Weaken Noisy Feature: Boosting Semi-Supervised Learning by Noise Estimation

Junjiang Liu, Dandan Sun✉, Hailun Xia, Jiangtao Bai, Xinyue Fan
Beijing University of Posts and Telecommunications
Beijing, China
{liujunjiang, sdd661, xiahailun, baijiangtao, fxy1108}@bupt.edu.cn

Abstract

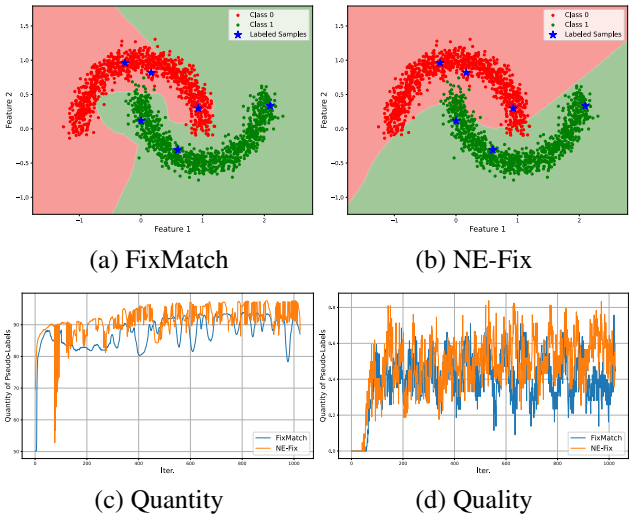
Due to poor bias calibration, current semi-supervised learning (SSL) is over-confident in false predictions. Modeling data noise helps alleviate confirmation bias. Specifically, we consider a posteriori knowledge that the data embedding follows a Gaussian distribution. We let the model predict and learn the mean (classification features) and variance (noise features) for each sample. In this way, the network can estimate and isolate noisy information in the latent space. Additionally, we propose Noise Estimation Curriculum Learning (NECL) to fit differences in noise level between classes. NECL can encourage the model to learn from hard classes actively. Extensive experiments demonstrate the effectiveness of our method, which is also orthogonal to FixMatch-based frameworks. InfoMatch with our components (NE-Info) showed state-of-the-art (SOAT) performance on several benchmarks such as CIFAR-10/100, SVHN, STL-10 and ImageNet.

Keywords: Semi-supervised learning, Image classification, Noise estimation, Multivariate Gaussian distribution

1. Introduction

Deep learning has great success because of large labelled datasets. However, labelling is boring and expensive. How to make model learn from wild images becomes a research hotspot. Among them, semi-supervised learning (SSL) is a main method to reduce the reliance on manual labelling. It gets better generalizability by joint training on unlabeled and labelled datasets.

Recently, mainstream SSL methods have combined pseudo-labeling [23] with consistency regularization (Laine and Aila 2016). For example, FixMatch [29] encourages the model to output the same prediction for two differently perturbed images. Although many methods improved this baseline from the perspectives of thresholding [40, 34, 37], negative learning [9, 38], entropy estimation [8, 17], etc., none of them weaken noisy feature actively. Noise makes pseudo-labels false even have a high confidence [16, 2]. Semi-supervised has poor tolerance of noise because very few labeling, limiting its generalizability on real-world data.



(a) FixMatch (b) NE-Fix

(c) Quantity (d) Quality

Figure 1: Experiments on Two-Moon Dataset with only 6 labeled samples (pentagram blue points) with others as unlabeled samples in training a 3-layer MLP classifier

To solve the problem, we introduce noise estimation (NE) (Kendall and Gal 2017) into SSL, which has been proven effective in other computer vision tasks. A similar work to ours is UPS [27], which pre-trained a model with uncertainty-aware on labeled datasets. The quantified uncertainty is used as part of the thresholding for filtering low quality pseudo-labels. However, in embedding space the distributions of labeled and unlabeled data are not completely consistent. Furthermore, UPS has not good performance when labeled data is low. We consider the embeddings in latent space to be a combination of classification features and other noisy features. And these noises follow a Gaussian distribution. We let the network predict the parameters of the distribution for each sample, thereby estimating noisy information while learning classification features. This estimation acts as a constraint on confidence of pseudo-labels. Additionally, there are also distribution and noise differences between classes, the sensitivity of classification features to noise interference also differs. We flex-

ibly adjust thresholds for different categories at each time step based on NE. If a category has bad noise estimate, it will get a lower threshold by this method called NECL. It can achieve a more balanced inter-class learning by encouraging the model to learn more from challenging categories.

We conducted experiments on Two-Moon Dataset with only 6 labeled samples (3‰ of the total sample). Fig. 1 (a) and (b) shows that FixMatch no longer overfitting with NE. NE-Fix can find a decision boundary that has lower data density. And it produced higher quantity and quality pseudo-labels during training as shown in Fig. 1 (c) and (d). In summary, our contributions are as follows:

- We find that there is negative info in mapping space affecting SSL. To the best of our knowledge, this is the first time that SSL has been modeled from the view of noise.
- Modelling noise estimation of data in SSL, proposed a simple yet effective method by weakening noise features. NE-method achieved tighter intra-class spacing and reduced data density at decision boundaries.
- Experiments showed that our approach is fully orthogonal to the FixMatch-based framework that can be deployed in current SSL algorithms easily and improve performance. Based on our components, we achieved state-of-the-art (SOTA) performance in multiple dimensions.

2. Related Work

Semi-Supervised Learning is an important direction in deep learning. Its success has enabled networks to greatly reduce dependence on labels, thereby reducing data costs. SSL has a range of applications in several fields of computer vision such as image classification [29, 17], semantic segmentation [30, 19], object detection [35, 33], facial recognition [31, 13] and so on. We will focus on the SSL for image classification, which is most relevant to our work.

Noise estimation in deep learning has been widely studied [1, 39, 24]. For quantization of noise, it can be broadly categorized into internal and external methods. Internal methods capturing the noise of the parameters in model [15, 20]. External meth-ods measure the noise of training data, e.g., error predictors, extra hidden units [7, 28]. Our approach is an external one.

Noise estimation in semi-supervised learning is a novel direction. Most previous methods try to keep high quality of pseudo-labels by a high threshold [29, 40]. However, it has no ability to correct hard samples. UPS pretrains an uncertainty-aware model on labelled datasets [27], but it does not work when there are very few labels. In contrast, we estimate uncertainty for pseudo-labels from the perspective of learning noise, which implements an end-to-end framework. It is more direct and effective by training on all samples.

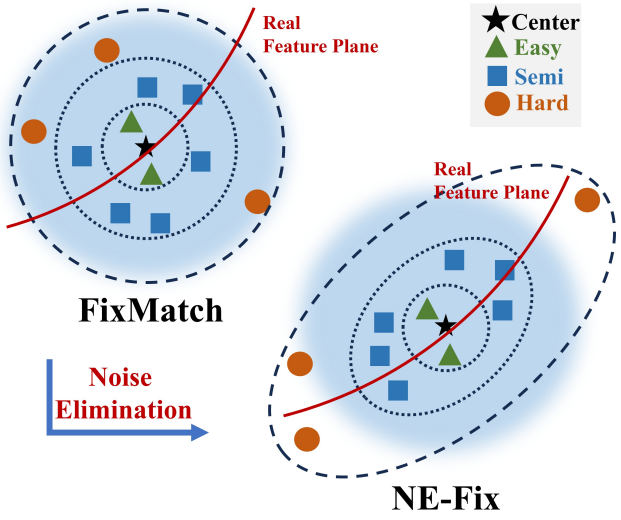


Figure 2: Embedding distribution before and after applying Noise Estimation on FixMatch. Embeddings of easy and semi samples become tighter. Embeddings of hard samples will be close to the plane but away from center.

3. Methodology

In this section, we review pseudo-labeling and consistency regularization in SSL for image classification. Then, we re-veal noisy estimates of data in continuous mapping spaces. Finally, we propose NE and NECL to improve FixMatch-based frameworks [29, 17].

3.1. Preliminaries

In SSL framework for C-class classification problem, we set $\mathcal{D}_L = \{x_i^l, y_i^l\}_{i=1}^{N_L}$ and $\mathcal{D}_U = \{x_i^u\}_{i=1}^{N_U}$ denote the labeled and unlabeled datasets, respectively. N_L and N_U is the number of datasets, $N_L \ll N_U$. We use $x_i^l, x_i^u \in \mathbb{R}^{H\times W\times 3}$ to represent labeled and unlabeled image, and y_i^l is the ground-truth label which be one-hot. In SSL, there are both supervised and unsupervised objective, i.e., $\mathcal{L}_S$ and $\mathcal{L}_U$. For the supervised loss, it is computed by cross-entropy loss ($\mathcal{H}$) on $\mathcal{D}_L$ in a B_L-sized batch:

$$\mathcal{L}_S = \frac{1}{B_L}\sum_{i=1}^{B_L} \mathcal{H}\left(y_i^l, p(y|x_i^l)\right), \tag{1}$$

where $p(y|\mathbf{x}) \in \mathbb{R}^C$ denotes the model's prediction on $\mathbf{x}$. For the unsupervised loss, FixMatch-based methods are use pseudo-labeling with consistency regularization, which mask out unconfident pseudo-labels by

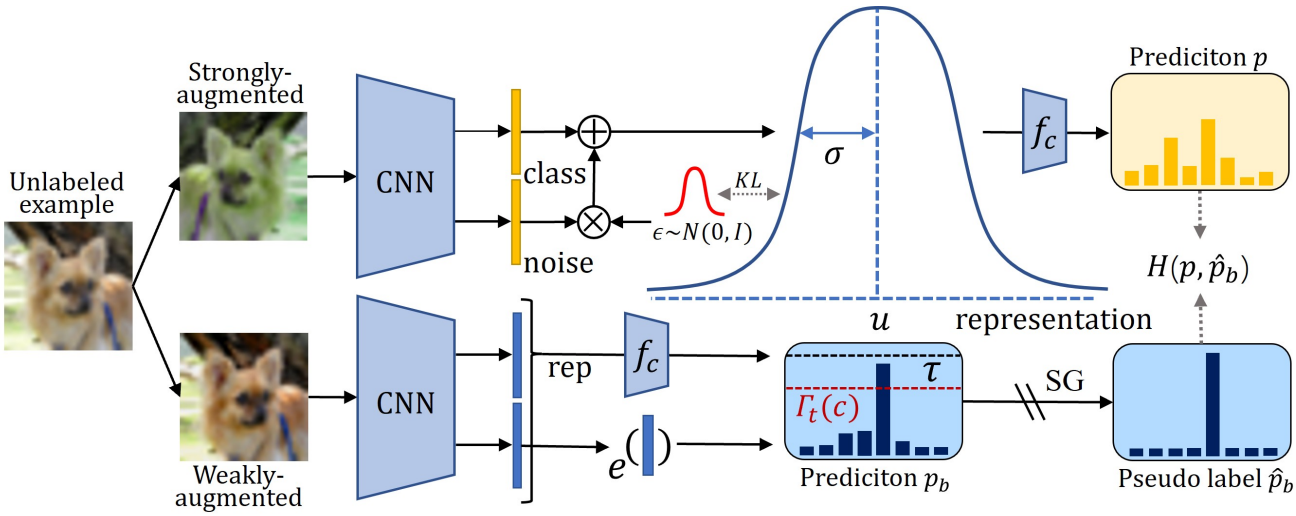


Figure 3: Overview of the proposed NE-Fix. The model outputs class (μ) and noisy (ψ) feature for a sample. The parameters are reparametrized before being fed into the fully-connected layer. The local threshold $\Gamma_t(c)$ for class c is determined by its credibility C_i. Our method is fully orthogonal to FixMatch [29].

a high threshold. Specifically, FixMatch formulates $\mathcal{L}_U$ as the weighted cross-entropy loss between pseudo-label from weakly-augmented sample $\omega(x_i^u)$ and prediction of strongly-augmented sample $\Phi(x_i^u)$:

$$\mathcal{L}_U = \frac{1}{B_U} \sum_{i=1}^{B_U} \mathbb{I}(\max(p_i) > \tau) \mathcal{H}\left(\hat{p}_i, p(y|\Phi(x_i^u))\right), \quad (2)$$

where $\hat{p}_i$ denotes the one-hot pseudo-label from $\arg\max(p_i)$, and $p_i = p(y|\omega(x_i^u))$; B_U is the batch size for unlabeled data; τ is the setting threshold to mask out $\hat{p}$ which has low $\max(p_i)$. In summary, total loss of SSL is $\mathcal{L} = \mathcal{L}_S + \mathcal{L}_U$.

3.2. Noise estimation in mapping space

The classification task is actually a mapping from $\mathcal{X}$ to $\mathcal{Y}$. We use a deep learning model to fit this mapping relationship through training. But $x_i \in \mathcal{X}$ contain more than pure category information $c(x_i)$ also noise $n(x_i)$. Consider a simple situation where $c(x_i)$ and $n(x_i)$ follow additive multivariate Gaussian distribution with mean of zero and x-dependent variance. Then in the mapping space: $y_i = f(c(x_i)) + f(c(n(x_i))), \epsilon \sim \mathcal{N}(0, I \in \mathbb{R}^d)$. The $f(\cdot)$ is the embedding function we want to get which can extract classification features and noise features separately. The two parameters of the Gaussian distribution: mean and variance then fit $c(x_i)$ and $\epsilon(n(x_i))$ [5, 14, 6, 26]. The problem with previous SSL has been lack of estimation of noise. They cannot quantify how reliable the pseudo-labels are. A significant portion of pseudo-labels during training are wrong, i.e., confirmation bias [2].

Also, consider a noiseless case, then the distribution of the sample's feature in mapping space is a linear plane. Samples embedded close to the center are usually more normalized. Color, variety, etc. make the embedding far from the center but still on this linear plane. Adding additive noise moves feature points away towards the direction of the plane's normal vector. Deep learning fits this feature plane through training. But it can estimate and weaken noise of data while fitting the feature plane with Noise Estimation, which could achieve tighter intra-class spacing as shown in Fig. 2.

3.3. Noise estimation in SSL

The overall architecture of our proposed method is shown in Fig. 3, called NE-SSL. We first define s_i be the represented of x_i in latent space, then we model the distribution of s_i by using Gaussian distribution as posteriori knowledge.

$$p(s_i|x_i) \sim \mathcal{N}(\mu_i, \sigma_i^2 \mathbf{I}) \quad (3)$$

Now, we let the model learning Gaussian embedding for each sample. The two parameters $\mu_i, \psi_i = \log(\sigma_i^2)$ extracts classification features and noise features, separately, which are obtained by adding two independent fully connected layers after the backbone. Let the network predict the logarithmic of variances ψ_i for a stable training, since negative and over large variances are difficult to interpret. We reparametrize (Kingma and Welling 2013) μ_i and ψ_i for an end-to-end optimization:

$$z_i = \mu_i + \epsilon \sqrt{e^{\psi_i}}, \epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad (4)$$

Algorithm 1: NE-Fix

Input: Labeled data x_l, y_l and unlabeled data x_u
Parameter: Initialize Model's weights θ
1: while i not reach the maximum iteration do.
2: Sample labeled batch x_l, y_l, unlabeled batch x_u.
3: Forward inference for networks N.
4: Calculate $\sigma_l^2 = e^{\psi_l}$, $\sigma_{wu}^2 = e^{\psi_{wu}}$ and $\sigma_{su}^2 = e^{\psi_{su}}$.
5: Calculate $L_{S_{cls}}$ using Equation (1).
6: Calculate $L_{L_{kl}}$ using Equation (5).
7: Calculate Γ_t using Equation (8).
8: Calculate mask: $l(max(s_{wu}) \geq \Gamma_t(argmax(s_{wu})))$.
9: Calculate L_U using Equation (2) $\cdot$ mask.
10: Calculate $L_{sU_{kl}}$ using Equation (5).
11: Update the θ of N via loss L.
12: end while
13: return μ for test

where ϵ is noise from a normal distribution by random sampling. It is worth noting that Eq. (4) is a reparametrization trick making the random sampling of Eq. (3) differentiable. Now, z_i is the final embedding of each image x_i, we input it into Equation (1) and (2) to participate in the cross-entropy. New supervised and unsupervised classification loss replaced with $\mathcal{L}_S^{cls}$ and $\mathcal{L}_U^{cls}$.

In addition to make s_i match our posteriori hypothesis, we introduce a regularization as an explicitly constraint. Concretely, we calculate the KL-dispersion between s_i and the normal distribution $\mathcal{N}(0, \mathbf{I})$:

$$\begin{aligned} \mathcal{L}^{kl} &= KL\left[\mathcal{N}(s_i; \mu_i, \sigma_i^2) \| \mathcal{N}(\epsilon|0, \mathbf{I})\right] \\ &= -\frac{1}{2}\left(1 + \psi_i - \mu_i^2 - e^{\psi_i}\right) \end{aligned} \tag{5}$$

It's an unsupervised process. We apply $\mathcal{L}^{kl}$ on labeled $\mathcal{L}_L^{kl}$ and strongly-augmented $\mathcal{L}_{sU}^{kl}$ data as a regularization term in optimization. Thus, the final loss of our method:

$$\mathcal{L} = \mathcal{L}_S^{cls} + \mathcal{L}_U^{cls} + \lambda(\mathcal{L}_L^{kl} + \mathcal{L}_{sU}^{kl}), \lambda \in [0, \lambda_{\max}] \tag{6}$$

In summary, $\mathcal{L}^{cls}$ and $\mathcal{L}^{kl}$ are responsible for the gradient descent of classification and noise features, separately. As a sample's latent distribution closer to the multivariate standard normal distribution, i.e., the average of variances closer to one, the noise features are estimated better. Hence when the model converges, the mean variance of the simple sample will be closer to 1. When a sample is more difficult, the KLD is larger. The model will actively adjust the distribution by changing μ_i and ψ_i to consider whether x_i belongs to another category or not. This is also why the embeddings of hard ones is far from the center in latent space as shown in Fig. 2. The two losses form a good balance to encourage the model to make more careful predictions for challenging categories. In testing, we feed μ of data into the classification layer. The pseudo-code of our method based on FixMatch (NE-Fix) is shown in Algorithm 1.

3.4. Noise estimation curriculum learning

One situation to consider is that in real data distributions, each class is not uniformly noise. And applying same noise to samples from different classes does not result in the same degree of destruction to their classification features. For example, a flipped airplane might be easier to classify than a flipped cat. Inspired by curriculum learning [3, 40], we introduce local dynamic thresholds for each class, aiming for the model to learn separately according to the difficulty of each class. We name it NECL. Specifically, we use the degree of proximity of σ_{wi}^2 to one quantifies the credibility C_i of $\omega(x_i^u)$,

$$C_i = \frac{1}{d}\sum_{i=1}^{d}(\mathbf{I} - |\mathbf{I} - \sigma_{wi}^2|), \tag{7}$$

where d is the dimension of the ψ_i, which we use 128. The threshold for class c at time step t:

$$\Gamma_t(c) = \frac{\tau}{N}\sum_{i=1}^{B} C_i \mathbb{I}(\hat{p}_i = c), \tag{8}$$

where N is the number of pseudo-labels that are predicted to be class c in a batch. We substitute $\Gamma_t(c)$ for τ in $\mathcal{L}_U^{cls}$. With NECL, the model will actively learn those categories with poor noise weakening.

4. Experiments

We conduct extensive experiments on five benchmark datasets, including CIFAR-10/100 [21], SVHN [25], STL-10 [10] and ImageNet [12], which are widely used in SSL. We select varying amounts of labeled data for experiments.

We evaluate our method based on FixMatch [29] and InfoMatch [17], the former is classical and the latter is recent. For a fair comparison, we follow the experimental setup exactly as in original work. Specifically, WideResNet (Zagoruyko and Komodakis 2016) is used as backbone on Cifar10/100, STL-10, SVHN and ResNet50 [18] on ImageNet. Consistent with previous works, the optimizer is standard stochastic gradient descent (SGD) with cosine decay learning rate at momentum 0.9, which the initial learning rate set to 0.3. We use RandAugment [11] as augmentation approach for data. All experiments are iterated a total of 2^20 times and test using an EMA model with a momentum of 0.999.

4.1. Main results

Qualitative analysis: As shown in Fig. 4, using a trained model to inference on test set. There are few simple and

Dataset	CIFAR-10			CIFAR-100		
# Label	40	250	4,000	400	2,500	10,000
PseudoLabel [23]	74.61±0.26	46.49±2.20	15.08±0.19	87.45±0.85	57.74±0.28	36.55±0.24
MeanTeacher [32]	70.09±1.60	37.46±3.30	8.10±0.21	81.11±1.44	45.17±1.06	31.75±0.23
MixMatch [4]	36.19±6.48	13.63±0.59	6.66±0.26	67.59±0.66	39.76±0.48	27.78±0.29
ReMixMatch [22]	9.88±1.03	6.30±0.05	4.84±0.01	42.75±1.05	26.03±0.35	20.02±0.27
UDA [36]	10.62±3.75	5.16±0.06	4.29±0.07	46.39±1.59	27.73±0.21	22.49±0.23
Dash [37]	7.47±0.28	4.86±0.05	4.21±0.08	44.82±0.96	27.15±0.22	22.20±0.12
FlexMatch [40]	4.97±0.06	4.98±0.09	4.19±0.01	39.94±1.62	26.49±0.20	21.90±0.15
UPS [27]	5.26±0.29	5.11±0.08	4.25±0.05	41.07±1.66	27.14±0.24	21.97±0.23
SimMatch [41]	5.60±1.37	4.84±0.39	3.96±0.01	37.81±2.21	25.07±0.32	20.58±0.11
FreeMatch [34]	4.90±0.04	4.88±0.18	4.10±0.02	37.98±0.42	26.47±0.20	21.68±0.03
SoftMatch [8]	4.91±0.12	4.82±0.09	4.04±0.02	**37.10±0.77**	26.66±0.25	22.03±0.03
FullMatch [9]	5.89±1.01	4.64±0.12	3.75±0.08	40.58±1.40	26.94±0.40	21.44±0.10
FixMatch [29]	7.47±0.28	4.86±0.05	4.21±0.08	46.42±0.82	28.03±0.16	22.20±0.12
NE-Fix*(Ours)*	**5.29±0.31**	**4.61±0.06**	**3.83±0.08**	**41.37±0.85**	**25.01±0.15**	**20.76±0.13**
InfoMatch [17]	4.22±0.14	4.01±0.07	3.29±0.08	-	-	19.47±0.56
NE-Info*(Ours)*	**3.98±0.17**	**3.52±0.09**	**3.16±0.10**	-	-	**19.25±0.59**
Fully-Supervised		4.62±0.05			19.30±0.09	

Table 1: Top-1 error rates (%) on CIFAR-10/100 datasets. **Bold** indicates the best result and underline indicates the second best result.

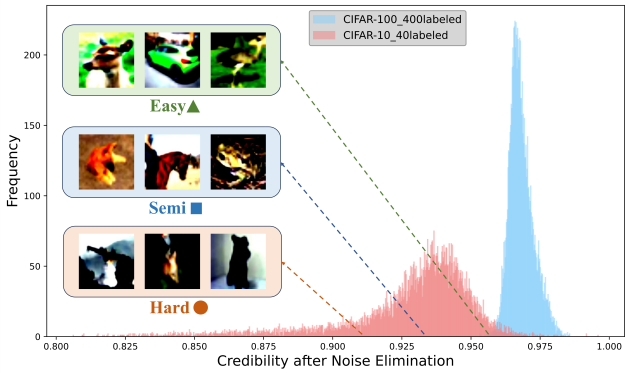


Figure 4: The credibility level C_i distribution of CIFAR-10 and CIFAR-100. The models are trained on 40 labeled and 400 labeled, respectively.

hard samples, and most of samples are semi. It's consistent with the distribution of data in real world. In the la-tent space, as noise features are removed as much as possible, the feature points are closer along the direction of normal vector of the feature plane.

However, ambiguous ones will be far away from the center because the model is not overconfident. More de-tails can be found in the T-SNE analysis in *4.3. Embedded space visualization via t-SNE.*

Quantitative analysis: We evaluate NE-Fix and NE-Info to compare with fully-supervised learning method and a range of representative semi-supervised learning methods.

We calculate the mean and variance of top-1 error rates when training on 5 different "folds" of labeled data. Results for CIFAR-10/100, SVHN, STL-10 with various labeled data size in section 4 and section 4.1. The results show that our method achieves the superior performance across all benchmarks. In particular, NE-Fix enhance performance by 5.05% for baseline on CIFAR-100 with just 400 labeled data. It also evident that our components are orthogonal to the FixMatch-based approach.

Method	Top-1	Top-5
FlexMatch [40]	43.66	21.80
FreeMatch [34]	40.57	18.77
SoftMatch [8]	40.52	18.70
FixMatch [29]	43.66	21.80
NE-Fix*(Ours)*	**41.05**	**19.90**
InfoMatch [17]	36.21	15.91
NE-Info*(Ours)*	**35.79**	**15.30**

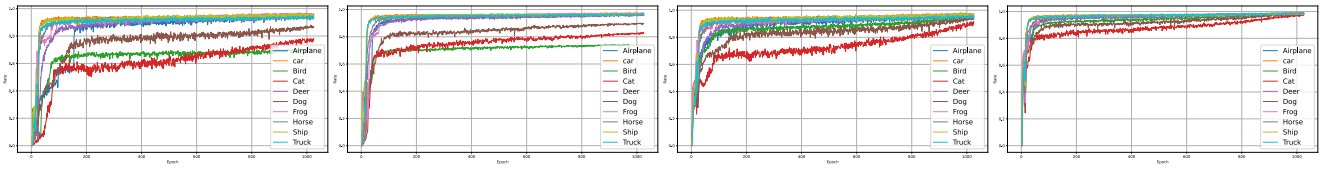
Table 2: Error rates (%) on ImageNet with 100 labels per class. **Bold** indicates the best result.

4.2. Results on ImageNet

ImageNet [12] is a large and complex dataset that more closely fits the real data situation. We re-port the performance of NE-Fix and NE-Info trained on ImageNet using

Dataset	SVHN			STL-10	
# Label	40	250	1,000	40	1,000
PseudoLabel [23]	64.61±5.60	15.59±0.95	9.40±0.32	74.68±0.99	32.64±0.71
MeanTeacher [32]	36.09±3.98	3.45±0.03	3.27±0.05	71.72±1.45	33.90±1.37
MixMatch [4]	30.60±8.39	4.56±0.32	3.69±0.37	54.93±0.96	21.70±0.68
ReMixMatch [22]	24.04±9.13	6.36±0.22	5.16±0.31	32.12±6.24	6.74±0.14
UDA [36]	5.12±4.27	1.92±0.05	1.89±0.01	37.42±8.44	6.64±0.17
Dash [37]	2.19±0.18	2.04±0.02	1.97±0.02	34.52±4.30	6.39±0.56
FlexMatch [40]	<u>8.19±3.20</u>	<u>6.59±2.29</u>	<u>6.72±0.30</u>	<u>29.15±4.16</u>	<u>5.77±0.18</u>
UPS [27]	-	-	-	-	6.02±0.40
SimMatch [41]	7.60±2.11	2.48±0.61	2.05±0.05	16.98±4.24	5.74±0.31
FreeMatch [34]	1.97±0.02	1.97±0.01	1.96±0.01	15.56±0.55	5.63±0.15
SoftMatch [8]	2.33±0.25	2.09±0.05	2.01±0.01	21.42±3.48	5.73±0.24
FullMatch [9]	2.35±0.10	-	1.99±0.03	-	5.74±0.09
FixMatch [29]	3.81±1.18	2.02±0.02	1.96±0.03	35.97±4.14	6.25±0.33
NE-Fix*(Ours)*	**3.01±1.17**	**1.95±0.03**	**1.93±0.03**	**32.78±4.19**	**5.71±0.36**
InfoMatch [17]	1.84±0.07	1.79±0.01	1.75±0.03	9.86±1.13	5.27±0.09
NE-Info*(Ours)*	**1.78±0.10**	**1.72±0.02**	**1.69±0.04**	**9.67±1.11**	**5.15±0.10**
Fully-Supervised		1.20±0.01		-	

Table 3: Top-1 error rates (%) on SVHN and STL-10 datasets. **Bold** indicates the best result and <u>underline</u> indicates the second best result.



(a) w/o NECL Quantity (b) w NECL Quantity (c) w/o NECL Quality (d) w NECL Quality

Figure 5: Ablation study of NECL. (a) and (b) Quantity of pseudo-labels by NE-Fix w/o NECL and w NECL (c) and (d) Quality of pseudo-labels by NE-Fix w/o and w NECL.

just 100 labels per class. NE-Fix reduces by 2.61% in Top-1 error rate and 1.9% in Top-5 error rate compared to FixMatch. The results in section 4.1 confirm the effectiveness of NE-SSL on complex data with uneven distribution.

4.3. Ablation study

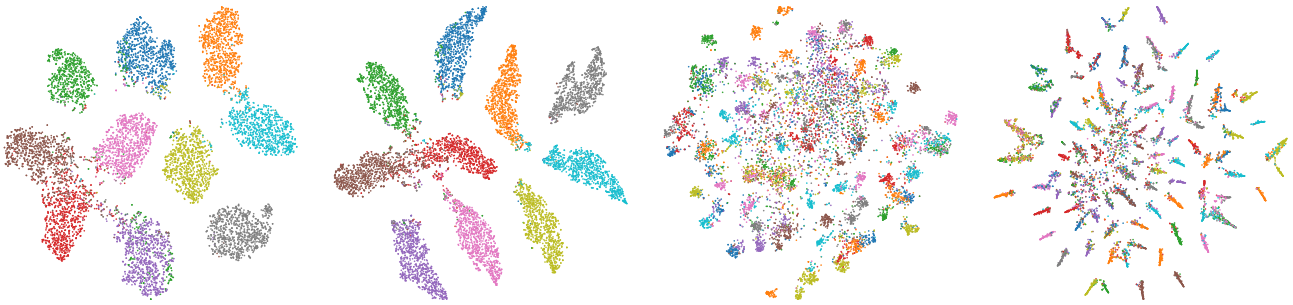
For ablation study of NE, the model degrades to original FixMatch when NE is not used which has been reported in Main Results. We test different combinations of NE on $\mathcal{D}_L$, $\omega(\mathcal{D}_L)$ and $\phi(\mathcal{D}_L)$. The results in section 4.3 show that NE works best when it is used on $\mathcal{D}_L$ and $\phi(\mathcal{D}_U)$. We analyze that this is because the gradient of $\omega(\mathcal{D}_U)$ does not update and the model cannot converge on noise. It is also shown that the $\mathcal{L}^{cls}$ is an important restriction for $\mathcal{L}^{kl}$.

For ablation study of NECL, we conduct experiments on CIFAR-10 with 40labels for NE-Fix w or w/o NECL. Fig. 5 shows the quantity and quality of pseudo labels of each category during training. NECL offers higher quantity and quality of pseudo-labels for NE-Fix and the model converges faster especially in the first 100 epochs.

$\mathcal{D}_L$	$\omega(\mathcal{D}_U)$	$\phi(\mathcal{D}_U)$	**Error Rate**
✓			7.40
	✓		10.56
		✓	6.57
✓	✓		9.67
✓	✓	✓	8.58
✓		✓	**5.22**
Baseline			7.47

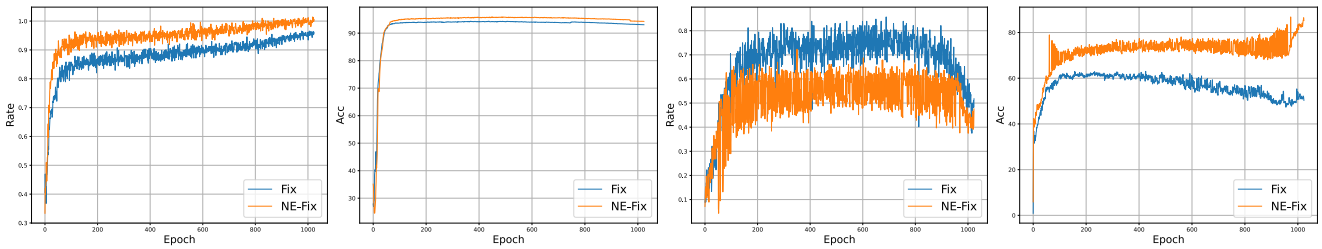
Table 4: Ablation study of different combinations of NE on D_L, $\omega(D_U)$ and $\phi(D_U)$. **Bold** indicates the best result.

For ablation study of λ, We explore impact of hyperparameter of $\mathcal{L}^{kl}$. When λ is zero, the performance is close to baseline. As the λ is increased, the model begins to es-



(a) w/o NECL Quantity (b) w NECL Quantity (c) w/o NECL Quality (d) w NECL Quality

Figure 6: t-SNE of representations obtained for the test set of CIFAR-10/100 using FixMacth and NE-Fix. (a) and (b) FixMatch and NE-Fix on CIFAR-10 with 40 labels. (c) and (d) FixMatch and NE-Fix on CIFAR-100 with 400 labels. **Best viewed in color.**



(a) Cifar10 Quantity (b) Cifar10 Quality (c) Cifar100 Quantity (d) Cifar100 Quality

Figure 7: Quantity and quality of pseudo-labels for FixMatch and NE-Fix on CIFAR-10/100. (a) and (b) Quantity and quality on CIFAR-10 with 40 labels. (c) and (d) Quantity and quality on CIFAR-100 with 400 labels.

λ	**Error Rate**
0.0	7.46
0.01	6.32
0.1	**5.22**
0.5	8.95
1.0	15.98
Baseline	7.47

Table 5: Results of L_{kl} trained on 40 labeled CIFAR-10 with different trade-off λ.

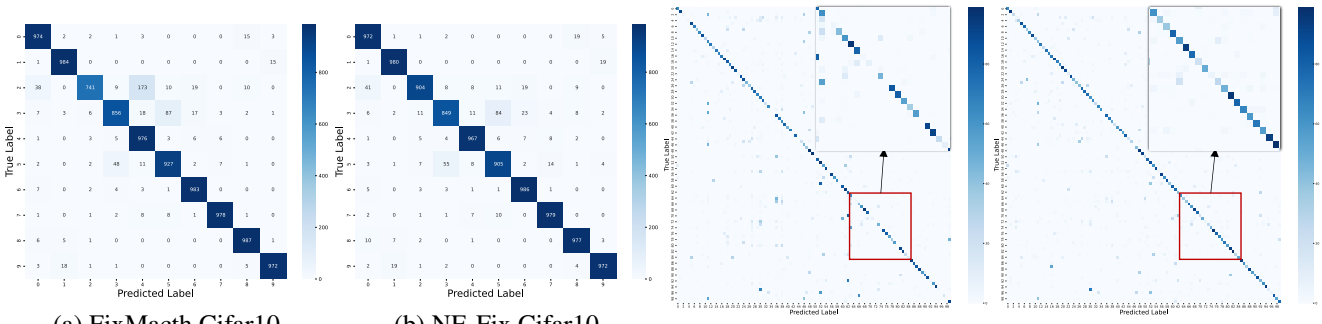
timate noise of data. When λ is too large, the model over-adjusts the predicted parameters of embeddings. Thus, the performances deteriorate rapidly because classification loss is difficult to converge and low quantity of pseudo-labels. The results are reported in section 4.3.

4.4. Embedded space visualization via t-SNE

Fig. 6 compares t-distributed stochastic neighbor embedding (t-SNE) (Van and Hinton, 2020) of representations obtained for the test set of CIFAR-10/100 using FixMatch and NE-Fix. We can observe that the representations obtained using NE-Fix has clearer boundaries. The shapes of category embedding become tighter and narrower. This is probably because the data embeddings show a linear distribution after weakening noise by NE. The challenge points are located more at the edge of category area, close to the decision boundary. The results are consistent with the theory in *3. Methodology* and the qualitative analysis in *4. Experiments*.

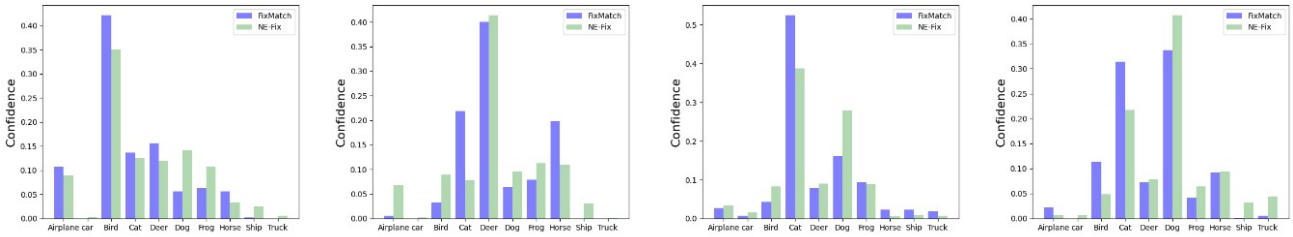
4.5. Quantity and Quality of Pseudo-Labels

In Fig. 7, we visualize the quantity and quality of pseudo-labels generated by FixMatch and NE-Fix on the CIFAR-10 with 40 labels, CIFAR-100 with 400 labels. Our comparison reveals that NE-Fix outperforms FixMatch in both the quantity and quality of pseudo-labels during the training process on CIFAR-10. This indicates that NE-Fix is more effective at producing reliable labels in this scenario. Since CIFAR-100 is more complex, NE-Fix adopts a more cautious prediction, leading to a lower overall number of pseudo-labels. But the quality of the pseudo-labels generated by NE-Fix on CIFAR-100 is significantly higher than that of FixMatch, highlighting the trade-off between quantity and quality in challenging environments.



(a) FixMacth Cifar10 (b) NE-Fix Cifar10 (c) FixMacth Cifar100 (d) NE-Fix Cifar100

Figure 8: Confusion matrices of the class predictions on the test set of CIFAR-10/100 using FixMacth and NE-Fix. (a) and (b) FixMatch and NE-Fix on CIFAR-10 with 40 labels. (c) and (d) FixMatch and NE-Fix on CIFAR-100 with 400 labels.



(a) Bird (b) Deer (c) Cat (d) Dog

Figure 9: Prediction distribution of FixMatch w and w/o NE for low confidence samples.

Method	Bird	Deer	Cat	Dog
FixMatch [29]	0.0158	0.0176	0.0134	0.0157
NE-Fix*(Ours)*	0.0104	0.0168	0.0155	0.0152
InfoMatch [17]	0.0295	0.0183	0.0182	0.0157
NE-Info*(Ours)*	0.0124	0.0148	0.0138	0.0174

Table 6: Variance of the mean confidence for ambiguous samples trained on CIFAR-10 40 labels.

Method	Precision	Recall	F1 Score	AUC
FixMatch [29]	0.9410	0.9378	0.9369	0.9928
NE-Fix*(Ours)*	**0.9510**	**0.9508**	**0.9505**	**0.9957**
InfoMatch [17]	0.9544	0.9545	0.9544	0.9946
NE-Info*(Ours)*	**0.9573**	**0.9573**	**0.9572**	**0.9969**

Table 7: Precision, recall, f1 score and AUC results on CIFAR-10 with 40 labels.

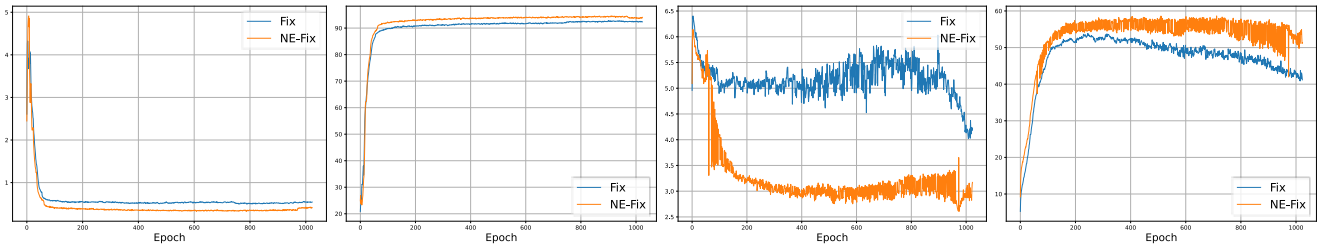
4.6. Class-wise Balance

We compared confusion matrix on CIFAR-10/100 with 40/400 labels using FixMatch and NE-Fix in Fig. 8. The results show that NE has better learning ability for difficult categories, especially "Bird" and "Dog", which has an impressive improvement over FixMatch.

As shown in section 4.5 and Fig. 9, NE-Fix's variance of the mean confidence for ambiguous samples (argmax(p)<0.6) is lower than baseline, which NE actively considers the possibility of otherness. FixMatch tends to confuse "Bird" with "Deer" and "Cat" with "Dog" due to overconfident iterations. This means that more cautious forecasting can produce more accurate results in SSL.

4.7. Detailed results

To comprehensively evaluate the performance of the method in classification, we further reported the precision, recall, F1 score, and AUC (area under the curve) on the CIFAR-10 dataset. As shown in section 4.6, we observed that in addition to reduced error rate, the NE method also achieved the best performance in terms of precision, recall, F1 score, and AUC. These metrics, along with the error rate (accuracy), demonstrate the powerful performance of our proposed method. We also re-ported the median error rate of the last 20 checkpoints, with all methods running the same number of iterations. There were 1024 iterations between every two checkpoints. The results shown in section 4.6 indicate that our method can significantly improve the performance of existing SSL algorithms. These conclu-



(a) Cifar10 Loss (b) Cifar10 Top-1 Acc (c) Cifar100 Loss (d) Cifar100 Top-1 Acc

Figure 10: Quantity and quality of pseudo-labels for FixMatch and NE-Fix on CIFAR-10/100. (a) and (b) Quantity and quality on CIFAR-10 with 40 labels. (c) and (d) Quantity and quality on CIFAR-100 with 400 labels.

Dataset	**CIFAR-10**			**CIFAR-100**		
Method	40	250	4,000	400	2,500	10,000
FixMatch [29]	7.76±0.59	5.22±0.33	4.42±0.11	47.87±1.19	28.10±0.25	22.96±0.12
NE-Fix*(Ours)*	**5.32±0.41**	**4.97±0.31**	**4.18±0.10**	**42.42±0.91**	**26.16±0.21**	**21.01±0.09**
InfoMatch [17]	4.43±0.15	4.51±0.08	3.78±0.09	-	-	20.36±0.54
NE-Info*(Ours)*	**3.76±0.16**	**3.82±0.08**	**3.35±0.10**	-	-	**20.19±0.55**

Table 8: Median error rates of the last 20 checkpoints on CIFAR-10/100.

sions are consistent with the results shown in section 4 and section 4.1 in the main text, demonstrating the effectiveness of our proposed NE algorithm.

4.8. Convergence speed

Another notable advantage of NE-Fix is its superior convergence speed, which is evidenced by the comparison of loss and Top-1 accuracy between FixMatch and NE-Fix on the CIFAR-10 dataset with a 40-label split, as illustrated in Fig. 10. The results show that the loss for NE-Fix decreases at a faster rate compared to FixMatch throughout the training process. This rapid decline in loss indicates that NE-Fix is able to learn more efficiently and effectively, leading to quicker improvements in performance. Consequently, NE-Fix not only achieves better accuracy more rapidly but also suggests a more effective utilization of the training data. This enhanced convergence speed is a significant advantage for practitioners looking to optimize their model performance in a shorter timeframe.

4.9. Noise estimation analysis

We graded the difficulty level based on the C_i computed from the images after model inference, with C_i in the top 20% considered "Easy", C_i in the bottom 20% considered "Hard", and the rest "Semi". Some of the images are shown in Fig. 11 and Fig. 12. We invited 20 volunteers who had never seen CIFAR-10/100 to score randomly selected images. There are 2 images of each of the 3 difficulty levels for each class, totaling 60 images. Each person sees the same set of images. Segment descriptions and scoring results are displayed in table 9 and table 10. The mean score for the "Hard" are about 3.2 points lower than the "Easy". The result demonstrates that our model's judgment a picture is basically in line with visual intuition.

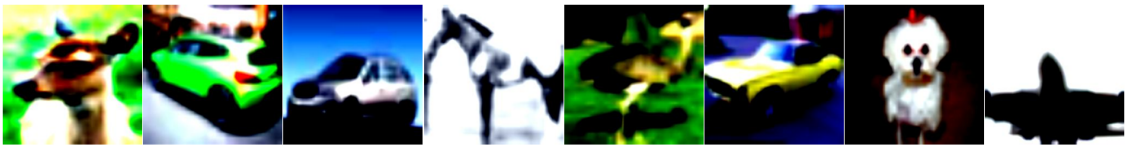
Score	**Description**
1	Highly Uncertain
2	Somewhat Uncertain
3	Somewhat Certain
4	Certain
5	Highly Certain

Table 9: Score descriptions

Dataset	**Level**	**Mean Score**
CIFAR-10	Easy	4.76
	Semi	3.88
	Hard	1.43
CIFAR-100	Easy	4.49
	Semi	3.55
	Hard	1.28

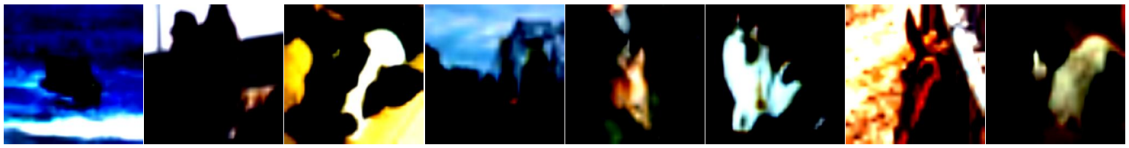
Table 10: Mean scores for different dataset levels

Fig. 13 shows category embedding representation and difficulty level embedding representation. The darker color of the embeddings, the harder they are. The results show that most of the difficult samples are distributed at the decision boundary, far from the feature center. And the error samples are basically difficult. This shows that NE-Fix has good estimation ability for noise.



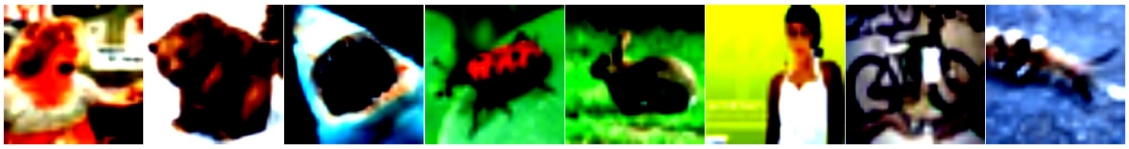
(a) Easy

(b) Semi

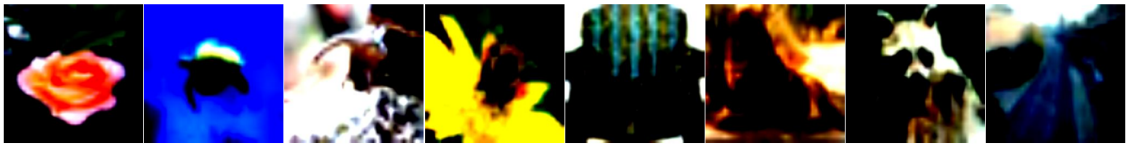


(c) Hard

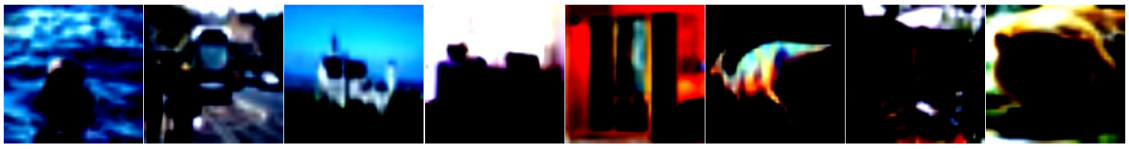
Figure 11: Model's judgment of the difficulty for Cifar10. The models are trained on 40 labeled.



(a) Easy



(b) Semi



(c) Hard

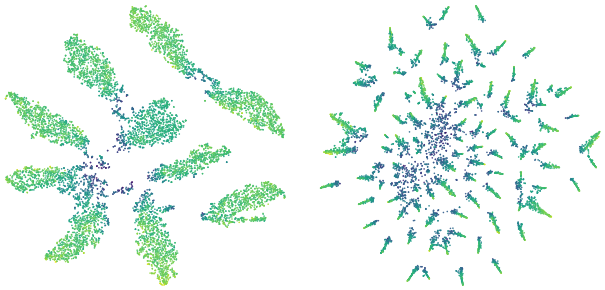
Figure 12: Model's judgment of the difficulty for Cifar100. The models are trained on 400 labeled.

4.10. Performance of NE on other benchmarks

We conducted a comparative analysis of various semi-supervised learning (SSL) benchmark methods, examining their performance alongside NE-enhanced counterparts. This evaluation was performed using the CIFAR-10, which featured a split of 40 labeled examples, as well as the CIFAR-100 dataset, which included 400 labeled examples. The results of our comparative analysis are summarized in table 11. This table provides a clear and detailed illustration of the performance metrics for each method evaluated. Notably, it demonstrates that the incorporation of the Neighborhood Enhancement (NE) component leads to consistent improvements across all baseline SSL methods.

5. Conclusion

In this work, we propose a novel method that takes an orthogonal approach to existing semi-supervised learning (SSL) techniques by focusing on the issue of data noise. Our method, termed NE-SSL, aims to extract purer classification features by effectively estimating and mitigating the influence of noisy features within the data. This process not only enhances the quality of the features used for classification but also ensures more reliable learning outcomes. Additionally, we introduce a curriculum learning framework that is based on noise estimation, which serves to refine the learning process and address the biases that can arise in traditional methods. Comprehensive experiments con-



(a) Cifar10 (b) Cifar100

Figure 13: t-SNE of representations obtained for the test set of CIFAR-10/100 using NE-Fix. **Best viewed in color.**

Table 11: More SSL benchmarks with NE. The ↓ darkgreen is the error rate that has decreased compared to the original version.

Dataset	**CIFAR-10**	**CIFAR-100**
Method	Method w NE	Method w NE
PseudoLabel [23]	61.86 (↓ 12.75)	75.16 (↓ 12.09)
MeanTeacher [32]	52.08 (↓ 18.01)	69.02 (↓ 12.09)
MixMatch [4]	27.40 (↓ 8.79)	57.70 (↓ 9.89)
FlexMatch [40]	4.56 (↓ 0.41)	37.62 (↓ 2.32)
FullMatch [9]	4.83 (↓ 1.06)	36.13 (↓ 0.97)
FreeMatch [34]	4.61 (↓ 0.29)	36.18 (↓ 1.80)

ducted across multiple benchmark tests illustrate that our proposed approach has achieved significant improvements in performance metrics, showcasing its efficacy. Furthermore, we delve into an analysis of how our noise estimation (NE) and noise estimation-based curriculum learning (NECL) strategies influence the training dynamics of SSL. This is explored through both qualitative observations and quantitative results, providing a thorough understanding of their impact.

Looking ahead, our future work will focus on extending the application of NE to other semi-supervised tasks, including text, audio, and video classification, thereby broadening the scope and potential of our proposed methodology.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 61976022). Thanks to the Advanced Information Network Beijing Laboratory and OPPO AI Center for providing computing resources. Thanks to Mingyu Mao, Jianlong Gao, Ruyi Wang and Ning Zhang for their valuable opinions and help.

References

[1] E. Amrani, R. Ben-Ari, D. Rotman, and A. Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2021. 2

[2] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020. 1, 3

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, 2009. 4

[4] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 5, 6, 11

[5] C. Bishop and C. Quazaz. Regression with input dependent noise: A bayesian treatment. In *Advances in Neural Information Processing Systems*, volume 9, 1996. 3

[6] A. Brando, J. A. Rodríguez-Serrano, M. Ciprian, R. Maestre, and J. Vitrià. Uncertainty modelling in deep networks: Forecasting short and noisy series. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 325–340, 2018. 3

[7] J. Chang, Z. Lan, C. Cheng, and Y. Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5710–5719, 2020. 2

[8] H. Chen, R. Tao, Y. Fan, Y. Wang, J. Wang, B. Schiele, X. Xie, B. Raj, and M. Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023. 1, 5, 6

[9] Y. Chen, X. Tan, B. Zhao, Z. Chen, R. Song, J. Liang, and X. Lu. Boosting semi-supervised learning by exploiting all unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7548–7557, 2023. 1, 5, 6, 11

[10] A. Coates, A. Y. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011. 4

[11] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 4

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4, 5

[13] D. Ferman, P. Garrido, and G. Bharaj. Facelift: Semi-supervised 3d facial landmark localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1781–1791, 2024. 2

[14] P. Goldberg, C. Williams, and C. Bishop. Regression with input-dependent noise: A gaussian process treatment. In *Advances in Neural Information Processing Systems*, 1997. 3

[15] J. Goldberger and E. Ben-Reuven. Training deep neural networks using a noise adaptation layer. In *International Conference on Learning Representations*, 2017. 2

[16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017. 1

[17] Q. Han, Z. Tian, C. Xia, and K. Zhan. Infomatch: Entropy neural estimation for semi-supervised image classification. *arXiv preprint arXiv:2404.11003*, 2024. 1, 2, 4, 5, 6, 8, 9

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[19] X. Hu, L. Jiang, and B. Schiele. Training vision transformers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4007–4017, 2024. 2

[20] 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*, pages 2304–2313, 2018. 2

[21] A. Krizhevsky, G. Hinton, and et al. Learning multiple layers of features from tiny images. In *Citeseer*, 2009. 4

[22] A. Kurakin, C. Raffel, D. Berthelot, E. D. Cubuk, H. Zhang, K. Sohn, and N. Carlini. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *ICLR*, 2020. 5, 6

[23] D.-H. Lee and et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 896, Atlanta, 2013. 1, 5, 6, 11

[24] K. Liu, K. Ok, W. Vega-Brown, and N. Roy. Deep inference for covariance estimation: Learning gaussian noise models for state estimation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1436–1443, 2018. 2

[25] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, and et al. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, volume 2011, page 4, 2011. 4

[26] D. A. Nix and A. S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60, 1994. 3

[27] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021. 1, 2, 5, 6

[28] Y. Shi and A. K. Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6902–6911, 2019. 2

[29] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 1, 2, 3, 4, 5, 6, 8, 9

[30] B. Sun, Y. Yang, L. Zhang, M.-M. Cheng, and Q. Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3097–3107, 2024. 2

[31] H. Sun, C. Pi, and W. Xie. Semi-supervised facial expression recognition by exploring false pseudo-labels. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 234–239, 2023. 2

[32] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017. 5, 6, 11

[33] H. Wang, Z. Zhang, J. Gao, and W. Hu. A-teacher: Asymmetric network for 3d semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14978–14987, 2024. 2

[34] Y. Wang, H. Chen, Q. Heng, W. Hou, Y. Fan, Z. Wu, J. Wang, M. Savvides, T. Shinozaki, B. Raj, and et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2023. 1, 5, 6, 11

[35] W. Wu, H.-S. Wong, S. Wu, and T. Zhang. Relational matching for weakly semi-supervised oriented object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27800–27810, 2024. 2

[36] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. V. Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020. 5, 6

[37] Y. Xu, L. Shang, J. Ye, Q. Qian, Y.-F. Li, B. Sun, H. Li, and R. Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pages 11525–11536, 2021. 1, 5, 6

[38] L. Yang, Z. Zhao, L. Qi, Y. Qiao, Y. Shi, and H. Zhao. Shrinking class space for enhanced certainty in semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16187–16196, 2023. 1

[39] Y. Yao, T. Liu, B. Han, M. Gong, J. Deng, G. Niu, and M. Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in Neural Information Processing Systems*, 33:7260–7271, 2020. 2

[40] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 1, 2, 4, 5, 6, 11

[41] M. Zheng, S. You, L. Huang, F. Wang, C. Qian, and C. Xu. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14471–14481, 2022. 5, 6