Scaling Up Semi-supervised Learning with Unconstrained Unlabelled Data

Shuvendu Roy, Ali Etemad
Queen’s University, Canada
{shuvendu.roy, ali.etemad}@queensu.ca

Abstract
We propose UnMixMatch, a semi-supervised learning framework which can learn effective representations from unconstrained unlabelled data in order to scale up performance. Most existing semi-supervised methods rely on the assumption that labelled and unlabelled samples are drawn from the same distribution, which limits the potential for improvement through the use of free-living unlabelled data. Consequently, the generalizability and scalability of semi-supervised learning are often hindered by this assumption. Our method aims to overcome these constraints and effectively utilize unconstrained unlabelled data in semi-supervised learning. UnMixMatch consists of three main components: a supervised learner with hard augmentations that provides strong regularization, a contrastive consistency regularizer to learn underlying representations from the unlabelled data, and a self-supervised loss to enhance the representations that are learnt from the unlabelled data. We perform extensive experiments on 4 commonly used datasets and demonstrate superior performance over existing semi-supervised methods with a performance boost of 4.79%. Extensive ablation and sensitivity studies show the effectiveness and impact of each of the proposed components of our method. The code for our work is publicly available.

1 Introduction
Semi-supervised learning (SSL) uses large amounts of unlabelled data along with small amounts of labelled data to reduce the reliance on fully-labelled datasets. Most existing semi-supervised methods rely on the assumption that the labelled and unlabelled data belong to the same distributions, an assumption that is not necessarily true in real-world scenarios. Moreover, this assumption prohibits us from leveraging free-living unlabelled data with different distributions. In fact, it has been shown in previous studies that incorporating out-of-distribution data with the unlabelled set for SSL impairs performance (Oliver et al. 2018).

To adopt a less constrained approach regarding unlabelled data, open set SSL has been proposed (Saito, Kim, and Saenko 2021; Yu et al. 2020), which allows the unlabelled training set to contain samples from classes which are not necessarily present in the labelled set. Nevertheless, this setting still places certain restrictions on the unlabelled data, necessitating the inclusion of samples from every known class and ensuring that its data distribution is similar. These constraints create two important challenges. First, collecting an unlabelled dataset that necessarily includes samples from certain classes can be challenging in real-world settings. Second, this approach severely restricts the scalability of SSL to large, web-scale, unconstrained, unlabelled data since such data do not hold the aforementioned constraints. Most existing semi-supervised methods are not suitable for learning from unconstrained data as they rely on pseudo-label predictions, which require the ‘unlabelled’ set to have the same classes as the ‘labelled’ set.

In this paper, we propose a novel SSL approach called UnMixMatch, which can learn effective representations from unconstrained unlabelled data and effectively enable SSL to scale up using web-scale unlabelled data. UnMixMatch comprises three main components, which have some similarities to previous SSL methods but have been specifically modified and tailored toward the ‘scalability’ criteria: (1) A supervised learner with hard augmentations: We introduce a new hard augmentation module that combines RandAug with MixUp to prevent overfitting on the small labelled set. The convention of using soft augmentations of the existing literature does not perform well when learning from unconstrained unlabelled data. (2) A contrastive consistency regularizer: The primary unsupervised learning component of our method involves a contrastive regularizer, which learns the underlying data representations by enforcing the model to produce consistent predictions under strong perturbations.
In contrast to existing SSL methods, we do not regularize the class predictions, as the unlabeled set contains unknown classes. Instead, we enforce consistency in the predicted embedding space. (3) A self-supervised pre-text learning module: To further enhance the learned representations, we include a pre-text task called rotation prediction on the unlabeled data, where the model learns by predicting the degree of rotation applied to each sample.

We conduct extensive research on four common datasets: CIFAR-10, CIFAR-100, SVHN, and STL-10. First, we re-implement and benchmark 13 recent semi-supervised methods with unconstrained unlabeled data, using ImageNet-100 as the unlabeled set. We find that existing methods experience performance degradation in unconstrained settings. In comparison, UnMixMatch outperformed existing methods by an average of 4.79\%. Additionally, UnMixMatch exhibits robust scaling capabilities regarding the size of the unlabeled datasets, as we observe an additional 5.61\% improvement when we increase the unlabeled data size by a factor of 10 (see Fig. 1). Furthermore, we achieved state-of-the-art results in the open set SSL. Finally, we ablate each component of UnMixMatch and demonstrate the crucial role that each component plays in the performance.

In summary, we make the following contributions:

• We propose a novel semi-supervised method that can learn effective representations from unconstrained unlabeled data for the first time.
• We conduct an extensive study to benchmark the performance of existing semi-supervised methods when the unlabeled data are not constrained to match the distribution of the labelled data.
• We demonstrate that our method outperforms previous methods by a large margin in unconstrained learning and sets a new state-of-the-art for open set SSL.
• We also show that the performance of our method scales up by increasing the amount of unconstrained unlabeled data.
• To facilitate reproducibility, we release the code at github.com/ShuvenduRoy/UnMixMatch.

2 Related Work

2.1 Constrained Semi-supervised Learning

Prior works on semi-supervised can be broadly divided into two main categories: pseudo-labeling and consistency regularization. Pseudo-labeling techniques (Lee et al. 2013; Xie et al. 2020b) mainly rely on the strategy of predicting pseudo-labels for the unlabeled data using the encoder being trained and learns with a combination of the labeled data and unlabeled data plus their pseudo-labels, and iterating on those predictions to gain gradual improvements. Consistency regularization techniques (Tarvainen and Valpola 2017; Sajjadi, Javanmardi, and Tasdizen 2016; Miyato et al. 2018) learn by forcing the embeddings of different augmented unlabeled samples to be similar. This takes place while the model also simultaneously learns via a supervised loss, which is optimized on the labeled samples. In the pseudo-labeling category, Lee et al. (2013) first introduced the overall approach, and subsequent works improved the technique by adding various interesting elements. For example, in Noisy Student (Xie et al. 2020b), a pre-trained teacher was introduced to generate the pseudo-labels, and a student learned from the pseudo-labels along with the labeled data. In the consistency regularization category, Pi-model (Sajjadi, Javanmardi, and Tasdizen 2016) was one of the earliest works which used a consistency loss on the predictions of two augmentations of a sample. Later, Mean Teacher (Tarvainen and Valpola 2017) improved the performance by enforcing consistency between the predictions of an online encoder and an exponential moving average (EMA) encoder. VAT (Miyato et al. 2018) is another modification of Mean Teacher, which replaced the augmentations with adversarial perturbations. Later, Unsupervised Domain Adaptation (UDA) (Xie et al. 2020a) showed large improvements by using hard augmentations.

It should be noted that one of the key weaknesses of pseudo-label-based methods is the confidence-bias problem, which arises when the model generates confident wrong pseudo-labels. Yet, their ability to virtually increase the labeled set size by generating pseudo-labels for the unlabeled data has motivated researchers to combine them with consistency regularization methods within the same framework. For instance, MixMatch (Berthelot et al. 2019b) predicts pseudo-labels for unlabeled samples while enforcing consistency across augmented images. ReMixMatch (Berthelot et al. 2019a) improved upon MixMatch with several implementation tricks, such as augmentation anchoring and distribution alignment. FixMatch is another popular hybrid method known for its simplicity and performance. It predicts the pseudo-label of a sample from a weakly augmented image and treats it as a label for a heavily augmented sample when the confidence of the pseudo-label is above a certain threshold. Subsequently, several works have attempted to improve different aspects of FixMatch. FlexMatch (Zhang et al. 2021) employs an adaptive curriculum threshold for each class based on that class’s performance. CoMatch (Li, Xiong, and Hoi 2021) improves upon FixMatch by introducing a graph-based contrastive loss that learns both class representations and low-dimensional embeddings. ConMatch (Kim et al. 2022) also employs a similar concept, using pseudo-labels as supervision in a contrastive loss. Similarly, SimMatch (Zheng et al. 2022) improves FixMatch by introducing an instance similarity loss in addition to the semantic similarity loss imposed by pseudo-labels. ScMatch (Gu et al. 2022) utilizes the concept of clustering with SSL by dynamically forming super-classes.

2.2 Open Set Semi-supervised Learning

Open set SSL is a type of SSL where the unlabeled set includes samples from unknown classes. In prior work (Oliver et al. 2018), it was demonstrated that the presence of unknown classes in the unlabeled dataset has a severe negative impact on the performance of semi-supervised methods. Similar findings were also reported by Su et al. (2021) that analyzed the performance of more recent semi-supervised methods in open set settings. Nonetheless, some recent methods have attempted to effectively address the detri-
mental effect of unknown classes while learning from open
set unlabelled data. For example, Yoshihashi et al. (2019)
learned to distinguish known classes from unknown ones,
effectively avoiding samples from unknown classes in the
learning process. A similar approach was taken in Guo et
al. (2020), which proposed a novel scoring function called
energy discrepancy to detect and remove instances of
unknown classes. OpenMatch (Saito, Kim, and Saenko 2021)
used the concept of out-of-distribution to mitigate the nega-
tive impact of unknown classes. CCSSL (Yang et al. 2022)
introduced a class-aware contrastive learning approach to
improve performance in open set settings. In this study, we
tackle a more challenging scenario where the unlabelled set
contains instances of unknown classes and does not neces-
sarily include all the known classes. Consequently, our goal
is not to detect and remove the images of unknown classes,
but rather to learn from them.

3 Method

3.1 Preliminaries and Overview

Let, \( X_U = (x_i)^N_{i=1} \) be an unlabelled set, and \( X_L = (x_i, y_i)^n_{i=1} \) be a labelled set where \( n \ll N \). In general,
semi-supervised methods learn from \( X_L \) and \( X_U \) in supervi-
sed and unsupervised settings, respectively. Formally, SSL
is formulated as:

\[
\min_\theta \sum_{(x,y) \in X_L} \mathcal{L}_S(x, y, \theta) + \alpha \sum_{x \in X_U} \mathcal{L}_U(x, \theta),
\]

where, \( \theta \) represents the learnable model, \( \mathcal{L}_S \) is the sup-
ervised loss, and \( \mathcal{L}_U \) is the unsupervised loss. Trivially, it is as-
sumed that \( X_L \) and \( X_U \) come from the same data and class
 distributions. Let \( Y_l \) and \( Y_u \) be the set of classes for labelled
and unlabelled data. In the constrained setting, \( Y_l = Y_u \),
and in the open set setting, \( Y_l \) is a proper subset of \( Y_u \), i.e.,
\( Y_l \subset Y_u \). Regarding data distribution, both settings assume
that the data comes from the same source and hence have
similar distributions. These assumptions are hard to satisfy in a
real-world task.

In this study, our objective is to learn from unconstrained
data that does not have any particular constraints and can
come from different data or class distributions, or both. The
unlabelled set may consist of images of unknown classes,
where \( Y_u \setminus Y_l \neq \emptyset \). Additionally, the unlabelled set may
not contain all the known classes, and in the extreme case,
\( Y_u \cap Y_l = \emptyset \). To address this challenge, we propose UnMix-
Match, a method that can learn effective visual representations
from unconstrained unlabelled data. UnMixMatch comprises three modules, which we discuss in detail in the
following subsections.

3.2 Supervised Module

As mentioned earlier, SSL requires \( X_L \) to be learned using a
supervised component. The first contribution of our method
is, therefore, to create a supervised learner suitable for our
purpose of scalable SSL. Here, we hypothesize that given
large amounts of unlabelled data in an unconstrained setting
and relatively very small amounts of labelled data, the su-
pervised module may outper the small labelled set. Thus, un-
like FixMatch (Sohn et al. 2020), MixMatch (Berthelot et al.
2019b), and ReMixMatch (Berthelot et al. 2019a), which
use weak augmentations for their supervised modules, we
apply hard augmentations on the labelled samples in our su-
pervised module. This acts as a regularizer for supervised
learning and is better able to deal with overfitting compared
to weaker augmentations. We utilize RandAug (Cubuk et al.
2020) as the hard augmentation followed by the MixUp op-
eration (Berthelot et al. 2019b,a). We denote RandAug plus
MixUp as the RandMixUp operation. Finally, a supervised loss is applied to a batch of samples.

**RandAug.** This is a hard augmentation technique for gen-
ergating diverse samples by employing a sequence of transforma-
tions (Cubuk et al. 2020). More specifically, it applies \( R_n \in [1..13] \) transformations randomly chosen from a list of
13 augmentations, including rotation, translation, and colour
distortion. The magnitude of each transformation is sampled
randomly from a pre-defined range. We denote the augmenta-
tion operation as \( \hat{x} = \text{RandAug}(x) \).

**MixUp.** Let \((x_1, y_1)\) and \((x_2, y_2)\) be two pairs of samples and their class labels. MixUp operation interpolates between
the data points to generate mixed samples and labels as:

\[
\bar{x} = \lambda \cdot x_1 + (1 - \lambda) \cdot x_2 \\
\bar{y} = \lambda \cdot y_1 + (1 - \lambda) \cdot y_2,
\]

where \( \lambda \) is the mixing coefficient. Following MixMatch (Berthelot et al. 2019b), we sample \( \lambda \) from a beta distribu-
tion \((\lambda \sim \text{Beta}(\alpha, \alpha))\) with hyper-parameter \( \alpha \).

**Supervised Loss.** For a batch of unlabelled samples \( X_u = ((\hat{x}_i); i \in \{1, \ldots, b\}) \), with batch size \( b \), we first generate the
pseudo-label for each sample \( X_i \) as \( p_i = P\theta(x_i) \), where \( P\theta \) is
the encoder with a classification head.

Next, for a batch of labelled samples \( X_l = ((\bar{x}_i, y_i); i \in \{1, \ldots, b\}) \) and the unlabelled samples with pseudo-labels
\( X_p = ((\bar{x}_i, p_i); i \in \{1, \ldots, b\}) \), we augment all the samples
using \( \tilde{X} = \text{RandAugMix}(X_l, X_p) \). Accordingly, we de-
fine the supervised loss of our method as:

\[
\mathcal{L}_{sup} = \frac{1}{b} \sum_{\tilde{x}, \tilde{y} \in \tilde{X}} \mathcal{H}(\tilde{y}, P\theta(y|\tilde{x}))
\]

where \( \mathcal{H} \) is the cross-entropy loss.

3.3 Consistency Regularization Module

To deal with the unconstrained nature of existing unlabelled
data and learn effective representations, we apply a consis-
tency regularizer as our second contribution. A consis-

![Figure 2: Overview of our proposed method.](image-url)
tency regularizer learns from the unlabelled data by enforcing consistency on its predictions under different augmentations. Prior works that have used consistency regularization for SSL (Sajjadi, Javanmardi, and Tasdizen 2016; Berthelot et al. 2019b,a) enforce consistency on the class predictions under different perturbations. However, regularization over class predictions is not useful for learning in unconstrained settings where unlabelled data do not necessarily come from the same classes as the labelled data. To address this, we enforce consistency in the low-dimensional embedding space using a contrastive loss. Using contrastive loss on the embedding space enables the model to learn class-agnostic representations from unconstrained unlabelled data. In UnMix-Match, we adopt the Noise Contrastive Estimation loss, a.k.a InfoNCE (Chen et al. 2020).

Contrastive learning learns from positive (perturbations of the same sample) and negative samples (all other samples) by bringing embeddings of the positive samples together and pushing them away for the negatives. For each sample, $x_i \in X_U$, two augmentations are applied to generate two augmented images $\tilde{x}_i = \text{RandAug}(x_i)$. These are first passed through the encoder and a projection head (shallow linear layers with non-linearity and batch normalization) to obtain embeddings $z_i = P_{\theta_e}(z|x_i)$. The contrastive loss is accordingly defined as:

\[
L_{\text{con}} = -\frac{1}{2b} \sum_{i=1}^{2b} \log \frac{\exp(z_i, z_{\kappa(i)}/\tau)}{\sum_{k=1}^{2b} \mathbb{1}_{[k \neq i]} \exp(z_i, z_k/\tau)},
\]

where, $\kappa(i)$ is the index of the second augmented sample, $\mathbb{1}_{[k \neq i]}$ is an indicator function which returns 1 when $k$ is not equal to $i$, and 0 otherwise. $\tau$ is a temperature parameter.

3.4 Self-supervised Module

Finally, we intend to enhance the quality of the representations extracted from the unconstrained unlabelled data using the consistency regularizer. It has been recently shown that self-supervised techniques can be employed to learn underlying domain-invariant representations for unlabelled data (Gidaris, Singh, and Komodakis 2018; Zhang, Isola, and Efros 2016). Moreover, this idea has already been demonstrated to be useful in conjunction with SSL (Gidaris, Singh, and Komodakis 2018; Zhai et al. 2019; Berthelot et al. 2019a). As a result, we integrate a straightforward yet effective self-supervised pre-text task called rotation prediction, which learns by predicting the degree of rotations applied to unlabelled images. In practice, a rotation module randomly samples one of the following rotations and applies it to an unlabelled image: $0^\circ$, $90^\circ$, $180^\circ$, $270^\circ$. As a result, the rotation prediction task can be viewed as a four-way classification problem, represented as:

\[
L_{\text{rot}} = \frac{1}{b} \sum_{u \in U} \mathcal{H}(r, P_{\theta_e}(r; \text{Rotate}(u)))
\]

Here, $P_{\theta_e}$ is the encoder with a rotation head that predicts the rotation, and $\mathcal{H}$ is the cross-entropy loss.

3.5 Total Loss

Finally, we incorporate the loss functions for the three modules above to create the total loss:

\[
L_{\text{UnMixMatch}} = L_{\text{sup}} + \beta L_{\text{con}} + \gamma L_{\text{rot}}.
\]

Here, $\beta$ and $\gamma$ are hyper-parameters that balance the significance of the contrastive and rotation losses. An overview of our proposed method is presented in Fig. 2.

4 Experiments and Results

4.1 Datasets and Implementation Details

For our main experiments, we follow the standard semi-supervised evaluation protocol from prior works (Sohn et al. 2020), and present the results for four datasets: CIFAR-10 (Krizhevsky, Hinton et al. 2009), CIFAR-100 (Krizhevsky, Hinton et al. 2009), SVHN (Netzer et al. 2011), and STL-10 (Coates, Ng, and Lee 2011). We present the results for different numbers of labelled samples, averaged over three runs. We use ImageNet-1K (Deng et al. 2009), and ImageNet-100 (a subset of ImageNet-1K) as the unconstrained unlabelled datasets since it has different class and data distribution in comparison to the four aforementioned datasets.

Our implementation and hyper-parameters closely follow FlexMatch (Zhang et al. 2021). For the encoder, following existing literature such as (Sohn et al. 2020; Zhang et al. 2021), we use Wide ResNet-28-2 (Zagoruyko and Komodakis 2016) for CIFAR-10 and SVHN, WRN-28-8 for CIFAR-100, and WRN-37-2 for STL-10. We train the method for $2^{20}$ iterations with a batch size of 64, a learning rate of 0.03, and an SGD optimizer with a momentum of 0.9 and weight-decay of 0.0005. The code is implemented with Pytorch and built using TorchSSL (Zhang et al. 2021).

4.2 Results

Here, we present the main results of our method, including the performance on the four datasets (with ImageNet-100 for unlabeled data) in unconstrained settings, analysis of UnMixMatch’s performance with increasing the number of unlabelled data, its effectiveness in open set settings, and its performance in a barely supervised setting.

Unconstrained Settings. Table 1 presents the main results of our work on the four datasets. Here, we first reimplement 13 semi-supervised methods and report the results with unconstrained unlabelled data. We report the average accuracies and standard deviations across three individual runs for each setting. We also report the average accuracy across all settings for overall comparison. It should be noted that the performance of prior methods is considerably lower than what has been reported in the original papers, where the unlabeled and labelled samples came from the same datasets (unlabelled data were not unconstrained). Next, we observe that UnMixMatch demonstrates superior performance compared to other methods, with an average improvement of 4.79%. We obtain considerable improvement across all datasets and splits, except using CIFAR-100 with 2500 labels, where CCSSL achieves a better result.

When considering the number of labelled samples, we notice that the differences between UnMixMatch and other
Table 1: Comparison of our method against other SSL methods with unconstrained data on 4 different datasets. The baseline methods are: Pi-Model (Sajjadi, Javanmardi, and Tasdizen 2016), MeanTeacher (Tarvainen and Valpola 2017), VAT (Miyato et al. 2018), Pseudo-label (Lee et al. 2013), UDA (Xie et al. 2020a), MixMatch (Berthelot et al. 2019b), ReMixMatch (Berthelot et al. 2019a), FixMatch (Sohn et al. 2020), FlexMatch (Zhang et al. 2021), CoMatch (Li, Xiong, and Hoi 2021), CCSSL (Yang et al. 2022), SimMatch (Zheng et al. 2022), and ScMatch (Gu et al. 2022).

Table 2: The impact of unlabeled set size. Here, Subsets 1 & 2 are two random subsets of ImageNet-1K (IN-1K).

Table 3: Performance comparison on barely supervised learning. Only 1 sample per class is used for training.

Table 4: Performance comparison on open set SSL. Only 1 sample per class is used for training.
trastive loss in our method with the class-aware contrastive loss of CCSSL (Yang et al. 2022). This method first predicts the pseudo-labels for the unlabelled samples and uses them with contrastive loss to learn clusters of known classes in the embedding space. For this experiment, we follow the experimental setup of OpenMatch (Saito, Kim, and Saenko 2021), which reports the results for CIFAR-10 with a 6/4 split. This split means that the labelled set contains images of 6 classes from CIFAR-10, while the unlabelled set includes images of 6 known and 4 unknown classes. Like OpenMatch, we take 6 animal classes as the known classes and 4 object classes as the unknown classes. We perform this experiment using three different splits with 50, 100, and 400 labelled samples per class in the known set. The results are presented in Table 3, where we find that our method outperforms the existing methods and sets a new state-of-the-art for open set SSL. Once again, UnMixMatch better demonstrates its effectiveness when the amount of labelled data is limited. With 50 labelled samples per class, our approach provides a 6.1% improvement over the second-best method, OpenMatch. For 100 and 400 labelled samples per class, UnMixMatch shows 3.9% and 3.1% improvements, respectively.

Barely Supervised Setting. In this section, we aim to test the performance of SSL in the extreme scenario where only one labelled sample per class is available. This barely supervised setting is considered to be very challenging, even with conventional SSL techniques that use constrained unlabelled data. Given the extremely low number of labelled samples, the results in such settings generally exhibit high variance, and therefore, we increase the number of folds to 5 to account for this variability. This is due to the fact that the quality of labelled data greatly influences the performance in such low data settings (Sohn et al. 2020; Roy and Etemad 2022). The results of this experiment are presented in Table 4. It shows the performance of CIFAR-10 for the three best methods identified previously in Table 1.

In this experiment, UnMixMatch achieves an accuracy of 27.54% and outperforms other methods by 5.58%. As before, CCSSL struggles to learn in low data settings, achieving a near chance-level accuracy of only 15.63%. However, FlexMatch shows relatively better performance with an accuracy of 21.96%. In general, we find that it is quite difficult to learn effective representations with just one labelled sample per class while using unconstrained unlabelled data. However, as previously shown in Table 1, UnMixMatch quickly gains significant improvements with a small increase in the number of labelled data and achieves an accuracy of 47.93% with four labelled samples per class.

4.3 Ablation Study

Main Components. Table 5a presents the main ablation results by removing each of the three main components of the proposed method: RandMixUp augmentation, consistency regularization, and rotation prediction. Note that we can not remove two components simultaneously, since semi-supervised learners require a minimum of one supervised and one unsupervised loss. Therefore, we can not perform ablation studies with only one component at a time. These experiments are done on CIFAR-10 with 40 samples. The table demonstrates that all three components have a significant impact on the final performance of the model, with the removal of any one component resulting in a large drop in accuracy. In the first ablation experiment, we remove the RandMixUp augmentation module, effectively learning from the labelled samples with weak augmentations only (random resized crop and horizontal flip). This experiment results in the highest drop in accuracy across the ablation settings, with a 9.21% decrease. The second largest drop in accuracy is observed when removing consistency regularization, resulting in a 6.68% decrease. Similarly, removing the rotation prediction component results in a 6.1% drop in performance.

RandMixUp. As demonstrated in the main ablation study, the RandMixUp augmentation is a crucial component of our approach. In Table 5b, we present the accuracy of CIFAR-10 with 40 samples for different alternatives to RandMixUp. Recall that RandMixUp combines RandAug (Cubuk et al. 2020) and MixUp (Berthelot et al. 2019b). We first test the accuracy for RandAug and MixUp individually and observe a decrease in performance for both settings alone, with a 1.92% and 2.7% decrease, respectively. Here, using RandAug exhibits a lower drop than MixUp, suggesting the higher importance of RandAug in the RandMixUp. This analysis also shows that, unlike MixMatch and ReMixMatch, learning using MixUp does not have the most significant impact on the supervised component. Instead, the role of hard augmentation as a regularizer is the key to the impact of RandMixUp in our method.

We also examine the performance of other well-known hard augmentation techniques, namely CutMix and AugMix. Like MixUp, CutMix combines two samples, but in this case, a part of the second sample is cut and inserted into the first sample to create a new one. Despite the conceptual similarity, CutMix yields 2.11% worse accuracy than MixUp, which is 4.81% lower than the accuracy achieved by the proposed RandAugMix. On the other hand, AugMix shares a similar concept as RandAug. While RandAug applies multiple augmentations in sequence, AugMix applies multiple augmentations separately and generates the final sample by interpolating between them. AugMix achieves an accuracy of 44.75%, which is 1.26% lower than that of RandAug and 3.18% lower than that of RandAugMix. The study’s overall findings demonstrate the critical importance of RandAugMix in our method, with other conceptually similar augmentation techniques failing to achieve the same level of performance.

Consistency Regularizer. As shown in the main ablation study, the contrastive regularizer is the second most crucial component of our method. Recall that the proposed contrastive loss for UnMixMatch is a noise contrastive estimation loss that has gained popularity in the self-supervised learning literature (Oord, Li, and Vinyals 2018; Chen et al. 2020). However, other variants of contrastive loss have also been introduced in the literature (Kim et al. 2022; Lee et al. 2022; Li, Xiong, and Hoi 2021; Yang et al. 2022) and have shown improvements in different problem settings, including SSL. In this study, we investigate four variants of con-
Next, we examine another aspect of our regularization: regularizing the embedding space instead of the class prediction. In Experiment 6, where we observe a 5.48% decrease in accuracy, the proposed method in one setting (CIFAR-100 with 400 labelled samples) only.

We explore the contrastive variants of ConMatch (Kim et al. 2022) and Contrastive Regularization (Lee et al. 2022), as well as graph contrastive learning (Li, Xiong, and Hoi 2021), and class-aware contrastive learning (Yang et al. 2022). Contrastive (Kim et al. 2022) is based on a variant of contrastive loss that involves two hard augmentations and one weak augmentation. In Contrastive Regularization (Lee et al. 2022), high-confidence pseudo-labels were used for supervision while learning with a contrastive loss. Two other methods proposed slightly different versions of utilizing the pseudo-labels in contrastive learning settings: CoMatch (Li, Xiong, and Hoi 2021) and CCSSL (Yang et al. 2022), where CoMatch used a graph contrastive learning, and CCSSL utilized pseudo-labels with a contrastive regularizer.

We show the results for variants of contrastive loss in Table 5c. Using the contrastive variant of both ConMatch and Contrastive Regularizer result in a large drop in accuracy with 2.16% and 2.11%, respectively. Using the graph contract concept of CoMatch with UnMixMatch results in a 47.12% accuracy, dropping by 0.81%. Finally, the class-aware contrastive loss of CCSSL gets the closest accuracy of the proposed UnMixMatch with 47.88% accuracy. In Table 6, we present expanded results across all datasets for the graph contrast from CoMatch (Li, Xiong, and Hoi 2021) and the class-aware contrast from CCSSL (Yang et al. 2022). Results in this table show that on average the proposed contrastive regularizer achieves 2.99% and 4.45% improvements over the graph contrast and class-aware contrast alternatives. The class-aware contrast obtains better results than the proposed method in one setting (CIFAR-100 with 400 labelled samples) only.

Next, we investigate different regularization strategies. First, we study the choice of applying the contrastive loss on the low-dimensional embedding space rather than the final prediction. In Table 5d, we present the results for this experiment, where we observe a 5.48% decrease in accuracy, indicating the importance of learning representations by regularizing the embedding space instead of the class predictions. Next, we examine another aspect of our regularization approach by replacing two strong augmentations with one weak and one strong augmentation (similar to FixMatch). This variant results in a 3.78% drop in accuracy.

Finally, we show the performance of our method by completely replacing the contrastive loss with other important viable losses, namely, BYOL (Grill et al. 2020), SimSiam (Chen and He 2021), and VICReg (Bardes, Ponce, and LeCun 2021). BYOL (Grill et al. 2020) learns by predicting the representation of a target encoder (exponential moving average of the online encoder), using the online encoder for different augmentations of the same sample. SimSiam (Chen and He 2021) also learns by matching the representations of two augmented samples, but unlike previous methods, it doesn’t need negative samples or the target encoder. VICReg (Bardes, Ponce, and LeCun 2021) learns from unlabelled data by combining three terms: variance, invariance, and covariance regularizations. The results of this experiment are presented in Table 5e, where we see that the accuracy of BYOL and SimSiam is considerably lower than the proposed contrastive loss with 46.89% and 46.50%. However, we find VICReg to have a competitive performance. To fully understand the performance of VICReg, we present the results for all datasets in Table 7. This table shows that VICReg achieves better results for a few settings (CIFAR-10 with 250 labelled samples, SVHN with 250 and 1000 labelled samples, and STL-10 with 1000 labelled samples). However, the overall average accuracy with VICReg loss is 0.92% lower than the contrastive regularizer, and hence, we choose this as the default for the proposed UnMixMatch.

**Rotation Loss.** We also investigate a few other self-supervised approaches instead of the rotation prediction task, including colorization, jigsaw solving, and image inpainting. The results of this experiment are included in Table 5f, where rotation prediction shows a clear advantage over other pre-text tasks. Apart from the fact that rotation prediction performs well, it has two additional benefits over other methods. First, it can be integrated into our semi-supervised framework without adding considerable computational overhead. In contrast, tasks like colorization or in-
Table 6: Comparison of our method against different variants of contrastive regularizers on 4 different datasets.

<table>
<thead>
<tr>
<th>Methods</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
<th>SVHN</th>
<th>STL-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>VICReg</td>
<td>47.93±1.1</td>
<td>68.72±0.6</td>
<td>89.58±0.2</td>
<td>84.73±0.6</td>
</tr>
<tr>
<td>UnMixMatch</td>
<td>47.93±1.1</td>
<td>68.72±0.6</td>
<td>89.58±0.2</td>
<td>84.73±0.6</td>
</tr>
<tr>
<td>Graph Con.</td>
<td>47.12±0.5</td>
<td>67.67±0.7</td>
<td>87.30±0.3</td>
<td>82.59±0.4</td>
</tr>
<tr>
<td>Class Con.</td>
<td>47.88±2.4</td>
<td>68.45±0.4</td>
<td>88.39±0.2</td>
<td>82.59±0.4</td>
</tr>
<tr>
<td>w/ Match loss</td>
<td>50.25±0.7</td>
<td>71.73±0.2</td>
<td>90.13±0.3</td>
<td>91.03±0.3</td>
</tr>
<tr>
<td>w/ KL-div. loss</td>
<td>43.99±12.2</td>
<td>78.05±0.2</td>
<td>91.03±0.3</td>
<td>91.03±0.3</td>
</tr>
</tbody>
</table>

Table 7: Comparison of our method for different alternate regularization strategies on 4 different datasets.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UnMixMatch</td>
<td>47.93</td>
</tr>
<tr>
<td>MixUp only</td>
<td>45.23</td>
</tr>
<tr>
<td>w/ Match loss</td>
<td>44.25</td>
</tr>
<tr>
<td>w/ KL-div. loss</td>
<td>47.02</td>
</tr>
</tbody>
</table>

Table 8: Study on the relation of our method to prior works.

Figure 3: Sensitivity study on important hyper-parameters.

4.4 Sensitivity Study

Our method involves three important hyper-parameters, namely the $\alpha$ value in MixUp, the contrastive loss weight $\beta$, and the rotation loss weight $\gamma$. Here, we present a sensitivity analysis of different values for these hyper-parameters. As illustrated in Fig. 3a, the best accuracy is achieved with an alpha value of 0.1, while very large or small values of $\alpha$ lead to a drop in accuracy. Fig. 3b indicates that the optimal performance is obtained with a $\beta$ of 1.0. Furthermore, Fig. 3c reveals that a higher value of $\gamma$ leads to better performance, highlighting the role of the self-supervised loss.

5 Conclusion

Existing semi-supervised methods struggle to learn when the assumption that the unlabeled data comes from the same distribution as the labelled data, is violated. This work proposes a new semi-supervised method called UnMixMatch for learning from unconstrained unlabelled data. Our method shows large improvements over existing methods and even larger improvements under low-labelled data settings. Our approach also outperforms existing methods on open set settings. Most importantly, UnMixMatch scales up in performance when the size of unlabelled data increases. We hope this research will draw attention to this more challenging and realistic SSL setting with unconstrained unlabelled data.

Limitations. While our comprehensive study on common benchmark datasets demonstrates the effectiveness of UnMixMatch across diverse image domains, some aspects require further investigation for universal applicability. Specifically, the use of hard augmentations and the rotation prediction task might necessitate additional tuning or minor modifications for optimal performance in specialized domains.
Acknowledgements
This work was supported by Mitacs, BMO, and Ingenuity Labs Research Institute. We are also thankful to SciNet HPC Consortium for helping with the computation resources.

References


