CONTRAST TO DIVIDE: SELF-SUPERVISED PRE-TRAINING FOR LEARNING WITH NOISY LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Advances in semi-supervised methods for image classification significantly boosted performance in the learning with noisy labels (LNL) task. Specifically, by discarding the erroneous labels (and keeping the samples), the LNL task becomes a semi-supervised one for which powerful tools exist. Identifying the noisy samples, however, heavily relies on the success of a warm-up stage where standard supervised training is performed using the full (noisy) training set. This stage is sensitive not only to the noise level but also to the choice of hyperparameters. In this paper, we propose to solve this problem by utilizing self-supervised pre-training. Our approach, which we name *Contrast to Divide*, offers several important advantages. First, by removing the labels altogether, our pre-trained features become agnostic to the labels' amount of noise, allowing accurate noisy separation even under high noise levels. Second, as recently shown, semi-supervised methods significantly benefit from self-supervised pre-training. Moreover, compared with standard pre-training approaches (e.g., supervised training on ImageNet), self-supervised pre-training does not suffer from a domain gap. We demonstrate the effectiveness of the proposed method in various settings with both synthetic and real noise. Our results indicate that Contrast to Divide brings a new state-of-the-art by a significant margin to both CIFAR-10 and CIFAR-100. For example, in the high-noise regime of 90%, we get a boost of more than 27% for CIFAR-100 and more than 17% for CIFAR-10 over the previous state-of-the-art. Moreover, we achieve comparable performance on Clothing-1M without using ImageNet pre-training. 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. This is especially true in cases when the annotation requires domain expertise such as medical training. 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 a large portion of mislabeled data. Examples of such processes include web crawling (Xiao et al., 2015; Li et al., 2017), automatic annotation based on meta-data (Mahajan et al., 2018), and un-curated crowdsourcing (Kuznetsova et al., 2020). 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 can be seen as noisy labels. From the other end, converting an LNL setting to a semi-supervised one can be done by identifying and discarding the noisy labels.

Based on these insights, a recently introduced LNL method named DivideMix (Li et al., 2020) has achieved impressive results. Specifically, DivideMix addresses the LNL problem as a semi-supervised one by (a) identifying the samples with noisy labels and (b) learning from the resulting partially labeled dataset. These two procedures alternate repeatedly, benefiting each other. Thanks to the

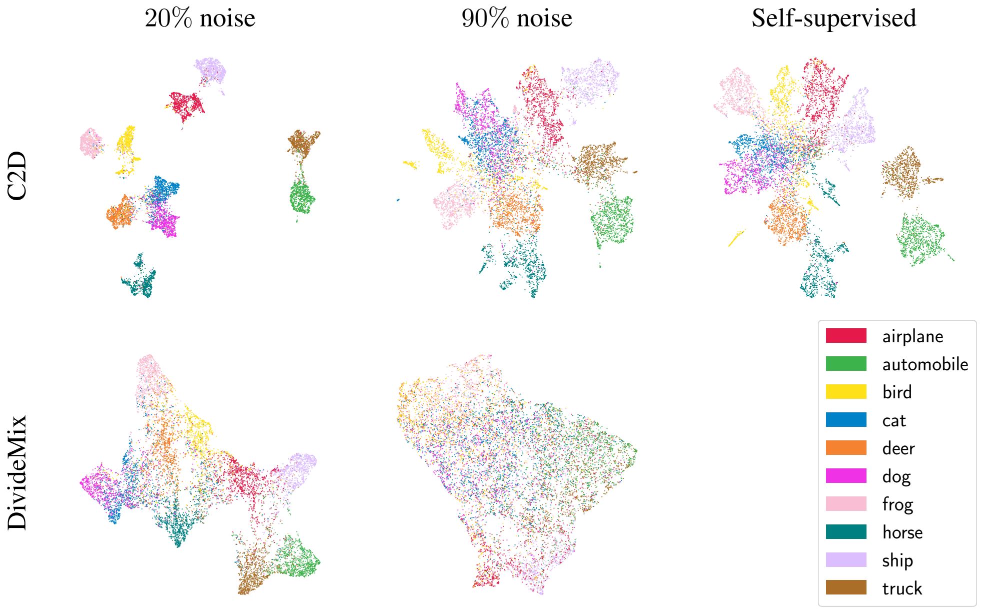

Figure 1: A UMAP (McInnes et al., 2018) of features extracted from CIFAR-10 using C2D (upper row) vs. DivideMix (lower row) for 20% and 90% noise at the end of warm-up stage, as well as self-supervised pre-training. Colors show ground-truth label.

powerful semi-supervised methods, the second stage can be solved efficiently. Yet, to achieve good separation, DivideMix relies on a warm-up stage where standard supervised training is performed on the full noisy dataset. As such, this stage is sensitive to training time, prone to overfitting to noise, and may generate inferior representations. To overcome this, some researchers have suggested using *supervised* pre-trained features (Xiao et al., 2015; Jiang et al., 2019). However, as will be seen later, it may bring little performance boost.

Instead, we propose to solve this issue by using *unsupervised* pre-training. Building on the recent success of contrastive learning (Hénaff et al., 2019; Chen et al., 2020a; Tian et al., 2020), we generate high-quality pre-trained features by training on the unlabeled train set samples. Thus, we benefit simultaneously from three effects: by ignoring the labels, we eliminate noise influence on the pre-training stage; we utilize self-supervised pre-training, which has been shown to enhance semi-supervised training (Chen et al., 2020b); and finally, by operating on the training set, we avoid a domain gap. Combining contrastive learning with DivideMix, we observe better noise detection and superior initialization for the semi-supervised stage. Altogether, we achieve a significant boost over DivideMix, with much better consistency across different noise levels. In particular, our approach stands out at high noise rates. For example, with 90% symmetric noise, we achieve a more than 27% accuracy boost for CIFAR-100 and more than 17% for CIFAR-10 with PreAct ResNet-18.

Below, we outline our main contributions and the organization of the remaining sections.

- We present "Contrast to Divide" (C2D), an LNL algorithm that utilizes pre-trained self-supervised features for learning with noisy labels.
- Building on the success of DivideMix, we demonstrate that C2D improves both of its components: noise detection and semi-supervised learning, making the warm-up stage 2–6× shorter and more robust to different noise rates.
- C2D significantly outperforms state-of-the-art results on standard benchmarks across various noise levels. Remarkably, on CIFAR-100 with up to 90% noise, C2D almost matches the performance of the equivalent *semi-supervised* method trained without noise. We provide an ablation study and qualitative analysis of the effect of the pre-trained features.

2 Related work

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 (Lee, 2013; Arazo et al., 2019b; Xie et al., 2019b; Yalniz et al., 2019). In its basic form, this solution uses a network trained on the labeled subset to predict labels for the unlabeled set. These, 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 (Tarvainen & Valpola, 2017), adversarial examples (Miyato et al., 2017), and augmentations (Xie et al., 2019a; Berthelot et al., 2019; Sohn et al., 2020; French et al., 2020). Recent methods have shown competitive results, with as little as 1% of labels for CIFAR-100. Most relevant to this work is MixMatch (Berthelot et al., 2019) which processes a batch of augmented labeled and unlabeled examples together with their guessed labels via a MixUp procedure (Zhang et al., 2018).

Self-supervised learning. The goal of self-supervised learning is to learn representations that are meaningful in some general sense, without 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 (Vincent et al., 2008; Zhang et al., 2016; Pathak et al., 2016; Zhang et al., 2017) Instead, follow-up methods opted for a classification task based on context prediction (Doersch et al., 2015; Noroozi & Favaro, 2016; Gidaris et al., 2018; Kolesnikov et al., 2019b), or clustering (Caron et al., 2018). Nevertheless, these impose an inherent problem when facing a particular downstream task which may not be well correlated with the self-supervised objective. Thus, there is no guarantee that the key information is kept and can be extracted from the features (Misra & van der Maaten, 2020). Some methods have proposed remedying the problem by making the self-supervised task aware of the downstream one (Zhai et al., 2019; Khosla et al., 2020).

Recently, a revival in self-supervised techniques based on contrastive loss (Hadsell et al., 2006) has shown markedly improved performance in large-scale computer vision tasks (Hénaff et al., 2019; Chen et al., 2020a; Tian et al., 2020; Xie et al., 2020). Most relevant to our method is the result reported by Chen et al. (2020b): self-supervised features obtained by contrastive learning can improve semi-supervised classification tasks after fine-tuning.

Learning with noisy labels. There are many variants to the problem of learning with noisy labels. While some methods (Veit et al., 2017; Litany & Freedman, 2018; Zhang et al., 2019) 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 (Wang et al., 2018; Zhang & Sabuncu, 2018).

Existing methods for LNL can be divided into two broad categories: loss modification and noise detection. The former includes techniques that account for noise distribution (Patrini et al., 2016; Xia et al., 2019; Yao et al., 2020). Alternatively, the loss itself may be replaced by a more robust version, such as mean absolute error (Ghosh et al., 2017), generalized cross-entropy (Zhang & Sabuncu, 2018), determinant-based mutual information (Xu et al., 2019), or a meta-learning objective (Li et al., 2019a). Differently, 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 (Jiang et al., 2018) or mutual teaching (Han et al., 2018; Yu et al., 2019) framework, geometry (Han et al., 2019), mixture models (Arazo et al., 2019a; Li et al., 2020), and quantiles of counterfactual loss distribution of samples (Song et al., 2020). These are often based on the observation that samples with noisy labels converge slower than those with clean ones (Arpit et al., 2017; Cicek et al., 2018; Li et al., 2019b; Pleiss et al., 2020). Hybrid methods that try to mix both noise detection and loss modification were also proposed (Song et al., 2019; Liu et al., 2020).

DivideMix. While noise detection methods have shown success in LNL problems, the two ways to deal with the detected noise both have disadvantages. Discarding the noisy samples means that valuable data may be lost, leading to poor sample variability (selected samples are usually easier than discarded ones, which leads to slow learning (Chang et al., 2017)) and overfitting to the small training set (Song et al., 2019). On the other hand, in the case of relabeling, all the samples are given the same weight, ignoring the fact that some of them are more likely to be noisy than the others. Instead, DivideMix proposed to keep the samples detected as noisy discarding only their labels, thereby effectively converting the LNL task into a semi-supervised one. The method can be broken down into three stages: (a) a warm-up phase, where standard supervised training is performed using the noisy set; (b) a division stage, where clean and noisy samples are split based on a mixture of Gaussians fitted to the loss distribution and a threshold value τ; and (c) a semi-supervised stage, where the noisy labels are discarded and MixMatch (Berthelot et al., 2019) is applied to the partially labeled data. Stages (b) and (c) are complementary since better splits would lead to improved semi-supervised performance, which would increase the accuracy of the noise classification, resulting in better separation.

3 METHOD

3.1 THE WARM-UP TRADE-OFF

Albeit crucial to the algorithm's success, the warm-up stage of DivideMix did not receive much attention. In particular, we identify two goals that need to be achieved by this stage: separable loss values and feature extraction. The former is clearly of utmost importance to the labeled subset's successful choice to be used by the semi-supervised stage. Despite a few measures for increasing resilience to noise (e.g., label smoothing), the algorithm assumes a relatively clean labeled subset. Regarding the latter, as was recently shown by Chen et al. (2020b), good features can significantly boost the performance of semi-supervised learning. These two competing goals create a trade-off between training longer to generate better features and early stopping to allow better separation between the clean and noisy samples. This limitation calls for other means to generate good features, which we propose to achieve via self-supervised pre-training. Notably, our proposed solution of self-supervised pre-trained features benefits both tasks: separation and semi-supervised learning, achieving a significant overall performance boost.

3.2 CONTRAST TO DIVIDE

As discussed above, strong pre-trained features can further boost the already powerful semi-supervised stage of DivideMix. In this section, we focus on the effect of these features on the critical warm-up phase. Specifically, we argue that by initializing the warm-up with good features, only mild adaptation is required. This effectively breaks the trade-off between feature extraction and noisy label detection, allowing faster and more accurate separation. Furthermore and most importantly, since the most demanding part of learning feature extraction is done without labels, the warm-up stage becomes much more robust to the noise level, saving the need for careful training time fine-tuning per problem.

Semi-supervised performance. Since DivideMix repeatedly converts the LNL task into a semi-supervised learning task, much of its strength comes from the underlying semi-supervised method, MixMatch. This implies that any improvement made to this stage would potentially boost its overall performance. In particular, recently, there has been encouraging evidence that self-supervised methods can help semi-supervised learning. We adopt the same approach and use contrastive learning to pre-train the network on the dataset at hand. Indeed, we observe that this network adapts very quickly to the task (as can be seen from Fig. 1), boosting the classification accuracy at every epoch, which in turn helps the separation and the overall performance.

Supervised vs. self-supervised pre-training. Since utilizing pre-trained features for improved performance is very common, it is important to distinguish between supervised pre-training on a large cleanly labeled data source and self-supervised pre-training on the given training set. In the context of LNL, we highlight the following aspects: dataset size and domain vs. task gaps.

It is far from trivial to assume access to a large dataset for a same task from a similar domain. A common example of such cases is using ImageNet for natural image classification (Kolesnikov et al., 2019a). Often domains do not have an ImageNet equivalent, necessitating a compromise on either quantity or domain similarity and resulting in lesser quality pre-trained features. Using self-supervision, on the other hand, eliminates the domain gap. Additionally, since noisy labels are often a result of large-scale annotation, this suggests the availability of massive amounts of data,

Table 1: C2D achieves consistently high classification accuracy (%, in form mean ± std over five runs) on CIFAR-10 under different noise rates and types, with markedly improved performance under very-high noise conditions. Meta-Learning results provided by Li et al. (2020).

Method	Architecture		Noise rate					
			20%	50%	80%	90%	95%	Asym. 40%
Meta-Learning (Li et al., 2019a)	PreAct ResNet-32	Peak	92.9	89.3	77.4	58.7	-	89.2
		Final	92.0	88.8	76.1	58.3	-	88.6
ELR+ (Liu et al., 2020)	ResNet-34	Peak	94.6	93.8	91.1	75.2	-	92.7
		Final	-	-	-	-	-	-
DivideMix (Li et al., 2020)	PreAct ResNet-18	Peak	96.1	94.6	93.2	76.0	-	**93.4**
		Final	95.7	94.4	92.9	75.4	-	**92.1**
C2D (our)	PreAct ResNet-18	Peak	**96.43 ± 0.07**	**95.32 ± 0.12**	**94.40 ± 0.04**	**93.57 ± 0.09**	**89.24 ± 0.75**	**93.45 ± 0.07**
		Final	**96.23 ± 0.09**	**95.15 ± 0.16**	**94.30 ± 0.12**	**93.42 ± 0.09**	**87.72 ± 2.21**	90.75 ± 0.35

Table 2: Peak and final classification accuracy (%, in form mean ± 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. (2020). * denotes results acquired by us based on published code.

Method	Architecture		Noise rate					
			20%	50%	80%	90%	95%	Asym. 40%
Meta-Learning (Li et al., 2019a)	PreAct ResNet-32	Peak	68.5	59.2	42.4	19.5	-	-
		Final	67.7	58.0	40.1	14.3	-	-
- ELR+ (Liu et al., 2020)	ResNet-34	Peak	77.5	72.4	58.2	30.8	-	76.5
		Final	-	-	-	-	-	-
ODD (Song et al., 2020)	WRN-28-10	Peak	79.1 ± 0.1	-	-	-	-	-
		Final	-	-	-	-	-	-
DivideMix (Li et al., 2020)	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*
C2D (our)	PreAct ResNet-18	Peak	78.69 ± 0.17	76.43 ± 0.25	67.78 ± 0,30	58.70 ± 0,31	**37.39 ± 3.80**	75.48 ± 0.16
		Final	78.32 ± 0.35	76.07 ± 0.41	67.43 ± 0,30	58.45 ± 0,30	**36.83 ± 4.29**	75.06 ± 0.16
C2D (our)	ResNet-50	Peak	**81.60**	**79.54**	**71.65**	**64.30**	-	**77.92**
		Final	**80.89**	**79.20**	**71.53**	**63.91**	-	**77.78**

which fits the data-hungry self-supervised setup well. On the other hand, contrastive pre-training is agnostic to the downstream task, inevitably creating a larger task gap. However, albeit being similar to CIFAR in both task and domain, ImageNet pre-training was unable to give the expected performance improvement. For further discussion we refer the reader to Section 4.1.

4 EXPERIMENTAL RESULTS

We perform an extensive evaluation of our method both on synthetic and real noise. We follow common practice in synthetic noise benchmarks and use CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) varying the amount of injected noise. For the real noise setting, we use Clothing1M (Xiao et al., 2015), a dataset of ∼1 million images of 14 classes of clothing acquired by web crawling.

4.1 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. 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. (2016).

Implementation details. We used two variants of ResNet (He et al., 2015): PreAct ResNet-18 and ResNet-50. Following the setup proposed by DivideMix, we used an SGD optimizer with a

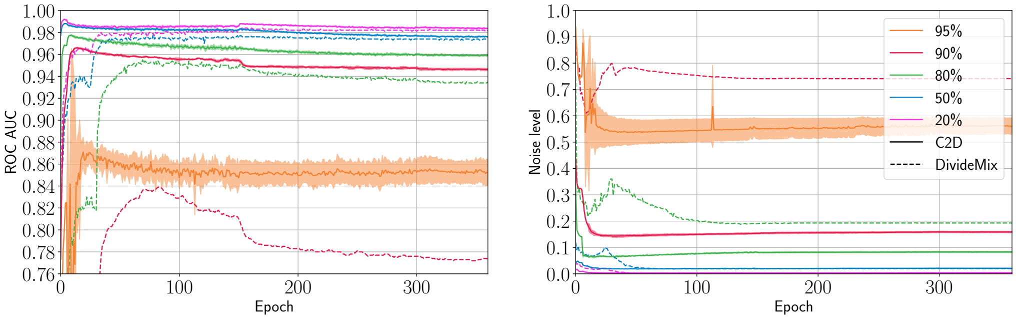

Figure 2: Training time ROC-AUC scores (left) and effective noise rates. C2D demonstrates higher initial score, faster rise, and more stable decrease in effective noise level.

momentum of 0.9, a weight decay of 0.0005, and an initial learning rate of 0.02, which is reduced by a factor of 10 after 150 epochs. The only modifications to the training hyper-parameters are: (a) to accommodate ResNet-50 in GPU memory, we reduced the batch size from 128 to 64 and (b) we observed that our network kept improving after 300 epochs and thus increased train length to 360 epochs. For self-supervised pre-training, we used a SimCLR implementation[1] in PyTorch (Paszke et al., 2019). The self-supervised model was trained for 1000 epochs on 4 NVIDIA 2080 Ti GPUs.

Using a small subset of the training set, we tuned the unlabeled loss weight $\lambda_{\mathcal{U}}$, the number of warm-up epochs, and the threshold for noisy label prediction τ. We chose a value of $\lambda_{\mathcal{U}}$ out of $\{0, 25, 50, 150, 500, 1000\}$. We observed that increasing λ 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 3.2, strong pre-trained features are expected to reduce the required warm-up length. We found that five epochs were sufficient, both for CIFAR-10 and CIFAR-100 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. Following Li et al. (2020), we present accuracy at the end of training together with the highest one 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. When presented with asymmetric noise, both DivideMix and C2D have a degradation between peak and final accuracy. Even though C2D shows stronger degradation, it performed on-par with previous art in terms of peak accuracy.

Table 2 shows 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 80% 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 5 noise realization), C2D still achieved a final accuracy of above 35% (and at least 30% in each individual run), surpassing the performance achieved by DivideMix at a noise rate of 90% with the same architecture (PreAct ResNet-18). In asymmetric noise, C2D performed similarly to prior art, with the smaller network, and achieved about 1.5% improvement over ELR+ (Liu et al., 2020) with ResNet-50.

Supervised vs. self-supervised pre-training. Transferring supervised pre-trained features from a source to a target domain (transfer learning) is widely used in deep learning. The large increase in performance due to self-supervised features raises a natural question: Can similar behavior be achieved by supervised pre-training on a different dataset?

[1] https://github.com/HobbitLong/SupContrast

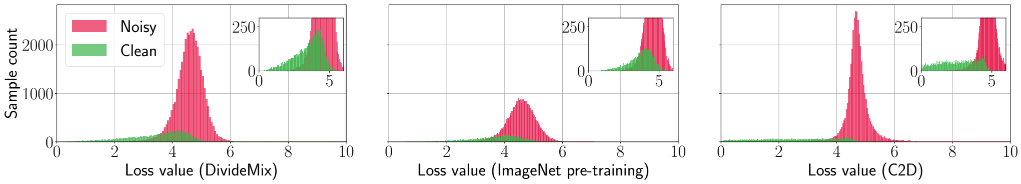

Figure 3: Loss distribution of clean and noisy samples after warm-up on CIFAR-100 with 80% noise for DivideMix, DivideMix with ImageNet pre-training, and C2D. As seen in the zoom-in, ImageNet pre-training damages the seperability whereas self-supervised pre-training (C2D), improves it.

Table 3: C2D nearly closes the gap with semi-supervised training on the same clean set size.

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

Table 4: Comparison with state-of-the-art methods in test accuracy (%) on Clothing1M. Results for baselines are copied from original papers.

Method	**Test accuracy**
Cross-Entropy	69.21
F-correction (Patrini et al., 2016)	69.84
Meta-Learning (Li et al., 2019a)	73.47
Self-learning (Han et al., 2019)	74.45
DivideMix (Li et al., 2020)	74.76
ELR+ (Liu et al., 2020)	**74.81**
C2D (our)	74.30

To answer this question, we ran DivideMix on CIFAR-100 with PreAct ResNet-18 initialized with ImageNet pre-trained weights. In light of the discussion in Section 3.2, one expects a small domain gap and no task gap, making this an almost ideal setup. Indeed, in addition to an expected shortening in the required warm-up length from 30 to 10 epochs, at the end of warm-up, on 80% noise we observed an increase both in the ROC-AUC score from 82% to 88% and classification accuracy from 26% to 36%. Although impressive, these improvements are quite far from the striking 97% ROC-AUC score and 59.4% classification accuracy achieved by C2D after warm-up.

Yet, most concerning was the almost immediate failure of DivideMix when entering the second stage of training. More specifically, after the warmup 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 mend this behavior, this phenomenon persisted across various threshold values ranging from 0.03 (C2D) to 0.5 (DivideMix), using either fixed or linearly increased values. While a full analysis of supervised transfer learning under noise conditions is outside the scope of this work, we suspect that the fast-adaptation property of the pre-trained network may instead have damaged the network resilience to noise.

4.2 Initial AUC and accuracy after warm-up

As discussed in Section 3.2, the self-supervised pre-training serves a dual purpose: boosting the separability between clean and noisy samples and providing a better initialization for the classification task. In the following, we analyze these properties. First, we qualitatively compare the features learned on the CIFAR-10 data by C2D at the end of the warm-up with features learned from scratch (as done in DivideMix). We visualize both in Fig. 1 using the dimensionality-reduction technique UMAP (McInnes et al., 2018). Colored using the ground-truth labels, C2D features (upper row) are clearly better clustered and easier to separate than the baseline (bottom row) at both noise levels. Furthermore, at high noise rates the baseline features suffer from acute degradation, while C2D features maintain some fidelity.

To evaluate the quality of noise detection, in Fig. 2 we present the ROC-AUC score and the effective noise rate, defined as the share of noisy samples in the labeled part of the dataset. C2D demonstrates multiple desired properties including a higher initial score, a much faster rise in separability score as well as a more stable decrease in effective noise level, and eventually a higher overall score and lower noise level. Moreover, even though C2D and the baseline both suffer from decrease in the ROC-AUC score due to overfitting, C2D demonstrated a lower gap between the peak and final scores than the baseline. Difference between loss histograms shown in Fig. 3 supports those claims.

4.3 Gap between LNL and semi-supervised learning

Having significantly strengthened the noise separation ability along with the improved initialization, one may ask what is the remaining gap between LNL and semi-supervised learning – the effective upper bound on the performance for this family of methods. To answer this question, we compared the performance of C2D 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 DivideMix noise separation procedure with an oracle. The result for 80% and 90% noise levels in CIFAR-100 are reported in Table 3. 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, those results indicate that good features can compensate for this lack of information even under severe noise conditions.

4.4 Clothing1M

We conclude the experimental section by testing our method on real-life noise present in the Clothing-1M dataset (Xiao et al., 2015). As some of the manually labeled images have both clean and noisy labels, we can estimate the noise level as approx. 38.5%. This also allows computing noise-related metrics such as the ROC-AUC of noise detection.

Implementation details. As most previous works, we used ResNet-50 architecture, but did not utilize ImageNet pre-training. For self-supervised pre-training, we used a SimCLR implementation[2] in PyTorch (Paszke et al., 2019), trained on 8 NVIDIA 2080 Ti GPUs for 750 epochs. We trained the network using the AdamW optimizer (Loshchilov & Hutter, 2017) with a weight decay of 0.001, and a batch size of 32. As in the case of CIFAR, the warm-up period is five epochs. We trained the network for 120 epochs, with initial learning rate of 0.002, reduced by a factor of 10 after 40 epochs. For each epoch, we sampled 1000 mini-batches from the training data with same amount of samples of every class (according to noisy label). We set $\lambda_{\mathcal{U}} = 0$. Since a large amount of data is available, we found that increasing value of the threshold to $\tau = 0.7$ improves the performance of the network.

Results. A comparison with state-of-the-art methods is reported in Table 4. C2D achieves a 0.5% accuracy gap from the current state-of-the-art. Importantly, all the compared methods use ImageNet pre-trained features. As discussed in Section 3.2, supervised pre-training may compensate for the (already minor) domain gap by eliminating the task gap, which may explain why C2D observes no additional gain. Moreover, C2D excels at high noise rates, which is not the case for Clothing1M. Interestingly, though, C2D did show a 3% improvement in the ROC-AUC score compared to baseline (81% vs. 78%). This suggests that the self-supervised features did help in separation but are less suited for classification than the supervised ones explicitly trained for that purpose on a richer dataset. We also emphasize that C2D is unique in that it did not require any additional external data.

5 Conclusion

In this paper, we proposed Contrast to Divide (C2D), a simple yet powerful modification to DivideMix – a method for learning with noisy labels – which leverages high-quality self-supervised features. In particular, we have shown that contrastive-learning-based pre-training can boost the crucial warm-up stage, dramatically improving noise detection. Along with providing strong initialization for the semi-supervised stage, C2D demonstrates consistently high performance across various noise levels. 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. In particular, we believe C2D has high potential in domains where no large-scale annotated datasets exist, such as medical images.

[2] https://github.com/HobbitLong/SupContrast

References

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin Mcguinness. Unsupervised label noise modeling and loss correction. volume 97 of *Proceedings of Machine Learning Research*, pp. 312–321, Long Beach, California, USA, 09–15 Jun 2019a. PMLR. URL http://proceedings.mlr.press/v97/arazo19a.html. (cited on p. 3)

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*, 2019b. URL https://arxiv.org/abs/1908.02983. (cited on p. 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. volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/arpit17a.html. (cited on p. 3)

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 (eds.), *Advances in Neural Information Processing Systems 32*, pp. 5049–5059. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8749-mixmatch-a-holistic-approach-to-semi-supervised-learning. (cited on pp. 3 and 4)

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)*, pp. 132–149, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Mathilde_Caron_Deep_Clustering_for_ECCV_2018_paper.html. (cited on p. 3)

Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 1002–1012. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6701-active-bias-training-more-accurate-neural-networks-by-emphasizing-high-varian (cited on p. 3)

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a. URL https://arxiv.org/abs/2002.05709. (cited on pp. 2 and 3)

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*, 2020b. URL https://arxiv.org/abs/2006.10029. (cited on pp. 2, 3, and 4)

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. URL http://openaccess.thecvf.com/content_ECCV_2018/html/Safa_Cicek_SaaS_Speed_as_ECCV_2018_paper. (cited on p. 3)

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. URL https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Doersch_Unsupervised_Visual_Representation_ICCV_2015_paper.html. (cited on p. 3)

Geoff French, Avital Oliver, and Tim Salimans. Milking cowmask for semi-supervised image classification. *arXiv preprint arXiv:2003.12022*, 2020. URL https://arxiv.org/abs/2003.12022. (cited on p. 3)

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, pp. 1919–1925. AAAI Press, 2017. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14759/14355. (cited on p. 3)

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. URL https://arxiv.org/abs/1803.07728. (cited on p. 3)

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, pp. 1735–1742. IEEE, 2006. URL https://ieeexplore.ieee.org/document/1640964. (cited on p. 3)

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 (eds.), *Advances in Neural Information Processing Systems 31*, pp. 8527–8537. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8072-co-teaching-robust-training-of-deep-neural-networks-with-extremely-noisy-labe (cited on p. 3)

Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. URL http://openaccess.thecvf.com/content_ICCV_2019/html/Han_Deep_Self-Learning_From_Noisy_Labels_ICCV_2019_paper. (cited on pp. 3 and 7)

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

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aäron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019. URL https://arxiv.org/abs/1905.09272. (cited on pp. 2 and 3)

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. volume 80 of *Proceedings of Machine Learning Research*, pp. 2304–2313, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/jiang18c.html. (cited on p. 3)

Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. *arXiv preprint arXiv:1911.09781*, 2019. URL https://arxiv.org/abs/1911.09781. (cited on p. 2)

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. URL https://arxiv.org/abs/2004.11362. (cited on p. 3)

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*, 2019a. URL https://arxiv.org/abs/1912.11370. (cited on p. 4)

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 2019b. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Kolesnikov_Revisiting_Self-Supervised_Visual_Representation_Learning_CVPR_2019_paper.html. (cited on p. 3)

Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. (cited on p. 5)

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*, pp. 1–26, 2020. URL https://link.springer.com/article/10.1007/s11263-020-01316-z. (cited on p. 1)

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. URL http://deeplearning.net/wp-content/uploads/2013/03/pseudo_label_final.pdf. (cited on p. 3)

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*, pp. 5051–5059, 2019a. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Li_Learning_to_Learn_From_Noisy_Labeled_Data_CVPR_2019_paper.html. (cited on pp. 3, 5, and 7)

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. URL https://openreview.net/forum?id=HJgExaVtwr. (cited on pp. 1, 3, 5, 6, and 7)

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*, 2019b. URL https://arxiv.org/abs/1903.11680. (cited on p. 3)

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. URL https://arxiv.org/abs/1708.02862. (cited on p. 1)

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

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. URL https://arxiv.org/abs/2007.00151. (cited on pp. 3, 5, 6, and 7)

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. URL https://arxiv.org/abs/1711.05101. (cited on p. 8)

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. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Dhruv_Mahajan_Exploring_the_Limits_ECCV_2018_paper.html. (cited on p. 1)

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. URL https://arxiv.org/abs/1802.03426. (cited on pp. 2 and 7)

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. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Misra_Self-Supervised_Learning_of_Pretext-Invariant_Representations_CVPR_2020_paper.html. (cited on p. 3)

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. URL https://arxiv.org/abs/1704.03976. (cited on p. 3)

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 69–84, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46466-4. URL https://link.springer.com/chapter/10.1007/978-3-319-46466-4_5. (cited on p. 3)

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 (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8026–8037. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library. (cited on pp. 6 and 8)

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. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Pathak_Context_Encoders_Feature_CVPR_2016_paper.html. (cited on p. 3)

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. URL https://arxiv.org/abs/1609.03683. (cited on pp. 3, 5, and 7)

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. URL https://arxiv.org/abs/2001.10528. (cited on p. 3)

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. URL https://arxiv.org/abs/2001.07685. (cited on p. 3)

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: Refurbishing unclean samples for robust deep learning. volume 97 of *Proceedings of Machine Learning Research*, pp. 5907–5915, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/song19b.html. (cited on pp. 3 and 4)

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. URL https://arxiv.org/abs/2003.10647. (cited on pp. 3 and 5)

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 (eds.), *Advances in Neural Information Processing Systems 30*, pp. 1195–1204. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6719-mean-teachers-are-better-role-models-weight-averaged-consistency-targets-impr (cited on p. 3)

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. URL https://arxiv.org/abs/2005.10243. (cited on pp. 2 and 3)

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. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Veit_Learning_From_Noisy_CVPR_2017_paper.html. (cited on p. 3)

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, pp. 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390294. URL https://doi.org/10.1145/1390156.1390294. (cited on p. 3)

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. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Iterative_Learning_With_CVPR_2018_paper.html. (cited on p. 3)

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 (eds.), *Advances in Neural Information Processing Systems 32*, pp. 6838–6849. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8908-are-anchor-points-really-indispensable-in-label-noise-learning. (cited on p. 3)

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. URL http://openaccess.thecvf.com/content_cvpr_2015/html/Xiao_Learning_From_Massive_2015_CVPR_paper.html. (cited on pp. 1, 2, 5, and 8)

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

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

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. URL https://arxiv.org/abs/2007.10985. (cited on p. 3)

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 (eds.), *Advances in Neural Information Processing Systems 32*, pp. 6225–6236. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8853-l_dmi-a-novel-information-theoretic-loss-function-for-training-deep-nets-robust-to-l
(cited on p. 3)

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. URL https://arxiv.org/abs/1905.00546. (cited on p. 3)

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. URL https://arxiv.org/abs/2006.07805. (cited on p. 3)

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? volume 97 of *Proceedings of Machine Learning Research*, pp. 7164–7173, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/yu19b.html. (cited on p. 3)

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4L: self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. URL https://openaccess.thecvf.com/

content_ICCV_2019/html/Zhai_S4L_Self-Supervised_Semi-Supervised_Learning_ICCV_2019_paper.html. (cited on p. 3)

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb. (cited on p. 3)

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *European Conference on Computer Vision*, pp. 649–666. Springer, 2016. URL https://link.springer.com/chapter/10.1007/978-3-319-46487-9_40. (cited on p. 3)

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. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Zhang_Split-Brain_Autoencoders_Unsupervised_CVPR_2017_paper.html. (cited on p. 3)

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 (eds.), *Advances in Neural Information Processing Systems 31*, pp. 8778–8788. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8094-generalized-cross-entropy-loss-for-training-deep-neural-networks-with-noisy-l (cited on p. 3)

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. URL https://arxiv.org/abs/1910.00701. (cited on p. 3)