This WACV 2023 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Bootstrapping the Relationship Between Images and Their Clean and Noisy Labels

Brandon Smart
Australian Institute for Machine Learning
University of Adelaide, Australia
a1743623@adelaide.edu.au

Gustavo Carneiro
Centre for Vision, Speech and Signal Processing
University of Surrey, United Kingdom
g.carneiro@surrey.ac.uk

Abstract

Many state-of-the-art noisy-label learning methods rely on learning mechanisms that estimate the samples' clean labels during training and discard their original noisy labels. However, this approach prevents the learning of the relationship between images, noisy labels and clean labels, which has been shown to be useful when dealing with instance-dependent label noise problems. Furthermore, methods that do aim to learn this relationship require cleanly annotated subsets of data, as well as distillation or multi-faceted models for training. In this paper, we propose a new training algorithm that relies on a simple model to learn the relationship between clean and noisy labels without the need for a cleanly labelled subset of data. Our algorithm follows a 3-stage process, namely: 1) self-supervised pre-training followed by an early-stopping training of the classifier to confidently predict clean labels for a subset of the training set; 2) use the clean set from stage (1) to bootstrap the relationship between images, noisy labels and clean labels, which we exploit for effective relabelling of the remaining training set using semi-supervised learning; and 3) supervised training of the classifier with all relabelled samples from stage (2). By learning this relationship, we achieve state-of-the-art performance in asymmetric and instance-dependent label noise problems[1]. *Code is available at* `https://github.com/btsmart/bootstrapping-label-noise`.

1. Introduction

Supervised deep learning has had great success generating effective classification models from sets of labelled training data [23, 25]. Modern deep learning models require large-scale datasets to achieve state-of-the-art (SOTA) results [37, 38]. However, real-world large-scale datasets, such as those collected from search engines or available from hospitals and clinics, tend to have a non-negligible amount of instance-dependent label noise (IDN) [31, 51]. Existing methods often attempt to address instance-independent label noise (IIN), such as symmetric or asymmetric noise [14, 56, 63]. Handling the IDN present in large-scale real-world datasets has become one of the main research problems in the field.

When naively trained with noisy-labelled data, deep learning models generalise poorly because they can easily overfit the incorrectly labelled samples [60]. Many methods have been developed for handling label noise, with SOTA approaches relying on sample relabelling mechanisms. These strategies are based on techniques to estimate the relationship between images and clean labels, and after relabelling, the old noisy labels are discarded [28, 46, 63]. However, to model how different image features and noisy labels affect the mislabelling process in IDN, we need to estimate the relationship between images, clean labels and noisy labels [16, 63]. Some methods have attempted to model this relationship with noise-transition matrices and corrective layers for asymmetric noise [14, 36, 55] or part-dependant noise in place of instance-dependant noise [54], but they failed to achieve SOTA results.

Rather than the usual noisy-label learning setting, where a set of cleanly annotated samples is not available, some methods assume the existence of a subset of training data containing images, clean labels *and* noisy labels [16, 20, 50]. By training a model that predicts clean labels from both images and noisy labels (see right of Figure 1), these methods are able to learn the relationship between image features, noisy labels and clean labels, allowing them to model IDN and more effectively relabel noisy samples. However, it can be expensive, difficult and time-consuming to obtain a clean subset of data with noisy labels *and* clean labels that is representative of the instance-dependant noise in the dataset. Furthermore, these methods require distillation to a more standard model (such as the one on the left of Figure 1) for evaluation on samples without labels.

[1] Supported by Australian Research Council through grants DP180103232 and FT190100525.

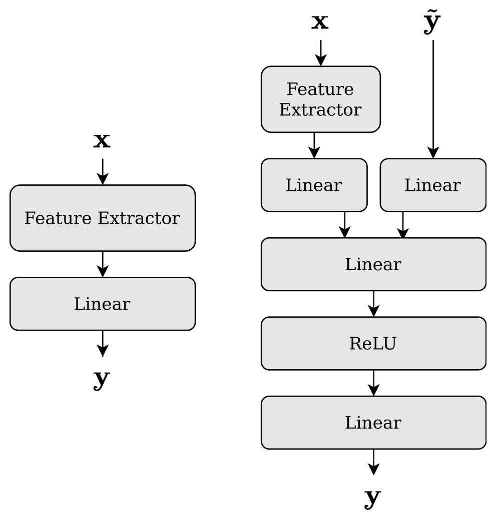


Figure 1. On the left, we show a 'normal' deep neural network model used for noisy label learning tasks. On the right, we present a 'modified' model that can learn the relationship between images $\mathbf{x}$, noisy labels $\tilde{\mathbf{y}}$ and clean labels $\mathbf{y}$, similar to those used by methods that have access to a clean set of data [16, 20, 50].

In this paper, we introduce a new algorithm to learn the relationship between images and their clean and noisy labels without using any clean-label set. Our algorithm follows a 3-stage process (see Fig. 2): 1) *Bootstrapping:* self-supervised pre-training followed by an early-stopping training [60] of the classifier that receives images and 'null' labels as input and predicts the noisy labels as output – this stage forms a subset of predicted clean labels for the second stage of training; 2) *Semi-supervised Learning:* use this predicted clean subset to learn the relationship between images, noisy labels and clean labels, which we exploit for an effective, explicit relabelling of the remaining training set; and 3) *Final training:* supervised training of the classifier using the relabelled samples. The main contributions of this paper are:

- An effective three-stage training algorithm designed to address instance-dependent label noise by learning the relationship between images and their clean and noisy labels – using a noise-transition sample balancing scheme, explicitly relabelling training samples and without requiring a cleanly annotated training set;
- A method that reaches SOTA asymmetric and instance dependent label noise results using a simple single-model architecture, unlike DivideMix [28] (and its derivatives such as [10, 21, 34, 41, 66]) that require a more complex 2-model architecture.
- A 'label dropping' strategy that removes the need for distillation to a standard model and allows predictions to be made on samples with *and* without noisy labels;

2. Related Work

2.1. Noisy Label Learning Based on Semi-supervised Learning (SSL)

Many SOTA methods in noisy label learning use SSL techniques to perform label correction and consistency regularization. DivideMix [28] and variants [10, 41, 66] perform sample re-labelling with co-teaching and *MixMatch* data augmentation [6]. ELR+ [32] and PES [5] similarly use MixUp [61] based SSL on top of a regularising loss functions. These techniques are effective for symmetric and asymmetric noise, but are dependent on carefully tuned hyperparameters and cautious integration of the sample re-labelling and SSL techniques used.

To identify the incorrectly labelled samples for forming the unlabelled set, many methods depend on loss-separation techniques, relying on the ability of deep networks to learn clean samples faster than noisy samples [2], which leads to lower loss values for clean samples after a few stages of training [1, 7, 17, 28, 57]. FINE alternatively uses eigen-decomposition to separate samples in feature space [21]. However, such automatic and dynamic identification of noisy-label samples is a brittle process that tends to fail in challenging noisy-label learning scenarios, such as instance-dependent label noise, because the differences between hard clean-label and noisy-label samples can be subtle during early training stages.

2.2. Label Transition Estimation Methods

Many methods attempt to model the class-dependent asymmetric noise, such as with a label transition matrix [55], by learning noise-adaption layers and performing loss-correction [14, 36], or by using reconstruction error as a consistency objective [39]. However, these methods are not as competitive as SSL approaches in Sec. 2.1 because they generally do not address instance-dependent noise and have limited mechanisms to make use of mislabeled samples. Methods that attempt to handle semantic noise by estimating instance-based transition matrices can in principle deal with semantic noise [14] and can provide guarantees on convergence and bounds on generalization error [63], but they do not provide SOTA results in practice.

2.3. Methods Based on Clean Validation Sets

Alternatively, researchers have explored learning methods that require the existence of a small, additional clean set of data to learn from. For instance, many meta-learning strategies require clean validation samples to adjust the weights of each training sample [40], to simulate regular training with synthetic noise labels [29], to learn an explicit weighting function [43], or to estimate the noise transition matrix [52]. Other noisy label learning algorithms rely on clean sets of data for which noisy labels

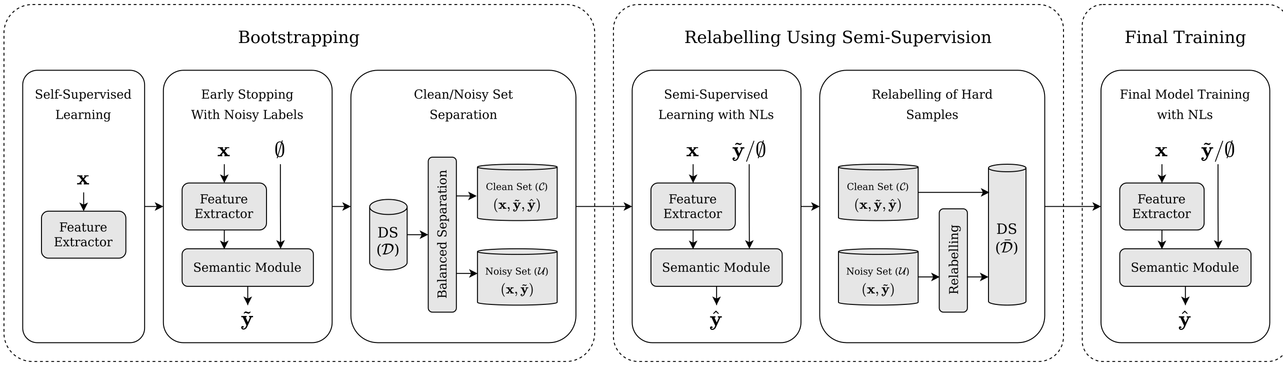


Figure 2. Our proposed algorithm. The bootstrapping stage consists of self-supervised pre-training, followed by early-stopping classification training to identify a small set of confidently relabelled samples (i.e., clean labelled samples). This set is then used to learn the relationship between images, noisy labels and clean labels during semi-supervised learning, and at the end of this stage, we relabel the samples classified as noisy during bootstrapping. The last stage is the final training that trains the classifier using the clean and noisy samples identified in the SSL stage.

and clean labels exist for samples, so that the relationship between image features, noisy labels and clean labels can be learned [16, 20, 50], using fully-connected neural networks to predict true labels from images and their noisy labels. Together, these methods show the utility of representative clean sets of data to the noisy label problem, but they rely on manual labelling, which can be expensive and time-consuming to collect, particularly in domains that require a high degree of expertise for labellers, such as medical imaging [33, 65].

Other methods aim to dynamically construct a pseudo-clean set out of high-confidence samples, such as by using K-Nearest Neighbours to identify related samples in the feature space [4, 35], or by using meta-learning to identify a dictionary of dynamically updating valuable training samples [64].

2.4. Background Material

Our method relies on many techniques previously developed in the field, such as self-supervision, SSL and data augmentation. Recently, self-supervised methods such as SimCLR [8, 9] and SCAN [49] have been used for pre-training, or as auxilliary objectives in noisy-label learning tasks [41, 66], due to their ability to learn high-level features from noisy data without the risk of overfitting incorrect labels. In this paper, we utilize the SSL method FixMatch [45], which uses pseudo-labelling thresholds and consistency between strong and weak augmentations to regularize training through consistency regularization [3, 24, 48] and entropy minimization [15, 26]. Strong data augmentation strategies, such as RandAugment [12], AutoAugment [11] and MixUp [61] have been shown to be effective for regularising training, preventing overfitting and dramatically improving the tolerance of algorithms to higher noise levels [1, 28, 34].

3. Methodology

Our algorithm (see Fig. 2) is motivated by the objective to train a model that can accurately relabel samples by predicting true labels from images *and* noisy labels without requiring clean-labelled data. The stages of the proposed algorithm are: 1) Bootstrapping: perform self-supervised pre-training and early-stopping training to identify a representative, clean subset of samples, 2) SSL: learn the instance-dependant noise relationship from the clean set (from stage 1) and use it to relabel the remaining noisy samples, and 3) Final Training: use the relabelled samples from stage 2 to train a final, regularized classifier.

For the methods described below, assume the availability of a training set $\mathcal{D} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{H \times W \times R}$ denotes an image of size $H \times W$ with R colour channels, and $\tilde{\mathbf{y}} \in \mathcal{Y} \subset \{0,1\}^{|\mathcal{Y}|}$ represents the one-hot noisy label. Our model, referred to as 'modified', receives an image and the noisy label at the input and outputs a clean label classification distribution, with $f_\theta : \mathcal{X} \times \mathcal{Y} \to \Delta_{|\mathcal{Y}|-1}$, where $\Delta_{|\mathcal{Y}|-1}$ represents the $|\mathcal{Y}| - 1$ probability simplex, and $\theta \in \Theta$ denotes the model parameters. Note that we also consider a 'normal' model, which is a model that takes an image input and outputs a classification, with $f_\theta : \mathcal{X} \to \Delta_{|\mathcal{Y}|-1}$.

3.1. Bootstrapping

The goal of this first stage is to train a model that accepts images and noisy labels and predicts clean labels, however at the beginning of training we only have access to images and noisy labels from $\mathcal{D}$. Following [49, 66], we start with SimCLR pre-training [8], which allows us to learn an initial feature representation from $\mathcal{D}$ without the risk of overfitting.

Next, we take the pre-trained model above to learn a classifier with early-stopping and small learning rate with

$$\theta^* = \arg\min_\theta \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i,\tilde{\mathbf{y}}_i)\in\mathcal{D}} \mathbb{E}_{a(.)\sim\mathcal{A}_S}\left[\ell_{CE}(\tilde{\mathbf{y}}_i, f_\theta(a(\mathbf{x}_i), \mathbf{0}_{|\mathcal{Y}|}))\right], \tag{1}$$

where $a(.)$ is a strong data augmentation sampled from the set of strong data augmentation functions $\mathcal{A}_S$, $\ell_{CE}(.)$ denotes the cross-entropy loss function, and $\mathbf{0}_{|\mathcal{Y}|}$ is a 'null' label vector with $|\mathcal{Y}|$ zeros.

Then, we use our trained model to generate a prediction distribution for all training samples using test-time weak augmentation, as follows:

$$\hat{\mathbf{y}}_i = \mathbb{E}_{a(.)\sim\mathcal{A}_W}\left[f_{\theta^*}(a(\mathbf{x}_i), \mathbf{0}_{|\mathcal{Y}|})\right], \tag{2}$$

where $a(.)$ is a weak augmentation sampled from the set of weak data augmentation functions $\mathcal{A}_W$. We also have dropout enabled during this evaluation process. By using dropout and multiple weak augmentations to evaluate samples, we penalise samples with highly confident but inconsistent predictions [63]. The confidence prediction for $\mathbf{x}_i$ is given by $\max_{c\in\mathcal{Y}} \hat{\mathbf{y}}_i(c)$.

We then split the dataset into a confident clean set and a noisy set. However, if we naively select the most confident samples, we will disproportionally select samples from easy classes, and samples whose original noisy labels were correct. We want the clean set to contain representative samples from all classes and noise transitions in the dataset, as the upcoming SSL process can only learn noise transitions which are present in the clean set. To achieve this, we propose noise-transition sample balancing that first estimates the dataset's noise transition matrix $\mathbf{T}$ by using the noisy labels and predicted labels for the 90% of most confident predictions per class, where $\mathbf{T}_{ij}$ represents the estimated proportion of samples in the dataset with the noisy label i and clean label j. We then select the $K \times |\mathcal{Y}| \times \mathbf{T}_{ij}$ most confident samples from each noise transition, as well as any other samples $\mathbf{x}_i$ where $\max_{c\in\mathcal{Y}} \hat{\mathbf{y}}_i(c) > \tau$, where K is a hyperparameter controlling the minimum fraction of samples from each subset we want to select, and τ is a hyperparameter controlling how confident a prediction needs to be before it is guaranteed to be selected.

We note that this process does not guarantee that all instance-dependent relationships between image features and noisy label transitions are captured, but in practice this ensures a large coverage of the different noise transitions present in the dataset. This 'noise-based' balancing approach can be contrasted against the usual class-based balancing typically seen in noisy-label learning methods, where samples are selected to balance the number of samples per class. The initial clean set will contain the samples with both the noisy and estimated clean labels with $\mathcal{C} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i, \hat{\mathbf{y}}_i) | (\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D}\}$, and the initial noisy set will contain the samples and noisy labels as $\mathcal{U} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) | (\mathbf{x}_i, \tilde{\mathbf{y}}_i, \hat{\mathbf{y}}_i) \notin \mathcal{C}, \text{ and } (\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D}\}$.

3.2. Semi-Supervised Learning (SSL) for Noisy Label Correction

The next stage of our framework takes the initial clean set $\mathcal{C}$ and the initial noisy set $\mathcal{U}$ to train the SSL model, where images and noisy labels (the model inputs) are present for all samples, and the true labels (the model output) are present for samples in $\mathcal{C}$. Our SSL algorithm is based on FixMatch (see Fig. 3), which achieves competitive performance by focusing on consistency regularization and entropy minimization [45]. We do not reinitialize the network before semi-supervised learning, instead using the bootstrapping process as a form of warmup.

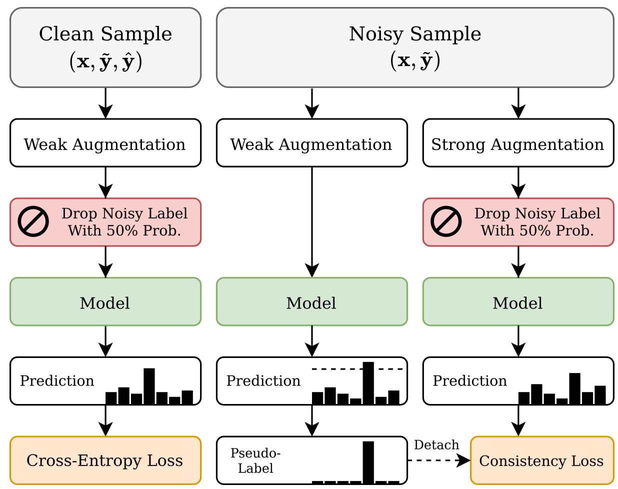


Figure 3. The noisy-label FixMatch algorithm.

Because we are learning to predict the 'true' labels of samples from images and noisy labels, our model is able to learn a joint distribution between the three, similar to works that require a clean set [16, 20, 50]. However, if we were to train our model to always make predictions from images *and* noisy labels, our model would no longer be able to make meaningful predictions on samples without associated noisy labels, such as those in a 'test' set. To remedy this, we randomly 'drop' the one-hot noisy label from samples 50% of the time, replacing it with a 'null' label. By doing this, our model also learns a direct relationship between images and true labels without depending on the noisy labels, allowing us to evaluate test samples by passing them into the model alongside a 'null' label. In our implementation of FixMatch, we drop the noisy label of supervised samples and strong augmentations of unsupervised samples 50% of the time. However, we keep the noisy labels for the weakly augmented unsupervised samples as we want to always use these noisy labels to predict higher accuracy pseudo-labels, given that the loss is not backpropagated along these samples (see Fig. 3). We experiment with this decision in the

Supplementary Material. The SSL training is based on the following optimisation:

$$\theta^* = \arg\min_{\theta\in\Theta} \frac{1}{|\mathcal{C}|}\sum_{(\mathbf{x}_i,\hat{\mathbf{x}}_i,\hat{\mathbf{y}}_i)\in\mathcal{C}} \mathbb{E}_{a(.)\in\mathcal{A}_W}[\ell_{CE}(\hat{\mathbf{y}}_i, f_\theta(a(\mathbf{x}_i), \iota_{50\%}(\tilde{\mathbf{y}}_i)))]+ \frac{1}{|\mathcal{U}|}\sum_{(\mathbf{x}_i,\tilde{\mathbf{x}}_i)\in\mathcal{U}} \mathbb{I}(\max\bar{\mathbf{y}}_i > \kappa)\mathbb{E}_{a(.)\in\mathcal{A}_S}[\ell_{CE}(\lceil\bar{\mathbf{y}}_i\rceil, f_\theta(a(\mathbf{x}_i), \iota_{50\%}(\tilde{\mathbf{y}}_i)))], \tag{3}$$

where $\mathbb{I}(.)$ denotes an indicator function, $\bar{\mathbf{y}}_i = \mathbb{E}_{a(.)\sim\mathcal{A}_W}[f(a(\mathbf{x}_i), \tilde{\mathbf{y}}_i)]$, $\iota_{50\%}(\tilde{\mathbf{y}}_i)$ randomly returns $\tilde{\mathbf{y}}_i$ or $\mathbf{0}_{|\mathcal{Y}|}$, each with 50% chance, and $\lceil\bar{\mathbf{y}}_i\rceil$ is an operator to transform $\bar{\mathbf{y}}_i$ into a binary vector, with 1 for the class with largest probability and 0 otherwise. After this SSL stage, we re-label the whole training set to form

$$\bar{\mathcal{D}} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i, \bar{\mathbf{y}}_i)|(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D}\}, \tag{4}$$

with $\bar{\mathbf{y}}_i$ defined in (3). As in the bootstrapping phase of training, we enable dropout when averaging over the predictions of multiple weak augmentations to calculate $\bar{\mathbf{y}}_i$. Note in (4) that our model uses the learned relationship between images, noisy labels and 'true' labels to relabel the remaining noisy samples.

3.3. Final Model Training

After forming $\bar{\mathcal{D}}$, we train a final model with strong augmentations and MixUp, due to their robustness to any noise that may remain in the re-labelled training set [1, 28]. MixUp is applied on both images and their noisy labels together. After applying Mixup, we randomly replace the noisy label with a 'null' label in 50% of samples.

4. Experiments

4.1. Data sets

To investigate our method, we perform experiments using the CIFAR-10, CIFAR-100 [22], Animal10N [46], and Webvision [31] datasets. CIFAR-10 and CIFAR-100 consist of 50,000 training and 10,000 testing images of size 32×32 pixels, with 10 and 100 classes respectively. As CIFAR-10 and CIFAR-100 do not contain label noise, we follow the literature and perform experiments with different types of controlled, synthetic label noise. The first type of noise is the Polynomial Margin Diminishing (PMD) semantic noise [62], where confusing samples near decision boundaries are mislabelled at higher rates than samples far from decision boundaries. The second type is the semantic noise introduced by Lee et al. [27], where incorrect predictions from trained VGG [44], DenseNet [19], and ResNet [18] models are used to generate mislabelled samples (which we refer to as 'RoG' noise). We also test our system with symmetric noise rates of $\{20\%, 50\%, 80\%, 90\%\}$ and asymmetric noise using the mapping from [28, 36] with 40% rate.

The Animal10N dataset [46] consists of 50,000 training images and 5,000 testing images of size 64×64, consisting of five pairs of semantically similar classes. Images are collected for each label using online search engines, resulting in incorrect classifications for an estimated 6% to 10% of samples. Finally, we test with Mini-Webvision, which consists of the 65,944 samples which make up the first 50 classes of the Webvision dataset that contains images collected from the internet. Images are resized to 256×256, and the corresponding 50 classes in the ILSVRC12 dataset [13] are also used for validation.

4.2. Implementation

Following contemporary work [10, 28, 32], we use a PreAct-ResNet-18 (PRN18) network [18] as our backbone model for CIFAR10 and CIFAR100 experiments, use a VGG19 model [44] as our backbone for Animal10N, and use a Inception-ResnetV2 model [47] as our backbone for Webvision. For weak augmentations, we use horizontal flipping and random cropping, and for strong augmentations we used AutoAugment [11], followed by horizontal flipping and random cropping. For all experiments, we perform bootstrapping with strong augmentations and perform pseudo-labelling with with 25 weak augmentations. Final model training is done using Mixup [61] (with $\alpha = 1$). For all stages of training, we use stochastic gradient descent, with additional information about the optimizer and training schedule hyperparameters provided in the supplementary material. For Webvision, we additionally use label smoothing (with $\epsilon = 0.1$). Following existing implementations of FixMatch (such as by *TorchSSL* [59]), we use Exponential Moving Average (EMA) models to perform temporal ensembling [24]. For fair comparison with existing noisy-label learning methods, we shorten the training schedules typically used by FixMatch implementations (we use 100,000 training iterations with $\mu = 4$ rather than 1,000,000 iterations with $\mu = 8$).

We report two sets of results for our experiments in order to understand how the use of noisy labels during semi-supervised learning can improve results: 'Normal Model' and 'Modified Model'. In both cases, we utilise the training procedure outlined in Figure 2, however:

1) In 'Normal Model', we use a standard model which only accepts image inputs. Because this type of model does not aim to learn the relationship between noisy labels and clean labels, we perform class-based balancing rather than the proposed noise-based balancing.

2) In 'Modified Model', we use noise-based balancing, as well as the model that accepts images and noisy labels, as described in Sec. 3.

For all our experiments, we also report the accuracy obtained from test-time augmentation, where 25 weakly aug-

Method	Top-1
Cross Entropy	79.4
SELFIE [46]	81.8
PLC [62]	83.4
NCT [42]	84.1
Ours (Normal Model)	85.84
+ *Test-Time Aug.*	*86.98*
Ours (Modified Model)	**88.48**
+ *Test-Time Aug.*	*89.38*

Table 1. Test accuracy (%) for Animal10N. Top methods within 1% in **bold** (Results other than ours are from [42]).

Method	Webvision		ILSVRC2012	
	Top-1	Top-5	Top-1	Top-5
ELR [32]	76.26	91.26	68.71	87.84
ELR+ [32]	77.78	91.68	70.29	89.76
PropMix [10]	78.84	90.56	/	/
NGC [53]	79.16	91.84	74.44	91.04
FaMuS [58]	79.40	**92.80**	**77.00**	**92.76**
RRL [30]	76.30	91.50	73.30	91.20
Ours (Normal Model)	**80.24**	90.84	76.44	90.00
+ *Test-Time Aug.*	*81.52*	*92.52*	*78.36*	*91.80*
Ours (Modified Model)	**80.88**	**92.76**	75.96	**92.20**
+ *Test-Time Aug.*	*83.16*	*94.28*	*79.64*	*94.20*

Table 2. Test accuracy (%) for Webvision. Top methods within 0.5% in **bold**

mented versions of each testing sample are generated, and the model's predictions (with dropout enabled) are averaged to generate a final prediction. Because this extends the inference time for our model, we report these results separately in italics.

4.3. Results

In Tables 1 and 2, we show the accuracy of our model on the Animal10N and Webvision benchmarks, respectively. In both cases, it can be seen that the modified model improves results, demonstrating that on these instance-dependant noise datasets, learning the relationship between images, noisy labels and 'true' labels has allowed more accurate relabelling, and an improvement in final model accuracy. Animal10N, which consists of five pairs of commonly confused image classes, sees a significant improvement of 2.64% from using our modified training approach, due to the relationship between commonly confused classes being easy to learn. Webvision sees a more modest but still meaningful improvement of 0.64% accuracy when using our modified model. It can be seen that on both benchmarks, we report state-of-the-art results, and that using dropout-based test-time augmentation further improves our accuracy.

We next turn our attention to the synthetic *'Polynomial Margin Diminishing (PMD)'* [62] and *'RoG'* [27] instance-dependent noisy label benchmarks based on CIFAR10 and CIFAR100 in Tables 3 and 4. In all cases considered in these two tables, our accuracy results are substantially higher than by other approaches, even without our modified model. When we do use a modified model for these synthetic instance-dependent noisy-label benchmarks, we find mixed results, with performance often decreasing, potentially due to learned noise transitions not generalizing as well as they do in real-world instance-dependent noise datasets.

Finally, in Table 5 we show the results of our method on the synthetically constructed symmetric and asymmetric noise for CIFAR10 and CIFAR100. These noise types are rare in practice, but they are common noisy-label benchmarks so we include them here for completeness. We see that our method is competitive with the state-of-the-art on CIFAR10 symmetric and asymmetric noise, despite featuring fewer mechanisms designed to address these types of noise. We particularly note 40% asymmetric noise, which benefits from the modified model due to noisy labels greatly limiting the set of feasible samples for each image, allowing us to exceed the state-of-the-art. In contrast to this, we report our results on symmetric label noise in CIFAR100, where existing methods perform better than ours. In all cases, our modified model is able to take advantage of the noise to provide more accurate relabelling, but the regularisation strategies that other methods use work well under the assumption of symmetric noise and provide stronger results. Perfectly symmetric noise over 100 classes is rare in practice though, and our results show universally strong performance on real-world instance-dependent datasets.

4.4. Predictions with Noisy Labels

A unique feature of our method is that our final model can be used to predict the labels of samples with *and* without noisy labels. In some applications, e.g., tagged image classification, images at test-time may also have noisy labels associated with them, which our model can use to improve classification performance. To show this, we generate artificial noisy labels for all the samples in the CIFAR10 test set using the same procedure as we used for 40% Asymmetric noise, and in Table 6 we show the accuracy of our model on these samples when these noisy labels are and are not used. We see that our model performance increases when using the noisy labels, because our model has learned to use noisy labels (when they are available) to improve prediction accuracy.

We further demonstrate this learned relationship in Table 7, where we show how the prediction of our model changes depending on what noisy label is provided to the model. In this 40% Asymmetric noise dataset, the image of the dog shown would only ever be labelled as a dog or a cat, and in both cases it can be seen that our model makes the correct prediction. Our model also makes the correct prediction when provided with no noisy label, although with lower

Dataset	CIFAR-10			CIFAR-100		
Noise Type	Type-I 35%	Type-II 35%	Type-III 35%	Type-I 35%	Type-II - 35%	Type-III - 35%
Cross-Entropy	78.11	76.65	76.89	57.68	57.83	56.07
PLC [63]	82.80	81.54	81.50	60.01	63.68	63.68
Ours (Normal Model)	**94.06**	**93.25**	**93.35**	65.87	65.80	66.36
+ *Test-Time Aug.*	*94.72*	*93.79*	*93.97*	*66.83*	*66.48*	*67.42*
Ours (Modified Model)	**94.00**	**93.76**	**94.23**	**68.25**	**68.14**	**68.22**
+ *Test-Time Aug.*	*94.39*	*94.19*	*94.23*	*70.13*	*69.35*	*70.13*

Table 3. Test accuracy (%) for Polynomial Margin Diminishing Noise [62]. Top methods are in **bold**.

Data set	CIFAR-10			CIFAR-100		
Method/ Noise Ratio	DenseNet (32%)	ResNet (38%)	VGG (34%)	DenseNet (34%)	ResNet (37%)	VGG (37%)
D2L + RoG [27]	68.57	60.25	59.94	31.67	39.92	45.42
CE + RoG [27]	68.33	64.15	70.04	61.14	53.09	53.64
PropMix [10]	84.25	82.51	85.74	60.98	58.44	60.01
Ours (Normal Model)	**93.26**	**92.05**	**93.29**	62.47	**64.91**	**64.98**
+ *Test-Time Aug.*	*93.87*	*92.66*	*93.86*	*63.40*	*65.74*	*66.10*
Ours (Modified Model)	89.46	90.97	89.77	**63.68**	63.09	63.89
+ *Test-Time Aug.*	90.25	*91.85*	*90.42*	*65.15*	*64.70*	*65.30*

Table 4. Test accuracy (%) for the RoG label noise benchmark [27], where baseline results are from [27]. Top methods are in **bold**.

Dataset	CIFAR-10					CIFAR-100			
Noise type	Sym.				Asym.	Sym.			
Method / Noise Ratio	20%	50%	80%	90%	40%	20%	50%	80%	90%
Cross-Entropy [28]	82.7	57.9	26.1	16.8	72.3	61.8	37.3	8.8	3.5
ELR [32]	**95.8**	**94.8**	93.3	78.7	93.0	77.6	73.6	60.8	33.4
DivideMix [28]	**95.7**	94.4	92.9	75.4	92.1	76.9	74.2	59.6	31.0
AugDesc [34]	**96.3**	**95.4**	**93.8**	91.9	94.6	**79.5**	**77.2**	66.4	41.2
ContrastToDivide [66]	**96.4**	**95.3**	**94.4**	**93.6**	93.5	**78.7**	**76.4**	**67.8**	**58.7**
PropMix [10]	**96.09**	**95.53**	**93.77**	**93.20**	94.64	76.99	73.71	66.75	**58.32**
Ours (Normal Model)	95.04	**95.13**	**94.51**	91.72	94.91	69.02	68.87	64.09	55.91
+ *Test-Time Aug.*	*95.47*	*95.39*	*94.90*	*92.32*	*95.14*	*70.17*	*69.57*	*65.10*	*56.81*
Ours (Modified Model)	**95.99**	**95.59**	**94.48**	**93.52**	**95.85**	75.09	70.86	57.03	39.95
+ *Test-Time Aug.*	*96.75*	*96.16*	*94.98*	*93.93*	*96.42*	*76.65*	*72.95*	*57.83*	*40.31*

Table 5. Test accuracy (%) for all competing methods on CIFAR-10 and CIFAR-100 under symmetric and asymmetric noises. Results from related approaches are as presented in [28] and [53]. Top methods within 1% are in **bold**.

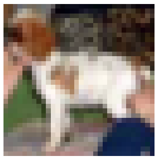
Test Set	Accuracy
Without Noisy Labels	95.85
With Noisy Labels	97.59

Table 6. Accuracy of our model tested with and without noisy labels on CIFAR10 Asym. 40% noise. Noisy labels are generated for training and testing samples using the same procedure.

confidence because it does not have access to the noisy label. Because our model has learned that none of the other classes are ever confused with dogs, it makes other predictions when provided with these labels. These results show that our model has learned the noise relationship, and that it can use both the image features and the noisy label (when it is available) to generate higher confidence predictions.

4.5. Clean Set Selection

On the left of Figure 4, we show the distribution of confidences of our model after bootstrapping, and see that the highest confidence predictions are almost entirely for samples whose predicted 'true' label ($\hat{\mathbf{y}}$) is correct, which allows us to select highly accurate clean sets for SSL. On the right, we show the percentage of the selected samples for the clean set that were originally clean (i.e. samples for which $\tilde{\mathbf{y}}$ matches the true label) if no form of noise balancing is performed. We see that if we did not perform noise balancing, the selected clean set would disproportionally consist of clean samples, which would cause a degenerate relationship between noisy labels and clean labels to be learned during the SSL stage of training, preventing accurate relabeling of samples from other noise transitions.



Noisy Label	Prediction (Confidence)
-	Dog (90.27%)
Airplane	Bird (65.39%)
Automobile	Automobile (62.75%)
Bird	Bird (98.30%)
Cat	Dog (94.51%)
Deer	Deer (96.06%)
Dog	Dog (96.29%)
Frog	Frog (96.49%)
Horse	Horse (75.84%)
Ship	Ship (95.46%)
Truck	Truck (96.88%)

Table 7. Predictions made by our model with different noisy labels for a testing sample (showing a dog) in Asym 0.4 noise for CIFAR10. '-' represents using a null label in place of a noisy label.

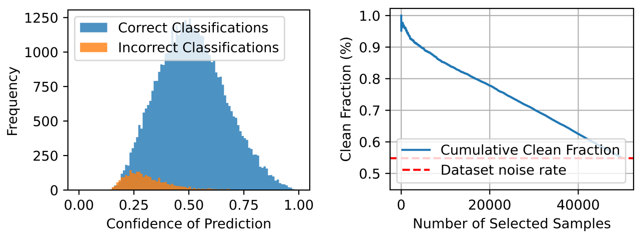


Figure 4. Histograms showing the distribution of confidences after bootstrapping for correct/incorrect classifications (left), and how the highest confidence samples are disproportionally clean (right) for 50% Symmetric Noise on CIFAR10.

4.6. Ablations and Training Time

In Tables 8, 9 and 10, we perform a number of ablation studies on the CIFAR10 40% Asymmetric noise. In Table 8, we show the accuracy of our trained model after each stage of training. We see that the model accuracy improves after each stage of training, with the final training's use of MixUp and strong augmentations providing an additional 0.87% accuracy over the semi-supervised learning stage.

In Table 9, we see the large impact that the choice of augmentations has on the number of errors in the clean set after the bootstrapping phase. We see that using strong augmentations for training greatly reduces the number of errors, likely due to their regularising effect and its ability to prevent overfitting to the noisy labels. During the evaluation stage however, averaging the prediction of the model over multiple weak augmentations performs best. We see that this matches the findings by Nishi et al. [34], who find that using strong augmentations for training and weaker augmentations for loss modelling works best.

In Table 10, we see the similarly large impact that using self-supervised pretraining has on the number of errors in the clean set after bootstrapping. In the supplementary material, we show additional experiments with 'null labels' and different model architectures.

Training Stage	Accuracy
After Bootstrapping	91.41
After Semi-Supervised Learning	94.98
After Final Training	95.85

Table 8. Model accuracy after each stage of training on CIFAR10 Asym. 40% noise.

		Evaluation Aug.		
		None	Weak	Strong
Train	None	579	361	456
	Weak	265	56	300
	Strong	28	**21**	31

Table 9. Effect of different training/testing augmentations on the number of errors in a clean set of 10,000 samples selected after bootstrapping. Test performed on CIFAR10 Asym. 40% noise.

Training Strategy	No. of Errors
No Self-Supervision	369
With Self-Supervision	21

Table 10. Effect of self-supervision on the number of errors in a clean set of 10,000 samples selected after bootstrapping. Test performed on CIFAR10 Asym. 40% noise

As for training time on CIFAR10 problems, our method takes on average 13.8h for SimCLR pretraining, 0.5h for bootstrapping, 7.5h for SSL, and 2.5h for final training (total of 24.3h) on an Nvidia RTX 2080. In comparison, DivideMix [28] takes on average 5h, and the more recent method of PropMix [10] can take up to 10h.

5. Conclusions

In this paper, we proposed a new method that predicts labels from images and their noisy labels. Unlike other methods, our training procedure does not require access to a clean set of data, which we achieve by introducing bootstrapping and a careful noise-based balancing procedure. By utilising the relationship between images, noisy labels and 'clean' labels to accurately relabel samples, we find that we can achieve SOTA results. By simply changing the model used, we further unify the noisy label learning and semi-supervised learning domains, resulting in a simplified architecture that can improve performance for challenging instance-dependent noisy-label tasks. Additionally, we find that by randomly replacing noisy labels with 'null' labels during training, we can remove the need for model distillation, allowing practitioners of our method to perform predictions with and without noisy labels.

References

[1] 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, 2019.

[2] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

[3] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014.

[4] Dara Bahri, Heinrich Jiang, and Maya Gupta. Deep k-nn for noisy labels. In *International Conference on Machine Learning*, pages 540–550. PMLR, 2020.

[5] Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24392–24403, 2021.

[6] 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.

[7] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *International Conference on Machine Learning*, pages 1062–1070. PMLR, 2019.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[10] F. R. Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Propmix: Hard sample filtering and proportional mixup for learning with noisy labels. *The 32nd British Machine Vision Conference*, 2021.

[11] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[12] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc 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.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017.

[15] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

[16] Keren Gu, Xander Masotto, Vandana Bachani, Balaji Lakshminarayanan, Jack Nikodem, and Dong Yin. An instance-dependent simulation framework for learning with label noise, 2021.

[17] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[20] Naoto Inoue, Edgar Simo-Serra, Toshihiko Yamasaki, and Hiroshi Ishikawa. Multi-label fashion image classification with minimal human supervision. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2261–2267, 2017.

[21] Taehyeon Kim, Jongwoo Ko, JinHwan Choi, Se-Young Yun, et al. Fine samples for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24137–24149, 2021.

[22] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Canadian Institute for Advanced Research, 2009.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[24] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[26] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[27] Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. In *International Conference on Machine Learning*, pages 3763–3772. PMLR, 2019.

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

[29] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5051–5059, 2019.

[30] Junnan Li, Caiming Xiong, and Steven CH Hoi. Learn-

ing from noisy data with robust representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9485–9494, 2021.

[31] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

[32] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

[33] Ricardo V Lloyd, Lori A Erickson, Mary B Casey, King Y Lam, Christine M Lohse, Sylvia L Asa, John KC Chan, Ronald A DeLellis, H Ruben Harach, Kennichi Kakudo, et al. Observer variation in the diagnosis of follicular variant of papillary thyroid carcinoma. *The American journal of surgical pathology*, 28(10):1336–1340, 2004.

[34] Kento Nishi, Yi Ding, Alex Rich, and Tobias Hollerer. Augmentation strategies for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8022–8031, June 2021.

[35] Diego Ortego, Eric Arazo, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6606–6615, 2021.

[36] 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.

[37] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[39] Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *ICLR 2015*, 2015.

[40] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343. PMLR, 2018.

[41] 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 preprint arXiv:2103.11395*, 2021.

[42] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Noisy concurrent training for efficient learning under label noise. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3159–3168, 2021.

[43] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019.

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[45] 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.

[46] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: Refurbishing unclean samples for robust deep learning. In *ICML*, 2019.

[47] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[48] Antti Tarvainen and Harri 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.

[49] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*, pages 268–285. Springer, 2020.

[50] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 839–847, 2017.

[51] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[52] Zhen Wang, Guosheng Hu, and Qinghua Hu. Training noise-robust deep neural networks via meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4524–4533, 2020.

[53] Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. Ngc: A unified framework for learning with open-world noisy data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 62–71, 2021.

[54] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33, 2020.

[55] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points

really indispensable in label-noise learning? *Advances in Neural Information Processing Systems*, 32, 2019.

[56] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.

[57] Youjiang Xu, Linchao Zhu, Lu Jiang, and Yi Yang. Faster meta update strategy for noise-robust deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 144–153, 2021.

[58] Youjiang Xu, Linchao Zhu, Lu Jiang, and Yi Yang. Faster meta update strategy for noise-robust deep learning. In *CVPR*, 2021.

[59] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34, 2021.

[60] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[61] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[62] Le Zhang, Ryutaro Tanno, Kevin Bronik, Chen Jin, Parashkev Nachev, Frederik Barkhof, Olga Ciccarelli, and Daniel C Alexander. Learning to segment when experts disagree. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 179–190. Springer, 2020.

[63] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. In *International Conference on Learning Representations*, 2021.

[64] Zizhao Zhang and Tomas Pfister. Learning fast sample re-weighting without reward data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 725–734, 2021.

[65] Zizhao Zhang, Han Zhang, Sercan O Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9294–9303, 2020.

[66] Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alex M Bronstein, and Or Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1657–1667, 2022.