.1	0.1
	GMM Threshold (τ_2)	0.3	0.3
	$\lambda_{\mathcal{C}}$	1.0	1.0
	$\lambda_{\mathcal{U}}$	100	100
	Sharpening Tempreture (T)	0.5	0.5
	κ-Neirest Neighbor	100	100
	MixUp,Manifold MixUp α	4	4

Table 1. Hyperparameter settings for our proposed method.

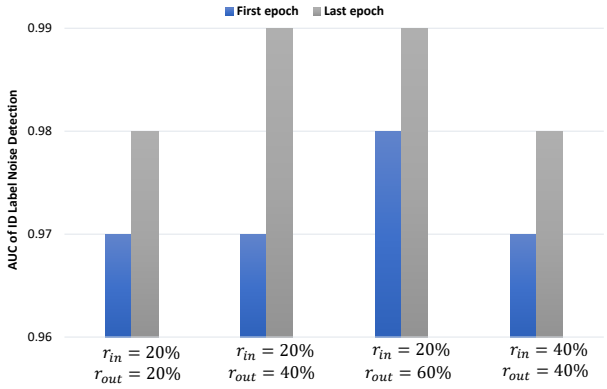


Figure 1. AUC of clean/ID labeled noise detection using GMM on training data of CIFAR-100 corrupted with ImageNet32. First epoch means the first epoch of our Semi-Supervised learning step (first epoch after warmup training) and last epoch means the final last epoch of semi-supervised learning.

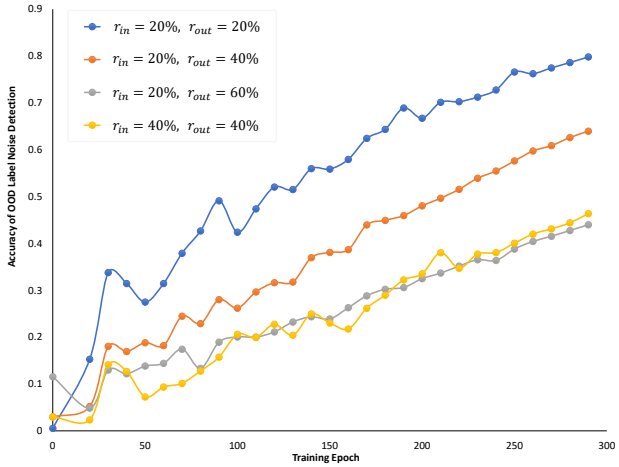


Figure 2. Accuracy of OOD labeled noise detection during warmup and semi-supervised learning steps for different level of in- and out-of-distribution labeled noise.

mance, we use $k = 100$ for all datasets regardless of the number of classes, samples, noise type, and noise rate.

Forth, a t-SNE visualization[11] of embedding space of training images is presented in Figure 5. We show the visualization for three stages of our algorithm: (1) self-supervised training (first column of Figure 5), (2) warmup training (second column of Figure 5), and (3) semi-supervised learning (third column of Figure 5). During self-supervised training step, the visualization shows two separate clusters: one for ID labeled noise and clean images and another for OOD labeled noise samples. In the warmup training stage, the algorithm starts to learn how to classify the clean images. This causes the ID and OOD labeled noise to scatter throughout the embedding space. In the final stage of training, the semi-supervised learning step, the algorithm is trained with some labeled examples to improve its classification performance. At this stage, the OOD labeled noise samples are seen to scatter along the cluster boundary, while the ID labeled noise samples form compact clusters with the clean

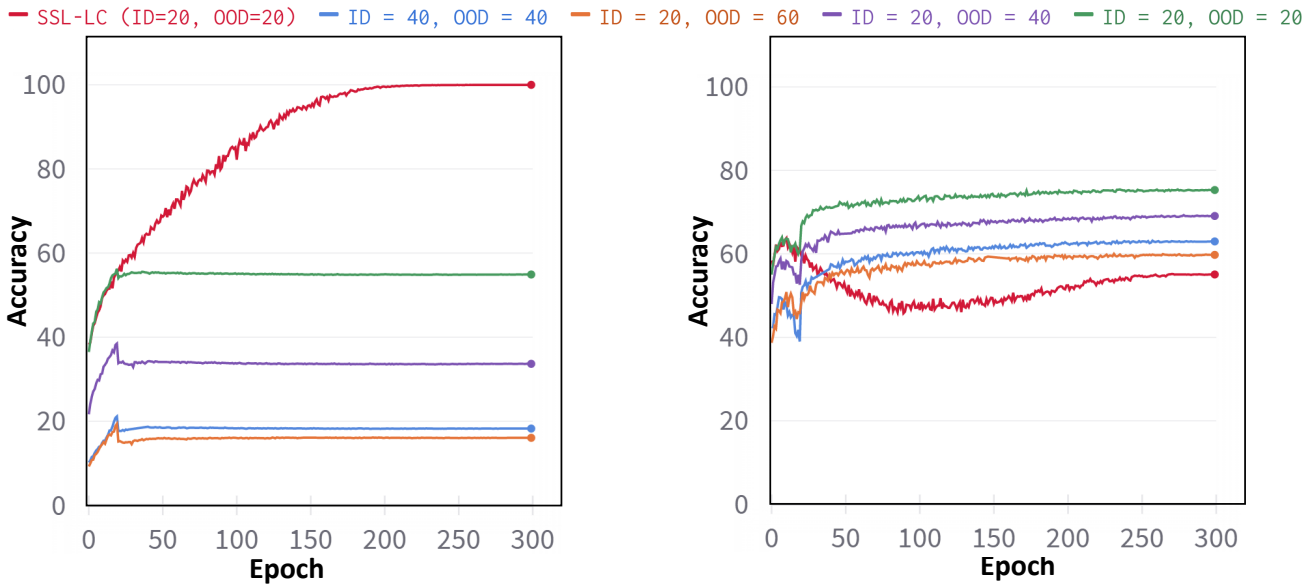


Figure 3. Top1 Accuracy of CIFAR-100 corrupted with ImageNet32 at different noise rate for (a) training and (b) test data.

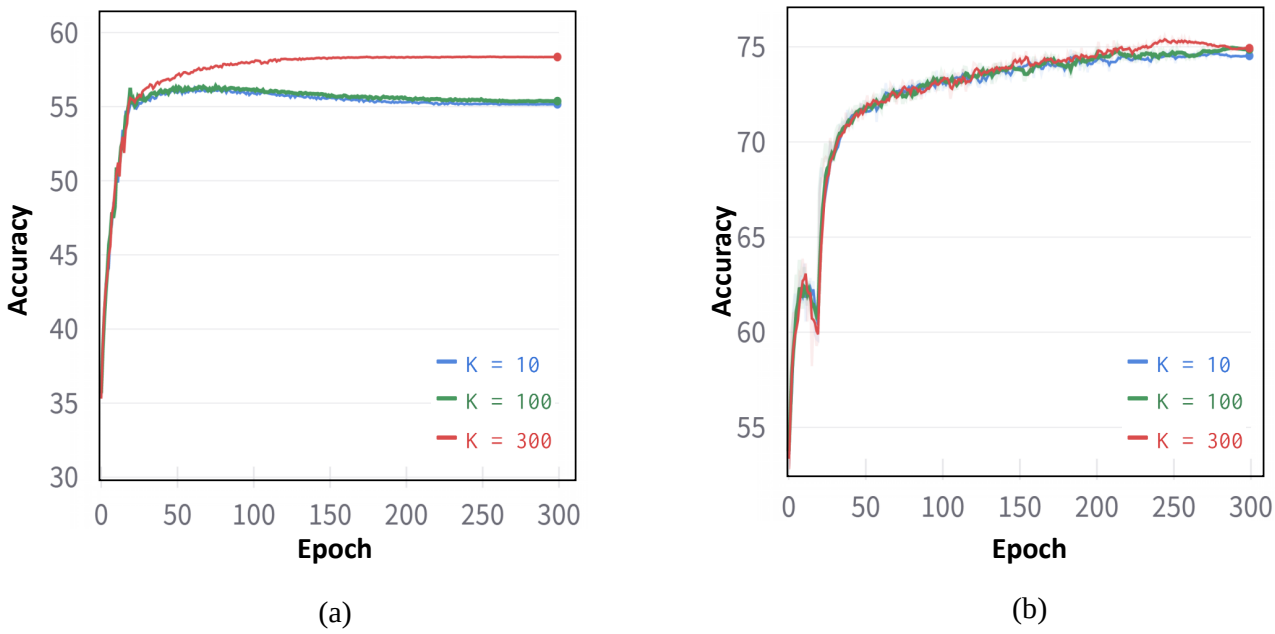


Figure 4. Top1 Accuracy of CIFAR-100 corrupted with ImageNet32 ($r_{in} = 20\%$ and $r_{out} = 20\%$) with three different K for (a) training and (b) test data.

samples (Figure 5a). The separation between the clusters becomes better as the OOD noise level decreases, but it becomes more difficult to separate the clean samples from the ID and OOD labeled noise samples at higher levels of OOD noise (Figure 5b and 5c). Overall, the t-SNE visualization helps understand how the algorithm is learning during different stages of training and how it is able to separate clean samples from ID and OOD labeled noise samples.

Method	Top-1	Top-5
Mixup [13]	75.4	90.1
MentrorNet [5]	63.0	81.4
*Co-Teaching [4]	63.6	85.2
*DivideMix [7]	77.3	91.6
*ELR [8]	77.8	91.7
*UNICON [6]	77.6	**93.4**
ScanMix [9]	**80.0**	93.0
*DSOS [2]	78.8	92.3
SNCP [1]	78.2	92.6
Ours	78.4	92.0

Table 2. Comparison of classification accuracy with the state-of-the-art methods on (mini)Webvision. "⋆" denotes algorithms using an ensemble of networks to predict.

References

[1] Paul Albert, Eric Arazo, Noel E O'Connor, and Kevin McGuinness. Embedding contrastive unsupervised features to cluster in-and out-of-distribution noise in corrupted image datasets. In *Computer Vision–Eccv 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 402–419. Springer, 2022. 2, 5

[2] Paul Albert, Diego Ortego, Eric Arazo, Noel O'Connor, and Kevin McGuinness. Addressing Out-of-Distribution Label Noise in Webly-Labelled Data. In *Winter Conference on Applications of Computer Vision (WACV)*, 2022. 5

[3] Filipe R Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. PropMix: Hard Sample Filtering and Proportional MixUp for Learning With Noisy Labels. *arXiv: 2110.11809*, 2021. 2

[4] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Co-Teaching: Robust Training of Deep Neural Networks With Extremely Noisy Labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 5

[5] L. Jiang, Z. Zhou, T. Leung, L.J. Li, and L. Fei-Fei. MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels. In *International Conference on Machine Learning (ICML)*, 2018. 5

[6] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9676–9686, 2022. 5

[7] Junnan Li, Hoi Socher, Richard, Steven CH, and Steven CH Hoi. DivideMix: Learning With Noisy Labels as Semi-Supervised Learning. In *International Conference on Learning Representations (ICLR)*, 2020. 5

[8] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda. Early-Learning Regularization Prevents Memorization of Noisy Labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 5

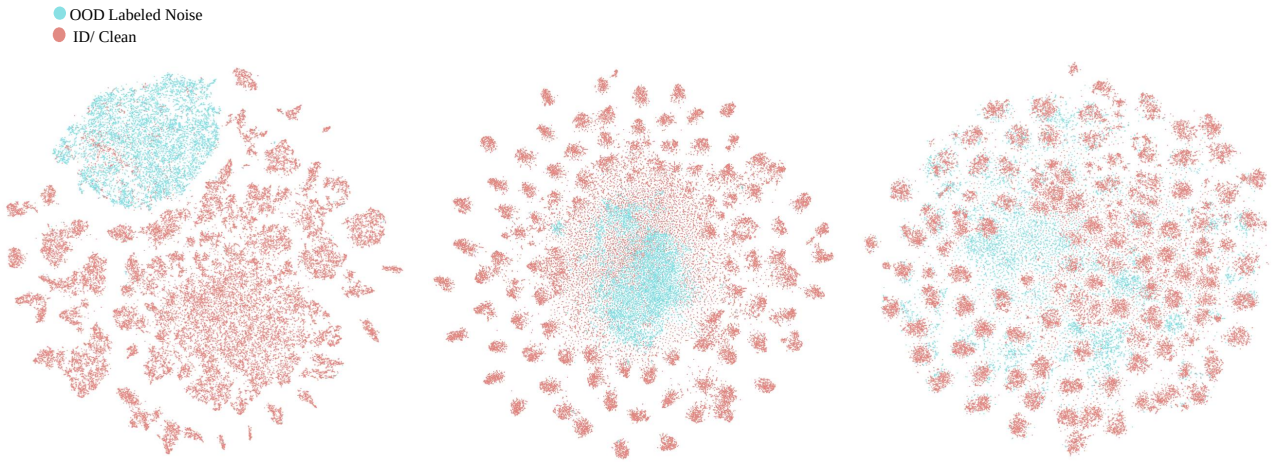
[9] Ragav Sachdeva, Filipe R Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. ScanMix: Learning From Severe Label Noise via Semantic Clustering and Semi-Supervised Learning. *arXiv: 2103.11395*, 2021. 5

[10] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-V4, Inception-Resnet and the Impact of Residual Connections on Learning. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2016. 2

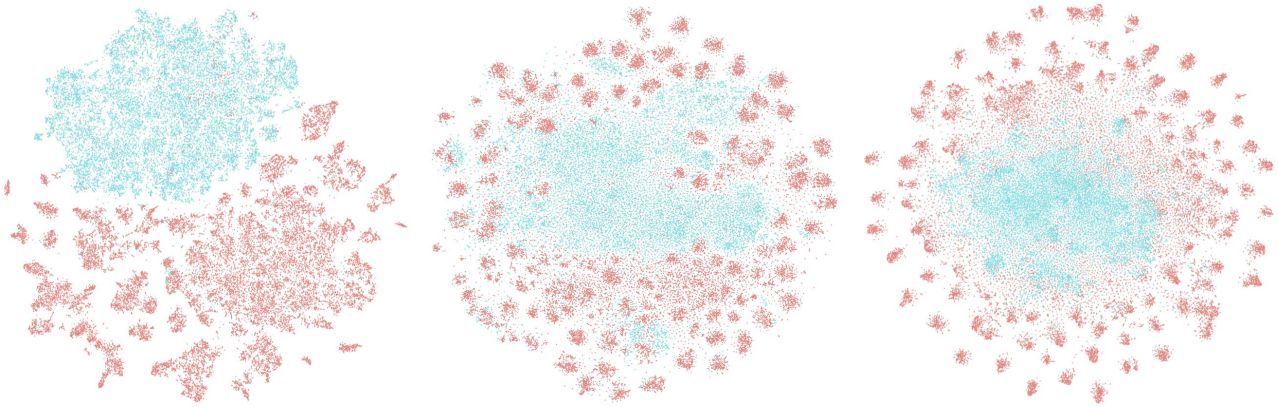
[11] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (11), 2008. 3

[12] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey. Symmetric Cross Entropy for Robust Learning With Noisy Labels. In *IEEE International Conference on Computer Vision (ECCV)*, 2019. 1

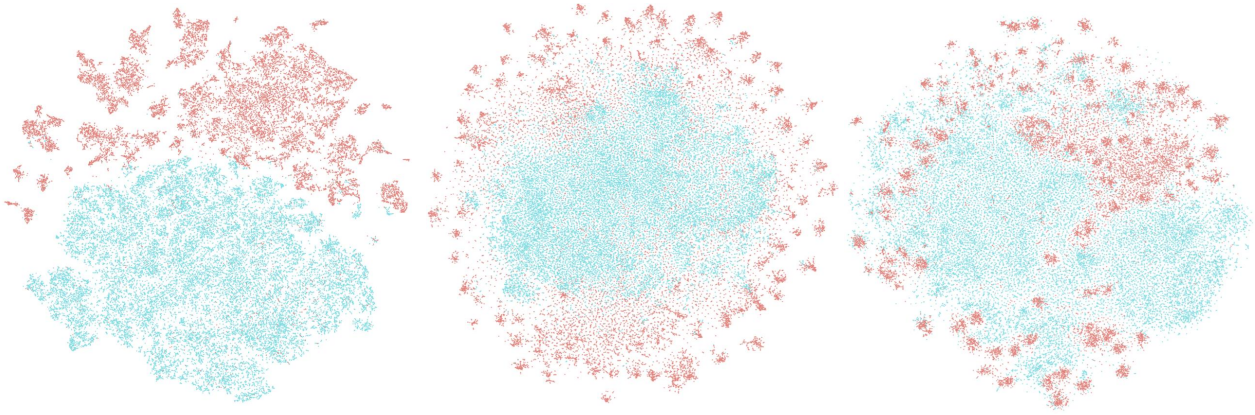
[13] H. Zhang, M. Cisse, Y.N. Dauphin, and D. Lopez-Paz. Mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 5



(a) $r_{in} = 20\%, r_{out} = 20\%$



(b) $r_{in} = 20\%, r_{out} = 40\%$



(c) $r_{in} = 20\%, r_{out} = 60\%$

Figure 5. T-SNE visualizations of embedding space of training images (CIFAR-100 corrupted with ImageNet32) with different ID and OOD noise rate. From left to right, the graphs show distribution of samples after SSL training, after Warmup training and after Semi-SL training(final model), respectively. During training, the distribution of samples mapped into 2D representation space changes in the way that simple KNN model can detect majority of OOD labeled noise samples and remove them from training. The color "Blue" and "Red" demonstrate the out-of-distribution samples and in-distribution samples (both clean and ID noise), respectively.