Semi-Supervised Learning under Class Distribution Mismatch

Yanbei Chen[1]**, Xiatian Zhu**[2]**, Wei Li**[1]**, Shaogang Gong**[1]
[1]Queen Mary University of London, [2]Vision Semantics Ltd.
yanbei.chen@qmul.ac.uk, eddy.zhuxt@gmail.com, w.li@qmul.ac.uk, s.gong@qmul.ac.uk

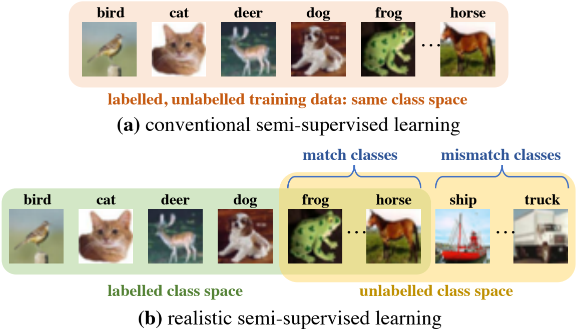
Abstract

Semi-supervised learning (SSL) aims to avoid the need for collecting prohibitively expensive labelled training data. Whilst demonstrating impressive performance boost, existing SSL methods artificially assume that small labelled data and large unlabelled data are drawn from the *same* class distribution. In a more realistic scenario with class distribution mismatch between the two sets, they often suffer severe performance degradation due to error propagation introduced by irrelevant unlabelled samples. Our work addresses this under-studied and realistic SSL problem by a novel algorithm named *Uncertainty-Aware Self-Distillation* (UASD). Specifically, UASD produces soft targets that avoid catastrophic error propagation, and empower learning effectively from unconstrained unlabelled data with out-of-distribution (OOD) samples. This is based on joint *Self-Distillation* and *OOD filtering* in a unified formulation. Without bells and whistles, UASD significantly outperforms *six* state-of-the-art methods in more realistic SSL under class distribution mismatch on *three* popular image classification datasets: CIFAR10, CIFAR100, and TinyImageNet.

Introduction

Deep neural networks (DNNs) often require supervised learning on large-scale labelled training data (LeCun, Bengio, and Hinton 2015). This greatly restricts their generalisability in the limited supervision regime – where labelled data is scare due to high annotation cost and/or prohibitive collecting process. As a remedy to mitigate the labelling overload, semi-supervised learning (SSL) aims for model optimisation with limited labelled and abundant unlabelled training data (Chapelle, Scholkopf, and Zien 2009). This learning paradigm hence brings enormous potential value to a variety of real-world applications, such as medical data analysis (Papernot et al. 2017), image search (Fergus, Weiss, and Torralba 2009), genetics and genomics (Libbrecht and Noble 2015).

Recent developments (Kingma et al. 2014; Rasmus et al. 2015; Miyato et al. 2016; Sajjadi, Javanmardi, and Tasdizen 2016; Laine and Aila 2017; Tarvainen and Valpola 2017; Athiwaratkun et al. 2019; Berthelot et al. 2019) have pushed the limit of SSL drastically, leading to increasingly generalisable DNNs. With a substantial fraction of labels discarded, recent advanced methods (Laine and Aila 2017; Tarvainen and Valpola 2017; Athiwaratkun et al. 2019) can even approach the performance of fully supervised learning. Built upon a *de facto* artificial assumption that labelled and unlabelled training data are drawn from an identical class space (i.e. every unlabelled sample must belong to one of the known classes), existing SSL methods are not practically deployable and scalable. This is because, unlabelled data are unlikely to be manually purified beforehand in many real-world applications for meeting the *de facto* assumption. More probably, unlabelled data are sampled from a class distribution with an *unknown* class mismatch rate against the labelled class distribution. Lacking algorithmic consideration to deal with the class distribution mismatch between labelled and unlabelled data, state-of-the-art SSL algorithms generally suffer severe performance degradation when deployed to such realistic settings, as identified by (Oliver et al. 2018).



(a) conventional semi-supervised learning

(b) realistic semi-supervised learning

Figure 1: **(a)** In conventional semi-supervised learning, both labelled and unlabelled training data come from an identical class distribution. **(b)** In real-world scenario, however, class distribution mismatch often exits between the labelled and unlabelled data.

In this work, we investigate the more realistic and under-studied semi-supervised learning scenario with *class distribution mismatch* between limited labelled and abundant unlabelled data sets. In particular, unlike the conventional SSL setting, we consider the unlabelled data is drawn from a mixture of known and unknown classes (Figure 1). This new problem poses a unique research question mostly ignored in existing SSL literature: *How can we maximise the value of any relevant unlabelled data, given no prior knowledge about whether an unlabelled sample belongs to a known class?*

Compared to conventional SSL, the challenge of realistic SSL is partly due to lacking the separation of known and unknown classes on the unlabelled training data. Together with the notorious ***overconfidence*** issue of deep neural networks (DNNs) (Nguyen, Yosinski, and Clune 2015), it is not surprised that contemporary SSL methods can easily produce corrupted, *overconfident* unsupervised learning signals that incur catastrophic error propagation. For instance, in *entropy minimisation* for SSL (Grandvalet and Bengio 2005; Lee 2013), model predictions are blindly enforced to be "*confident*" (i.e. low-entropy) on unlabelled samples, despite these samples may be unrelated to the target learning task at hand. In *consistency regularisation* (Tarvainen and Valpola 2017), the inherent *overconfident* tendency in DNNs can also reinforce the wrong class assignments of those irrelevant unlabelled samples to the known classes. Therefore, to exploit unconstrained unlabelled data effectively, we address this realistic SSL problem based on two essential algorithmic considerations: (1) self-discover and discard irrelevant unlabelled data on-the-fly; and (2) formulate reliable learning signals that avoid overconfident class assignments.

Specifically, we formulate a generic and novel SSL deep learning algorithm, named **Uncertainty-Aware Self-Distillation** (UASD), which addresses the aforementioned challenge in a systematic end-to-end formulation. UASD specially forms the *soft targets* that serve as *regularisers* to empower more robust semi-supervised learning under class distribution mismatch. Critically, UASD prevents the tendency of *overconfidence* in DNN, a fundamental limitation that existing SSL methods commonly suffer – consequently causing their error propagation and catastrophic degradation in the more realistic SSL setting. This is achieved by formulating a sequence of ensemble models aggregated accumulatively on-the-fly for joint *Self-Distillation* and *OOD filtering*. Unlike existing SSL methods that derived their overconfident learning signals based on the *de facto* assumption, our formulation is aware of the uncertainty of whether an unlabelled sample likely lies in- or out-of-distribution, and selectively learns from the unconstrained unlabelled data.

In summary, our **contribution** is three-fold:

- We study semi-supervised learning under class distribution mismatch – a realistic SSL scenario largely ignored in existing SSL literature. To our knowledge, this work is the first attempt to systematically address this new problem.
- We formulate a novel algorithm, *Uncertainty-Aware Self-Distillation* (UASD), for solving the unique SSL challenge involved in class distribution mismatch. UASD overcomes the *overconfident* issue of DNNs and enables robust SSL under class distribution mismatch.
- We provide extensive benchmarking results in this realistic SSL scenario, including our proposed UASD and *six* representative state-of-the-art SSL methods on *three* image classification datasets: CIFAR10, CIFAR100 and TinyImageNet. Remarkably, UASD outperforms all the strong competitors often by large margins, and demonstrates great potential to exploit the unconstrained unlabelled data.

Related Work

This work is closely connected to three threads of research: **(1)** semi-supervised deep learning, **(2)** OOD detection, and **(3)** knowledge distillation, as briefly introduced below.

Semi-supervised deep learning has attracted increasing attention in recent years (Sajjadi, Javanmardi, and Tasdizen 2016; Laine and Aila 2017; Tarvainen and Valpola 2017; Miyato et al. 2018; Chen, Zhu, and Gong 2018; Athiwaratkun et al. 2019; Wang, Li, and Van Gool 2019; Berthelot et al. 2019), due to its promising aim to alleviate the requirement of large-scale expensive labelled data. The best results on image classification benchmarks are mostly achieved by consistency regularisation, which generally enforces the distributional smoothness under randomisation in input images or model weights. For instance, Temporal Ensembling (Laine and Aila 2017), Mean Teacher (Tarvainen and Valpola 2017) are two representative techniques that keep an *exponential moving average* (EMA) in output or weight space to derive a consistency cost, which let the network produce the likely class assignments on unlabelled data. However, these methods tend to enforce *overconfident* class assignments, regardless of the underlying class distribution. Thus, they are likely to spread the wrong class labels to unlabelled samples lying out-of-distribution. To resolve this issue, we propose to accumulatively aggregate the predictions from a growing amount of networks by *equal averaging*, which leads to much *softer* class assignments for more reliable SSL under class distribution mismatch.

Out-of-distribution (OOD) detection is a task of detecting OOD samples. An intuitive approach is to identify OOD samples based on confidence scores estimated as the maximum softmax probabilities (Hendrycks and Gimpel 2017). However, softmax-based confidence estimate by a *single* DNN can be problematic, as DNNs generally suffer from *overconfidence* (Nguyen, Yosinski, and Clune 2015), e.g. Feeding random noise to a DNN can give rise to a maximal probability score over 99.6%. To address this, one line of research focuses on *confidence calibration* (Liang, Li, and Srikant 2018; Lee et al. 2018; DeVries and Taylor 2018; Hendrycks, Mazeika, and Dietterich 2019) to form *softer* predictive distributions that encompass uncertainty. Rooted in similar spirit, we derive *soft targets* that can serve as an indicator for OOD detection. In particular, we introduce a simple OOD filter to automatically discard OOD samples on-the-fly, without requiring heavy computation cost to train an OOD filter, nor the need of auxiliary OOD training samples.

Knowledge distillation (Hinton, Vinyals, and Dean 2015) aims to transfer the knowledge from an ensemble of multiple DNNs into a single DNN. Typically, a *student* network is supervised by the soft targets generated by averaging the network outputs from a cumbersome whole ensemble of *teacher* networks. Inspired by the recently proposed *online distillation* (Anil et al. 2018; Lan, Zhu, and Gong 2018), our approach particularly aims to exploit the knowledge discovered *on-the-fly* by accumulating predictions of all historic stochastic forward passes. This yields *soft targets* that encode uncertainty for OOD data filtering; and more importantly, produce *less overconfident* and softer class assignments on unlabelled data to avoid catastrophic error propagation.

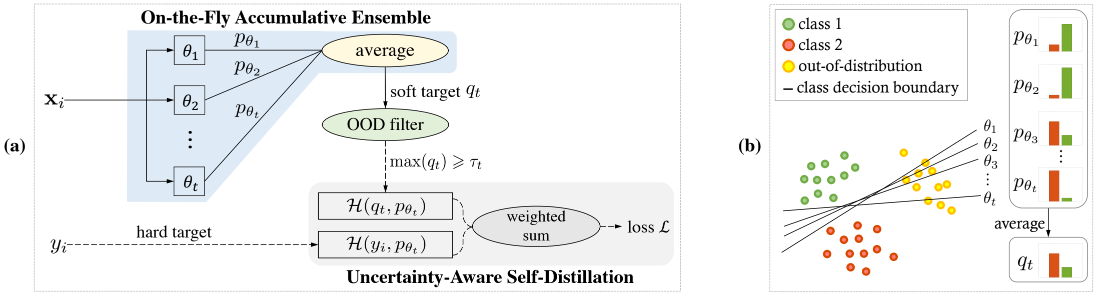


Figure 2: **(a) Approach overview**: The predictions from historic stochastic passes on each sample are accumulatively *averaged* to derive a smooth predictive distribution q_t. For robust SSL, q_t is used for unlabelled training data filtering and uncertainty-aware self-distillation. **(b) Schematic illustration**: If a sample is not consistently assigned to a class by an ensemble of classifiers, it is likely to lie out-of-distribution.

Uncertainty-Aware Self-Distillation

Problem Statement. We consider a realistic semi-supervised learning (SSL) scenario with class distribution mismatch, where we have access to a limited amount of labelled samples $\mathcal{D}_l = \{\mathbf{x}_{i,l}, y_i\}_{i=1}^{N_l}$, and abundant unlabelled samples $\mathcal{D}_u = \{\mathbf{x}_{i,u}\}_{i=1}^{N_u}$. Each labelled sample $\mathbf{x}_{i,l}$ belongs to one of K known classes $\mathcal{Y} = \{y_k\}_{k=1}^{K}$, while any unlabelled sample $\mathbf{x}_{i,u}$ is **not** guaranteed to be one of these K known classes. The class distribution mismatch proportion between $\mathcal{D}_l$ and $\mathcal{D}_u$ is also **unknown**. Our ultimate goal is to exploit useful unlabelled data to boost the learning task at hand. Compared to conventional SSL (Chapelle, Scholkopf, and Zien 2009), this scenario raises a unique challenge on how to mitigate the risk of error propagation mostly incurred by overconfidently assigning out-of-distribution samples to the known classes.

Approach Formulation. To perform SSL under class distribution mismatch, we need to achieve two goals concurrently: (1) minimise the negative influence incurred by irrelevant unlabelled data; and (2) maximise the exploitation of relevant unlabelled data to improve the target learning task. To this end, we propose **Uncertainty-Aware Self-Distillation** (UASD), a unified SSL algorithmic framework that jointly perceives data ambiguity with predictive uncertainty, and produces soft targets as effective regularisers to selectively learn from unlabelled data with mismatched classes.

On-the-Fly Accumulative Ensemble

To formulate UASD, we exploit the generic model ensemble principle (Schapire 1990; Breiman 2001), with an aim to sufficiently reduce model misspecification and yield soft targets as regularisers. The rationale is that a committee of models can cover different regions of the version space (Mitchell 1982), as different models tend to make predictions and mistakes differently. Thus, by aggregating predictions from multiple models, we can not only derive smoother predictive distributions (a.k.a. *soft targets*) that encompass predictive uncertainty, but also produce a stronger model that offers richer knowledge beyond the available class label information. However, training a cumbersome ensemble is computationally expensive. To address this issue, we construct the ensemble model on-the-fly by an accumulation strategy.

Specifically, we exploit a sequence of ensemble models that accumulatively grows the ensemble size on-the-fly. Formally, at the t-th epoch we build an ensemble by aggregating all the historic networks $\{\theta_j\}_{j=0}^{t}$. Given a sample $\mathbf{x}_i$, we derive its ensemble prediction $q_t(y|\mathbf{x}_i)$ by averaging over all the preceding network predictions:

$$q_t(y|\mathbf{x}_i) = \frac{1}{t}\sum_{j=0}^{t-1} p(y|\mathbf{x}_i;\theta_j) \tag{1}$$

where $p(y|\mathbf{x}_i;\theta_j)$ denotes the network prediction at the j-th epoch. It is worth noting that, the stochasticity induced by various data augmentation, batch norm, and network weights in different stochastic passes enables building an ensemble out of an increasing amount of models with diverse decision boundaries. Joining with the training process, we can easily scale up the ensemble size by modulating the network aggregating frequency. Crucially, averaging all historic predictions helps to reduce the bias by cancelling out mistaken and overconfident class assignments made by individual networks. This effectively produces smoother ensemble predictive distributions – a type of *soft targets* that naturally encompass both the predictive uncertainty (Lakshminarayanan, Pritzel, and Blundell 2017) and regularities (Hinton, Vinyals, and Dean 2015) discovered by a very large ensemble of models.

Unlabelled Training Data Filtering

For robust SSL under class distribution mismatch, we leverage the soft targets q_t derived in Eq. (1) as an indicator to discard the potentially irrelevant unlabelled training data. Since q_t reflects the agreement among the historic networks in a frequentist perspective (Dawid 1982), its maximal class probability indicates the best consensus a sample is assigned to a specific class. Accordingly, we define the predictive confidence score on each sample as:

$$c_t(\mathbf{x}_i) = \max(q_t(y|\mathbf{x}_i)) \tag{2}$$

where a lower confidence score $c_t(\mathbf{x}_i)$ reflects higher *predictive uncertainty*, indicating the sample is *likely* to lie out-of-distribution (OOD) and uncorrelated to the core learning task at hand. To minimise the harmful effect incurred by irrelevant unlabelled samples, we define an OOD filter to discard the samples with low confidence scores:

$$f(\mathbf{x}_i;\tau_t) = \begin{cases} 1, & \text{if } c_t(\mathbf{x}_i) \geqslant \tau_t, \quad \text{selected} \\ 0, & \text{if } c_t(\mathbf{x}_i) < \tau_t, \quad \text{rejected} \end{cases} \tag{3}$$

where $f(\mathbf{x}_i; \tau_t)$ specifies a batch-wise binary sample filtering criterion to select samples for model learning based upon a confidence threshold τ_t. The threshold τ_t is often heuristically set, which however, is unsuitable in our context, as it depends heavily on the in-training model with high dynamics. Thus, we dynamically estimate τ_t in a data-driven manner by using the validation set (10% of training data) of known classes as reference. Formally, we compute τ_t as the *average* confidence score on the in-distribution validation samples, and refresh τ_t iteratively per epoch during the course of training.

SSL by Uncertainty-Aware Self-Distillation

To enable semi-supervised learning, we employ the soft target q_t to derive a self-supervision signal, which is imposed as a regulariser to learn additionally from the relevant unlabelled data, i.e. samples likely to be correlated to the learning task. Specifically, we capitalise the rich information encoded in soft targets for model learning, including (1) the regularities among known classes, and (2) the predictive uncertainty. All such information are discovered by on-the-fly accumulative ensembling *without* the need of class labelling. Therefore, the soft targets naturally serve to propagate soft class assignments on the unlabelled data in an *unsupervised* manner.

Formally, motivated by the generic distillation principle (Hinton, Vinyals, and Dean 2015), we consider the soft targets as a kind of teaching signal and formulate the final SSL objective with uncertainty-aware self-distillation as:

$$\mathcal{L} = \mathcal{H}(y_{\text{true}}, p_\theta) + w(t) f(\cdot; \tau_t) \cdot \mathcal{H}(q_t, p_\theta) \qquad (4)$$

where the first term refers to the standard *supervised* cross-entropy loss, computed between the network prediction p_θ and the ground-truth labels y_{true}. The second term is the *unsupervised* uncertainty-aware self-distillation loss, computed as the cross-entropy between p_θ and the soft targets q_t. The OOD filter $f(\cdot; \tau_t)$ (Eq (3)) is aware of uncertainty and used to discard the potentially irrelevant samples of low confidence scores. In the beginning of training, the soft targets may not be sufficiently informative due to lacking diversity in ensembling and reliability in predictions. Thus, for robust model optimisation, we utilise a ramp-up weighting function $w(t)$ to gradually increase the importance of the self-distillation loss. An algorithmic overview is summarised in Algorithm 1.

Algorithm 1 Uncertainty-Aware Self-Distillation (UASD)

```
Require: Labelled data $\mathcal{D}_l = \{\mathbf{x}_{i,l}, y_i\}_{i=1}^{N_l}$. Unlabelled data $\mathcal{D}_u = \{\mathbf{x}_{i,u}\}_{i=1}^{N_u}$.
Require: Trainable neural network $\theta$. Ramp-up weighting function $w(t)$.
for $t = 1$ to max_epoch do
  Refresh confidence threshold $\tau_t$ per epoch.
  for $k = 1$ to max_iter_per_epoch do
    Forward propagation to accumulate network prediction $q_t(y|\mathbf{x}_i)$ (Eq (1)) for every in-batch sample.
    Apply OOD filtering (Eq (2), (3)).
    Update network parameters $\theta$ with loss function Eq (4).
  end for
end for
```

Remarks. Overall, our approach has several unique merits to benefit SSL under class distribution mismatch: **(I)** Instead of computing the ensemble predictions based on *exponential moving average* in the *logit space* as Temporal Ensembling (Laine and Aila 2017), our ensemble predictions are acquired by accumulative ensembling, which keeps an *equal average* in the *prediction space* to derive *less overconfident* and softer class assignments. **(II)** Rather than filtering the unlabelled data using an OOD detector pre-trained on OOD samples, we leverage the *soft targets* as an indicator to discard OOD samples on-the-fly, therefore eschewing the requirement of OOD training data and additional training cost for an OOD filter. **(III)** We integrate *Self-Distillation* and *OOD filtering* in a unified end-to-end training framework, which allows the model to benefit learning from the unconstrained unlabelled data in a more reliable way.

Experiments

Implementation details. For a comprehensive and fair comparison, our experiments are built upon the open-source Tensorflow implementation by Oliver et al. (Oliver et al. 2018). It uses the standard Wide ResNet (Zagoruyko and Komodakis 2016), i.e. *WRN-28-2*, as the base network and Adam optimiser (Kingma and Ba 2014) for training. We revise the default 10-dimensional classification layer to K-dimension, where K is the number of known classes in the labelled data. Unless stated otherwise, all hyper-parameters, the ramp-up function, and training procedures are the same as that of (Oliver et al. 2018). For all comparisons, we report the supervised baseline results using only labelled data for training.

Datasets. We use three image classification benchmark datasets. **(1)** CIFAR10: A natural image dataset with 50,000/10,000 training/test samples from 10 object classes. **(2)** CIFAR100: A dataset of 100 fine-grained classes, with the same amount of training/test samples as CIFAR10. **(3)** TinyImageNet: A subset of ImageNet (Deng et al. 2009) with 200 classes, each of which has 500/50 training/validation images. For all datasets, we resize the images to 32×32.

Compared methods. We compare our method with six representative state-of-the-art SSL methods, including **(1)** pseudo-labels (Lee 2013), **(2)** VAT (Miyato et al. 2016), **(3)** Π-model (Sajjadi, Javanmardi, and Tasdizen 2016; Laine and Aila 2017), **(4)** Temporal Ensembling (Laine and Aila 2017), **(5)** Mean-Teacher (Tarvainen and Valpola 2017), and **(6)** SWA (Athiwaratkun et al. 2019). All these methods introduce an additional unsupervised supervision signal, originally proposed and tested under the conventional SSL setting *without* class distribution mismatch. To preserve the nature of these methods, we replicate them following the procedures as (Oliver et al. 2018) – all share the same network architecture, data augmentation, optimiser and training time. In all experiments, we test each method for five runs with a same set of random seeds to choose the labelled samples. We report the averaged error rate with mean and standard deviation over five runs.

Evaluation on CIFAR10

Evaluation protocol. To simulate more realistic SSL with class distribution mismatch, we construct the unlabelled data

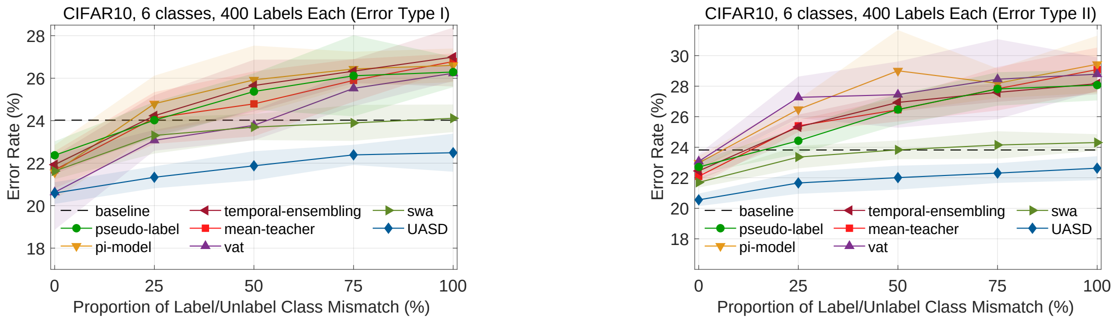


Figure 3: Experiment results of different SSL methods on CIFAR10 under varying class distribution mismatch proportion. **Left** (I): Test error rates are reported at the point of lowest validation error. **Right** (II): Test error rates are reported as the median of last 20 epochs. Shaded area indicates the standard deviation over five runs.

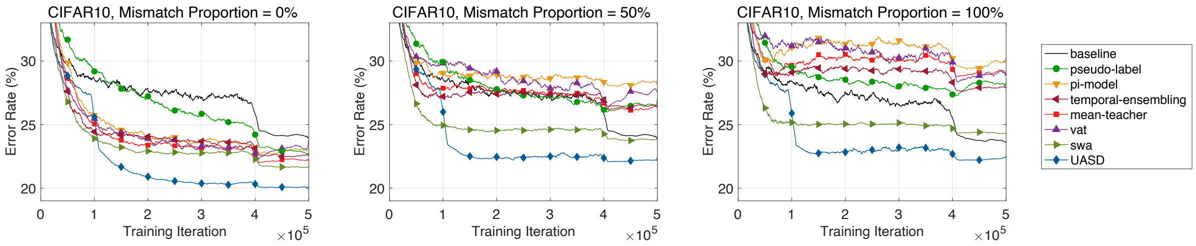


Figure 4: Smoothed learning curves averaged over five runs of different SSL methods on CIFAR10. **Left/Middle/Right** correspond to learning curves under class distribution mismatch proportion of 0/50/100%.

with unknown classes not present in the labelled data. Following (Oliver et al. 2018), we perform experiments on CIFAR10 for a 6-class classification task, using *400 labels per class*. The labelled set contains 6 classes of animals: *bird, cat, deer, dog, frog, horse*; while the unlabelled data comes from 4 classes, with a varying class distribution mismatch proportion from 0% to 100%. For instance, for a mismatch proportion of 50%, the unlabelled data contains classes of *airplane, automobile, frog, horse*. The test errors are reported on the 6 known classes.

Evaluation results. Figure 3 shows experiment results on CIFAR10, including six SSL methods and our UASD under varying class distribution mismatch proportion. The two diagrams (left and right) show the test error rates in two ways: (I) test error rate at the point of the lowest validation error; (II) median test error rate of last 20 epochs. It can be observed that when increasing the amount of unlabelled samples from unknown classes, the performance of most state-of-the-art SSL methods degrade drastically, except SWA (Athiwaratkun et al. 2019), a very recent SSL method that performs weight averaging during training.

Compared to SWA, UASD surprisingly improves the error rates and suffers much less degradation under high mismatch proportions – which indicates its capability to exploit unlabelled data in a more reliable way. Moreover, the error rates of UASD stay consistently in two ways of test error calculation, whilst other methods show more severe performance degradation when reporting the median error rate in the last 20 epochs. This means the other methods commonly suffer unstable degradation at the end of training, while UASD exhibits much more robust convergence.

Learning dynamics analysis. To understand the learning dynamics, we visualise the learning curves in terms of test error rate during training in Figure 4. It is evident that UASD remains superior learning performance compared to other SSL methods under different class distribution mismatch proportions, demonstrating more stable convergence and more reduction in error rate. The relative benefits compared to other SSL methods are more significant under higher class mismatch proportions, e.g. 50%, 100%. This suggests that UASD does yield more reliable supervision signals to guarantee the effectiveness and robustness of SSL under class distribution mismatch.

Evaluation on CIFAR100 and TinyImageNet

Evaluation protocols. We conduct experiments on CIFAR100 and TinyImageNet to evaluate SSL under large class distribution mismatch in larger class space, including three settings as described next.

- On CIFAR100, we use the first half classes (1-50) as labelled classes, and the 25-75 classes as unlabelled classes, leading to a class distribution mismatch proportion of **50%** between labelled and unlabelled data.
- On TinyImageNet, we use the 1-100 classes as labelled classes and the 50-150 classes as unlabelled classes, which results in a mismatch proportion of **50%**.
- We further test in a **cross-dataset** scenario using 100 classes from CIFAR100 as the labelled classes, and 200 classes from TinyImageNet as the unlabelled classes, which gives a mismatch proportion of **86.5%**.

Method	CIFAR100	TinyImageNet	CIFAR100 + TinyImageNet
baseline	39.79 ± 1.19	61.64 ± 0.59	48.31 ± 0.63
pseudo-label	43.30 ± 0.57	62.41 ± 0.57	53.3 ± 0.73
VAT	43.78 ± 1.15	63.75 ± 0.69	50.55 ± 0.55
Π-Model	42.96 ± 0.46	61.79 ± 0.67	53.05 ± 2.21
Temporal Ensembling	41.27 ± 0.76	60.69 ± 0.31	47.88 ± 0.64
Mean-Teacher	40.98 ± 0.98	$\mathbf{60.54 \pm 0.31}$	49.67 ± 1.95
SWA	$\mathbf{37.66 \pm 0.48}$	$\mathbf{57.97 \pm 0.42}$	$\mathbf{44.61 \pm 0.52}$
Ours	$\mathbf{35.93 \pm 0.60}$	$\mathbf{57.15 \pm 0.76}$	$\mathbf{42.83 \pm 0.25}$

Table 1: Results on CIFAR100 and TinyImageNet averaged over 5 runs. Results with reduction in error rate compared to baseline are highlighted in **bold**. Best results are highlighted in **red**.

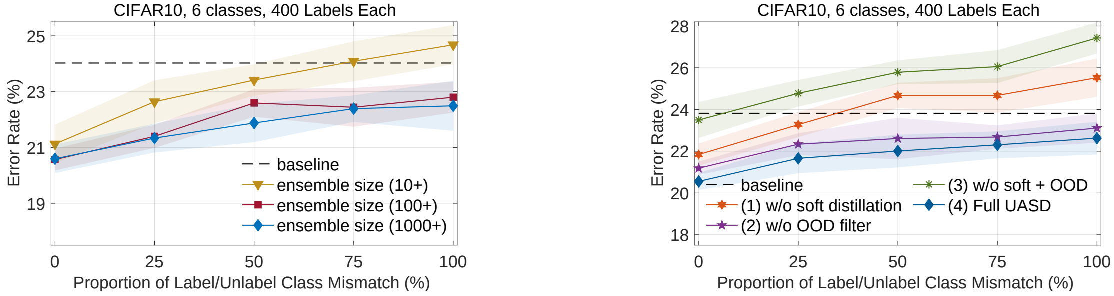


Figure 5: Ablative evaluation (test error rates on CIFAR10). **Left**: Ensemble size analysis. **Right**: Loss formulation analysis.

For all experiments in this section, we use *100 labels per class* and report the test error rate as the median of last 20 epochs to reflect the final convergence.

Evaluation results. Table 1 shows UASD remains remarkably better than other methods when learning under large class distribution mismatch in the finer-grained classification tasks on CIFAR100 and TinyImageNet. While most SSL methods suffer model degradation, UASD consistently outperforms all of them in all settings. It improves upon the supervised baseline with test error reduction of $3.86\%, 4.49\%, 5.48\%$. Crucially, it succeeds even when a large class distribution mismatch proportion (i.e. 86.5%) exists across two datasets (CIFAR100 + TinyImageNet). This shows the efficacy of UASD in exploiting unconstrained unlabelled data coming from *unknown but related classes*, or even *unseen distribution of another dataset*.

Ablative Analysis

To assess different aspects in our algorithmic formulation, we conduct ablative evaluation by changing one individual factor at a time whilst keeping others fixed.

(I) Ensemble size. As aforementioned, the ensemble size is accumulatively growing on-the-fly, which results in a very large ensemble in the end of training (e.g. 1000+ on CIFAR10). To evaluate how the ensemble size affects the model performance, we modulate the ensembling frequency from *per epoch* to *10 epochs* and *100 epochs*, which results in an ensemble size of 100+ and 10+. As Figure 5 (left) shows, the smaller ensemble sizes lead to overall worse performance. This suggests that ensemble size does matter, and aligns with our motivation of building a stronger ensemble model out of an increasing number of networks for deriving smoother and more reliable supervision signals.

(II) Uncertainty-Aware Self-Distillation loss. To evaluate how the loss formulation brings positive benefits, we conduct three ablative experiments: (1) w/o soft targets, which replaces soft targets with one-hot hard targets; (2) w/o OOD filter, which removes the OOD filter and takes all unlabelled data for training; (3) w/o soft + OOD, which uses one-hot hard targets and removes the OOD filter. Figure 5 (right) shows the ablative evaluation on CIFAR10, from which we analyse in two aspects as follows.

(i) Effect of soft targets in Self-Distillation. When replacing the soft targets q_t in Eq.(4) as hard targets, i.e. argmax(q_t), we observe the ablative baseline (1) suffers large performance drops compared to the full model (4). How about learning from all unlabelled data with the soft targets? We find the performance still keeps in a reasonable range – see ablative baseline (2). This means that the soft targets yielded by UASD do provide rich information beyond the label supervision, which enables the network to learn from the unlabelled data in a self-supervised fashion. On one side, rather than blindly fitting to overconfident class assignments, the network is encouraged to align with smoother predictive distributions, thus preserving the predictive uncertainty. On the other side, the soft targets serve to communicate the regularities of smoother class decision boundaries discovered by the preceding networks, therefore allowing to distill the knowledge from a committee of networks.

(ii) Effect of OOD filter. When removing OOD filter, we observe the ablative baseline (2) suffer consistent performance drops compared to the full model (4). When removing the soft

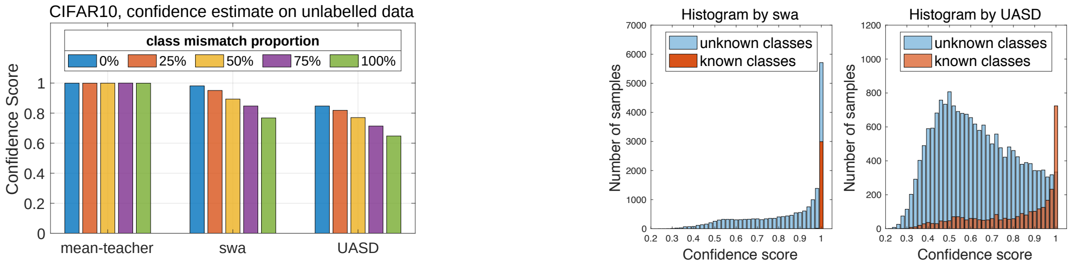


Figure 6: **Left**: Average confidence score on unlabelled data estimated by mean-teacher, SWA, UASD under varying mismatch proportion. **Right**: Histogram of confidence score by SWA, UASD.

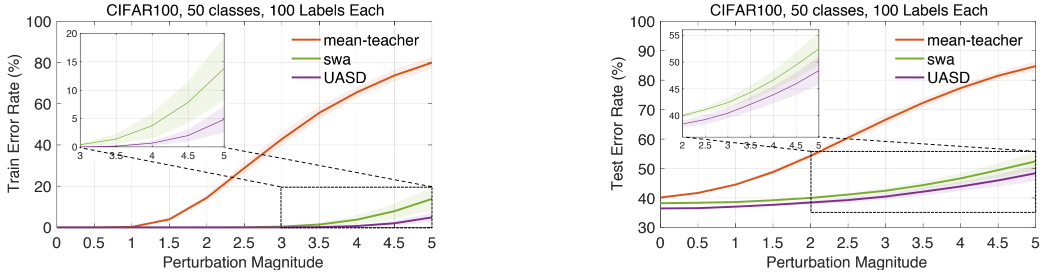


Figure 7: Model robustness under varying magnitude of perturbation on training data (**Left**) and test data (**Right**). Shaded area indicates the standard deviation over five randomly sampled perturbations.

targets and OOD filter concurrently, we observe the worst performance drops – see ablative baseline (3). This indicates that discarding irrelevant unlabelled samples is important, and is especially vital when the unsupervised teaching signals are prone to overconfident, e.g. when using hard targets.

Further Analysis

To further understand why UASD is effective in SSL under class distribution mismatch. We compare three most competitive SSL methods for a more in-depth analysis, namely, (1) mean-teacher, (2) SWA and (3) UASD. We analyse in two different aspects as below.

(I) Confidence calibration. We compare the average confidence score (i.e. *maximum probability*) on the unlabelled data, estimated by the teaching signals in the end of training. In Figure 6 (left), it is evident that teaching signals given by mean-teacher are most *overconfident* – with the same level of *high confidence scores* under varying class distribution mismatch proportion. In contrast, confidence scores estimated by SWA, UASD are stratified to reflect uncertainty of the underlying class distribution. We further compare the histogram of confidence scores by SWA and UASD in Figure 6 (right). It shows UASD can better delimit between data from known and unknown classes. This indicates UASD yields *softer targets* (less overconfident), which are essential to guarantee robust SSL under class distribution mismatch.

(II) Model generalisation. To evaluate the model generalisation, we quantify the model robustness as the shifts of training and test error rates by adding perturbations on the networks. This is based on a well-known finding that convergence to a *wider optimum* typically leads to better *model generalisation* (Keskar et al. 2017; Chaudhari et al. 2017), while the width of optima can be approximately reflected as the model robustness under small perturbation (Izmailov et al. 2018): $\theta(k,p) = \theta + k \cdot p$, where p is the perturbation added on the model θ – a direction vector with unit length drawn from a uniform distribution. We vary the scaling factor k between $[0,5]$ to control the magnitude of perturbation. Figure 7 shows the model robustness against perturbations lie in an order as UASD>SWA>mean-teacher on both training and test data. This suggests UASD does find the widest optimum among the three and thus gives better model generalisability.

Conclusion and Discussion

In this work, we systematically studied the more realistic semi-supervised learning (SSL) under class distribution mismatch, which poses a new challenge of how to maximise the value of unconstrained unlabelled data. To address this challenge, we proposed *Uncertainty-Aware Self-Distillation* (UASD), a novel SSL algorithm that utilises an on-the-fly accumulative ensemble to produce *soft targets* for joint *Self-Distillation* and *OOD filtering*. UASD consistently outperforms *six* state-of-the-art SSL methods on *three* image classification datasets. Although UASD has shown effectiveness in SSL under class distribution mismatch and suggests great value in practical use, we consider there are still several potential research directions to be further explored for addressing this new challenge: (1) tackle the class imbalance induced by class distribution mismatch; and (2) integrate class-incremental learning to cope with the unknown classes. Overall, our new problem setting along with the proposed approach open up many avenues for future research in SSL.

Acknowledgements This work is supported by Vision Semantics Limited, the China Scholarship Council, the Alan

Turing Institute, and Innovate UK Industrial Challenge Project (98111-571149).

References

Anil, R.; Pereyra, G.; Passos, A.; Ormandi, R.; Dahl, G. E.; and Hinton, G. E. 2018. Large scale distributed neural network training through online distillation. In *ICLR*.

Athiwaratkun, B.; Finzi, M.; Izmailov, P.; and Wilson, A. G. 2019. There are many consistent explanations of unlabeled data: Why you should average. In *ICLR*.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. 2019. Mixmatch: A holistic approach to semi-supervised learning.

Breiman, L. 2001. Random forests. *Machine learning*.

Chapelle, O.; Scholkopf, B.; and Zien, A. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*.

Chaudhari, P.; Choromanska, A.; Soatto, S.; LeCun, Y.; Baldassi, C.; Borgs, C.; Chayes, J.; Sagun, L.; and Zecchina, R. 2017. Entropy-sgd: Biasing gradient descent into wide valleys. In *ICLR*.

Chen, Y.; Zhu, X.; and Gong, S. 2018. Semi-supervised deep learning with memory. In *ECCV*.

Dawid, A. P. 1982. The well-calibrated bayesian. *Journal of the American Statistical Association* 77(379):605–610.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

DeVries, T., and Taylor, G. W. 2018. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*.

Fergus, R.; Weiss, Y.; and Torralba, A. 2009. Semi-supervised learning in gigantic image collections. In *NIPS*.

Grandvalet, Y., and Bengio, Y. 2005. Semi-supervised learning by entropy minimization. In *NIPS*.

Hendrycks, D., and Gimpel, K. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. G. 2019. Deep anomaly detection with outlier exposure. In *ICLR*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D.; and Wilson, A. G. 2018. Averaging weights leads to wider optima and better generalization.

Keskar, N. S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; and Tang, P. T. P. 2017. On large-batch training for deep learning: Generalization gap and sharp minima. In *ICLR*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. In *ICLR*.

Kingma, D. P.; Mohamed, S.; Rezende, D. J.; and Welling, M. 2014. Semi-supervised learning with deep generative models. In *NIPS*.

Laine, S., and Aila, T. 2017. Temporal ensembling for semi-supervised learning. In *ICLR*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*.

Lan, X.; Zhu, X.; and Gong, S. 2018. Knowledge distillation by on-the-fly native ensemble. In *NIPS*.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature*.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NIPS*.

Lee, D.-H. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*.

Libbrecht, M. W., and Noble, W. S. 2015. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*.

Mitchell, T. M. 1982. Generalization as search. *Artificial intelligence*.

Miyato, T.; Maeda, S.-i.; Koyama, M.; Nakae, K.; and Ishii, S. 2016. Distributional smoothing with virtual adversarial training. In *ICLR*.

Miyato, T.; Maeda, S.-i.; Ishii, S.; and Koyama, M. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*.

Oliver, A.; Odena, A.; Raffel, C. A.; Cubuk, E. D.; and Goodfellow, I. 2018. Realistic evaluation of deep semi-supervised learning algorithms. In *NIPS*.

Papernot, N.; Abadi, M.; Erlingsson, U.; Goodfellow, I.; and Talwar, K. 2017. Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*.

Rasmus, A.; Berglund, M.; Honkala, M.; Valpola, H.; and Raiko, T. 2015. Semi-supervised learning with ladder networks. In *NIPS*.

Sajjadi, M.; Javanmardi, M.; and Tasdizen, T. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NIPS*.

Schapire, R. E. 1990. The strength of weak learnability. *Machine learning*.

Tarvainen, A., and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*.

Wang, Q.; Li, W.; and Van Gool, L. 2019. Semi-supervised learning by augmented distribution alignment. In *ICCV*.

Zagoruyko, S., and Komodakis, N. 2016. Wide residual networks. In *BMVC*.