An Efficient Semi-supervised Learning Method with Noisy Labels

JiHee Kim
Hanyang University
Seoul, Korea
im.jiheek@gmail.com

Sangki Park
Hanyang University
Seoul, Korea
skpark1101@hanyang.ac.kr

Si-Dong Roh
Hanyang University
Seoul, Korea
sdroh1027@naver.com

Ki-Seok Chung
Hanyang University
Seoul, Korea
kchung@hanyang.ac.kr

ABSTRACT

Even though deep learning models make success in many application areas, it is well-known that they are vulnerable to data noise. Therefore, researches on a model that detects and removes noisy data or the one that operates robustly against noisy data have been actively conducted. However, most existing approaches have limitations in either that important information could be left out while noisy data are cleaned up or that prior information on the dataset is required while such information may not be easily available. In this paper, we propose an effective semi-supervised learning method with model ensemble and parameter scheduling techniques. Our experiment results show that the proposed method achieves the best accuracy under 20% and 40% noise-ratio conditions. The proposed model is robust to data noise, suffering from only 2.08% of accuracy degradation when the noise ratio increases from 20% to 60% on CIFAR-10. We additionally perform an ablation study to verify net accuracy enhancement by applying one technique after another.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**.

KEYWORDS

Noisy label, Noisy data, Semi-supervised learning, Classification

ACM Reference Format:
JiHee Kim, Sangki Park, Si-Dong Roh, and Ki-Seok Chung. 2018. An Efficient Semi-supervised Learning Method with Noisy Labels. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Since data labeling process is commonly done by human, labeling mistakes are inevitable. Therefore, deep learning models in the real world often are trained based on incorrect labels, which may degrade the model's generalization performance [1–3]. To resolve this issue, many studies have attempted either to come up with algorithms that detect and remove noisy data or to propose models that perform robustly against noisy data [4–8]. However, the former has a limitation that important information could be left out while noisy data are cleaned up. The latter requires prior information on the dataset such as clean validation data while such information may not be easily available.

Semi-supervised learning [9] is a learning technique that can train a model with a small portion of labeled data and a larger amount of unlabeled data. In this paper, instead of removing detected noisily-labeled data, the noisy data is utilized as unlabeled data for semi-supervised learning. Our robust semi-supervised learning method with noisy labels consists of the following components:

- Prediction ensemble: To achieve better predictions, an ensemble of the prediction of the model and the Exponential Moving Average (EMA) is utilized. We use two weak augmentation strategies for better consistency regularization. Further details will be addressed in Section 3.1.
- Loss weight (λ) scheduling: Labeled losses at the start of training dominate unsupervised losses, but they may be unintentionally regarded as noisy data. Therefore, it may be desirable to increase the weight of unsupervised losses in the early stage to increase the convergence speed of learning. Further details will be discussed in Section 3.2.
- Sharpening: As proved in [10], sharpening may be effective to reduce the entropy of label distribution. In this study, we apply sharpening to modify the temperature of categorical distribution.

The rest of this paper is organized as follows. In Section 2, we address noisy labels, semi-supervised learning, and related works. Section 3 discusses the proposed method. In Section 4, we discuss implementation details, experimental results, and analysis of the results. Section 5 will conclude the paper.

2 RELATED WORKS

2.1 Noisy Label

A noisy label is defined as an incorrectly-assigned data label and it is regarded as one type of data noises. Noisy labels are known to be more harmful to learning than feature noises [3], another type of data noises. For example, when it comes to an image, even if some pixels are distorted, a model can get useful information

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

from other pixels to avoid significant deterioration. However, a significant information loss occurs when a sample is incorrectly labeled because the label cannot be replaced by other information. There are two main approaches to train a model with noisy data. The first one is to train a model with the remaining clean data after detecting and removing data with noisy labels. The second is to train a robust model that works well even with noisy labels by utilizing regularization or loss functions to prevent overfitting due to noisy labels [4].

Some of the well-known existing works to remove data with noisy labels are as follows. Co-teaching [5] detected noisy data by allowing two networks to be trained together where one network is used to select data for training the other. In Huang et al.'s O2U-Net [6], the learning progress from underfitting to overfitting is repeated several times just by changing the learning rate, which is a parameter, and the noisy label is corrected by calculating the average loss of data.

On the other hand, Li et al.'s DivideMix [7] used co-refinement and co-guessing to improve MixMatch [10]. MixMatch guessed a low-entropy label for unlabeled data and used MixUp [11] to mix labeled and unlabeled data. They used a dynamic Gaussian Mixture Model (GMM) to divide the training set into labeled and unlabeled data based on the loss of each sample. In Zhou et al. [8], the noisy label was detected more reliably than a method using the simple average by calculating the loss using the EMA. In this paper, we also employ the EMA model to improve the classification performance, focusing on generating a prediction with a smaller variance compared with the instantaneous prediction.

2.2 Semi-Supervised Learning

Semi-supervised learning is a machine learning technique that trains a model with a small amount of labeled data and a larger amount of unlabeled data [9]. It is designed to overcome the limitations of supervised learning and unsupervised learning. Supervised learning requires only the labeled data, where data labeling takes lots of effort and time. Unsupervised learning uses only the unlabeled data, but its application area is very limited. In general, there are two different approaches in semi-supervised learning: Pseudo-labeling methods and Hybrid methods [12]. The first approach refers to a method of creating pseudo labels with a weakly supervised model. Self-supervised learning of Zhai et al.[13] trains the base model with a supervised method with a small amount of labeled data, and the prediction value of this model is used as a label for the remaining unlabeled data. Blum et al.'s Co-Training [14] trains two separate classifiers based on two views of data. Classifiers train each other with the pseudo-label with the highest confidence. Finally, a single classification result is obtained by combining the predictions of the two classifiers.

The other approach, the hybrid method refers to a combined method of various semi-supervised ideas. Berthelot et al.'s MixMatch created a single loss by combining the following three approaches: entropy minimization, consistency regularization, and generic regularization. Entropy minimization is a method to induce a model to generate confident output predictions for unlabeled data. Consistency regularization is a method to make a model generate the same output distribution even for distorted inputs. Finally, generic regularization is a method to ensure that a model generalizes well and does not overfit the training data.

$$\begin{aligned} l_s &= H(y, p(\bar{y}|x_l)) \\ l_u &= H(y, q(\bar{y}|A(x_u))) \\ l &= l_s + \lambda l_u \end{aligned} \tag{1}$$

Sohn et al.'s FixMatch [15] is a semi-supervised learning method using consistency regularization and pseudo-labeling. Equation 1 is a loss term of unlabeled data in FixMatch. Labeled data (x_l) are trained by supervised learning to generate the supervised loss (l_s). Unlabeled data (x_u) are first weakly augmented and are then inferred to generate a pseudo label (q), which is used as a target value when learning strongly-augmented data. So, the unsupervised loss(l_u) is a cross entropy loss between the pseudo label and the target value. The final loss is the sum of the two losses(l_s, l_u). The proposed method of this paper is a modified FixMatch method in the sense that labels for noisy label data are removed to make unlabeled data and clean data are used for supervised learning.

3 PROPOSED METHOD

Algorithm 1: Learning with Noisy Data using Semi-supervised Learning

Input: model, noisy dataset D, remain rate p, initial loss weight λ_s, final loss weight λ_f, augmentation strategies (Augment1-3), sharpening temparature T

Step 1: Select clean data from dataset D

for $epoch \leftarrow 0$ ***to*** *total_epoch* **do**
    compute loss on every sample of D
    record loss per sample l_t
    $l_t = l_{t-1} + l_t$
remove the label of top (1-p)% of samples with high loss
get a new labeled (X, Y) and a new unlabeled dataset (X_u)

Step 2: Train model with ensembled FixMatch

for $epoch \leftarrow 0$ ***to*** *max_epoch* **do**
    for $j \leftarrow 0$ ***to*** *batches* **do**
        compute loss on X_l: l_s
        $logit_1, logits_{EMA1}$ = model(Augment1(X_u)), EMA(Augment1(X_u))
        $logit_2, logits_{EMA2}$ = model(Augment2(X_u)), EMA(Augment2(X_u))
        ensemble logit $l_e = (logit_1 + logits_{EMA1} + logit_2 + logits_{EMA2})/4$
        compute loss on (Augment3(X_u), l_e): l_u
        $\lambda \leftarrow Cosine\_decay(\lambda_s, \lambda_f, \text{epoch})$
        $l = l_s + \lambda * l_u$
        update model weights

This paper proposes a semi-supervised learning method that efficiently trains the dataset with noisy labels. The proposed method is based on FixMatch, and it is described in Algorithm 1. First, we

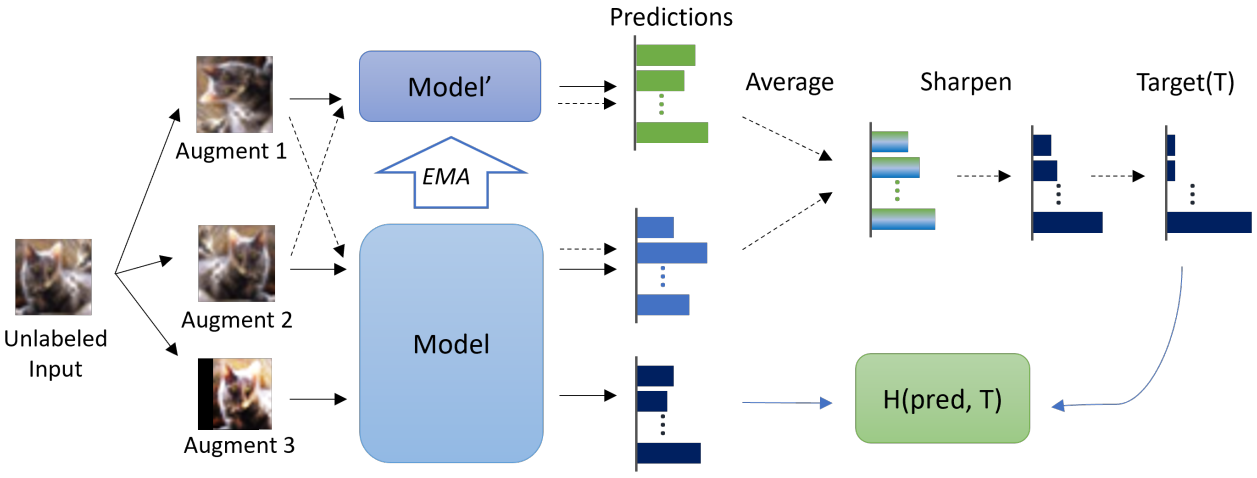

Figure 1: Learning algorithm for unlabeled data. Unlabeled images that are weakly augmented by two weak augmentation strategies (Augment1 and Augment2) are fed into the original model (`Model`) and the EMA model(`Model'`). Then, prediction ensembles are obtained and sharpened with temperature T. Prediction ensembles with the maximum value that exceeds the threshold are converted to pseudo labels. Finally, the pseudo labels are used to compute the cross-entropy loss (H) with strongly-augmented (Augment3) images and update weights.

train all data and select p% of data with the smallest average loss. We choose p as 10 in the experiments of this study. The selected data are regarded as clean data and a labeled dataset (X_l, Y) is formed. The remaining unselected data are considered as noisy data and their labels are removed to create an unlabeled dataset (X_u).

The labeled (X_l, Y) and the unlabeled (X_u) datasets are the inputs to the model in every iteration. Labeled data generates the supervised loss (l_s) through supervised learning, and unlabeled data generates the unsupervised loss (l_u). The learning method with unlabeled data of the proposed algorithm is illustrated in Figure 1. First, the same image is fed into the model through different augmentation techniques. The weakly augmented image is the input to the model and the EMA model, respectively, and the two predict ensembles are used to create a pseudo label. In the experiments of this study, α was chosen to be 0.999, which is generally used in the EMA method. Equation 2 shows how parameter $\Theta_t^{'}$ of the EMA model is calculated. $\Theta_t^{'}$ at the current iteration t is computed as a weighted sum of Θ_t, the parameter of the original model and $\Theta_{t-1}^{'}$, the previous parameter of the EMA model. α is the weight.

$$\Theta_t^{'} = \alpha\Theta_{t-1}^{'} + (1-\alpha)\Theta_t \tag{2}$$

$$sharpen(l, T) = \frac{l_i^{\frac{1}{T}}}{\sum_{j=0}^{L} l_j^{\frac{1}{T}}} \tag{3}$$

After calculating the loss between the generated pseudo label and the predicted value of the strongly-augmented image, sharpening is applied to the prediction to reduce the entropy of the label distribution, as Equation 3. Hyperparameter T was set to 0.5 in the experiments of this study. When temperature T is lower than 1, the model will output low entropy predictions. It is added to the supervised loss to calculate the final loss value that will be used for training and the model weights are updated accordingly.

Table 1: Different types of Augmentation strategies used

Augmentations	Intensity	Augmentation Strategies
Augment1	Weak	RandomHorizontalFlip, RandomCrop
Augment2	Weak	RandomRotation, RandomCrop
Augment3	Strong	Augment1+RandAugmentation

The loss weight λ that determines the ratio of the labeled and the unsupervised losses, follows cosine decay, as in Equation 4.

$$Cosine\_decay(\lambda_s, \lambda_f, t) = 1 + \frac{1}{2}(\lambda_s - \lambda_f)(1 + cos(\frac{t}{t_f}\pi)) \tag{4}$$

Thus, the unsupervised loss will have a considerable influence in the early stage when the influence of unlabeled data is small. $\lambda, \lambda_s, \lambda_f, t_f$ are the current loss weight, the initial loss weight, the final loss weight and the total number of iterations, respectively.

3.1 Prediction Ensemble

Figure 1 shows the overall learning algorithm for unlabeled data. One of the important things in training a model using the unsupervised loss with unlabeled predictors is to make pseudo labels reliable. Therefore, in addition to getting better target values with the weight EMA model (`Model'` in Figure 1), a prediction ensemble is used to ensure the stability of the model predictions. In addition, the efficiency of the ensemble was improved by adding a second weak augmentation technique to FixMatch, which used only one weak augmentation technique. The specific augmentation technique used is shown in Table 1. The net improvement due to this method will be discussed in Section 4.

Table 2: Comparison of accuracy between the proposed method and the previous papers. The best results are in highlighted bold.

	CIFAR-10				CIFAR-100			
Noise Ratio	20%	40%	60%	80%	20%	40%	60%	80%
MentorNet [16]	92	91.2	74.2	60	73.5	68.5	61.2	35.5
Co-Teaching [5]	87.26	82.8	74.04	26.23	64.4	57.42	47.98	23.22
Arazo et al. [17]	94	92.8	90.3	74.1	73.7	70.1	59.5	39.5
O2U-net [6]	92.57	90.33	-	43.41	74.12	69.21	-	39.39
MentorMix [18]	95.6	94.2	91.3	81	**78.6**	71.3	64.6	41.2
DivideMix [7]	96.1	94.9	94.3	**93.2**	77.3	75.2	**72**	**60**
Ours	**96.43**	**95.14**	**94.35**	80.46	75.81	**75.49**	70.99	54.96

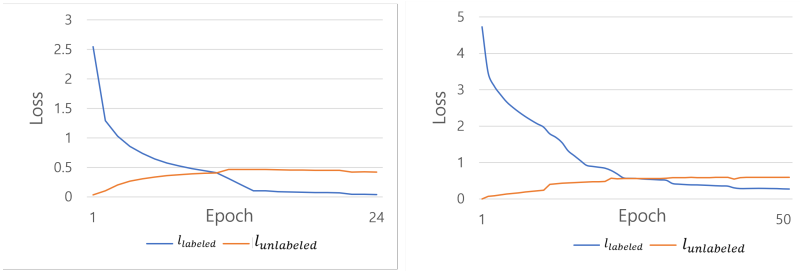

Figure 2: Training loss of labeled (blue line) and unlabeled (orange line) data. (a) Plot of loss on the CIFAR-10 dataset. (b) Plot of loss on the CIFAR-100 dataset.

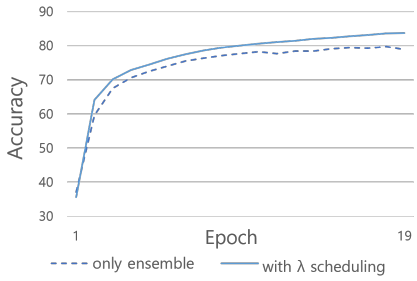

Figure 3: Comparison of accuracy between the case where only ensemble(blue dotted line) is applied and the case with the loss weight(λ) scheduling (blue line)

3.2 Loss Weight Scheduling

The loss weight λ determines the ratio of the labeled and the unsupervised losses. In the proposed algorithm, the loss weight multiplied by the unsupervised loss is amplified in the early stage, and the value gets smaller gradually. The advantages of such scheduling are: 1) At the beginning of training, the valuable properties of unlabeled data are stably learned by amplifying the ones with a small impact of the unsupervised loss. 2) This scheduling has a compensation effect for labeled loss caused by noisy data that cannot be completely removed. Figure 2 shows the increasing trend of the labeled and the unsupervised losses when training. At the beginning of training, the labeled loss dominates the unsupervised loss, but it is more likely to be calculated with noisy data. Therefore, it is possible to increase the proportion of the unsupervised loss at the beginning for faster convergence speed of learning.

Figure 3 compares the initial accuracy trends between the case where only the ensemble method is used and the other case where the loss weight scheduling is additionally applied. Experimental results confirm that the initial accuracy increases when the loss weight scheduling is used.

4 EVALUATION

4.1 Implementation Details

Experiments are carried out in 2 steps. First, clean data from noisy dataset D are selected. Next, the ensembled model is trained. We used ResNet50 [19] for both steps. We trained the model with Stochastic Gradient Descent (SGD) with a learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.00001. CIFAR-10 and CIFAR-100 were used with a batch size of 32 for a total of 500 epochs. As the start and the end values of the loss weight in Equation 4 (λ_s, λ_f), we chose 1.2 and 1, respectively. Sharpening temperature T was set to 0.5. The experiments were conducted on the dual NVIDIA GeForce RTX 3090 GPU.

4.2 Experiment Results and Analysis

Comparison with existing works. Table 2 summarizes the accuracy comparison with existing works on the CIFAR-10 and the CIFAR-100 datasets. As a result, higher accuracy was achieved than the compared existing methods with 20% to 60% noise ratios on CIIFAR-10 and 40% noise ratios on CIFAR-100.

The results show that the proposed method is more efficient at the noise ratio of 20% and 40%, while it shows less improvement at the noise ratio more than 60%. Especially for the noise ratio of 20% to 40%, our method shows only 1.29% of accuracy degradation on the CIFAR-10 dataset and the 0.32% on CIFAR-100 dataset.

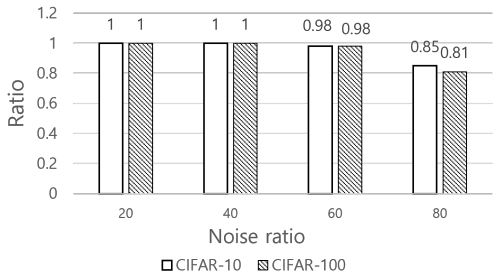

Figure 4: The accuracy of Step 1 from Algorithm 1. It shows the portion of real noisy label among detected noisy label data by the model.

The results for noise ratio 80% on both CIFAR-10 and CIFAR-100 show worse performance than the existing works. We analyze that it is because detected noise label from `Step 1` of Algorithm 1 was unstable. Figure 4 shows the percentage of truly noisy labels among detected noisy labels by the model. From this result, we can figure out that while `Step 1` can detect clean data for the noise ratio of 20 to 60% with high accuracy (higher than 98%), accuracy drops sharply when the ratio reaches 80%.

Ablation Study. Figure 5 shows the net accuracy enhancement when we incrementally apply each technique of the proposed method one by one. This experiment is conducted with the noise ratios of 40% and 60% on the CIFAR-10 dataset. The result shows that the EMA ensemble and the λ scheduling improve classification performance and shows the best accuracy when applying both methods together. The accuracy is improved by 1.59% and 2.77% for the noise ratio of 40% and 60%, respectively.

The discussions in this section can be summarized as follows:

- We proposed a prediction ensemble model using the EMA model and a loss weight(λ) scheduling. It has confirmed that two methods work well for classification task containing noisy label data from ablation study, showing 1.59% and 2.77% of accuracy improvement.
- The proposed method shows better accuracy for data with a relatively low noise ratio. We figure out that it is because the model can hardly detect clean data when the noise ratio is above 80%. The detection accuracy drops below 85% for the noise ratio of 80%.
- The trained model is robust to noise ratio changes. It suffers from only 2.08% of accuracy degradation when the noise ratio increases from 20% to 60% on CIFAR-10 and 0.32% of accuracy degradation when the noise ratio increases from 20% to 40% on CIFAR-100, respectively.

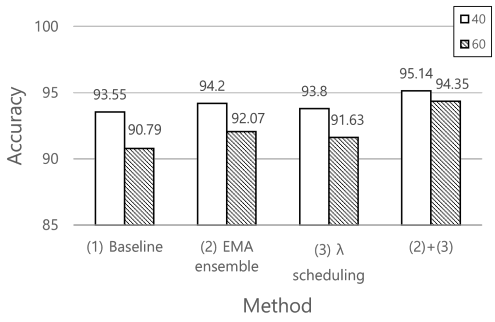

Figure 5: The accuracy comparison results of (1) Baseline (only FixMatch), (2) applying the EMA ensemble on Baseline, (3) applying the λ scheduling on Baseline and (4) applying both the ensemble and the scheduling on Baseline. The white bar indicates the result of noisy data for noise ratio of 40% and the hatched bar is that for the noise ratio of 60% on the CIFAR-10 dataset.

5 CONCLUSION

Since the data labeling process is usually done by humans, labeling mistakes are inevitable. Deep learning models are vulnerable to incorrect labels and the model's performance may drop drastically when trained with noisily-labeled data. Methods that detect and remove noisy data have been studied to solve the problem. However, removing lots of noisy data may cause significant information loss. In this paper, a method to convert noisy data into unlabeled data instead of removing the noisily-labeled data is proposed. When two enhancement techniques, ensemble, and λ scheduling are applied together with the proposed method, the accuracy is improved by 1.59% and 2.77% for the noise ratios of 40% and 60%, respectively on the CIFAR-10 dataset.

ACKNOWLEDGMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-00131, Development of Intelligent Edge Computing Semiconductor For Lightweight Manufacturing Inspection Equipment)

REFERENCES

[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Michael C Mozer, and Yoram Singer. Identity crisis: Memorization and generalization under extreme overparameterization. *arXiv preprint arXiv:1902.04698*, 2019.
[2] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Tin-Yau Kwok. Searching to exploit memorization effect in learning with noisy labels. In *International Conference on Machine Learning*, pages 10789–10798. PMLR, 2020.
[3] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
[4] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.
[5] Lynne Cook and Marilyn Friend. Co-teaching: Guidelines for creating effective practices. *Focus on exceptional children*, 28, 1995.
[6] Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. O2u-net: A simple noisy label detection approach for deep neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3326–3334, 2019.

[7] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

[8] Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Robust curriculum learning: from clean label detection to noisy label self-correction. In *International Conference on Learning Representations*, 2020.

[9] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

[10] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[11] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[12] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *arXiv preprint arXiv:2103.00550*, 2021.

[13] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019.

[14] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[15] 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*, 33:596–608, 2020.

[16] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR, 2018.

[17] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR, 2019.

[18] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *International Conference on Machine Learning*, pages 4804–4815. PMLR, 2020.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.