Contrast to Divide: Self-Supervised Pre-Training for Learning with Noisy Labels

Evgenii Zheltonozhskii[*1], Chaim Baskin[*1], Avi Mendelson[1], Alex M. Bronstein[1], Or Litany[2]

[1]Technion – Israel Institute of Technology, [2]NVIDIA

evgeniizh@campus.technion.ac.il; chaimbaskin@campus.technion.ac.il;

mendlson@technion.ac.il; bron@cs.technion.ac.il;

orlitany@gmail.com

Abstract

The success of learning with noisy labels (LNL) methods relies heavily on the success of a warm-up stage where standard supervised training is performed using the full (noisy) training set. In this paper, we identify a "warm-up obstacle": the inability of standard warm-up stages to train high quality feature extractors and avert memorization of noisy labels. We propose "Contrast to Divide" (C2D), a simple framework that solves this problem by pre-training the feature extractor in a self-supervised fashion. Using self-supervised pre-training boosts the performance of existing LNL approaches by drastically reducing the warm-up stage's susceptibility to noise level, shortening its duration, and improving extracted feature quality. C2D works out of the box with existing methods and demonstrates markedly improved performance, especially in the high noise regime, where we get a boost of more than 27% for CIFAR-100 with 90% noise over the previous state of the art. In real-life noise settings, C2D trained on mini-WebVision outperforms previous works both in WebVision and ImageNet validation sets by 3% top-1 accuracy. We perform an in-depth analysis of the framework, including investigating the performance of different pre-training approaches and estimating the effective upper bound of the LNL performance with semi-supervised learning. Code for reproducing our experiments is available at https://github.com/ContrastToDivide/C2D.

1. Introduction

Many deep-learning-based methods owe their success to the availability of large data sources with reliable labels. Quality annotation at scale, however, is often prohibitively expensive. Two common approaches that address this challenge are semi-supervised learning and learning with noisy labels (LNL). The former assumes the availability of a limited amount of high-quality labeled data as well as a large

amount of unlabeled data of the same distribution. The main challenge is to propagate the labels to the unlabeled samples to allow gleaning knowledge from them as well. In contrast, the latter approach suggests acquiring cheap annotations at scale at the cost of having some mislabeled data. Examples of such processes include web crawling [32, 55], automatic annotation based on meta-data [35], and uncurated crowdsourcing [27]. Though seemingly different, the two approaches are in fact closely related. Many semi-supervised learning approaches are based on predicting pseudo-labels for the unlabeled data, which are, effectively, noisy labels. From the other end, an LNL setting can be converted into a semi-supervised one by identifying and discarding the noisy labels. Separation of the noisy labeled samples from the clean ones is one of the key challenges in LNL.

To that end, multiple LNL methods utilize a "warm-up" stage [18, 29, 34, 42] – short supervised training on the full noisy dataset that precedes the more sophisticated algorithms designed to deal with label noise. During warm-up, the network's inherent robustness to noise tends to lower the classification loss of the cleanly labeled samples faster than that of the noisy ones [10, 34]. While this stage has not received much attention in previous works (possibly due to its algorithmic simplicity), it is in fact crucial to the success of LNL. Unfortunately, it is prone to memorizing noise, and thus its efficacy is contingent on the noise level as well as the amount of training iterations and other hyperparameters.

These limitations create a significant obstacle to improving the performance of LNL approaches. While supervised pre-training on a large clean dataset (e.g., ImageNet [44]) may seem to be a possible solution to this problem, the availability of such data may be limited in some domains (e.g., medical data). In addition, our experiments show that in some scenarios, ImageNet pre-training may degrade the performance of LNL algorithms.

We propose to overcome the warm-up obstacle by using *unsupervised* pre-training. Building on the recent success of self-supervised learning [7, 20, 50, 64], especially in closely related semi-supervised tasks [8], we generate high-quality
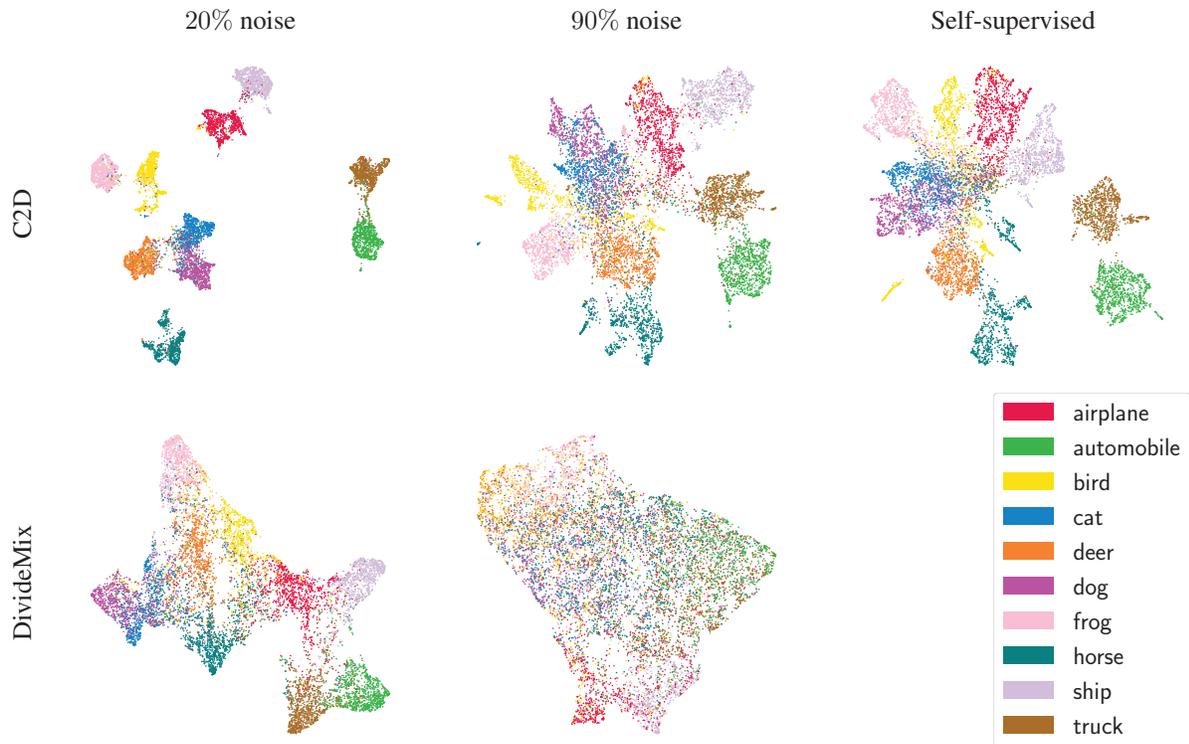

Figure 1: A UMAP [36] of features extracted from CIFAR-10 using C2D (top row) vs. DivideMix (bottom row) for 20% and 90% noise at the end of warm-up stage, as well as self-supervised pre-training. Colors indicate the ground-truth labels.

features by pre-training on the unlabeled train set samples. Thus, we benefit simultaneously from several effects. We do not require external data sources; by ignoring the labels, we eliminate the influence of noise on the pre-training stage and prevent noise memorization; finally, by operating on the training set, we avoid a domain gap. Importantly, this can be seamlessly combined with any LNL method.

Altogether, our framework provides a significant boost over previous LNL methods, with much better consistency across different noise levels. For example, with 90% symmetric noise, we achieve a more than 27% accuracy boost for CIFAR-100 with PreAct ResNet-18. In real-life settings, on mini-WebVision the proposed framework achieves an accuracy boost of more than 3% on top-1 accuracy both in WebVision and ImageNet validation sets; on Clothing1M it matches performance of the ImageNet pre-training without any external data.

Below, we outline our main contributions.

- First, we identify and characterize the warm-up importance for LNL. Our proposed framework, "Contrast to Divide" (C2D), improves warm-up performance by utilizing self-supervised pre-training.

- C2D significantly outperforms state-of-the-art methods on standard benchmarks that do not utilize external data:

CIFAR and WebVision. Moreover, on the challenging Clothing1M benchmark, C2D matches the state of the art that uses pre-training on ImageNet.

- We perform an extensive analysis of C2D, including loss separation and feature quality for different initialization schemes, loss distribution after the warm-up stage, and the performance gap between C2D and semi-supervised learning.

2. Related work

Self-supervised learning. Self-supervised learning aims to learn representations that are meaningful in some general sense, without using externally provided labels. Usually, this is done by solving a pretext task. One family of methods is based on reconstructing a corrupted version of the input [41, 52, 67, 68]. Other methods opt for using a classification task based on context prediction [11, 14, 25, 39] or clustering [5]. Nevertheless, all these methods impose an inherent problem when facing a particular downstream task that may not be well correlated with the self-supervised objective. Thus, there is no guarantee that the key information is retained and can be extracted from the features [37]. Some methods propose to remedy this problem by making the self-supervised

task aware of the downstream one [23, 65].

Recently, a revival in self-supervised techniques based on contrastive loss [16] has shown markedly improved performance in large-scale computer vision tasks [7, 20, 50, 58]. Subsequently, similar approaches without utilizing contrast between samples were proposed [9, 15, 64].

Semi-supervised learning. Given a partially labeled dataset, semi-supervised techniques aim at utilizing the unlabeled samples for boosting the learning procedure beyond what is achievable with just the labeled set. A simple yet efficient baseline for this problem is pseudo-labeling [2, 28, 57, 60]. In its basic form, this solution uses a network trained on the labeled subset to predict labels for the unlabeled set. These new labels, in turn, are used to refine the network (or a larger one) on the now fully labeled set. Another popular approach to semi-supervised learning is consistency regularization, where in addition to the cross-entropy loss, consistency is enforced between different perturbations of unlabeled (and possibly labeled) samples. Various implementations of those perturbation were studied, including predictions by different networks [49], adversarial examples [38], and augmentations [4, 12, 45, 56]. Recent methods have shown competitive results on CIFAR-100 using labels for as little as 1% of samples. Moreover, self-supervised pre-training can further improve semi-supervised classification [8].

Learning with noisy labels. There are many variants of the LNL problem. While some methods [33, 51, 71] assume the availability of a small subset of clean labels, we do not make those assumptions. We also consider closed-set noisy labels, i.e., where the mislabeled images belong to one of the training classes as opposed to the open-set setup [53, 70].

Existing methods for LNL can be divided into two broad categories: loss modification and noise detection. The former group includes techniques that account for noise distribution [42, 54, 61]. Alternatively, the loss itself may be replaced by a more robust version, such as mean absolute error [13], generalized cross-entropy [70], determinant-based mutual information [59], or a meta-learning objective [30]. On the other hand, noise detection methods aim to discover which samples are mislabeled to either relabel or discard them. Techniques for detecting noisy labels include utilizing multiple networks in a teacher-student [22] or mutual teaching [17, 63] framework, geometry [18], mixture models [1, 29], and quantiles of counterfactual loss distribution of samples [47]. These are often based on the observation that samples with noisy labels converge slower than those with clean ones [3, 10, 31, 43]. Hybrid methods that try to mix both noise detection and loss modification were also proposed [34, 46].
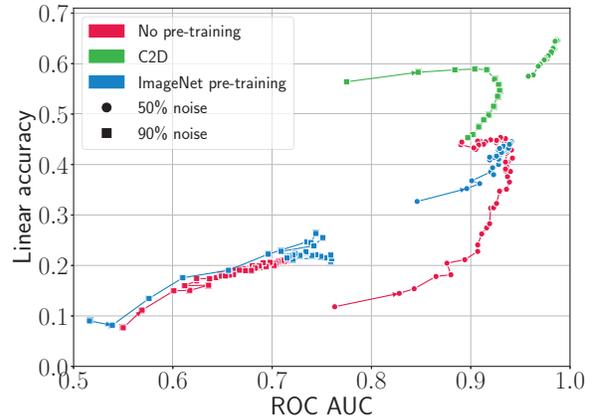

Figure 2: The ROC-AUC score of noise detection and the linear accuracy using clean labels under various noise levels for CIFAR-100 for standard warm-up, ImageNet pre-training, and C2D. Each point is one epoch of training, arrowhead denotes time direction. Colors denote the pre-training scheme and markers denote the noise level.

3. The warm-up obstacle

C2D is motivated by the observation of an inherent obstacle that is at the core of LNL methods. It has been shown that deep networks can perform meaningful learning in the presence of noise before they enter a memorization phase [43]. LNL methods utilize this behavior by performing a warm-up – supervised training on the full set of (noisy) labels for a short period of time. Most methods utilize either "hard" (starting an LNL procedure after a number of epochs [29]) or "soft" (gradually increasing the weight of additional regularization terms [34]) version of warm-up.

A warm-up stage has two main goals: loss separability and feature extraction. The former means that the model is still in the early learning phase, allowing the follow up stage to rely on noisy labeled samples having high loss values clean labeled ones having low loss values. The latter refers to the quality of representation learned by the model. Little, however, is understood about the *determinants* of network robustness to noise or how to boost it. As a result, the warm-up performance in LNL methods is consequential and bounded by an unavoidable memorization. Current practice in LNL is to merely adjust the warm-up length according to the observed robustness of the model under different noise levels. We identify this as a major obstacle in the ability to improve performance.

To demonstrate this phenomenon, we run a supervised training on CIFAR-100 with noise and measured the level of the aforementioned properties at each epoch. The results are visualized in Fig. 2. Separation is measured as the ROC AUC

Method	Architecture		Noise rate					
			20%	50%	80%	90%	95%	Asym. 40%
Meta-learning	PreAct ResNet-32	Peak	92.9	89.3	77.4	58.7	–	89.2
[30]		Final	92.0	88.8	76.1	58.3	–	88.6
ELR+	ResNet-34	Peak	–	–	–	–	–	–
[34]		Final	95.8	94.8	93.3	78.7	–	93.0
DivideMix	PreAct ResNet-18	Peak	96.1	94.6	93.2	76.0	–	93.4
[29]		Final	95.7	94.4	92.9	75.4	–	92.1
CE+mixup with SimCLR	ResNet-34	Peak	$94.98_{\pm 0.17}$	$93.83_{\pm 0.12}$	$91.33_{\pm 0.17}$	$87.54_{\pm 0.45}$	$73.97_{\pm 1.22}$	$91.63_{\pm 0.38}$
		Final	$94.56_{\pm 0.22}$	$93.20_{\pm 0.25}$	$90.77_{\pm 0.57}$	$85.77_{\pm 2.42}$	$72.71_{\pm 1.39}$	$90.39_{\pm 0.35}$
C2D (ELR+ with SimCLR)	ResNet-34	Peak	$\mathbf{96.83}_{\pm 0.10}$	$\mathbf{95.96}_{\pm 0.09}$	$93.67_{\pm 0.16}$	$89.94_{\pm 0.20}$	$83.23_{\pm 0.68}$	$\mathbf{94.32}_{\pm 0.48}$
		Final	$\mathbf{96.74}_{\pm 0.12}$	$\mathbf{95.55}_{\pm 0.32}$	$93.11_{\pm 0.70}$	$89.30_{\pm 0.21}$	$80.21_{\pm 1.91}$	$\mathbf{93.78}_{\pm 0.91}$
C2D (DivideMix with SimCLR)	PreAct ResNet-18	Peak	$96.43_{\pm 0.07}$	$95.32_{\pm 0.12}$	$\mathbf{94.40}_{\pm 0.04}$	$\mathbf{93.57}_{\pm 0.09}$	$\mathbf{89.24}_{\pm 0.75}$	$93.45_{\pm 0.07}$
		Final	$96.23_{\pm 0.09}$	$95.15_{\pm 0.16}$	$\mathbf{94.30}_{\pm 0.12}$	$\mathbf{93.42}_{\pm 0.09}$	$\mathbf{87.72}_{\pm 2.21}$	$90.75_{\pm 0.35}$

Table 1: Classification accuracy (%, mean ± std over five runs) on CIFAR-10. C2D achieves consistently high accuracy under different noise rates and types, with markedly improved performance under very-high noise conditions. Meta-learning results provided by Li et al. [29].

of the noise detection with Gaussian mixture model (GMM) applied to loss values [29], and feature quality is measured using the classification accuracy of a linear classifier trained without noise – a standard approach for feature quality assessment [7, 20, 67]. As can be seen, the two measures are strongly correlated, peaking jointly at some iteration, which, of course, is unknown unless clean labels are available. Critically, not only do the separation and feature quality values deteriorate quickly as the noise level increases, but also no known remedy exists. In other words, even if we knew the *optimal* warm-up length, the current LNL toolbox lacks the tools to improve the observed values. The effect of feature deterioration as the noise level increases can also be seen in the bottom part of Fig. 1, where we visualized the extracted features after the warm-up using UMAP [36].

To circumvent the deterioration issue, prior works have resorted to supervised pre-training on an external dataset, such as ImageNet. This solution suffers from two disadvantages. First, it necessitates a large, cleanly-labeled dataset of a similar domain, which is not always available. Second, as will be made apparent by our analyses, features generated via supervised pre-training may fall short in noisy label separation. Our solution to overcoming the warm-up obstacle is to use self-supervised pre-training.

4. Contrast to Divide

As discussed in Section 3, the warm-up phase performance in LNL pipelines is bounded by memorization. Encouraged by the recent success in semi-supervised learning [8], we study whether self-supervised pre-training could break this barrier. More specifically, given a dataset with contaminated labels, we propose a straight-forward two-phase framework. First, we perform self-supervised contrastive learning [7, 64] to obtain high-quality feature extractor (contrast phase). We then proceed with a standard LNL algorithm that can now better detect noisy labels (divide phase). Much like standard transfer learning, this framework can be used to boost virtually any existing LNL method. We do not, however, rely on a cleanly labeled external data source; pre-training is done directly on the target training set. Importantly, by discarding the labels, we avoid label noise influence on feature extractor and provide a robust initialization as can be seen in Fig. 2. Even under extreme noise level conditions, this initialization boosts the warm-up far beyond the memorization bound. As shown in the experimental section, the improved loss separation supports both explicit separation using a classification model as well as an implicit one based on regularization terms.

5. Experimental results

Our evaluation of the proposed framework uses two state-of-the-art LNL methods: ELR+ [34] and DivideMix [29], both on synthetic and real noise. We follow common practice in synthetic noise benchmarks and use CIFAR-10 and CIFAR-100 [26], varying the amount of injected noise. For the real noise setting, we use WebVision [32], a dataset of ~2.4 million images based on queries generated from the 1,000 ImageNet [44] classes, and Clothing1M [55], which contains ~1 million images of 14 classes of clothing. Both datasets are acquired by web crawling.

Common evaluation using Clothing1M includes utilizing the ImageNet pre-trained network, while for CIFAR and WebVision, networks are trained from scratch. Thus, for the former we provide a comparison with supervised pre-

Method	Architecture		Noise rate					
			20%	50%	80%	90%	95%	Asym. 40%
Meta-learning [30]	PreAct ResNet-32	Peak	68.5	59.2	42.4	19.5	–	–
		Final	67.7	58.0	40.1	14.3	–	–
ELR+ [34]	ResNet-34	Peak	–	–	–	–	–	–
		Final	77.6	73.6	60.8	33.4	–	77.5
ODD [47]	WRN-28-10	Peak	$79.1_{\pm 0.1}$	–	–	–	–	–
		Final	–	–	–	–	–	–
DivideMix [29]	PreAct ResNet-18	Peak	77.3	74.6	61.6^*	31.5	–	72.2^*
		Final	76.9	74.2	61.3^*	31.0	–	72.4^*
CE+mixup with SimCLR	ResNet-34	Peak	$76.46_{\pm 0.15}$	$69.14_{\pm 0.31}$	$61.39_{\pm 0.26}$	$55.51_{\pm 0.24}$	$43.59_{\pm 0.59}$	$65.19_{\pm 0.63}$
		Final	$76.34_{\pm 0.19}$	$67.97_{\pm 1.22}$	$60.81_{\pm 0.67}$	$54.64_{\pm 0.72}$	$42.11_{\pm 1.99}$	$54.75_{\pm 0.93}$
C2D (ELR+ with SimCLR)	ResNet-34	Peak	$79.18_{\pm 0.19}$	$76.33_{\pm 0.31}$	$64.72_{\pm 0.18}$	$55.08_{\pm 0.32}$	$44.06_{\pm 0.84}$	$\mathbf{77.87_{\pm 0.29}}$
		Final	$79.03_{\pm 0.20}$	$76.10_{\pm 0.36}$	$64.18_{\pm 0.13}$	$54.06_{\pm 1.50}$	$42.60_{\pm 1.67}$	$\mathbf{77.63_{\pm 0.27}}$
C2D (DivideMix with SimCLR)	PreAct ResNet-18	Peak	$78.69_{\pm 0.17}$	$76.43_{\pm 0.25}$	$67.78_{\pm 0.30}$	$58.70_{\pm 0.31}$	$38.89_{\pm 1.19}$	$75.48_{\pm 0.16}$
		Final	$78.32_{\pm 0.35}$	$76.07_{\pm 0.41}$	$67.43_{\pm 0.30}$	$58.45_{\pm 0.30}$	$38.03_{\pm 2.13}$	$75.06_{\pm 0.16}$
C2D (DivideMix with SimCLR)	ResNet-50	Peak	**81.60**	**79.54**	**71.65**	**64.30**	**49.11**	**77.92**
		Final	**80.89**	**79.20**	**71.53**	**63.91**	**48.50**	**77.78**

Table 2: Peak and final classification accuracies (%, mean $\pm$ std over five runs) on CIFAR-100. Unlike previous methods that suffer from rapid degradation, C2D was able to maintain good performance even under severe noise. Meta-learning results provided by Li et al. [29]. * denotes results acquired by us based on published code.

training, while for the latter we compare self-supervised pre-training with no pre-training whatsoever.

5.1. CIFAR-10 and CIFAR-100

We conducted experiments with two types of label noise: symmetric and asymmetric. Symmetric noise is generated by randomly replacing the labels in a percentage of the training data with a random label drawn from a uniform distribution over all labels. Following the common approach [1, 29, 66], the new label may be the real one. In this way, we are guaranteed that the clean label is the most frequent label for each class for any noise level. Thus, the real number of mislabeled examples is smaller by $1/n_{cl}$. Asymmetric noise is designed to mimic the structure of real-world label errors, where classes that are generally similar in appearance are more likely to switch labels. In this case, we follow a scheme proposed by Patrini et al. [42].

Implementation details. For both methods, we followed the setup of the original experiments as close as possible. We also used the original architectures, PreAct ResNet-18 [19] for DivideMix and ResNet-34 for ELR+. Since self-supervised training is known to benefit from increased network capacity [7, 20], for CIFAR-100 we performed experiments with ResNet-50 as well. For self-supervised pre-training, we used a SimCLR implementation[1] in PyTorch

[40]. The self-supervised model was trained for 1000 epochs on 4 NVIDIA 2080 Ti GPUs.

ELR+ required no hyperparameter tweaking. For DivideMix, we performed a number of minor modifications: (a) to accommodate ResNet-50 in GPU memory, we reduced the batch size from 128 to 64 and (b) for DivideMix we observed that our network kept improving after 300 epochs and thus increased training length to 360 epochs. In addition, we tuned the hyperparameters mentioned in the original paper [29]: the unlabeled loss weight $\lambda_{\mathcal{U}}$, the number of warm-up epochs, and the threshold for noisy label prediction τ. For $\lambda_{\mathcal{U}}$, we acquired similar results with and without C2D. Those results match the results of Li et al. [29], except that increasing $\lambda_{\mathcal{U}}$ also benefits the baseline DivideMix solution in high noise settings: for CIFAR-100 with 80% noise, increasing $\lambda_{\mathcal{U}}$ from 150 to 500 improved DivideMix accuracy from 60.2% to 61.3%. As discussed in Section 3, strong pretrained features are expected to reduce the required warm-up duration. We found that five epochs were sufficient for CIFAR at all noise levels. As a reference, DivideMix uses 10 epochs for CIFAR-10 and 30 epochs for CIFAR-100. Lastly, we set the GMM threshold to $\tau = 0.03$, which is significantly lower than the 0.5 used by DivideMix. This can be explained by the fact our model is able to determine most of the noisy examples with high confidence.

Results. Table 1 presents the comparison of our method with prior state of the art for symmetric and asymmetric

noisy labels on the CIFAR-10 dataset. "final" refers to the accuracy at the end of training for DivideMix, and the one with highest internal validation score for ELR+ as done in the original papers. "peak" refers to the highest validation score achieved during training.

In addition to maintaining consistently high classification accuracy across all noise levels, C2D significantly outperforms prior methods at high noise levels ($\geqslant 80\%$). We attribute this desired behavior to the fact that our pre-trained features are agnostic to the noise level.

Table 2 shows the classification accuracy on CIFAR-100. Compared with CIFAR-10, this task is more complex, resulting in a steeper drop in performance of prior methods as noise rates increase. In contrast, C2D demonstrates a graceful degradation, achieving a remarkable gain of more than 30% in accuracy at 90% noise level. We therefore decided to stress test C2D by subjecting it to an extreme noise level of 95%. Despite a higher variance in the results (measured across five noise realizations), C2D still achieved a final accuracy of above 38% (and at least 30% in each individual run), surpassing the performance achieved by previous approaches at a noise rate of 90%. In asymmetric noise, C2D performed similarly to prior art with ResNet-18, and achieved a minor improvement over ELR+ [34] with larger networks (ResNet-34 and ResNet-50).

We also provide an additional baseline which uses only first stage of C2D, i.e., self-supervised pre-training followed by vanilla cross-entropy training with mixup. For harder tasks ($\geqslant 90\%$ noise on CIFAR-100) the improvement provided by second stage (ELR+ training) is marginal, while for intermediate noise rates the gain is maximal (e.g., 8% difference for CIFAR-100 with 50% noise and CIFAR-10 with 95% noise).

5.2. Clothing1M

We tested our framework on the real-life noise present in the Clothing1M dataset [55]. As some of the manually labeled images have both clean and noisy labels, we can estimate the noise level as approximately 38.5%. We also use these double-labeled samples to compute noise-related metrics such as the ROC AUC of noise detection. Implementation details are specified in the appendix.

Results. The default approach [18, 29, 30, 34, 42, 48, 62, 69] for Clothing1M is to leverage a ResNet-50 pre-trained on ImageNet. The rich variety of visual concepts along with high-quality labels provides a strong initialization for network weights. C2D, on the other hand, uses only the dataset itself for pre-training. A comparison with state-of-the-art methods is reported in Table 3. The results highlight two interesting phenomena. First, by comparing the performance of the standard cross-entropy training, we confirm that our self-supervised pre-training is significantly better than Im-

Method	Test accuracy
Cross-entropy	69.21
F-correction [42]	69.84
Joint-Optim [48]	72.16
MetaCleaner [69]	72.50
Meta-Learning [30]	73.47
P-correction [62]	73.49
Self-learning [18]	74.45
DivideMix [29]	74.76
ELR+ [34]	**74.81**
Cross-entropy (SimCLR init.)	72.05
Cross-entropy (BT init.)	73.03
C2D (ELR+ with BT)	73.52
C2D (DivideMix with SimCLR)	74.30
C2D (ELR+ with SimCLR)	**74.58** $\pm$ **0.15**

Table 3: Comparison with state-of-the-art methods in test accuracy (%) on Clothing1M. The upper part of the table uses ImageNet pre-training, while the lower half does not.

ageNet pre-training, demonstrating the warm-up gain C2D brings. A second observation is that, similarly to harder instances of CIFAR, this advantage is not leveraged by the LNL methods, resulting in an overall performance similar to ImageNet pre-training. This may be attributed to the complicated noise structure and leaves room for research of the way the methods utilize the improved initialization. Finally, it is encouraging that nearly state-of-the-art results (74.58% vs. 74.81% accuracy) can be achieved without external data.

5.3. WebVision

Following previous work [6, 21, 29, 34, 47], we evaluate our framework on the mini-WebVision 1.0 dataset [32], which contains the first 50 classes of the Google image subset for a total of ~61,000 images. Implementation details are specified in the appendix.

Results As shown in Table 4, C2D outperforms previous works on both the WebVision and ImageNet validation sets by at least 3% top-1 accuracy. Since we used a different network architecture, we also evaluated vanilla DivideMix with ResNet-50, reaching ~1% degradation of top-1 accuracy when compared to Inception-ResNet-v2.

6. Analysis

6.1. Warm-up performance

In Section 3, we defined low feature quality and poor loss separation as a major obstacle to improving LNL perfor-

Method	Architecture	ILSVRC12		WebVision	
		Top-1	Top-5	Top-1	Top-5
MentorNet [22]	Inception-ResNet-v2	63.8	85.8	–	–
Iterative-CV [6]	Inception-ResNet-v2	61.6	85.0	65.2	85.3
MentorMix [21]	Inception-ResNet-v2	72.9	91.1	76.0	90.2
DivideMix [29]	Inception-ResNet-v2	75.20	90.84	77.32	91.64
ODD [47]	Inception-ResNet-v2	66.7	86.3	74.6	90.6
ELR+ [34]	Inception-ResNet-v2	70.29	89.76	77.78	91.68
DivideMix* [29]	ResNet-50	74.42 ± 0.29	91.21 ± 0.12	76.32 ± 0.36	90.65 ± 0.16
C2D (DivideMix with SimCLR)	ResNet-50	$\mathbf{78.57} \pm \mathbf{0.37}$	$\mathbf{93.04} \pm \mathbf{0.10}$	$\mathbf{79.42} \pm \mathbf{0.34}$	$\mathbf{92.32} \pm \mathbf{0.33}$

Table 4: Accuracy (%, mean $\pm$ std over five runs) on the WebVision validation set and the ILSVRC12 (ImageNet) validation sets, for the networks trained on (mini) WebVision dataset. * denotes results acquired by us based on published code.

mance. Here, we show that C2D improves both metrics. We start with a qualitative comparison of the features learned on the CIFAR-10 data with and without C2D at the end of the warm-up. We do this by visualizing the features in Fig. 1 using a dimensionality-reduction technique [36]. C2D features (upper row) are clearly better clustered and easier to separate than the baseline (bottom row) at both 20% and 90% noise levels. Furthermore, at high noise rates, the baseline features suffer from acute degradation (note the blending of the dog and deer categories into their surrounding classes), while C2D features maintain some fidelity.

To evaluate the impact of different pre-training approaches (a standard warm-up, ImageNet pre-training, and C2D) on the performance of the warm-up stage, in Fig. 2 we visualize the properties defined in Section 3: ROC-AUC score of noise detection and the linear accuracy using clean labels under different noise levels for CIFAR-100. C2D demonstrates significantly higher overall feature quality and loss separability at all noise levels: even after one epoch both values are higher than the peak ones for vanilla warm-up and ImageNet pre-training. Moreover, thanks to the pre-training on the same dataset and discarding the labels, C2D shows faster convergence and smaller performance degradation with increased noise. In addition, in Fig. 3 we visualize a histogram of loss values of clean and noisy samples for all three approaches. It can be seen that C2D does indeed provide a lower loss for most clean samples with the smallest overlap between clean and noisy samples, which allows both faster convergence and higher overall performance.

6.2. The drawbacks of supervised pre-training

Utilizing supervised pre-trained features from large cleanly labeled data source (e.g., ImageNet [24]) is standard procedure in noise-free learning tasks. In the context of LNL, much less is known about ImageNet pre-training benefits besides its high performance on Clothing1M. Therefore,

we chose to apply ImageNet pre-training to noisy CIFAR.

We ran ELR+ and DivideMix on CIFAR-100 with a network initialized with ImageNet pre-trained weights. One may expect that a small domain gap along with versatile high-quality pre-trained features will make this an almost ideal setup. Indeed, Fig. 2 shows, at least for 90% noise level, notable improvement in the warm-up phase, which is, however, far inferior to that of C2D.

Surprisingly, these improvements did not result in an improvement of the performance of LNL. On ELR+ [34], adding ImageNet pre-training reduced the accuracy from 60.8% to 48.58 ± 0.16% for CIFAR-100 with 80% noise and from 33.4% to 23% with 90% noise. On DivideMix, in addition to an expected shortening of the required warm-up length (from 30 to 10 epochs), at the end of the warm-up, on 80% noise we observed an increase both in the ROC-AUC score and the classification accuracy. Yet, most concerning was the almost immediate failure of DivideMix when entering the second stage of training. After the warm-up, the loss values of the clean and noisy samples were almost indistinguishable, which resulted in a severe decrease in classification accuracy as depicted in Fig. 3. Despite our attempts to rectify this behavior, this phenomenon persisted across various sets of hyperparameters.

Our results indicate that supervised pre-training has a generally unpredictable effect on LNL. We leave the influence analysis of different conditions (e.g. domain gap or noise level) to future work. On the contrary, using C2D resulted in consistent improvement across all our experimental settings. In addition, C2D does not require external data nor additional supervision.

6.3. Self-supervised pre-training method

Further verifying the universality of our approach, we examine how the self-supervised pre-training approach affects the performance of C2D. In addition to SimCLR, we
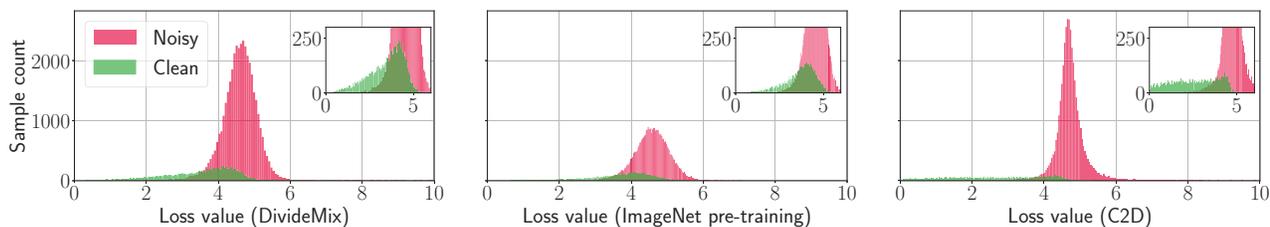
Figure 3: Loss distribution of clean and noisy samples after warm-up on CIFAR-100 with 80% noise for DivideMix, DivideMix with ImageNet pre-training, and C2D. As seen in the zoom-in, ImageNet pre-training damages the separability whereas self-supervised pre-training (C2D), improves it.

Method	Missing/noisy label rate	
	80%	90%
MixMatch	70.46	64.60
MixMatch (SimCLR init.)	71.86	66.10
C2D (DivideMix with SimCLR)	71.65	64.30

Table 5: C2D nearly closes the gap with semi-supervised training on the same clean set size.

used Barlow Twins [64], due to its high performance and the distinct differences with SimCLR. The results are presented in Table 3. Using cross-entropy fine-tuning (without accounting for noise) Barlow Twins outperforms SimCLR, achieving only 1.5% below the state-of-the-art. While this advantage did not translate into significantly better overall performance (0.5% improvement when using ELR+), high overall performance without supervised pretraining indicates that C2D can work well with other self-supervised feature extraction techniques.

6.4. Gap between LNL and semi-supervised learning

In the case of DivideMix [29] and other methods that utilize semi-supervised learning, semi-supervised accuracy is effectively the upper bound on the performance. A significantly better noise separation ability along with the improved initialization raises the question whether any performance gap remains between LNL and semi-supervised learning. To answer this question, we compared the performance of C2D (with DivideMix) with MixMatch – a semi-supervised method – provided with the same amount of labels as the clean portion of the C2D training set. This procedure is roughly equivalent to replacing the noise detection procedure with an oracle. The result for 80% and 90% noise levels in CIFAR-100 are reported in Table 5. Remarkably, C2D is on par with MixMatch and less than 2% below MixMatch with

self-supervised pre-training. Even though the LNL setup has *strictly less information* than the semi-supervised one, these results indicate that good features can compensate for this lack of information even under severe noise conditions.

7. Conclusion

In this paper, we proposed Contrast to Divide (C2D), a simple yet powerful framework for learning with noisy labels that do not rely on external labeled data and leverages self-supervised pre-training instead. We have identified and analyzed a major obstacle to LNL: due to memorization, loss separability and feature quality after warm-up are bounded, and deteriorate quickly with increasing noise level. Moreover, while the natural robustness of the neural networks allows us to acquire good results, little is known about the sources of this robustness and how to improve it. We have shown that self-supervised pre-training boosts both warm-up goals, which in turn dramatically improves the performance of existing LNL approaches.

C2D is straightforward to implement, does not require any external data, and works out of the box with multiple existing LNL approaches, demonstrating consistently high performance across various noise levels. In real-life settings, we tested C2D on mini-WebVision and achieved more than a 3% top-1 accuracy boost over the previous state of the art. In addition, C2D shows stable performance under severe noise, outperforming prior art by more than 20% for 90% noise on CIFAR-100 and nearly closing the gap with semi-supervised learning trained on the same amount of labeled samples as the clean portion.

Even though C2D provides significant performance improvement, studying the robustness property is still an open research question. Clothing1M results also suggest that the way existing methods utilize pre-trained features can be improved too. We leave those questions for the future work.

References

1. Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin Mcguinness. Unsupervised label noise modeling and

loss correction. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 312–321, Long Beach, California, USA, 09–15 Jun 2019. PMLR. (cited on pp. 3 and 5)

2. Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. *arXiv preprint arXiv:1908.02983*, 2019. (cited on p. 3)

3. Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. (cited on p. 3)

4. David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A. Raffel. MixMatch: A holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5049–5059. Curran Associates, Inc., 2019. (cited on p. 3)

5. Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. (cited on p. 2)

6. Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1062–1070, Long Beach, California, USA, 09–15 Jun 2019. PMLR. (cited on pp. 6 and 7)

7. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. (cited on pp. 1, 3, 4, and 5)

8. Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. (cited on pp. 1, 3, and 4)

9. Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. (cited on p. 3)

10. Safa Cicek, Alhussein Fawzi, and Stefano Soatto. SaaS: Speed as a supervisor for semi-supervised learning. In *The European Conference on Computer Vision (ECCV)*, September 2018. (cited on pp. 1 and 3)

11. Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. (cited on p. 2)

12. Geoff French, Avital Oliver, and Tim Salimans. Milking CowMask for semi-supervised image classification. *arXiv preprint arXiv:2003.12022*, 2020. (cited on p. 3)

13. Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 1919–1925. AAAI Press, 2017. (cited on p. 3)

14. Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. (cited on p. 2)

15. Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. (cited on p. 3)

16. Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. (cited on p. 3)

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. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8527–8537. Curran Associates, Inc., 2018. (cited on p. 3)

18. Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. (cited on pp. 1, 3, and 6)

19. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. (cited on p. 5)

20. Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aäron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019. (cited on pp. 1, 3, 4, and 5)

21. Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. *arXiv preprint arXiv:1911.09781*, 2019. (cited on pp. 6 and 7)

22. 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 Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2304–2313, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. (cited on pp. 3 and 7)

23. Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. (cited on p. 3)

24. Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 2019. (cited on p. 7)

25. Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. (cited on p. 2)

26. Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009. (cited on p. 4)

27. Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, pages 1–26, 2020. (cited on p. 1)

28. 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. (cited on p. 3)

29. Junnan Li, Richard Socher, and Steven C.H. Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020. (cited on pp. 1, 3, 4, 5, 6, 7, and 8)

30. Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5051–5059, 2019. (cited on pp. 3, 4, 5, and 6)

31. Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *arXiv preprint arXiv:1903.11680*, 2019. (cited on p. 3)

32. 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. (cited on pp. 1, 4, and 6)

33. Or Litany and Daniel Freedman. Soseleto: A unified approach to transfer learning and training with noisy labels. *arXiv preprint arXiv:1805.09622*, 2018. (cited on p. 3)

34. Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *arXiv preprint arXiv:2007.00151*, 2020. (cited on pp. 1, 3, 4, 5, 6, and 7)

35. Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. (cited on p. 1)

36. Leland McInnes, John Healy, and James Melville. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. (cited on pp. 2, 4, and 7)

37. Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. (cited on p. 2)

38. Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *arXiv preprint arXiv:1704.03976*, 2017. (cited on p. 3)

39. Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 69–84, Cham, 2016. Springer International Publishing. (cited on p. 2)

40. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: an imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc., 2019. (cited on p. 5)

41. Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. (cited on p. 2)

42. Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: a loss correction approach. *arXiv preprint arXiv:1609.03683*, 2016. (cited on pp. 1, 3, 5, and 6)

43. Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. *arXiv preprint arXiv:2001.10528*, 2020. (cited on p. 3)

44. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. (cited on pp. 1 and 4)

45. Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. (cited on p. 3)

46. Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: Refurbishing unclean samples for robust deep learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5907–5915, Long Beach, California, USA, 09–15 Jun 2019. PMLR. (cited on p. 3)

47. Jiaming Song, Lunjia Hu, Michael Auli, Yann Dauphin, and Tengyu Ma. Robust and on-the-fly dataset denoising for image classification. *arXiv preprint arXiv:2003.10647*, 2020. (cited on pp. 3, 5, 6, and 7)

48. Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5552–5560, 2018. (cited on p. 6)

49. Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. V.

Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1195–1204. Curran Associates, Inc., 2017. (cited on p. 3)

50. Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020. (cited on pp. 1 and 3)

51. 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 (CVPR)*, July 2017. (cited on p. 3)

52. Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery. (cited on p. 2)

53. Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. (cited on p. 3)

54. Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6838–6849. Curran Associates, Inc., 2019. (cited on p. 3)

55. Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. (cited on pp. 1, 4, and 6)

56. Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019. (cited on p. 3)

57. Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*, 2019. (cited on p. 3)

58. Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. PointContrast: unsupervised pre-training for 3d point cloud understanding. *arXiv preprint arXiv:2007.10985*, 2020. (cited on p. 3)

59. Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_DMI: a novel information-theoretic loss function for training deep nets robust to label noise. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6225–6236. Curran Associates, Inc., 2019. (cited on p. 3)

60. I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. (cited on p. 3)

61. Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *arXiv preprint arXiv:2006.07805*, 2020. (cited on p. 3)

62. Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. (cited on p. 6)

63. Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7164–7173, Long Beach, California, USA, 09–15 Jun 2019. PMLR. (cited on p. 3)

64. Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. (cited on pp. 1, 3, 4, and 8)

65. Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S^{4}L: self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. (cited on p. 3)

66. Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016. (cited on p. 5)

67. Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 649–666, Cham, 2016. Springer International Publishing. (cited on pp. 2 and 4)

68. Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. (cited on p. 2)

69. Weihe Zhang, Yali Wang, and Yu Qiao. MetaCleaner: learning to hallucinate clean representations for noisy-labeled visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. (cited on p. 6)

70. Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8778–8788. Curran Associates, Inc., 2018. (cited on p. 3)

71. Zizhao Zhang, Han Zhang, Sercan O. Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. *arXiv preprint arXiv:1910.00701*, 2019. (cited on p. 3)