ReLaB: Reliable Label Bootstrapping for Semi-Supervised Learning

Paul Albert, Diego Ortego, Eric Arazo, Noel O'Connor, Kevin McGuinness
School of Electronic Engineering,
Insight Centre for Data Analytics, Dublin City Univeristy (DCU)
paul.albert@insight-centre.org

***Abstract*—Reducing the amount of labels required to train convolutional neural networks without performance degradation is key to effectively reduce human annotation efforts. We propose Reliable Label Bootstrapping (ReLaB), an unsupervised preprossessing algorithm which improves the performance of semi-supervised algorithms in extremely low supervision settings. Given a dataset with few labeled samples, we first learn meaningful self-supervised, latent features for the data. Second, a label propagation algorithm propagates the known labels on the unsupervised features, effectively labeling the full dataset in an automatic fashion. Third, we select a subset of correctly labeled (reliable) samples using a label noise detection algorithm. Finally, we train a semi-supervised algorithm on the extended subset. We show that the selection of the network architecture and the self-supervised algorithm are important factors to achieve successful label propagation and demonstrate that ReLaB substantially improves semi-supervised learning in scenarios of very limited supervision on image classification benchmarks such as CIFAR-10, CIFAR-100 and mini-ImageNet. We reach average error rates of 22.34 with 1 random labeled sample per class on CIFAR-10 and lower this error to 8.46 when the labeled sample in each class is highly representative. Our work is fully reproducible: https://github.com/PaulAlbert31/ReLaB.**

I. Introduction

Convolutional neural networks (CNNs) are now the established standard for visual representation learning [1]–[3], yet one of their most prevalent limitations is the large quantity of labeled data they require. Although enormous quantities of unlabeled data are now accessible and can be collected with minimal effort, the annotation process remains limited by human intervention [4]–[7].

There are several alternatives in the literature, that reduce the need for the strong supervision required to train deep neural networks. These include transfer learning [8] or few-shot learning [9], where supervised pre-trained features are exploited; semi-supervised learning [10], where only a part of the dataset is labeled; self-supervised learning [11], where a pretext task is used to learn meaningful features from the data alone; label noise [12], where labels are inferred automatically; and oversampling [13], where extra image samples are generated from the existing pool.

There exists different approaches for semi-supervised scenarios in the state-of-the-art. In particular, consistency regularization [14], [15] and pseudo-labeling methods [16], [17] are the two dominating strategies. To learn from unlabeled data, consistency regularization encourages consistency in the predictions for the same sample under different perturbations, while pseudo-labeling generates pseudo-labels for unlabeled samples directly from the network predictions. Despite recent efforts in the semi-supervised learning literature aiming at reducing human supervision further, extreme label scarcity is still challenging [18], [19]. In the absence of labels, the self-supervised paradigm for unsupervised visual representation learning has recently gained traction [20]–[24]. Self-supervised learning constructs a supervisory signal using a pretext task where pretext labels are generated from the data. By solving pretext tasks such as colorization of greyscale images [25], prediction of image rotations [23], or contrasting different views of the same image [24], high quality features can be learned without human annotations. The success of self-supervised learning has motivated its adoption for semi-supervised learning, which improved performance in cases of very low label availability [18], [26]. Berthelot et al. [18] and Wang et al. [26] use self-supervision as a regularization which stabilizes network training, while Rebuffi et al. [27] make use of self-supervision [23] as an initialization strategy for semi-supervised training.

In this paper, we explore the idea of automatically annotating image data using label propagation. In particular, we use representations learned by self-supervised tasks together with a low amount of labels to apply label propagation and spread the available labels to the entirety of the samples. The resulting is a fully labeled dataset which contains numerous incorrect (noisy) annotations. We then select a trusted, clean subset from this noisy dataset that reliably extends the initially labeled data. The extended labeled dataset is then used to enhance the performance of any semi-supervised image classification algorithm when very few labeled samples are available. We name this label bootstrapping strategy ReLaB. When ReLaB is used to bootstrap labels for ReMixMatch [18] on CIFAR-10 with 10, 40, 100 labeled samples, we reduce the accuracy error by more than 36, 22, 15 absolute points respectively. ReLaB's unsupervised knowledge-bootstrapping pipeline makes use of self-supervised, image retrieval and label noise solutions to provide an approach for scenarios of extremely scare annotations in semi-supervised learning. This could include visual domains where annotations are either time-consuming and expensive to gather or when expert annotators are required. Our contributions are as follow:

1) We propose an unsupervised knowledge-bootstrapping

pipeline which enhances the performance of semi-supervised algorithms when very few labeled samples are available.

2) We propose a reliable sample selection method in the presence of label noise induced by label propagation. The method is robust to class and noise imbalance.
3) We evaluate the importance of good self-supervised features for label propagation, and demonstrate the superiority of our approach when dealing with feature-based label noise generated by label propagation.

II. Related Work

A. Semi-supervised learning

Semi-supervised learning seeks to reduce human supervision by jointly learning from sparsely labeled data and extensive unlabeled data. Semi-supervised learning has evolved rapidly in recent years by exploiting two distinct strategies [10]: consistency regularization and pseudo-labeling.

a) Consistency regularization: promotes consistency in the network's predictions for the same unlabeled sample altered by different perturbations. Notable examples of consistency regularization algorithms are VAT [28] where samples are perturbed by virtual adversarial attacks, Mean Teacher [15] where a teacher network is built from the exponential moving average of the student network weights to produce perturbed predictions, and ICT [29] which encourages predictions of interpolated samples to be consistent with the interpolation of the predictions. Berthelot et al. propose MixMatch [14], where perturbed predictions are generated by means of data-augmented, sharpened labels and where labeled and unlabeled examples are mixed together using Mixup [30]. MixMatch was extended in ReMixMatch [18] by exploiting distribution alignment [31] and an augmentation anchoring policy.

b) Pseudo-labeling: directly exploits the network predictions on unlabeled samples by using them as labels (pseudo-labels) to regularize training. Lee et al. [32] propose an early attempt at pseudo-labeling, limited to a finetuning stage on a pre-trained network. Shi et al. [33] derive certainty weights for unlabeled samples from their distance to neighboring samples in the feature space. Arazo et al. [17] have shown that a pure pseudo-labeling without using consistency regularization can reach competitive performance when addressing confirmation bias [34]. Interestingly, Iscen et al. [35] proposed a label-propagation based strategy for semi-supervised learning. In particular, they estimate pseudo-labels using both the network prediction and label-propagation on the current features of the network, producing two different supervised objectives.

B. Self-supervised learning

Self-supervised learning defines proxy or pretext tasks to learn useful representations without human intervention [11]. Context prediction [21], colorization [25], puzzle solving [36], instance discrimination [37], image rotation prediction [23], interactive clustering [38], optimal transport [20], image transformation prediction [39] and construction of local neighborhoods [40] are some examples of pretext tasks. Unsupervised contrastive learning has recently emerged as the new standard for representation learning [24], [41] where a given sample is encouraged to have similar features to augmented versions of itself and dissimilar representations to other samples in the dataset.

Recent contributions shows that coupling self-supervised and semi-supervised learning can increase the accuracy when few labels are available. Rebuffi et al. [27] use RotNet [23] as a network initialization strategy, ReMixMatch [18] exploits RotNet [23] together with their semi-supervised algorithm to achieve stability with few labels, and EnAET [26] leverage transformation encoding from AET [39] to improve the consistency of predictions on transformed images.

C. Label propagation for semi-supervised learning

Label propagation processes stem from the image retrieval literature. Diffusion [42]–[44] constructs a pairwise affinity matrix, relating images to each other using meaningful features before diffusing the affinity values to the entirety of the graph. In the case of label propagation, the image retrieval objective is reformulated as a label propagation objective which transfers the information from labeled data to an unlabeled dataset [45]. The diffusion result can be directly used to estimate labels and finetune pre-trained networks in few-shot learning [9] or to define pseudo-labels for semi-supervised learning [35]. Other attempts using label propagation for semi-supervised learning include dynamically capturing the manifold's structure and regularize it to form compact clusters which facilitate class separation [46] or to encourage random walks ending in the same class they started from, while penalizing different class endings [47].

D. Label noise

Label noise is a topic of increasing interest for the community [48], which aims at limiting the degradation of CNN representations when learning in label noise conditions [49]. Label noise algorithms can be categorized in four different approaches: loss correction [50]–[52], relabeling [53]–[55], semi-supervised [56], [57] and regularization [30], [58]. Loss correction algorithms reduce the contribution of the incorrect or noisy labels in the training objective by approximating true labels at sample [52], [58], [59] or class level [51], [60] or by weighing down noisy samples in the loss [50], [61], [62]. Relabeling methods [53], [55] iteratively update noisy labels to an estimation of the true label. Semi-supervised methods detect the noisy samples before discarding their labels and exploit the resulting unlabeled content in a semi-supervised setup [56], [57], [63], [64]. Finally, strong regularization such as Mixup [30] enables robustness to label noise without explicitly addressing it. A recurrent paradigm used to identify clean samples is the small loss trick [50], [53], [59], [64] where clean samples exhibit a lower loss early in the training since they are often easier to learn.

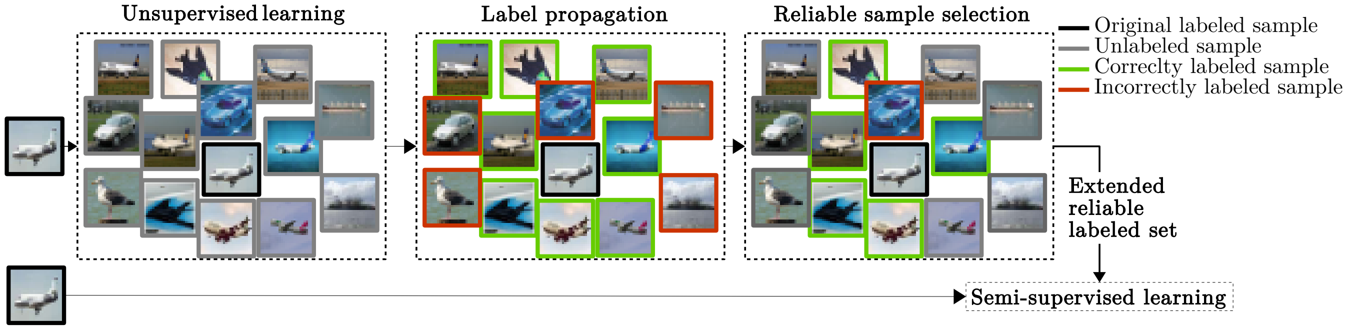

Fig. 1. Reliable Label Bootstrapping (ReLaB) overview (best viewed in color). Unlike traditional SSL (bottom) that directly uses the labeled examples provided (*airplane*), ReLaB (top) bootstraps additional labels before applying SSL. Unsupervised learning using labeled (black) and unlabeled (gray) samples is done to obtain discriminative representations. Label propagation on unsupervised representations propagates the few labeled examples to all samples. This leads to both correct (green) and incorrect (red) labels. A sample selection is finally performed to avoid noisy labels (some will unavoidably be selected) and create a reliable extended labeled set.

III. Reliable label bootstrapping for semi-supervised learning

We formulate a semi-supervised classification task for C classes as learning a model h_ψ given a training set $\mathcal{D}$ of N samples. The dataset consists of the labeled set $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ with corresponding one-hot encoded labels $y_i \in \{0,1\}^C$ and the unlabeled set $\mathcal{D}_u = \{x_i\}_{i=1}^{N_u}$, $N = N_l + N_u$ the total number of samples. We consider a CNN for $h_\psi : \mathcal{D} \rightarrow [0,1]^C$, where ψ denotes the model's parameters. The network is comprised of a feature extractor $h_{\psi_f} : \mathcal{D} \rightarrow \Phi$ with parameters ψ_f, mapping the input space to the feature space Φ, and a classifier $h_{\psi_c} : \Phi \rightarrow [0,1]^C$ with parameters ψ_c.

We address the case where $\mathcal{D}_l$ contains a low number of samples. We propose to extend $\mathcal{D}_l$ to a larger dataset $\mathcal{D}_r$ of size $N_r > N_l$ by automatically labeling samples from $\mathcal{D}_u$. To do so, we propagate labels from $\mathcal{D}_l$ to $\mathcal{D}_u$ using self-supervised features learned on $\mathcal{D}$. This results in an automatically labeled set $\hat{\mathcal{D}} = \mathcal{D}_l + \hat{\mathcal{D}}_u$ with the samples from $\hat{\mathcal{D}}_u$ having now been approximately annotated. We build $\mathcal{D}_r$ by aggregating $\mathcal{D}_l$ and by selecting correctly annotated (reliable) samples from the propagated labels from $\hat{\mathcal{D}}_u$ using label noise methodologies. Training on $\mathcal{D}_r$ greatly improves the performance of semi-supervised algorithms when very few labels are available. Figure 1 presents and overview of our proposed approach.

A. Label propagation on self-supervised features

Knowledge transfer from the labeled set $\mathcal{D}_l$ to the unlabeled set $\mathcal{D}_u$ is implicitly done by semi-supervised learning approaches as the network predictions for $\mathcal{D}_u$ can be seen as estimated labels. With few labeled samples however, it is difficult to learn useful initial representations from $\mathcal{D}_l$ and performance is substantially degraded [18] (see Subsection IV-E).

Conversely, we propose to learn a set of descriptors in an unsupervised manner and subsequently propagate the labels on the data manifold, in order to retrieve additional labels for the unlabeled data. We adopt the established graph diffusion algorithm [35], [42], [43], [65], [66] for label propagation. We formulate the label propagation problem in a similar fashion than [35] except that we study the estimation of $\tilde{y}$ as a label propagation task using unsupervised visual representations learned from all samples in $\mathcal{D}$. In particular, we learn a feature extractor h_{φ_f} using self-supervision to obtain class-discriminative image representations [11] and subsequently propagate labels from the N_l labeled images to estimate labels $\tilde{y}$ for the N_u unlabeled samples. We do so by solving a label propagation problem based on graph diffusion [35]. First, the set of descriptors $\{v_i\}_{i=1}^{N}$ are used to define the affinity matrix:

$$S = D^{-1/2} A D^{-1/2}, \tag{1}$$

where $D = \operatorname{diag}(A \mathbb{1}_N)$ is the degree matrix of the graph and the adjacency matrix A is computed as $A_{ij} = \left({}^{v_i^T v_j}/_{\|v_i\| \|v_j\|} \right)^\gamma$ if $i \neq j$ and 0 otherwise. γ weighs the affinity term to control the sensitivity to far neighbors and is set to 3 as in [35]. The diffusion process estimates the $N \times C$ matrix as:

$$F = (I - \alpha S)^{-1} Y, \tag{2}$$

where α denotes the probability of jumping to adjacent vertices in the graph and Y is the $N \times C$ label matrix defined such that $Y_{ic} = 1$ if sample $x_i \in \mathcal{D}_l$ and $y_i = c$ (i.e. belongs to the c class), where i (c) indexes the rows (columns) in Y. Finally, the estimated one-hot label $\tilde{y}_i$ is:

$$\tilde{y}_{ic} = \begin{cases} 1, & \text{if } c = \arg\max\limits_{c} \; F_{ic} \\ 0, & \text{otherwise} \end{cases},$$

for each unlabeled sample $x_i \in \mathcal{D}_u$. The estimated labels allow the creation of the extended dataset with estimated noisy labels $\tilde{\mathcal{D}} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$, where $\tilde{y}_i = y_i, \forall\, x_i \in \mathcal{D}_l$.

B. Reliable sample selection: dealing with noisy labels

Propagating existing labels using self-supervised representations as described in Section III-A, results in estimated labels $\tilde{y}_i$ that might be incorrect, i.e. label noise. Using noisy labels as a supervised objective on $\tilde{\mathcal{D}}$ leads to performance degradation due to label noise memorization [35], [49]. Since the label noise in $\tilde{\mathcal{D}}$ comes from features extracted from the data, noisy samples tend to be visually similar to the seed samples which poses a challenging scenario as noise-robust, state-of-the-art training

TABLE I

CLASS AND NOISE IMBALANCE AFTER APPLYING LABEL PROPAGATION

	CIFAR-10		CIFAR-100	
$\frac{N_l}{C}$	#sample	noise ratio	#sample	noise ratio
4	4249 ± 1726	24.14 ± 10.42	472 ± 161	50.52 ± 16.79
10	4888 ± 1367	24.28 ± 7.43	477 ± 180	39.92 ± 15.31
25	4990 ± 1036	9.50 ± 6.90	444 ± 233	33.39 ± 12.55

strategies [30], [58], [59] experience important limitations (see Table IV). Moreover, we find that this label noise is unbalanced in terms of number of samples and different levels of noise in each class. We report in Table I the median and standard deviation for the number of sample per class (#samples) and noise ratio over the classes of CIFAR-10 and CIFAR-100 for different amounts of labeled samples in N_l. Using the small loss trick to select a subset of clean samples is commonly used in the label noise literature [56], [63], [64], [67], but the issues specific to label noise resulting from label propagation are not addressed in the label noise literature and pose additional challenges, see Section IV-C.

In particular, we identify clean samples using the cross-entropy loss:

$$\ell_i = -\tilde{y}_i^T \log h_\psi(x_i), \tag{3}$$

with softmax-normalized logits $h_\psi(x_i)$ and training with a high learning rate (small loss) which helps prevent label noise memorization [59] on the extended dataset $\tilde{\mathcal{D}}$. The reliable set $\mathcal{D}_r = \{(x_i, \tilde{y}_i)\}_{i=1}^{N_r}$, with $N_r > N_l$, is then created by selecting for each class c the N_l^c originally labeled samples for that class c in $\mathcal{D}_l$ and the $N_r^c - N_l^c$ samples in class c from $\mathcal{D}_u$ with the lowest loss ℓ_i.

Differently from previous works tackling synthetic noise [64], we find that the noise present in $\tilde{\mathcal{D}}$ makes the clean sample retrieval using the loss ℓ_i during any particular epoch unstable and that the noise is class-unbalanced (see Table I), making it more challenging. We therefore impose the selection of a class-balanced clean subset and choose to average the network losses over the last T training epochs. This results in a clean, trusted subset which limits the label noise bias introduced to the semi-supervised algorithm. Table III shows that the knowledge we bootstrap in $\mathcal{D}_r$ is not overly sensitive to N_r.

C. Semi-supervised learning

Unlike traditional learning from $\mathcal{D}_l$ and $\mathcal{D}_u$, ReLaB provides semi-supervised algorithms with a (larger) reliable labeled set $\mathcal{D}_r$ extended from the original (smaller) labeled set $\mathcal{D}_l$. The extension from $\mathcal{D}_l$ to $\mathcal{D}_r$ is done in a completely unsupervised manner and promotes a significant reduction of the error rates of SSL algorithms when few labels are given, e.g. in Table VII the 50.62% error of ReMixMatch [18] in CIFAR-10 for one labeled sample per class ($N_l = 10$) is reduced to 8.46%.

IV. EXPERIMENTS

A. Datasets and implementation details

We experiment with three image classification datasets: CIFAR-10 [69], CIFAR-100 [69], and mini-ImageNet [70]. CIFAR (mini-ImageNet) data consists of 60K 32×32 (84×84) RGB images split into 50K training samples and 10K for testing. CIFAR-10 samples are organized in 10 classes, while CIFAR-100 and mini-ImageNet are in 100. We follow common practices for image retrieval [71], [72] and perform PCA whitening as well as L_2 normalization on the features v before applying diffusion. We construct the reliable set $\mathcal{D}_r$ by training for 60 epochs with a high learning rate (0.1) to prevent label noise memorization [59] and select the samples with the lowest loss per class at the end of the training. We average the per-sample loss over the last $T = 30$ epochs of training. For the semi-supervised learning experiments, we always use a standard WideResNet-28-2 [3] for fair comparison with related work. We combine our approach with state-of-the-art pseudo-labeling [17] and consistency regularization-based [18] semi-supervised methods to prove the stability of ReLaB when applied to different semi-supervised strategies. We use the default configuration for pseudo-labeling[1] except for the network initialization, where we make use of the Rotation self-supervised objective [23] and freeze all the layers up to the last convolutional block in a similar fashion to Rebufi et al. [27]. We find that this is necessary to preserve strong early features throughout the training. The network is warmed up on the labeled set for 200 epochs and then trained for 400 epochs on the whole dataset. For ReMixMatch[2] we train the network from scratch for 256 epochs. Experiments in Section IV-C for the supervised alternatives on dealing with label noise [30], [59] follow the authors's configurations, while cross-entropy and Mixup training in Table IV is done for 150 epochs with an initial learning rate of 0.1 that we divide by 10 in epochs 80 and 130.

B. Self-supervised representations for label propagation

Label propagation relies upon self-supervised representations extracted form the data, i.e. the quality of the propagation directly depends on these representations. We propose to explore different unsupervised learning alternatives to obtain these representations. Table II, presents the label noise percentage of the extended labeled set $\tilde{\mathcal{D}}$ in CIFAR-10 (100) formed after label propagation of the specified self-supervised representations with 1, 4 and 10 (4, 10 and 25) labeled samples per-class in $\mathcal{D}_l$. We select RotNet [23], NPID [37], UEL [68], AND [40] and iMix [41] as five recent self-supervised methods. We experiment training WideResNet-28-2 (WRN-28-2) [3], ResNet-18 (RN-18) and ResNet-50 (RN-50) [2] architectures. All the self-supervised methods are trained using the recommended configuration. We report average noise percentage and standard deviation for 3 different labeled subset $\mathcal{D}_l$. We confirm that the architecture has a key impact on the

[1] https://github.com/EricArazo/PseudoLabeling

[2] https://github.com/google-research/remixmatch

TABLE II

Label noise percentage in $\tilde{\mathcal{D}}$ using different amounts of labeled samples per class after label propagation using different self-supervised methods and network architectures. Lower is better.

		CIFAR-10			CIFAR-100		
		1	4	10	4	10	25
RotNet [23]	WRN-28-2	67.90 ± 8.51	51.68 ± 3.03	50.09 ± 2.55	83.08 ± 0.52	76.31 ± 0.33	67.81 ± 0.15
	RN-18	66.02 ± 5.98	53.58 ± 1.57	47.60 ± 3.51	80.83 ± 0.56	73.79 ± 0.42	65.58 ± 0.34
	RN-50	80.52 ± 30.08	77.58 ± 3.45	71.07 ± 1.05	80.75 ± 0.23	72.33 ± 0.15	62.78 ± 0.12
NPID [37]	WRN-28-2	68.72 ± 1.51	56.3 ± 2.42	51.35 ± 1.55	84.02 ± 0.30	76.91 ± 0.40	67.97 ± 0.13
	RN-18	59.34 ± 7.13	42.70 ± 2.32	37.14 ± 0.48	77.80 ± 0.55	69.54 ± 0.25	61.29 ± 0.67
	RN-50	59.44 ± 3.10	44.54 ± 2.32	38.13 ± 0.63	76.67 ± 0.58	68.54 ± 0.10	60.46 ± 0.16
UEL [68]	WRN-28-2	60.81 ± 6.41	45.84 ± 2.09	41.30 ± 2.00	79.21 ± 0.09	71.29 ± 0.39	62.89 ± 0.19
	RN-18	52.02 ± 7.24	34.51 ± 1.03	29.84 ± 0.78	71.9 ± 0.36	63.25 ± 0.41	56.51 ± 0.22
	RN-50	49.48 ± 7.66	32.81 ± 1.50	28.78 ± 1.08	69.62 ± 0.13	60.81 ± 0.48	54.08 ± 0.22
AND [40]	WRN-28-2	61.35 ± 0.57	46.12 ± 4.07	40.78 ± 0.27	79.38 ± 0.37	71.65 ± 0.03	63.29 ± 0.38
	RN-18	46.55 ± 5.64	28.82 ± 1.29	24.64 ± 1.44	67.48 ± 1.04	58.3 ± 0.26	51.47 ± 0.13
	RN-50	41.96 ± 8.74	24.34 ± 0.94	21.28 ± 0.75	66.25 ± 0.33	56.6 ± 0.52	46.31 ± 0.15
iMix [41]	WRN-28-2	53.75 ± 2.58	37.06 ± 2.40	31.27 ± 0.27	76.26 ± 0.60	64.92 ± 0.18	57.95 ± 0.45
+	RN-18	46.25 ± 6.11	18.55 ± 1.81	14.51 ± 2.35	49.74 ± 1.20	42.90 ± 0.39	39.17 ± 0.26
N-pairs	RN-50	$\mathbf{38.14 \pm 8.34}$	$\mathbf{16.93 \pm 1.73}$	$\mathbf{13.72 \pm 1.70}$	$\mathbf{45.49 \pm 1.04}$	$\mathbf{39.41 \pm 0.08}$	$\mathbf{35.75 \pm 0.26}$

TABLE III

Sensitivity of semi-supervised methods to different amounts of bootstrapped samples per class ($\frac{N_r}{c}$) considering an initial 4 labeled samples per class ($\frac{N_l}{c} = 4$). We report label noise percentage in $\mathcal{D}_r$ and final error rates after semi-supervised training.

	CIFAR-10		CIFAR-100	
$\frac{N_r}{C}$	Noise (%)	SSL error	Noise (%)	SSL error
25	**0.40**	12.12	**25.48**	51.90
50	0.60	9.18	30.20	51.43
75	1.07	**8.76**	33.51	**50.65**
100	1.30	8.79	35.69	51.14

TABLE IV

Learning from $\hat{\mathcal{D}}$ constructed from 4 labeled samples per class on CIFAR-10 ($N_l = 40$) and CIFAR-100 ($N_l = 400$). Error rates

	CIFAR-10	CIFAR-100
CE	22.64	59.88
M [30]	21.27	57.92
DB [59]	14.84	55.07
ELR [58]	17.39	47.95
Ret. score + PL [17]	17.55	54.19
ReLaB + PL [17]	12.38	53.58
ReLaB + RMM [18]	**6.68**	**43.53**

label noise percentage, which agrees with previous observations on the quality benefits of self-supervised features from larger architectures [11], [41]. We find that using diffusion on features learned using the iMix algorithm promotes the lowest amount of noise and adopt it together with a ResNet-50 in the subsequent experiments.

C. Dealing with noisy labels

We analyze the importance of the selected number of samples N_r over the label noise percentage in the extended reliable subset $\mathcal{D}_r$ and semi-supervised performance (using ReMixMatch (RMM) [18]). Table III shows how, a balance has to be found between a sufficient amount of bootstrapped samples and a low noise ratio. Increasing the number of samples in $\mathcal{D}_r$ is beneficial up to 100 samples per class, where adding more does not compensate the higher noise percentage. Based on this experiment and the typical amounts of labeled samples needed to perform successful SSL [14], [15], [17], [35], we choose a conservative $N_r = 500$ (4000) for CIFAR-10 (100) for further experiments.

Since $\tilde{\mathcal{D}}$ is corrupted with label noise, it is reasonable to expect that supervised alternatives on dealing with label noise [30], [59] could help combat this label noise. Table IV compares our proposed approach against training on $\tilde{\mathcal{D}}$ with standard cross-entropy (CE) and label noise robust methods such as Mixup (M) [30], the Dynamic Bootstrapping (DB) loss correction method [59] and the Early Regularization (ELR) strategy [58]. We also report using the retrieval score (Ret. score) from the label propagation ($\max_c F_{ic}$ in eq. 2) instead of ReLaB for selecting the trusted subset. In the presence of label noise in $\tilde{\mathcal{D}}$, we show in Table IV that for both CIFAR-10 and CIFAR-100, ReLaB + ReMixMatch (RMM) outperforms supervised label noise alternatives by reaching lower error rates when training on $\tilde{\mathcal{D}}$.

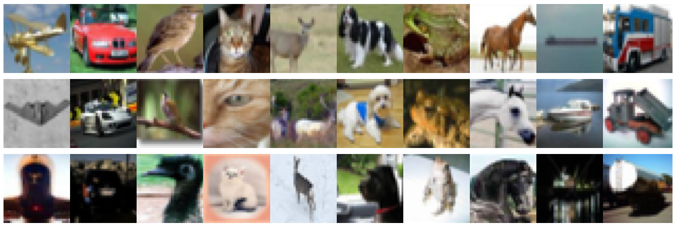

Fig. 2. Labeled samples used for the 1 sample per class study on CIFAR-10 and taken from [19], ordered from top to bottom from most representative to least representative.

TABLE V

ReLaB for semi-supervised learning on CIFAR-10 and CIFAR-100 with very limited amounts of labeled data. Error rates. We mark with † the methods we run ourselves. Other results are from [19] or [26]. Bold denotes best.

	CIFAR-10				CIFAR-100			
Labeled samples	10	40	100	250	100	400	1000	2500
π-model [73]	-	-	-	54.26 ± 3.97	-	-	-	57.25 ± 0.48
MT [15]	-	-	-	32.32 ± 2.30	-	-	-	53.91 ± 0.57
PL [17]†	55.61 ± 5.28	29.65 ± 5.71	12.83 ± 0.68	12.00 ± 0.32	88.23 ± 0.32	67.57 ± 0.58	55.20 ± 0.69	45.42 ± 0.68
MM [14]	-	47.54 ± 11.50	-	11.05 ± 0.86	-	67.61 ± 1.32	-	39.94 ± 0.37
UDA [74]	-	29.05 ± 5.93	-	8.82 ± 1.08	-	-	-	-
RMM [18]†	58.80 ± 1.98	31.36 ± 4.37	22.56 ± 2.58	7.80 ± 0.83	81.18 ± 2.36	57.44 ± 2.53	44.11 ± 1.51	36.66 ± 0.33
EnAET [26]	-	-	9.35	7.60 ± 0.34	-	-	58.73	-
ReLaB + PL†	29.89 ± 3.64	12.38 ± 0.78	11.38 ± 0.64	10.68 ± 0.66	68.04 ± 2.52	53.58 ± 1.20	48.79 ± 0.82	43.84 ± 0.72
ReLaB + RMM†	$\mathbf{22.34 \pm 4.92}$	$\mathbf{8.23 \pm 1.38}$	$\mathbf{6.89 \pm 0.18}$	$\mathbf{6.71 \pm 0.20}$	$\mathbf{62.02 \pm 2.77}$	$\mathbf{44.09 \pm 0.51}$	$\mathbf{39.58 \pm 0.70}$	$\mathbf{35.19 \pm 0.74}$

TABLE VI

Effect of ReLaB on mini-ImageNet with very limited amounts of labeled data and $N_r = 4000$. Error rates.

Labeled samples	100	400	1000	2500
PL [17]	90.89 ± 0.62	85.00 ± 0.94	75.47 ± 0.52	55.10 ± 1.52
ReLaB + PL	$\mathbf{76.25 \pm 0.80}$	$\mathbf{66.66 \pm 0.54}$	$\mathbf{60.82 \pm 1.04}$	$\mathbf{52.39 \pm 1.03}$

TABLE VII

Error rates for 1 sample per class on CIFAR-10 with different labeled sets. We run all the methods ourselves except for FixMatch [19]. Key: MR (Most Representative), LR (Less Representative), NR (Not Representative).

	MR	LR	NR
ReMixMatch [18]	50.62	62.57	90.00
FixMatch [19]	22.00	35.00	90.00
ReLaB + PL	19.86	32.38	79.9
ReLaB + RMM	**8.46**	**21.75**	**78.25**

D. Semi-supervised learning with ReLaB

Table V presents the benefits of ReLaB for semi-supervised learning, showing great improvements for both PL [17] and ReMixMatch (RMM) [18] when paired with ReLaB. Our focus is on very low levels of labeled samples as semi-supervised methods [18] already achieve very good performance with larger numbers of labeled samples. We further study the 1 sample per class scenario in Section IV-E.

Table VI demonstrates the scalability of our approach to higher resolution images by evaluating ReLaB + PL [17] on mini-ImageNet [70]. Due to GPU memory constrains, we use ResNet-18 instead of ResNet-50 to train iMix with an acceptable batch size for the mini-ImageNet experiments.

Further evaluation shows that the proposed method gives a significant performance improvement compared to the next best tested approach [18] (one sided t-test, $p < 0.05$), except for CIFAR10 in the 25 samples per class configuration where the low number of measurements did not allow us to demonstrate a significant difference ($p = 0.07$).

E. Very low levels of labeled samples

The high standard deviation using 1 sample per class ($N_l = 10$) in CIFAR-10 (Table V) motivates the proposal of a more reasonable method to compare against other approaches. To this end, Sohn et al. [19] proposed 8 different labeled subsets for 1 sample per class in CIFAR-10, ordered from more representative to less representative, we reduce the experiments to 3 subsets: the most representative, the least representative, and one in the middle. Figure 2 shows the selected subsets; the exact sample ids are available together with our code for easy reproduction.

Table VII reports the performance for each subset and compares against FixMatch [19] and ReMixMatch [18]. Note that the results obtained for the less representative samples reflect the results that can be expected on average when drawing labeled samples randomly. In the case of the not representative subset, ReLaB enables the semi-supervised learning algorithms to converge better than a random guess. We find that for CIFAR-100 and mini-ImageNet, runs accross different initial labeled samples are more consistent and a comparison to other methods can be made even when drawing the labeled samples at random.

V. Conclusion

ReLaB leverages methods from different vision tasks (image retrieval, self-supervised feature learning, label noise for image classification) to propose an unsupervised bootstrapping of additional labeled samples which can in term be used to enhance any semi-supervised learning algorithm. We demonstrate the direct impact of better unsupervised features for the performance of ReLaB and the relevance of our reliable sample selection. Using the extended amount of supervision of ReLaB's reliable set, we enable semi-supervised algorithms to reach remarkable and stable accuracies with very few labeled samples on standard datasets. The extremely low levels of labeled samples we consider in this paper ($<$ 25 per class) addresses a gap in the semi-supervised literature, which otherwise perform on par with supervised learning for moderate levels of labeled samples ($>$ 25 per class). Direct applications of ReLaB would include scenarios where the annotation of images is very time consuming or requiring expert annotators for example for medical imaging.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number [SFI/15/SIRG/3283] and [SFI/12/RC/2289_P2] as well as from the Department of Agriculture, Food and Marine on behalf of the Government of Ireland under Grant Number [16/RC/3835].

References

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *European Conference on Computer Vision (ECCV)*, 2018.

[2] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv: 1605.07146*, 2016.

[4] D. Damen, H. Doughty, G. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling Egocentric Vision: The EPIC-KITCHENS Dataset," in *European Conference on Computer Vision (ECCV)*, 2018.

[5] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. Shamma, M. Bernstein, and F.-F. Li, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *arXiv: 1602.07332*, 2016.

[6] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang, "YouTube-VOS: Sequence-to-Sequence Video Object Segmentation," in *European Conference on Computer Vision (ECCV)*, 2018.

[7] P. Zhu, L. Wen, X. Bian, L. Haibin, and Q. Hu, "Vision Meets Drones: A Challenge," *arXiv: 1804.07437*, 2018.

[8] A. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] M. Douze, A. Szlam, B. Hariharan, and H. Jegou, "Low-shot learning with large-scale diffusion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *Advances in Neural Information Processing Systems (NeuRIPS)*, 2018.

[11] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] W. Li, L. Wang, W. Li, E. Agustsson, and L. Van Gool, "WebVision Database: Visual Learning and Understanding from Web Data," *arXiv: 1708.02862*, 2017.

[13] "Generative Adversarial Networks and Markov Random Fields for oversampling very small training sets," *Expert Systems with Applications*, 2021.

[14] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "MixMatch: A Holistic Approach to Semi-Supervised Learning," in *Advances in Neural Information Processing Systems (NeuRIPS)*, 2019.

[15] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[16] W. Shi, Y. Gong, C. Ding, Z. MaXiaoyu Tao, and N. Zheng, "Transductive Semi-Supervised Deep Learning using Min-Max Features," in *European Conference on Computer Vision (ECCV)*, 2018.

[17] E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. McGuinness, "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning," *arXiv: 1908.02983*, 2019.

[18] D. Berthelot, N. Carlini, E. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring," in *International Conference on Learning Representations (ICLR)*, 2020.

[19] K. Sohn, D. Berthelot, C.-L. L, Z. Zhang, N. Carlini, E. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence," *arXiv: 2001.07685*, 2020.

[20] Y. M. Asano, C. Rupprecht, and A. Vedaldi, "Self-labelling via simultaneous clustering and representation learning," in *International Conference on Learning Representations (ICLR)*, 2020.

[21] C. Doersch, A. Gupta, and A. Efros, "Unsupervised Visual Representation Learning by Context Prediction," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[22] Z. Feng, C. Xu, and D. Tao, "Self-Supervised Representation Learning by Rotation Feature Decoupling," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[23] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised Representation Learning by Predicting Image Rotations," in *International Conference on Learning Representations (ICLR)*, 2018.

[24] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, 2020.

[25] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European Conference on Computer Vision (ECCV)*, 2016.

[26] W. Xiao, K. Daisuke, L. Jiebo, and Q. Guo-Jun, "EnAET: Self-Trained Ensemble AutoEncoding Transformations for Semi-Supervised Learning," *arXiv: 1911.09265*, 2019.

[27] S.-A. Rebuffi, S. Ehrhardt, K. Han, A. Vedaldi, and A. Zisserman, "Semi-Supervised Learning with Scarce Annotations," *arXiv: 1905.08845*, 2019.

[28] T. Miyato, S. Maeda, S. Koyama, and S. Ishii, "Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017.

[29] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation Consistency Training for Semi-Supervised Learning," in *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2019.

[30] H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in *International Conference on Learning Representations (ICLR)*, 2018.

[31] J. Bridle, A. Heading, and D. MacKay, "Unsupervised Classifiers, Mutual Information and'Phantom Targets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 1992.

[32] L. Dong-Hyun, "Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks," *International Conference on Machine Learning Workshops (ICMLW)*, 2013.

[33] W. Shi, Y. Gong, C. Ding, Z. Ma, X. Tao, and N. Zheng, "Transductive Semi-Supervised Deep Learning Using Min-Max Features," in *European Conference on Computer Vision (ECCV)*, 2018.

[34] Y. Li, L. Liu, and R. Tan, "Certainty-Driven Consistency Loss for Semi-supervised Learning," *arXiv: 1901.05657*, 2019.

[35] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Label propagation for deep semi-supervised learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[36] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision (ECCV)*, 2016.

[37] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[38] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[39] L. Zhang, G.-J. Qi, L. Wang, and J. Luo, "Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] H. Jiabo, D. Qi, G. Shaogang, and Z. Xiatian, "Unsupervised Deep Learning by Neighbourhood Discovery," in *International Conference on Machine Learning (ICML)*, 2019.

[41] K. Lee, Y. Zhu, K. Sohn, C.-L. Li, J. Shin, and H. Lee, "i-Mix: A Strategy for Regularizing Contrastive Representation Learning," in *International Conference on Learning Representations (ICLR)*, 2021.

[42] M. Donoser and H. Bischof, "Diffusion Processes for Retrieval Revisited," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[43] M. Szummer and J. Tommi, "Partially labeled classification with Markov random walks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2002.

[44] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with Local and Global Consistency," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2003.

[45] Y. Bengio, O. Delalleau, and N. Le Roux, "Label propagation and quadratic criterion," Carnegie Mellon University, Tech. Rep., 2006.

[46] K. Kamnitsas, D. Castro, L. Le Folgoc, I. Walker, R. Tanno, D. Rueckert, B. Glocker, A. Criminisi, and A. V. Nori, "Semi-Supervised Learning via Compact Latent Space Clustering," in *International Conference on Machine Learning (ICML)*, 2018.

[47] P. Husser, A. Mordvintsev, and D. Cremers, "Learning by Association - A versatile semi-supervised training method for neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[48] H. Song, M. Kim, D. Park, and J.-G. Lee, "Learning from noisy labels with deep neural network: A survey," *arXiv: 2007.08199*, 2020.

[49] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires re-thinking generalization," in *International Conference on Learning Representations (ICLR)*, 2017.

[50] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Realistic evaluation of deep semi-supervised learning algorithms," in *Advances in Neural Information Processing Systems (NeuRIPS)*, 2018.

[51] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[52] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," *arXiv: 1412.6596*, 2014.

[53] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint Optimization Framework for Learning with Noisy Labels," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[54] K. Yi and J. Wu, "Probabilistic End-To-End Noise Correction for Learning With Noisy Labels," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[55] N. Vyas, S. Saxena, and T. Voice, "Learning Soft Labels via Meta Learning," *arXiv: 2009.09496*, 2020.

[56] Y. Ding, L. Wang, D. Fan, and B. Gong, "A Semi-Supervised Two-Stage Approach to Learning from Noisy Labels," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[57] J. Li, R. Socher, and S. Hoi, "DivideMix: Learning with Noisy Labels as Semi-supervised Learning," in *International Conference on Learning Representations (ICLR)*, 2020.

[58] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, "Early-Learning Regularization Prevents Memorization of Noisy Labels," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[59] E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. McGuinness, "Unsupervised Label Noise Modeling and Loss Correction," in *International Conference on Machine Learning (ICML)*, 2019.

[60] D. Hendrycks, M. Mazeika, D. Wilson, and K. Gimpel, "Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[61] Y. Wang, W. Liu, X. Ma, J. Bailey, H. Zha, L. Song, and S.-T. Xia, "Iterative learning with open-set noisy labels," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[62] L. Jiang, D. Huang, M. Liu, and W. Yang, "Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels," in *International Conference on Machine Learning (ICML)*, 2020.

[63] Y. Kim, J. Yim, J. Yun, and J. Kim, "NLNL: Negative Learning for Noisy Labels," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[64] D. Ortego, E. Arazo, P. Albert, N. O'Connor, and K. McGuinness, "Towards Robust Learning with Different Label Noise Distributions," in *International Conference on Pattern Recognition (ICPR)*, 2020.

[65] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum, "Efficient Diffusion on Region Manifolds: Recovering Small Objects with Compact CNN Representations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[66] G. Tolias, Y. Avrithis, and H. Jégou, "To Aggregate or Not to aggregate: Selective Match Kernels for Image Search," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[67] D. Ortego, E. Arazo, P. Albert, N. E. O'Connor, and K. McGuinness, "Multi-Objective Interpolation Training for Robustness to Label Noise," *arXiv: 2012.04462*, 2020.

[68] Y. Mang, Z. Xu, Y. Pong, and C. Shih-Fu, "Unsupervised Embedding Learning via Invariant and Spreading Instance Feature," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[69] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[70] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching Networks for One Shot Learning," in *Advances in Neural Information Processing Systems (NeuRIPS)*, 2016.

[71] A. Babenko and V. S. Lempitsky, "Aggregating Deep Convolutional Features for Image Retrieval," in *European Conference on Computer Vision (ECCV)*, 2015.

[72] F. Radenovic, G. Tolias, and O. Chum, "Fine-tuning CNN Image Retrieval with No Human Annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[73] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, and T. Raiko, "Semi-Supervised Learning with Ladder Network," in *Advances in Neural Information Processing Systems (NeuRIPS)*, 2015.

[74] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. Le, "Unsupervised Data Augmentation for Consistency Training," *arXiv: 1904.12848*, 2019.