

# AuxMix: Semi-Supervised Learning with Unconstrained Unlabeled Data

Amin Banitalebi-Dehkordi, Pratik Gujjar, and Yong Zhang

Huawei Technologies Canada Co., Ltd.

Vancouver, BC, Canada

{amin.banitalebi, pratik.gujjar, yong.zhang3}@huawei.com

## Abstract

*Semi-supervised learning (SSL) has seen great strides when labeled data is scarce but unlabeled data is abundant. Critically, most recent work assume that such unlabeled data is drawn from the same distribution as the labeled data. In this work, we show that state-of-the-art SSL algorithms suffer a degradation in performance in the presence of unlabeled auxiliary data that does not necessarily possess the same class distribution as the labeled set. We term this problem as Auxiliary-SSL and propose AuxMix, an algorithm that leverages self-supervised learning tasks to learn generic features in order to mask auxiliary data that are not semantically similar to the labeled set. We also propose to regularize learning by maximizing the predicted entropy for dissimilar auxiliary samples. We show an improvement of 5% over existing baselines on a ResNet-50 model when trained on CIFAR10 dataset with 4k labeled samples and all unlabeled data is drawn from the Tiny-Imagenet dataset. We report competitive results on several datasets and conduct ablation studies.*

## 1. Introduction

Collecting and annotating large amounts of labeled data remains to be a fundamental barrier in exploiting the full potential of machine learning algorithms. Not only is it time consuming and costly, in some cases (e.g. rare medical conditions) it might not even be possible. To mitigate these issues, there are a number of semi-supervised learning (SSL) algorithms proposed in the literature which aim to supplement a small set of available labeled data with a larger set of unlabeled examples during training [5, 9, 37, 42, 47]. Unlabeled data, when properly used in conjunction with the available labeled data, can greatly improve the learning generalization [4, 41].

The majority of the existing semi-supervised learning approaches assume that the unlabeled data is drawn from the same distribution as the labeled training dataset. In many cases where the unlabeled data is cheaply available,

this does not impose any problems. In fact, many state-of-the-art (SOTA) SSL algorithms can be used to reduce the need for manual labeling of the data. However, this is a strong assumption, and will not necessarily always hold. In line with the observation made by Oliver et al., [30], we will show that existing SSL algorithms suffer a considerable performance drop if the unlabeled dataset is chosen from sources other than the same distribution as the labeled set. For example, FixMatch- with 95.7% top-1 classification accuracy when using 4K CIFAR10 labeled and 46K CIFAR10 unlabeled examples [37], drops down to only 58.48% accuracy when using 4K CIFAR10 labels but 100K Tiny-Imagenet [22] unlabeled examples (more evidence and results in Section 5). The significant drop is due to the unlabeled distribution mismatch in the two cases. Intuitively this is also expected as any information (such as pseudo-labels) about the unlabeled data queried from a model trained with a different training label distribution will not be reliable. In addition, in some applications where collecting unlabeled data in large portions may not be possible, such data are available for related applications (e.g. neck x-ray vs knee x-ray). This necessitates new solutions to address the aforementioned problems.

In this paper, we formalize this problem and introduce it as a new paradigm termed Auxiliary-SSL. We refer to the unlabeled data from unconstrained distributions as auxiliary data to distinguish it with the in-distribution unlabeled data settings used in existing SSL approaches (See Figure 1 for a visualization). In addition, we propose AuxMix, an algorithm of mixing the labeled and auxiliary data in an efficient way that reduces the degradation gap caused in Auxiliary-SSL.

The main contributions of this paper can be summarized as follows:

- We argue that the existing semi-supervised learning algorithms will under-perform, sometimes at big margins, if they use unlabeled data from unconstrained distributions (auxiliary data), and confirm this empirically on multiple datasets.

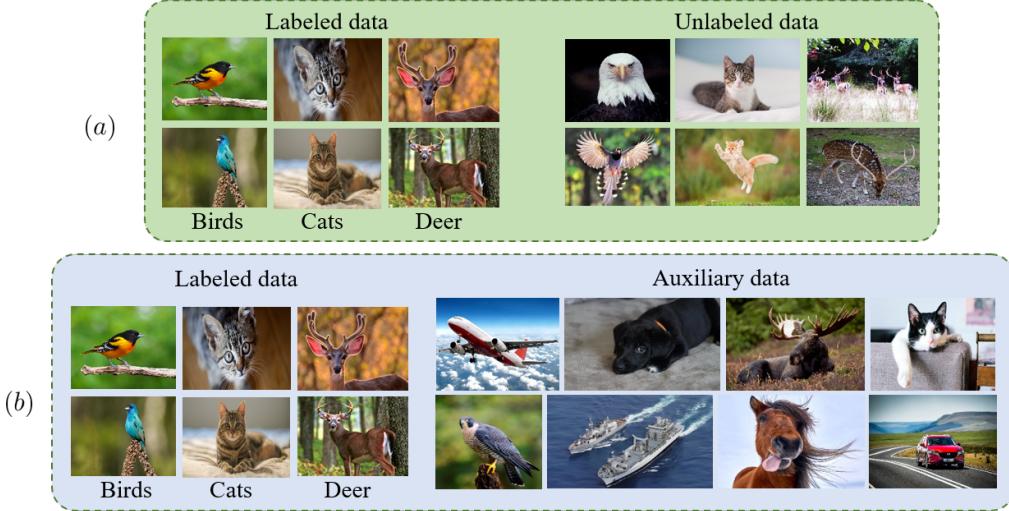


Figure 1. A visualization of unlabeled versus auxiliary data in the semi-supervised setting. (a) Unlabeled data is from the same distribution as the labeled data. (b) Unlabeled data contains out-of-distribution data.

- We establish this paradigm as Auxiliary-SSL, and provide a formal definition for it.
- We propose AuxMix, a method for tackling the Auxiliary-SSL. Our method makes use of self-supervised tasks to learn an initial representation of the auxiliary data, with which we then mask the auxiliary set. We also propose a technique of entropy maximization for the dissimilar negative auxiliary samples. Experimentally, we confirm that our method is effective in reducing the performance drop caused by the label distribution mismatch of the Auxiliary-SSL.

The rest of the paper is organized as follows. Section 2 provides a brief summary of the related literature. Section 3 formalizes the Auxiliary-SSL paradigm. Section 4 describes our AuxMix method in details. Section 5 contains the experiment results and discussions around the observations. Finally, Section 6 concludes the paper.

## 2. Related Work

In this section, we provide a brief summary of literature related to our work. Broadly, we look at state-of-the-art semi-supervised learning approaches, some recent works that investigate label distribution mismatch and open-set domain adaptation methods.

### 2.1. Semi-supervised learning

There is an extensive variety of literature on semi-supervised learning algorithms, especially after the boom of deep learning [41]. Among them, pseudo-label based methods [1, 6, 25, 34, 35, 38] train a model on the existing labeled

data and then use this model to generate pseudo-labels on the unlabeled data, which will later be used for additional training. Another emerging direction is to leverage self-supervised learning algorithms such as RotNet [14], Jig-Saw [29], SimCLR [10], or MOCO [16] for unsupervised pretraining and then fine-tune with the limited labeled set [11, 19]. Furthermore, a number of recent SOTA methods rely on regularizing consistency in predictions across different augmentations [3].  $\Pi$ -model retains an exponential-moving-average (EMA) of the predictions [21] whereas Mean Teacher proposes to average model parameters instead [39, 40]. In an other line of work, Virtual Adversarial Training (VAT) defines adversarial perturbations as input augmentations [27]. [43, 48] in their work enforce a consistency between convex interpolations of unlabeled samples and their similarly interpolated predictions. Using MixUp [48] as their augmentation procedure, MixMatch [9] and ReMixMatch [8] also perform consistency regularization. Our work is also closely related to UDA [45] and FixMatch [37]. Both of these methods use predictions from a weakly augmented sample to regularize prediction on a strongly augmented sample. ReMixMatch [8] also follows a similar procedure along with Distribution Alignment and Augmentation Anchoring. UDA further sharpens the predicted distribution whereas FixMatch uses pseudo-labels instead. These methods rely on confidence-based thresholding to filter spurious predictions. As mentioned in Section 1, in the presence of unconstrained auxiliary data, the traditional semi-supervised approaches suffer a considerable performance drop due to improper filtering of auxiliary data coupled with unreliable predictions for the out-of-

distribution unlabeled set.

## 2.2. Class distribution mismatch

As discussed in section 1, the problem of class distribution mismatch between labeled and unlabeled data is more practically observed. Efforts in this area however, are limited. Oliver et al., [30] in their paper draw forth several scenarios for evaluating semi-supervised learning. They show that adding unlabeled data from a mismatched set of classes can actually degrade the performance compared to not using them at all. RealMix [28] tackles this problem with consistency training and Entropy Minimization. They show their results using 400 labeled samples per class from CIFAR-10 Animals subset with the rest of CIFAR-10 classes as the mismatched unlabeled set. Chen et al., [12] propose Uncertainty-Aware Self-Distillation (UASD) that uses soft-targets and self-distillation to combat class mismatch. DS3L [15] provide a theoretical guarantee that their algorithm never learns worse than with only using labeled data. They showcase their results on MNIST and CIFAR10 datasets with upto 60% mismatch. In this paper, we formalize this problem as Auxiliary-SSL and refer to the unconstrained unlabeled data as auxiliary set. We report improved results when using Tiny-Imagenet, Caltech-UCSD Birds (CUB) [44] and noise images drawn from a uniformly random distribution as auxiliary datasets.

## 2.3. Open-set domain adaptation techniques

Domain adaptation (DA) methods also have a similar setup to Auxiliary-SSL. The goal of DA is to make accurate predictions on the target domain having learnt from labeled data in the source domain. In contrast, Auxiliary-SSL tries to enhance learning from limited labeled data in the source domain itself by leveraging related data from an unlabeled dataset. Also, traditional (closed-set) domain adaptation deals with the problem of a distribution shift between the source and the target domain but assumes that the labels remain fixed [7], which is different than our Auxiliary-SSL setting. That being said, open-set domain adaptation does consider the case of label distribution mismatch between

the source and target domain datasets [24, 31, 36]. However unlike the Auxiliary-SSL unconstrained paradigm, these methods still impose some conditions on the unlabeled dataset used for adaptation. For example, they usually assume a set of known target classes along with a set of unknown categories. In addition, the evaluation of such methods is usually in controlled scenarios where some classes of a dataset are considered only for the target dataset. On the contrary, in our method, we relax these constraints (even experiment with training with pure noise in Section 5), and allow the auxiliary set to be freely selected. To cope with the potential problems of this design choice, we propose a method to filter the auxiliary examples and regularize the negatively masked samples.

## 3. Problem Setting

We formulate our problem of learning from auxiliary unlabeled data following that of existing semi-supervised learning (SSL) work [30]. A key change in our work is the more general assumption that the data from the unlabeled set does not share the same class distribution as the labeled set. Therefore, a classifier trained on the labeled set cannot trivially generate labels for the unlabeled set. We denote this paradigm as Auxiliary-SSL, and propose AuxMix as a method to tackle this problem. We refer to the unlabeled set as the auxiliary set with samples drawn from an auxiliary distribution.

More formally, we are given a labeled set  $D_l = \{(x_i^l, y_i^l) \sim p\}_{i=1}^{n_l}$  sampled i.i.d. from an unknown data generating distribution  $p(X, Y_l)$  and an auxiliary set  $D_a = \{(x_i^a) \sim q_x\}_{i=1}^{n_a}$  sampled i.i.d from distribution  $q(X)$  which is a marginalization of the data distribution over labels in the auxiliary domain  $Y_a$ . The shared label set between the two distributions  $Y_s = Y_l \cap Y_a \neq \emptyset$  and the auxiliary set has private labels  $Y_p = Y_a \setminus Y_s$ .

It is worth noting that in general, the relationship between two label distributions  $Y_l$  and  $Y_a$  can be characterized in the following different ways as shown in Figure 2: (a) The label set  $Y_l$  could have a partial overlap with  $Y_a$ , (b) no overlap between  $Y_a$  and  $Y_l$ , and (c) a complete overlap

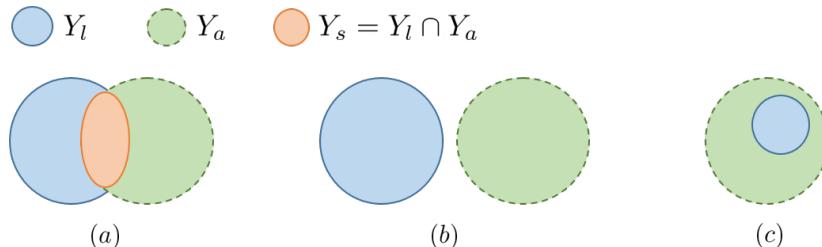


Figure 2. Venn diagram illustrating the relationship between two labeled and auxiliary distributions: (a) partial overlap, (b) no overlap, and (c) complete overlap.

between  $Y_l$  and  $Y_a$  reverting to standard semi-supervised learning setting. In this work, we consider the first two cases. The objective of our algorithm is therefore to learn a model  $f_\theta$  that maximizes the classification performance on the source labeled set  $Y_l$  by distinguishing between auxiliary samples in the shared label set  $Y_s$  and the auxiliary private label set  $Y_p$ .

## 4. Method

Our algorithm, dubbed AuxMix, consists of two phases. In phase one, we learn to score each auxiliary data sample  $\{(x_i^a) \sim D_a\}_{i=1}^{n_a}$  according to their semantic similarity with labeled data and separate them into a positive selection set  $D_a^+$  and a negative regularization set  $D_a^-$ . In phase two, the positive set  $D_a^+$  is used in semi-supervised learning based on consistency regularization across augmentations [37]. We further regularize learning from auxiliary data by maximizing the entropy of predictions on the negative set  $D_a^-$  to penalize confident output distributions [32]. Intuitively, the positive set encourages reliable predictions for a meaningful consistency regularization. On the other hand, the model is required to be unable to decide which class labels to assign the samples in the negative set. We do this by pushing the model towards high uncertainty for these data samples.

### 4.1. Scoring and pre-training based on self-supervised learning

We train a model  $f_\theta^p$  consisting of an encoder  $g^p(\cdot)$  and a classifier head  $c^p(\cdot)$ . In the limited labeled data setting of Auxiliary-SSL, we employ self-supervised learning to pre-train the model on a self-supervised prediction task. Self-supervised tasks allow the model to learn general representations from the data [18]. In the absence of ground-truth labels for the auxiliary samples, these representations can be exploited to gain useful insights about the samples. There are several self-supervised learning tasks in recent literature that have shown promising results in the complete absence of labels as discussed in section 2. In AuxMix, we leverage rotation prediction following the work of Gidaris et al., [14] that shows that by predicting the rotations for each image following their augmentation, the model learns to extract semantic features of objects contained in them. These semantic features become key to measuring similarity between the images of the auxiliary set and the labeled set at a later stage. However, other self-supervised approaches could also be utilized.

Given a set of training images  $\{x_i^p \sim D_l \cup D_a\}_{i=1}^{n_l+n_a}$  the model  $f_\theta^p$  is first trained to minimize the rotation prediction loss  $\mathcal{L}_{rot}$  defined as:

$$\mathcal{L}_{rot} = H(\phi, f_\theta(Rot(x_i^p, \phi))) \quad (1)$$

$$\forall \phi \in [0^\circ, 90^\circ, 180^\circ, 270^\circ],$$

where,  $Rot(x_i, \phi)$  is an augmentation operator that rotates the images by the specified angle  $\phi$  and  $H(y, f(x))$  is the standard cross-entropy loss between a target  $y$  and a prediction  $f(x)$ . We then calculate a benchmark semantic representation or a prototype for each class  $c$  in  $Y_l = [1, \dots, C]$  by using the labeled set  $(x_i^l, y_i^l)$ . Each prototype  $\{h_k\}_{k=1}^C$  is a mean vector of the output of the encoder  $g^p(\cdot)$  for all input samples  $x_i^l$  in the labeled set and class label  $y_i^l = k$ :

$$h_k = \frac{1}{N_k} \sum_{i=1}^{n_l} g^p(x_i^l) \quad (2)$$

Every sample in the auxiliary dataset is then compared with each prototype to yield an affinity score  $\{a_i\}_{i=1}^{n_a}$ . We use cosine similarity as the distance measure between the two vectors. Other distance measures such a Euclidean distance could also be used in place of cosine similarity.

$$a_i = \max_{k=[1, \dots, C]} \{sim(g^p(x_i^a), h_k)\} \quad (3)$$

These scores show how similar auxiliary examples are to the labeled data prototypes, or their *affinity* to the labeled data distribution.

### 4.2. Consistency regularization and entropy maximization

A threshold  $\tau$  on the affinity scores can be used to separate the auxiliary set into a selection set  $D_a^+$  and regularization set  $D_a^-$ . We set the threshold according to the equation  $\tau = \mu + \beta\sigma$  where  $\mu$  and  $\sigma$  are the mean and standard deviation of the affinity score distribution  $\{a_i\}_{i=1}^{n_a}$  and  $\beta$  is a hyperparameter:

$$D_a^+ = \{x_i^a\} \mid a_i \geq \tau \quad (4)$$

$$D_a^- = \{x_i^a\} \mid a_i < \tau$$

Enforcing consistency in predictions across different augmentations of the same input image is a commonly used method in semi-supervised learning [8, 9, 37]. It encourages a label-transfer from the limited labeled data to the vast amount of in-distribution unlabeled data in general semi-supervised learning settings. We follow the same principle of consistency regularization for the labeled set  $D_l$  and the selection set  $D_a^+$ . In particular, we implement the strategy used in methods such [8, 37, 45], which consider a weak and a strong augmentation of an input sample denoted by  $\alpha(\cdot)$  and  $\mathcal{A}(\cdot)$  respectively. The pre-trained encoder from the previous step  $g^p(\cdot)$  is now attached to a new  $C$ -class classifier head  $c^T(\cdot)$  to form our target model  $f_\theta^T$ . Weakly augmented samples from the labeled set are used to train  $f_\theta^T$  with the standard supervised cross-entropy loss  $\mathcal{L}_x$  over

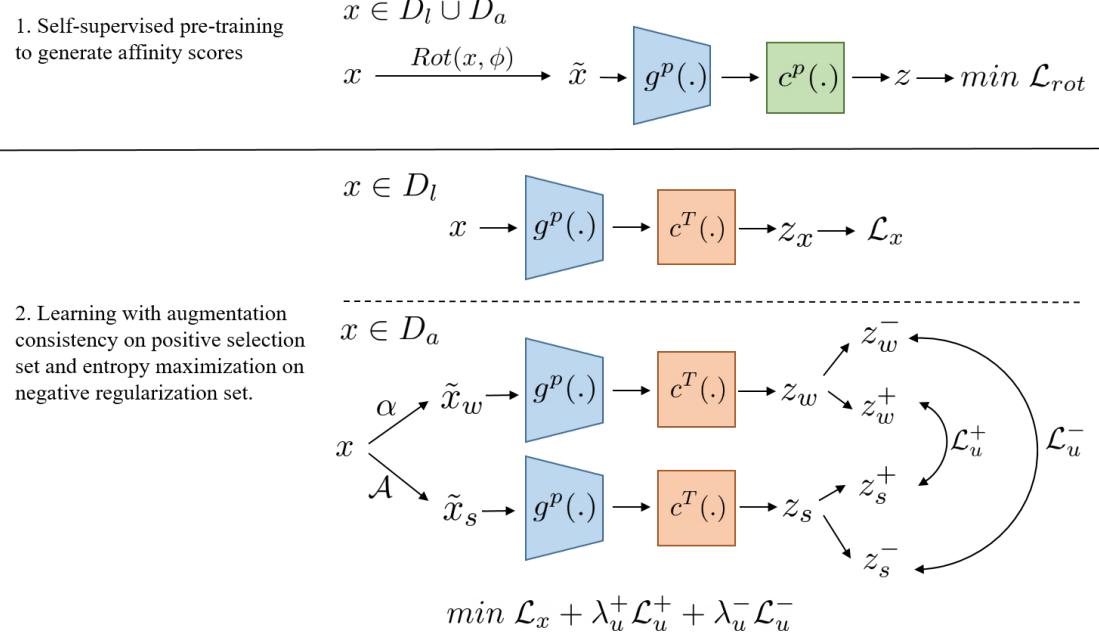


Figure 3. Visual description of the proposed AuxMix method. In Phase 1, we pre-train a model on self-supervised rotation and use the representations to generate affinity scores for the auxiliary dataset. In phase 2, we perform consistency regularization on the positive samples and entropy maximization on the negative regularization set at the same time as learning from the limited labeled data. The overall loss will therefore have three terms.

a batch  $B$  of labeled samples as:

$$\mathcal{L}_x = \frac{1}{B} \sum_{i=1}^B H(y_i^l, f_\theta^T(\alpha(x_i^l))) \quad (5)$$

Moreover, hard pseudo-labels are generated only for the weakly-augmented versions of samples for data from the selection set  $D_a^+$  rather than all of the unlabeled set. The pseudo-labels  $\hat{y}_i$  are used to enforce consistency on the strongly-augmented versions of the same input sample by minimizing the loss  $\mathcal{L}_u^+$ . Following [37] during training, for a batch  $B$  of labeled data,  $\mu B$  batches of unlabeled data are sampled and used.

$$\begin{aligned} \hat{y}_i &= \arg \max (f_\theta^T(\alpha(x_i^a))) \\ \mathcal{L}_u^+ &= \frac{1}{\mu B} \sum_{i=1}^{\mu B} H(\hat{y}_i, f_\theta^T(\mathcal{A}(x_i^a))) \quad \forall (x_i^a) \in D_a^+. \end{aligned} \quad (6)$$

Learning from the pseudo-labels of the selection set minimizes the entropy on the source label set  $Y_l$ . On the other hand, the data from the regularization set is used to regularize the model by maximizing the entropy of the output predictions. This min-max entropy optimization benefits the model by contrasting positive samples from the selection

set against the negative samples from the regularization set. Positive samples are relatively reliable, so they are used in the semi-supervised setting with consistency regularization. On the other hand, negative samples are not as reliable, so they are used to push the model to be unable to decide which class labels to assign to these samples. We penalize prediction confidences by assigning a uniform distribution over all classes as the target distribution for rejection samples. The loss  $\mathcal{L}_u^-$  is then defined as:

$$\begin{aligned} p(\hat{y}_i = k) &= \frac{1}{|Y_L|} \quad \forall k \in Y_L \\ \mathcal{L}_u^- &= \frac{1}{\mu B} \sum_{i=1}^{\mu B} \{H(\hat{y}_i, f_\theta^T(\alpha(x_i^a))) + H(\hat{y}_i, f_\theta^T(\mathcal{A}(x_i^a)))\} \\ &\quad \forall (x_i^a) \in D_a^-. \end{aligned} \quad (7)$$

Finally, the net loss minimized by the model is  $\mathcal{L} = \mathcal{L}_x + \lambda_u^+ \mathcal{L}_u^+ + \lambda_u^- \mathcal{L}_u^-$ , where  $\lambda_u^+$  and  $\lambda_u^-$  are loss weights for the selection set and the regularization set respectively. AuxMix maintains an Exponential Moving Average (EMA) [40] of the parameters  $\theta$  of the model for stable training and inference in both phases.

## 5. Experiments

In this section we review and discuss the experiment results, as well as ablation studies.

### 5.1. Experiments settings

We evaluate the performance of AuxMix on the standard semi supervised learning benchmark dataset, CIFAR10 [20]. The unlabeled data however, is drawn from entirely different datasets: (1) Tiny-ImageNet, (2) Caltech-UCSD Birds - 2011 (CUB) [44], and (3) Noise images sampled from a uniform random distribution, to constitute our auxiliary set. We compare our results for each of these datasets with the recent SOTA SSL methods: FixMatch [37] and MPL [33]. In addition, we report on several other SSL methods previously evaluated by [28] and [30] on the Animals-vs-Others class split of the CIFAR10 dataset. We also perform results of ablations to determine the contribution of each component of our algorithm.

### 5.2. Training details

We use RandAugment [13] with parameters as defined in [37] as the strong augmentation and random horizontal flips and random crops as the weak augmentation for AuxMix. In all of our experiments, we set the following hyper-parameters:  $\lambda_u^+ = 1$ ,  $\lambda_u^- = 1$ ,  $B = 64$ ,  $\mu = 7$ , ema-decay = 0.999, weight-decay=5e-4, rotation learning rate phase 1 = 0.1, learning rate phase 2 = 0.03, and cosine learning rate scheduling. All experiments including baselines are run for 300k iterations and we report the best score on 6k test set for CIFAR10-Animals-Others and 10k standard test set for all other experiments. The threshold  $\tau$  is varied by changing the hyperparameter  $\beta$  and the best results are reported in the tables.

### 5.3. Distribution mismatch results on CIFAR10 Animals-vs-Others

We follow the distribution mismatch experiment settings as described in [30] which splits the CIFAR10 dataset into 400 samples of the six animals classes as the labeled set and the rest of the data (20000 samples) from the four other classes as the auxiliary set. We report the results for our method and other baselines on Wide ResNet 28-2 [46] in Table 1. AuxMix at 14.12% shows 3.5% lower error rate than RealMix and over 8% lower than supervised training with the labeled set only. It is interesting to note that most methods show a higher error rate than supervised learning due to the unlabeled data being out-of-distribution. That means unsuitable unlabeled data may in fact damage the learning.

Method	Error rate (%)
Supervised	22.47
Temporal Ensembling [21]	27.02
Mean Teacher [40]	26.81
VAT [26]	26.19
Pseudo-Label [23]	25.94
SWA [2]	24.10
UASD [12]	22.47
RealMix [28]	17.62
AuxMix (Ours)	<b>14.12</b>

Table 1. Error rate on distribution mismatch between CIFAR10-Animals as labeled set and CIFAR10-Others as unlabeled set on Wide-ResNet-28-2. Lower is better.

### 5.4. Tiny-Imagenet, CUB and noise as auxiliary data

Next, we investigate the impact of changing the auxiliary dataset. For the auxiliary set, we use 100k images across 200 classes from the training set of Tiny-Imagenet, 11788 images from CUB, and 50k noise samples drawn from a random uniform distribution, in three sets of experiments. We use a 4K labeled random subset of CIFAR10 as the labeled set. Moreover, we use a ResNet-50 [17] model to compare with previous work. All baselines are trained for 300k iterations except pseudo-labelling and supervised training, which we run for 100k iterations. Other hyper-parameters for each method are set according to the descriptions in their respective papers. Table 2 contains results from this experiment. We observe from this table that AuxMix performs competitively in comparison to other baselines, across different auxiliary datasets. AuxMix is close to 10% better than FixMatch and close to 19% better than MPL when TinyImagenet is used as the auxiliary dataset. It is interesting to note that pseudo-labelling performs very well when compared to the recent SSL baselines in most cases. In addition, as expected, using auxiliary data from outside label distributions in some baselines results in a performance lower than that of the supervised model.

### 5.5. Ablation study

Table 3 shows an ablation study on AuxMix, where we study the contributions of the entropy-maximization and sample masking components of our method. For this study, we used the CIFAR10 Animals-vs-Others experiment settings. As shown in Table 3, each component on its own achieves a reasonable accuracy, but together the highest performance is achieved.

Method	Accuracy (%)		
	Tiny-Imagenet	CUB	Noise
Supervised-4K	59.91	59.91	59.91
Pseudo-Label [23]	63.04	61.80	64.89
MPL [33]	49.18	68.86	52.44
FixMatch [37]	58.48	49.23	65.88
<b>AuxMix (Ours)</b>	<b>68.38</b>	<b>73.95</b>	<b>69.34</b>

Table 2. Classification accuracy for the 4K CIFAR10 experiment with the test set from CIFAR10 test set, and using entirely different datasets as unlabeled auxiliary data.

Method	Accuracy (%)
Supervised	77.53
AuxMix	85.88
AuxMix without Entropy Maximization	85.10
AuxMix without Sample Masking	84.57

Table 3. Ablation study on AuxMix with CIFAR10 Animals-vs-Others experiment. We observe that masking and regularization both contribute to the end-to-end performance.

## 6. Conclusion

In this paper we first provided empirical evidence as well as intuitions that unlabeled data from unconstrained distributions can considerably damage the semi-supervised learning accuracy of existing SSL methods. Then we formalized this SSL setting as a new Auxiliary-SSL paradigm. Moreover, we proposed an algorithm called AuxMix to tackle the issues raised in Auxiliary-SSL. Experimental evaluations showed that AuxMix can achieve a competitive performance in recovering the performance drops occurred when using unconstrained unlabeled data.

## References

- [1] Eric Arazo, Diego Ortego, et al. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. [2](#)
- [2] Ben Athiwaratkun, Marc Finzi, et al. Improving consistency-based semi-supervised learning with weight averaging. *CoRR*, abs/1806.05594, 2018. [6](#)
- [3] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, pages 3365–3373, 2014. [2](#)
- [4] Amin Banitalebi-Dehkordi. Knowledge distillation for low-power object detection: A simple technique and its extensions for training compact models using unlabeled data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 769–778, 2021. [1](#)
- [5] Amin Banitalebi-Dehkordi. Revisiting knowledge distillation for object detection. *arXiv preprint arXiv:2105.10633*, 2021. [1](#)
- [6] Amin Banitalebi-Dehkordi and Yong Zhang. Repaint: Improving the generalization of down-stream visual tasks by generating multiple instances of training examples. In *the 32nd British Machine Vision Conference*, 2021. [2](#)
- [7] Shai Ben-David, John Blitzer, Koby Crammer, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007. [3](#)
- [8] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, et al. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *CoRR*, abs/1911.09785, 2019. [2, 4](#)
- [9] David Berthelot, Nicholas Carlini, Ian Goodfellow, et al. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. [1, 2, 4](#)
- [10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. [2](#)
- [11] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. [2](#)
- [12] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:3569–3576, 04 2020. [3, 6](#)
- [13] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical data augmentation with no separate search. *CoRR*, abs/1909.13719, 2019. [6](#)
- [14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. [2, 4](#)
- [15] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, et al. Safe deep semi-supervised learning for unseen-class unlabeled data. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3897–3906. PMLR, 13–18 Jul 2020. [3](#)
- [16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. [2](#)
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. [6](#)
- [18] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE*

*Transactions on Pattern Analysis and Machine Intelligence*, 2020. 4

- [19] Byoungjin Kim, Jinho Choo, Yeong-Dae Kwon, et al. Self-match: Combining contrastive self-supervision and consistency for semi-supervised learning. *arXiv preprint arXiv:2101.06480*, 2021. 2
- [20] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009. 6
- [21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2016. 2, 6
- [22] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7:7, 2015. 1
- [23] 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. 6, 7
- [24] Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2019. 3
- [25] Vishnu Suresh Lokhande, Songwong Tasneeyapant, et al. Generating accurate pseudo-labels in semi-supervised learning and avoiding overconfident predictions via hermite polynomial activations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11435–11443, 2020. 2
- [26] Takeru Miyato, Shin-ichi Maeda, et al. Virtual adversarial training: A regularization method for supervised and semi-supervised learning, 2018. 6
- [27] Takeru Miyato, Shin-ichi Maeda, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 2
- [28] Varun Nair, Javier Fuentes Alonso, and Tony Beltramelli. Realmix: Towards realistic semi-supervised deep learning algorithms. *CoRR*, abs/1912.08766, 2019. 3, 6
- [29] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. *CoRR*, abs/1603.09246, 2016. 2
- [30] Avital Oliver, Augustus Odena, Colin Raffel, et al. Realistic evaluation of deep semi-supervised learning algorithms. *CoRR*, abs/1804.09170, 2018. 1, 3, 6
- [31] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 754–763, 2017. 3
- [32] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. *CoRR*, abs/1701.06548, 2017. 4
- [33] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021. 6, 7
- [34] Rindra Ramamujison, Amin Banitalebi-Dehkordi, Xinyu Kang, Xiaolong Bai, and Yong Zhang. Simrod: A simple adaptation method for robust object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3570–3579, 2021. 2
- [35] Mamshad Nayeem Rizve, Kevin Duarte, et al. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021. 2
- [36] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–168, 2018. 3
- [37] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2, 4, 5, 6, 7
- [38] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 2
- [39] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1195–1204, 2017. 2
- [40] Antti Tarvainen and Harri Valpola. Weight-averaged consistency targets improve semi-supervised deep learning results. *CoRR*, abs/1703.01780, 2017. 2, 5, 6
- [41] Isaac Triguero, Salvador García, and Francisco Herrera. Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. *Knowl. Inf. Syst.*, 42(2):245–284, Feb. 2015. 1, 2
- [42] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020. 1
- [43] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI’19*, 2019. 2
- [44] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 3, 6
- [45] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 4
- [46] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016. 6
- [47] 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*, pages 1476–1485, 2019. 1
- [48] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017. 2