arXiv:2310.16459v1 [cs.CV] 25 Oct 2023

DualMatch: Robust Semi-Supervised Learning with Dual-Level Interaction

Cong Wang[1⋆], Xiaofeng Cao[1⋆] (✉), Lanzhe Guo[2], and Zenglin Shi[3]

[1] School of Artificial Intelligence, Jilin University, Changchun, 130012, China
cwang21@mails.jlu.edu.cn, xiaofengcao@jlu.edu.cn
[2] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
guolz@lamda.nju.edu.cn
[3] I2R, Agency for Science, Technology and Research (A*STAR)
shizl@i2r.a-star.edu.sg

Abstract. Semi-supervised learning provides an expressive framework for exploiting unlabeled data when labels are insufficient. Previous semi-supervised learning methods typically match model predictions of different data-augmented views in a single-level interaction manner, which highly relies on the quality of pseudo-labels and results in semi-supervised learning not robust. In this paper, we propose a novel SSL method called DualMatch, in which the class prediction jointly invokes feature embedding in a dual-level interaction manner. DualMatch requires consistent regularizations for data augmentation, specifically, 1) ensuring that different augmented views are regulated with consistent class predictions, and 2) ensuring that different data of one class are regulated with similar feature embeddings. Extensive experiments demonstrate the effectiveness of DualMatch. In the standard SSL setting, the proposal achieves 9% error reduction compared with SOTA methods, even in a more challenging class-imbalanced setting, the proposal can still achieve 6% error reduction. Code is available at https://github.com/CWangAI/DualMatch

Keywords: Semi-supervised learning· Dual-Level interaction.

1 Introduction

Machine learning, especially deep learning [12], has achieved great success in various tasks. These tasks, however, crucially rely on the availability of an enormous amount of labeled training data. In many real-world applications, the acquisition of labeled data is expensive and inefficient. On the contrary, there are usually massive amounts of unlabeled data. Therefore, how to exploit unlabeled data to improve learning performance is a hot topic in the machine learning community [18].

Semi-supervised learning (SSL) provides an expressive framework for leveraging unlabeled data when labels are insufficient. Existing SSL methods can be categorized into several main classes in terms of the use of unlabeled data, such as pseudo-labeling

[⋆] Equal contribution

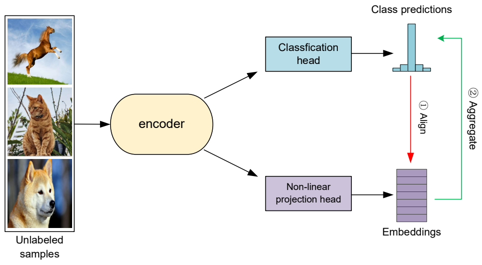


Fig. 1: An illustration of DualMatch with dual-level head interaction. Aligning the predictions of augmented data into their ground-truth labels is a single-level interaction manner (①) in semi-supervised learning, while the class predictions in such a manner may lack a stability guarantee for pseudo-labeling, even not robust. DualMatch reconsiders ① by aligning the feature embedding of one class and then considers a new interaction ② by aggregating class distribution with consistent feature embeddings.

methods [13], which assign pseudo-labels to unlabeled data based on the model prediction and train the model with labels and pseudo-labels in a supervised manner, and consistency regularization methods, which require that the output of the model should be the same when the model or data is perturbed. In much recent work, it has been reported that holistic SSL methods, e.g., MixMatch [3], ReMixMatch [2], and FixMatch [21], which consider the pseudo-labeling and consistency strategies simultaneously, have reached state-of-the-art (SOTA) performance. For example, in the image classification task, holistic SSL methods can achieve the performance of fully supervised learning even when a substantial portion of the labels in a given dataset have been discarded [21].

Although the holistic SSL methods have been reported to achieve positive results, they mainly adopt a single-level interaction manner between class prediction and the feature embedding, resulting in low quality of the pseudo-labels and weak SSL robustness performance. Take the SOTA FixMatch method as an example: FixMatch generates both weakly and strongly augmented views for unlabeled data, assigns high-confidence pseudo-labels predicted on the weakly augmented data to the strongly augmented one, and then optimizes the model by minimizing the cross-entropy loss between the prediction of the strongly augmented views and the corresponding pseudo-labels. This process is a single-level interaction since only different data augmentations are regulated by consistent class predictions. This results in the SSL performance being highly related to the correctness of the pseudo-label, and wrong pseudo-labels can lead to the confirmation bias of the model with error accumulation [1]. How to improve the robustness of SSL methods for pseudo-labels has emerged as a critical issue in SSL research.

In this paper, we propose a novel SSL algorithm called DualMatch. Compared with previous SSL methods that only consider the consistency between predictions for different augmentations, two consistency regularization factors are proposed in DualMatch, which derives more robust learning performance: 1) different augmented representations of training data should be regulated with consistent class predictions, and 2) different class predictions should be regulated with consistent feature representations. We illustrate the new manner of dual-level interaction in Figure 1. Specifically, in the first-level interaction, supervised contrastive learning is utilized for aligning the feature representations of one class with highly confident predictions. This requires that the features of strongly augmented views be clustered together in the low-dimensional embedding space, and then pseudo-labels are assigned from their weakly augmented views. In the second-level interaction, class distributions with consistent feature embeddings are aggregated to generate pseudo-labels for class prediction fine-tuning. Under this dual-level learning manner, the consistency of the same data represented in two heads is enhanced, and more reliable pseudo-labels are generated for matching strongly augmented class prediction. Compared with the FixMatch algorithm, the DualMatch achieves 9% error reduction in the CIFAR-10 dataset; even on a more challenging class-imbalanced semi-supervised learning task, the DualMatch can still achieve 6% error reduction compared with the FixMatch algorithm.

Our contributions can be summarized as follows:

- We point out that the single-level interaction that existing SSL methods commonly adopted will result in weak SSL robustness performance.
- We first propose the dual-level interaction between classification and feature embeddings and a novel DualMatch algorithm to improve the robustness of SSL.
- We rigorously evaluate the efficacy of our proposed approach by conducting experiments on standard SSL benchmarks and class-imbalanced semi-supervised learning. Our results demonstrate significant performance improvement.

2 Related Work

2.1 Semi-Supervised Learning

A prerequisite for SSL is that the data distribution should be based on a few assumptions, including smoothness, cluster, and manifold [4]. Technically, the smoothing assumption denotes that the nearby data are likely to share the same class label, and the manifold assumption denotes that the data located inner on low-dimensional streaming clusters are more likely to share the same labels. Recently, consistency regularization methods [3,26] have been widely applied and achieved outstanding results in SSL. An inherent observation is that the consistency regularization could be founded on the manifold or smoothness assumption, and requires that different perturbation methods for the same data hold consistent predictions against their employed diverse models.

From the perspective of consistency, there are two classical branches: model-level [20,11] and data-level consistency. As an early branch, [20] denotes the addition of random perturbation techniques (such as dropout [22] and random max-pooling methods) to the model should have consistent prediction results. To improve its robustness, [11] further aggregates the previous results of the model. Considering the pseudo-label cannot vary in iterative epochs, [24] then replace the aggregation with the exponential moving average (EMA) method. Data-level consistency is established by virtual adversarial training (VAT) [17] and unsupervised data augmentation (UDA) [26]. As an expressive consistency method, VAT produces optimally augmented views by adding random noise to the data and using an adversarial attack method. Differently, UDA utilizes the random augmentation (RA) [7] technique to produce strongly augmented views and minimizes the prediction disagreement between those views and their associated original data. Considering different levels of perturbations to the original input data, aligning different models' feedback to their early slightly perturbed inputs, i.e., anchoring, has been proven to be more effective. A series of strategies are then presented by taking this augmentation anchoring idea. In detail, MixMatch [3] adopts the mixup [28] trick to generate more augmented data by randomly pairing samples with each other and sharpening the average of multiple augmented data prediction distributions to generate pseudo-labels. Remixmatch [2] further improves the MixMatch approach by proposing a distribution alignment method (DA), which encourages the prediction of the marginal distribution of mixed data to be consistent with the true data distribution. FixMatch [21] simply considers weakly augmented view predictions with high confidence in unlabeled data as pseudo-labels for strongly augmented views and achieves SOTA performance.

2.2 Supervised Contrastive Learning

Self-supervised contrastive learning has been widely noticeable for its excellent performance by training models using unlabeled data and fine-tuning them for downstream tasks. MoCo [8] and SimCLR [5] establish the classical framework of self-supervised contrastive learning, which distinguishes the representations of each sample from the others. The contrastive learning frameworks consider different augmented views of the same sample as positive sample pairs and other samples as negative samples, by minimizing the info Noise Contrastive Estimation (InfoNCE) loss to pull the positive samples together and to push the negative samples away in the low-dimensional embedding space. For semi-supervised tasks, SimCLR v2 [6] indicates that a big self-supervised pre-trained model is a strong semi-supervised learner and simply fine-tunes the pre-trained model by using labeled samples to train a semi-supervised model. However, self-supervised contrastive learning only considers data features without focusing on class information and causes class conflicts by pushing far away samples, resulting in the inability to be directly combined with SSL. Supervised contrastive learning [9] extends the self-supervised contrastive learning methods by leveraging labeled data information to pull the samples of one class closer and push apart clusters of samples from different classes in a low-dimensional embedding space. Therefore, supervised contrastive learning mitigates the class collision phenomenon and it can be considered for application in SSL tasks.

3 Method

In this section, we introduce the preliminaries and present the two levels of interaction of DualMatch. Consisting of the new manner, the final objective is constructed.

3.1 Preliminaries

The semi-supervised classification setting is described as follows. For an C-class classification problem, given a batch of B labeled samples $\mathcal{X} = \{(x_b, y_b) : b \in (1, \ldots, B)\}$,

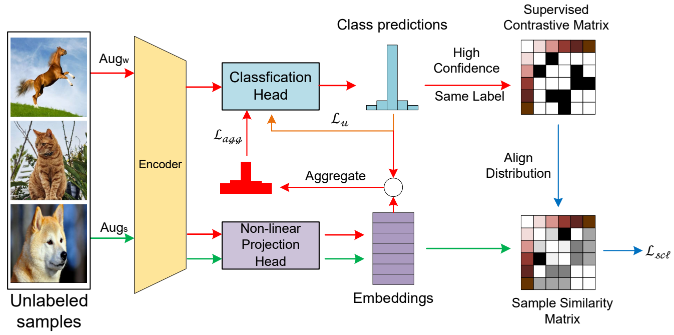


Fig. 2: The framework of the proposed DualMatch. Given a batch of unlabeled images, a class prediction of weakly augmented views is generated by the classifier head. The first-level interaction aligning distribution: pseudo-labels with high confidence are used to generate the supervised contrastive matrix and the sample similarity matrix is constructed by computing the similarity between strongly augmented embeddings to match the supervised contrastive matrix. The second-level interaction aggregating pseudo-labeling: the low-dimensional embedding similarity features of the weakly augmented view are combined with predictions to aggregate the class distribution of samples. (The red and green lines indicate the process lines of the weakly and strongly augmented views, respectively.)

where x_b denotes the training samples and y_b denotes its one-hot label, and a batch of unlabeled samples are denoted by $\mathcal{U} = \{u_b : b \in (1, \ldots, \mu B)\}$, where μ determines the relative sizes of $\mathcal{X}$ and $\mathcal{U}$. Given those settings, the next is to learn a convolutional encoder $f(\cdot)$ with labeled and unlabeled samples, a fully connected classification head $g(\cdot)$, and a non-linear projection head $h(\cdot)$. In particular, the labeled samples are randomly weakly augmented $\mathrm{Aug}_{\mathrm{w}}(\cdot)$ predicted by the classifier head $p_b = g(f(\mathrm{Aug}_{\mathrm{w}}(x_b)))$. Then the labeled samples can be optimized with cross-entropy loss which evaluates the ground-truth labels and the class predictions:

$$\mathcal{L}_x = \frac{1}{B} \sum_{b=1}^{B} \mathrm{H}(y_b, p_b), \tag{1}$$

where $\mathrm{H}(\cdot, \cdot)$ denotes the cross-entropy between two distributions.

Following FixMatch, [21] apply the weak augmentation operation and the Random Augmentation method as strong augmentation operation $\mathrm{Aug}_{\mathrm{s}}(\cdot)$ to the unlabeled samples to obtain weakly and strongly augmented views respectively. The unsupervised classification loss can be defined as the cross-entropy loss of the predictions of the two views:

$$\mathcal{L}_u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}(\max(\mathrm{DA}(p_b^w)) \geq \tau) \, \mathrm{H}(\hat{y}_b, p_b^s), \tag{2}$$

where $p_b^w = g(f(\mathrm{Aug}_{\mathrm{w}}(u_b)))$ and $p_b^s = g(f(\mathrm{Aug}_{\mathrm{s}}(u_b)))$ refer to the prediction distributions of the weakly augmented and strongly augmented classifications of unlabeled samples. $\hat{y_b} = \mathrm{argmax}(\mathrm{DA}(p_b^w))$ is the pseudo-label of the predicted weakly augmented view. τ is the pseudo-label confidence threshold. We only consider data pseudo-labels with maximum class probability greater than the threshold τ. By following [2], DA $(\cdot)$ denotes the distribution alignment (DA) trick that is applied to the model's class prediction for unlabeled samples. DA maintains the predicted marginal distribution of the data consistent with the true data distribution. We compute $\tilde{p}^w$ as the moving average of the model's predictions for unlabeled samples over the last 32 steps as the marginal distribution and adjust p_b^w with Normalize $(p_b^w/\tilde{p}^w)$.

3.2 The DualMatch Structure

Motivation of DualMatch. In DualMatch, we adopts dual-level interaction between class prediction p of classification head and feature embedding $z = h\left(f\left(x\right)\right)$ of non-linear projection head. Following the augmentation anchoring method [2], there are two augmented views used to represent the feature embedding, i.e., weakly augmented embedding z_b^w and strongly augmented embedding z_b^s. However, some strategies use multiple augmented views to capture the feature embedding and also obtain promised performance. To provide a fair comparison with classical SSL methods [21], we generate a weakly augmented view for labeled samples and also solicit another strongly augmented view for unlabeled samples.

Framework of DualMatch. Figure 2 illustrates the DualMatch framework with dual-level interaction. In the first-level interaction, we introduce the aligning distribution algorithm, which utilizes supervised contrastive learning to cluster the feature embedding with consistent predictions. Then, we show the aggregated pseudo-labeling method in the second-level interaction, which fine-tunes the class prediction by aggregating pseudo-labels of similar feature embeddings. We below explain the two interactions of DualMatch in detail.

3.3 First-level Interaction: Align

The first-level interaction aligning distribution aims to align the underlying distribution of the class prediction and feature embedding, where its inherent assumption is that different data of one class should have similar feature embedding. Theoretically, strongly augmented views of unlabeled samples should be clustered together in the low-dimensional embedding space, while their weakly augmented views should have the same confidence level on pseudo-labeling.

Protocol of Aligning. Our protocol of aligning the class prediction and feature embedding is generalized into their matrix match. Specifically, we construct a supervised contrastive matrix to solicit those predictions with high confidence from the weakly augmented views, which are required to match its associated embeddings of the sample similarity matrix from the strongly augmented views.

In short, we construct the set $\mathcal{Z} = \mathcal{Z}_x \cup \mathcal{Z}_u$ of the feature embeddings, including all labeled feature embeddings and partial unlabeled feature embeddings, where $\mathcal{Z}_x = \{(z_b^x, y_b) : b \in (1, \ldots, B)\}$ and $\mathcal{Z}_u = \{(z_b^s, \hat{y_b}) : \max(p_b^w) \geq \tau, b \in (1, \ldots, \mu B)\}$. Note that τ denotes the confidence level threshold.

Supervised Contrastive Matrix aims to obtain associations between samples from the class prediction information of the weakly augmented views. Inspired by the positive and negative sample pairs proposed by self-supervised contrastive learning [8], we consider the samples of one class as positive samples and samples of different classes as negative samples. In this way, we construct a supervised contrastive matrix $\mathbf{W}_{scl}$ to represent the category relationship between different samples, where the element located at the i-th row and j-th column is defined as follows:

$$w_{ij}^{scl} = \begin{cases} 0 & \text{if } \ i = j, \\ 1 & \text{if } \ y_i = y_j \ \text{ and } \ i \neq j, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Remark 1. Following contrastive learning, each sample is used as an anchor for the other samples, not as a positive sample. We thus set the samples with the same indices (i.e., elements on the diagonal) as 0 and the samples with the same labels as 1.

Sample Similarity Matrix aims to obtain the similarity between the low-dimensional feature embeddings of the samples. The sample similarity matrix $\mathbf{S}$ is constructed by computing the similarity between embeddings in the set $\mathcal{Z}$. For each element $s_{ij} \in \mathbf{S}$, it is characterized by the cosine similarity, i.e.,

$$\operatorname{sim}(z_i, z_j) = \frac{z_i^T z_j}{\|z_i\| \, \|z_j\|}. \tag{4}$$

where z_i and z_j are feature embeddings of $\mathcal{Z}$.

Recalling the protocol of aligning, improving the consistency between the class predictions and feature embeddings can be achieved by matching two matrices $\mathbf{W}_{scl}$ and $\mathbf{S}$. Due to the disagreement of metrics in the two matrices, we employ the InfoNCE loss of supervised contrastive learning [9] to align their elements:

$$\begin{aligned} \mathcal{L}_{scl} &= \sum_{i \in I} \mathcal{L}_i(z_i) \\ &= \sum_{i \in I} \frac{-1}{|\mathcal{J}(i)|} \sum_{j \in \mathcal{J}(i)} \log \frac{\exp\left(z_i \cdot z_j / t\right)}{\sum_{a \in A(i)} \exp\left(z_i \cdot z_a / t\right)}, \end{aligned} \tag{5}$$

where $i \in I$ denotes the indices of the embedding in Z, $A(i) = I \backslash \{i\}$ denotes the set of indices without i, $\mathcal{J}(i) = \{j \in A\left(i\right) : y_j = y_i\}$ is the indices of the set of positive instances of the same label as i, and t is a temperature parameter. Let $\mathbf{W}_i$ denote the i-th row of the matrix $\mathbf{W}_{scl}$. To facilitate computer calculations, the Eq. (5) can be simplified by the elements in matrix $\mathbf{W}_{scl}$ and $\mathbf{S}$ as follows:

$$\mathcal{L}_{scl} = -\sum_{i \in I} \frac{1}{\|\mathbf{W}_i\|} \sum_{j \in I} \log \frac{w_{ij}^{scl} \cdot \exp\left(s_{ij}/t\right)}{\sum_{a \in A(i)} \exp\left(s_{ia}/t\right)} \tag{6}$$

3.4 Second-level Interaction: Aggregate

The second-level interaction aggregating pseudo-labeling aims to aggregate class distributions with consistent feature embeddings to generate pseudo-labels for class prediction fine-tuning. Intuitively, samples with similar features embeddings in the low-dimensional embedding space should have the same labels, so that, for a batch of unlabeled samples, we can generate aggregated pseudo-labels by aggregating the class predictions of each sample's neighbors in the embedding space to improve pseudo-labeling robustness. To avoid the cumulative error caused by the class predictions of dissimilar samples, we select K neighbor samples with the most similar feature embeddings. Then the aggregated pseudo-label q_b^w of u_b in a batch of unlabeled samples can be defined as follows:

$$q_b^w = \frac{1}{K} \sum_{k=1}^{K} \mathrm{sim}(z_b^w, z_k^w) \cdot p_k^w, \tag{7}$$

where p_k^w and z_k^w denote the class prediction and feature embedding of weakly augmented unlabeled views, respectively. In particular, the class distribution is weighted by the similarity $\mathrm{sim}(z_b^w, z_k^w)$ of the samples to their neighbors. Since the weighted class distribution cannot directly represent the classification probabilities, we adjust $q_b^w = \text{Normalize}\left(q_b^w\right)$. Like the unsupervised classification loss, we only consider the samples with high confidence aggregated pseudo-labels. The difference is that aggregated pseudo-label is soft (a vector of probabilities) because we aim to adjust the class predictions. The aggregation loss can be optimized by cross-entropy as follows:

$$\mathcal{L}_{agg} = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}\left(\max\left(q_b^w\right) \geq \tau_1\right) \mathrm{H}\left(q_b^w, p_b^s\right), \tag{8}$$

where τ_1 is the confidence threshold of the aggregated label.

3.5 Final Objective

The overall loss of the semi-supervised DualMatch method consists of the supervised loss $\mathcal{L}_x$ (w.r.t. Eq. (1)) and unsupervised loss $\mathcal{L}_u$ (w.r.t. Eq. (2)). Meanwhile, to achieve the consistency of the classification prediction and feature embedding, we add the supervised contrastive loss $\mathcal{L}_{scl}$ (w.r.t. Eq. (6)), and aggregation loss $\mathcal{L}_{agg}$ (w.r.t. Eq. (8)). In such settings, our optimization objective is to minimize the overall loss:

$$\mathcal{L}_{overall} = \mathcal{L}_x + \lambda_u \mathcal{L}_u + \lambda_{scl} \mathcal{L}_{scl} + \lambda_{agg} \mathcal{L}_{agg}, \tag{9}$$

where λ_u, λ_{scl}, and λ_{agg} are hyperparameters used to control the weights of loss. DualMatch can be summarized as Algorithm 1.

Exponential Moving Average. From the perspective of consistent model regularization, we employ the Exponential Moving Average (EMA) strategy [24] to smooth the model parameters with an expectation of lower variation. Technically, the parameters of EMA are usually weighted by previously associated model parameters in the iterative updates:

$$\overline{\theta} = m\overline{\theta} + (1 - m)\,\theta, \tag{10}$$

where $\overline{\theta}$ denotes the parameters of the EMA model, θ denotes the parameters of the training model, and m denotes the EMA decay rate. Note that the experiments also employ the EMA model for testing.

Algorithm 1: DualMatch algorithm.

```
Input: Labeled batch $\mathcal{X} = \{(x_b, y_b) : b \in (1, \ldots, B)\}$, unlabeled batch
  $\mathcal{U} = \{u_b : b \in (1, \ldots, \mu B)\}$, encoder $f(\cdot)$, classification head $g(\cdot)$, non-linear
  projection head $h(\cdot)$.
for step=1 to total-step do
    $p_b = g\left(f\left(\mathrm{Aug_w}\left(x_b\right)\right)\right) \quad z_b^x = h\left(f\left(\mathrm{Aug_w}\left(x_b\right)\right)\right)$
    $p_b^w = g\left(f\left(\mathrm{Aug_w}\left(u_b\right)\right)\right) \quad z_b^w = h\left(f\left(\mathrm{Aug_w}\left(x_b\right)\right)\right)$
    $p_b^s = g\left(f\left(\mathrm{Aug_s}\left(u_b\right)\right)\right) \quad z_b^s = h\left(f\left(\mathrm{Aug_s}\left(x_b\right)\right)\right)$
    $\hat{y_b} = \mathrm{argmax}\left(\mathrm{DA}\left(p_b^w\right)\right)$
    Construct feature embedding set $\mathcal{Z} = \mathcal{Z}_x \cup \mathcal{Z}_u$
    $\mathcal{Z}_x = \{(z_b^x, y_b) : b \in (1, \ldots, B)\}$
    $\mathcal{Z}_u = \{(z_b^s, \hat{y_b}) : \max(p_b^w) > \tau, b \in (1, \ldots, \mu B)\}$
    for $i \in \{1, ..., \mu B\}$ and $j \in \{1, ..., \mu B\}$ do
        $w_{ij}^{scl} = \in \mathbf{W}_{scl}$ is constructed by Eq.( 3)
        $s_{ij} \in \mathbf{S}$ is constructed by Eq. (4)
    end
    $q_b^w = \frac{1}{K} \sum_{k=1}^{K} \mathrm{sim}(z_b^w, z_k^w) \cdot p_k^w$
    $q_b^w = \mathrm{Normalize}\left(q_b^w\right)$
    $\mathcal{L}_x = \frac{1}{B} \sum_{b=1}^{B} \mathrm{H}\left(y_b, p_b\right),$
    $\mathcal{L}_u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}\left(\max(\mathrm{DA}\left(p_b^w\right)) \geq \tau\right) \mathrm{H}\left(\hat{y_b}, p_b^s\right)$
    $\mathcal{L}_{scl} = -\sum_{i \in I} \frac{1}{\|\mathbf{W}_i\|} \sum_{j \in I} \log \frac{w_{ij}^{scl} \cdot \exp\left(s_{ij}/t\right)}{\sum_{a \in A(i)} \exp(s_{ia}/t)}$
    $\mathcal{L}_{agg} = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}\left(\max\left(q_b^w\right) \geq \tau_1\right) \mathrm{H}\left(q_b^w, p_b^s\right)$
    $\mathcal{L}_{overall} = \mathcal{L}_x + \lambda_u \mathcal{L}_u + \lambda_{scl} \mathcal{L}_{scl} + \lambda_{agg} \mathcal{L}_{agg}$
    Optimize $f(\cdot)$, $g(\cdot)$ ,and $h(\cdot)$ by minimizing $\mathcal{L}_{overall}$
end
Output: Trained model.
```

4 Experiment

In this section, we evaluate DualMatch on several semi-supervised tasks including semi-supervised classification and class-imbalanced semi-supervised classification. Our ablation studies the effect of dual-level interaction and hyperparameters on the framework.

4.1 Semi-supervised Classification

First, we evaluate DualMatch on the semi-supervised classification using the CIFAR-10, CIFAR-100 and STL-10 datasets. CIFAR-10 consists of 60,000 32×32 images divided into 10 classes, with 6,000 images in each class. There are 50,000 training images and 10,000 test images. Following the widely adopted setting in SSL studies Fixmatch [21], we randomly select 4, 25, and 400 samples per class from the training set as labeled data and then use the rest of the training set as unlabeled data, respectively. In this setting, CIFAR-100 has the same number of training set and test set images as CIFAR-10, while CIFAR-100 is divided into 100 classes with 600 images in each class. We thus randomly select 25, 100 samples per class as labeled data. STL-10 has 5,000

Table 1: Error rate (mean±std %) of semi-supervised classification of DualMatch vs. baseline methods over varying numbers of labeled samples (5 runs).

	CIFAR-10			CIFAR-100		STL-10
Method	40 labels	250 labels	4000 labels	2500 labels	10000 labels	1000 labels
Π-Model	74.34±1.76	54.26±3.97	41.01±0.38	57.25±0.48	37.88±0.11	32.78±0.40
Pseudo-Labeling	74.61±0.26	49.78±0.43	16.09±0.28	57.38±0.46	36.21±0.19	32.64±0.71
Mean Teacher	70.09±1.60	32.32±2.30	9.19±0.19	53.91±0.57	35.83±0.24	33.90±1.37
MixMatch	47.54±11.50	11.05±0.86	6.42±0.10	39.94±0.37	28.31±0.33	21.70±0.68
UDA	29.05±5.93	8.82±1.08	4.88±0.18	33.13±0.22	24.50±0.25	6.64±0.17
ReMixMatch	19.10±9.64	5.44±0.05	4.72±0.13	27.43±0.31	23.03±0.56	6.74±0.14
FixMatch	13.81±3.37	5.07±0.65	4.26±0.05	28.29±0.11	22.60±0.12	6.25±0.33
CoMatch	6.91±1.39	4.91±0.33	4.06±0.03	27.18±0.21	21.83±0.23	8.66±0.41
CR	**5.69**±0.90	5.04±0.30	4.16±0.13	27.58±0.37	21.03±0.23	6.96±0.42
DualMatch(Ours)	5.75±1.01	**4.89**±0.52	**3.88**±0.10	**27.08**±0.23	**20.78**±0.15	**5.94** ±0.08

labeled and 100,000 unlabeled 96×96 images in 10 classes for training, and 8,000 images for testing. We randomly select 100 samples per class from labeled images as labeled data. Please note that we evaluate the experiment with different random seeds for 5 runs.

Implementation Details. We use the Wide ResNet-28-2 [27] with a weight decay of 0.0005 for the CIFAR-10, Wide ResNet-28-8 with a weight decay of 0.001 for the CIFAR-100 , and Wide ResNet-37-2 with a weight decay of 0.0005 for the STL-10. The classification head is a softmax layer and the non-linear projection head is set as a two-layer MLP. Following the implementation of [21], the model uses the SGD optimizer with the Nesterov momentum [23] of 0.9. For the learning rate, we use the cosine learning rate decay and set the learning rate to $0.03 \cdot \cos\left(\frac{7\pi n}{16N}\right)$, where n denotes the current training steps and N denotes the number of the total training steps. For the rest of the hyperparameters, we set $\lambda_u = 1$, $\lambda_{scl} = 1$, $\lambda_{agg} = 0.5$, $\mu = 7$, $B = 64$, $\tau = 0.95$, $\tau_1 = 0.9$, and $t = 0.5$, $m = 0.999$. For the training steps, we set $N = 2^{20}$ for CIFAR-10, STL-10 and $N = 2^{19}$ for CIFAR-100. Moreover, we utilize the warm-up trick to train aggregation loss after the first 30×2^{10} training steps. For the neighbor settings, we set $K = 10$ for CIFAR-10, STL-10 and $K = 2$ for CIFAR-100. For data augmentation, we follow the implementation details of the weak and strong augmentation of FixMatch [21].

Compared Methods. We compare with the following baseline methods: 1) Model-level consistency methods including the Π-Model [19], Pseudo-labeling [13], and Mean Teacher [24], 2) Data-level methods including the UDA [26], MixMatch [3], ReMixMatch [2], FixMatch [21], CoMatch [15], CR [14].

Results. The SSL results are presented in Table 1, where DualMatch achieves SOTA performance at different number settings of labeled samples. For model-level consistency, we observe that the Π-model, Pseudo-Labeling, and Mean Teacher perform poorly with extremely few numbers of labeled samples, but the improvement in error rate becomes more significant after adding more labeled samples. It is thus the model-level consistency semi-supervised methods that are highly dependent on the number of labeled samples. For data-level consistency, we observed that the performance of UDA,

Table 2: Error rate (mean±std %) for CIFAR-10 with the labeled ratio $\beta = 10\%$ and imbalance ratio $\gamma = \{50, 100, 200\}$ (5 runs).

Method	$\gamma = 50$	$\gamma = 100$	$\gamma = 200$
Pseudo-Labeling	47.5±0.74	53.5±1.29	58.0±1.39
Mean Teacher	42.9±3.00	51.9±0.71	54.9±1.28
MixMatch	30.9±1.18	39.6±2.24	45.5±1.87
FixMatch	20.6±0.65	33.7±1.74	40.3±0.74
FixMatch w/ DA	19.8±0.45	30.3±1.27	38.0±0.84
CoMatch	19.7±0.68	28.6±1.85	40.0±1.56
DualMatch(Ours)	**19.0**±0.82	**28.3**±1.38	**37.3**±0.39

MixMatch, ReMixMatch, and FixMatch with the help of data augmentation methods improved significantly in extremely few labeled samples. Moreover, the performance of the semi-supervised methods using strong augmentation (e.g., Randaugment [7]) tricks exceeds that of simple that of the simple tricks for data augmentation, e.g., mixup. Therefore, the data-level consistency semi-supervised methods utilize various data augmentation tricks to overcome the shortcoming of insufficient labeled data volume. Compared to the above methods, CoMatch and DualMatch introduce feature embedding to further exploit the underlying distribution of classes, and the error rate reduction of training on 40 labeled samples of CIFAR-10 is much better than that of the data-level and model-level consistency methods. Furthermore, training on 250 and 4000 labeled samples of CIFAR-10 also achieves attractive results, but not so significantly as 40 labeled samples. Additionally, compared with FixMatch, DualMatch achieves a 9% error reduction in CIFAR-10. The potential result is that such a semi-supervised training manner with efficient feature embeddings performs closely to fully supervised training in CIFAR-10.

4.2 Class-imbalanced Semi-supervised Classification

Standard SSL assumes the class distribution is balanced, however, in real-world tasks, the data distribution is often class-imbalanced [16]. How to guarantee the performance robustness of SSL algorithms under class-imbalanced settings is an important problem that has attracted the great attention of SSL researchers [10,25]. Therefore, we also conduct experiments to evaluate the effectiveness of our proposal on class-imbalanced semi-supervised learning problems. DARP [10] denotes that class-imbalanced data biases SSL methods in generating pseudo-labels for the majority classes. To evaluate the effectiveness of the semi-supervised model in the class-imbalance task, we compare the results of Dualmatch and major semi-supervised methods under imbalanced data distribution.

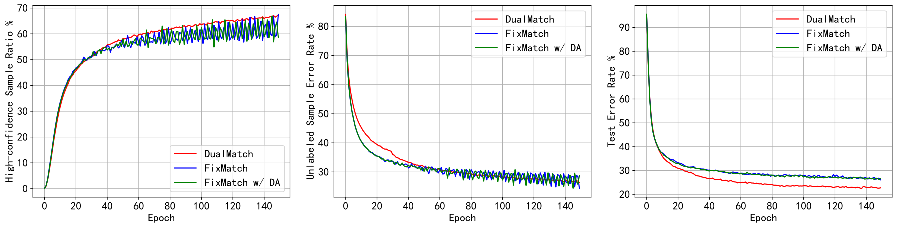
Problem Setup. By following [25], for an C-class classification problem, given a labeled set $\mathcal{X} = (x_m, y_m) : m \in (1, \ldots, M)$, where x_m are the training samples and y_m are one-hot labels. The number of class c in $\mathcal{X}$ is denoted as M_c and $\sum_{c=1}^{C} M_c = M$. [25] assume that the marginal class distribution of $\mathcal{X}$is skewed and the classes are ordered by decreasing order, i.e. $M_1 \geq M_2 \geq \cdots \geq M_C$. Class imbalance can be measured by the imbalance ratio $\gamma = \frac{M_1}{M_C}$. And given a unlabeled set $\mathcal{U} = (u_l : l \in (1, \ldots, L)$ with the same class distribution as $\mathcal{X}$. The labeled ratio $\beta = \frac{M}{M+L}$ denotes the percentage of labeled data to the training data. Specifically, the CIFAR-10 dataset consists of 5000 images in each class, and the imbalanced majority class employs 5000 images. The setting of our evaluation experiment is on CIFAR-10 with the labeled ratio β of 10%, i.e. 500 labeled images and 4500 unlabeled images in the majority class and the imbalance ratio γ of 50, 100, and 200, respectively. For the evaluation criterion of the experiment, the data of the test set is class-balanced.

Implementation Details. We use mostly the same parameter settings as for the semi-supervised classification task, except that the number of neighbor samples K is set to 2. For each experimental setting, the training steps are set to 2^{17} for MixMatch and 2^{16} for FixMatch and CoMatch. For a fair comparison, we set the total training steps to 2^{16} for DualMatch. For each experiment, we evaluate 5 times with different random seeds and report the mean and std of the test error rate. We report the performance using the EMA model.

Results. The results of class-imbalanced semi-supervised classification are presented in Table 2. Overall, the DualMatch achieves better performance than the typical semi-supervised baselines using different imbalance ratios. Moreover, all semi-supervised baselines are affected by class-imbalanced data, and their error rate increases with the increase of the imbalance ratio. For this ratio, we also observe that the data-level consistency baselines achieve the best performance if the imbalance ratio is set as 100, at least better than the setting of 50 and 200. The potential reasons are as follows. For the imbalance ratio of 200, there is only 1 labeled sample for the minority class, which leads difficult to learn the features of the minority class during model training. For the imbalance ratio of 50, the effect of imbalanced data is not significant in causing class bias, but rather in the increase in error rate due to the reduction of training samples. For the imbalance ratio of 100, the class bias caused by class-imbalanced data leads to instability of the model and increases the std of error rate. The results show that CoMatch is more affected by the imbalance ratio and performs poorly at the imbalance ratio of 200, and the improvement of DualMatch is effective. We can conclude that DualMatch aligns the feature embeddings of one class during the training period, which can separate the features of different classes and make the classification boundary clearer. It also adjusts the bias of class prediction by aggregating feature embeddings to enhance the robustness of the classification boundary and mitigate the influence of class with few samples from others. Additionally, DualMatch achieves a 6% error reduction at the imbalance rate of 100 compared to FixMatch, and a 6.5% error reduction at the imbalance rate of 200 compared to CoMatch.

4.3 Ablation Study

We study the unlabeled data error rate of FixMatch and DualMatch on the setting of training CIFAR-100 with 10000 labeled samples. This helps us to reveal the potential influence of pseudo-labeling on semi-supervised training. Then, we analyze the head interaction and parameter perturbation of each level of DualMatch on the setting of training CIFAR-10 with 250 labeled samples.



(a) High-confidence Sample Ratio (b) Unlabeled Sample Error Rate (c) Test Error Rate

Fig. 3: The training process of FixMatch and DualMatch on CIFAR-100 with 10000 labeled samples. (a) Ratio of samples with high-confidence pseudo-labels. (b) Error rate of all unlabeled sample. (c) Error rate of test samples.

Unlabeled Samples Error Rate. In Figure 3, we study the training process of FixMatch, FixMatch with DA, and DualMatch on the setting of training CIFAR-100 with 10000 labeled samples. The potential observation factors are 1) the unlabeled sample error rate, and 2) the sample ratio of high-confidence pseudo-labels. From the presented curves of Figures 3b and 3a, as the number of training epochs increases, both the unlabeled sample error rate and high-confidence sample ratio of FixMatch fluctuate dramatically and become increasingly unstable. It is worth noting that the DualMatch starts with a high unlabeled sample error rate in the first few epochs, however, as the number of training epochs increases, the unlabeled sample error rate decreases more smoothly to the FixMatch level. The DualMatch achieves a lower test error rate than FixMatch and FixMatch with DA throughout the training process. The results show that the pseudo-labels of the unlabeled samples of FixMatch vary continuously, which makes the learning model's poor stability even worse and affects the classification results. In contrast, DualMatch provides more robust and high-quality pseudo-labeling during training, which significantly improves the performance of the semi-supervised learning model.

Align Distribution. We vary the labeled and unlabeled augmentation views of the feature embedding set to perform the ablation study of Align Distribution (AD). The results are presented in Table 3. Note that when the feature embeddings are not used, the experiments fall back to DualMatch without AD. Furthermore, we also observe that simultaneously employing both labeled and unlabeled feature embeddings can effectively improve the model performance.

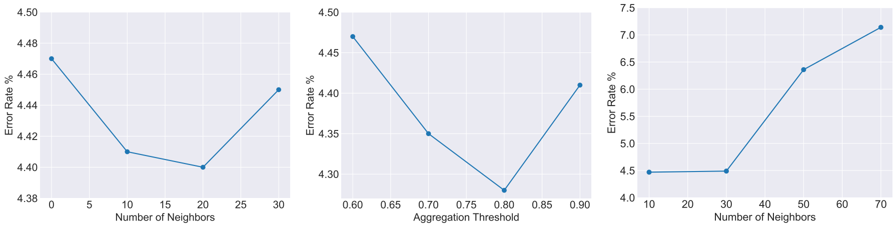
Number of Neighbors. Figure 4a illustrates the effect of different numbers of neighbors K of Eq. (7) on the aggregating pseudo-labeling. Note that $K = 0$ means that DualMatch only uses Aligning Distribution. We observe that aggregating neighbor information improves model performance, but the number of neighbors within a scope has less influence on the model with a high confidence threshold.

Aggregation Threshold. We vary the threshold τ_1 of Eq. (8) to control the confident level of aggregated labels. Figure 4b shows the effect on aggregation threshold. When $\tau_1 > 0.6$, the aggregated labels are less affected by the unreliable pseudo-labels.

We also investigate the effect of different aggregation thresholds combined with different numbers of neighbors on model performance. Figure 4c illustrates the effect

Table 3: Error rate(%) of varying the labeled and unlabeled augmentation views of the feature embedding set.

Ablation	Labeled	Unlabeled		Error Rate
	Weak	Weak	Strong	
DualMatch	1	0	1	4.41
w/o AD	0	0	0	4.77
w/o AD ($\tau_1 = 0.6$)	0	0	0	5.49
w/o labeled	0	0	1	4.68
w/o unlabeled	1	0	0	4.93
w/ multi	2	1	1	4.48



(a) Number of Neighbors.

(b) Aggregation Threshold.

(c) Number of Neighbors with $\tau_1 = 0.6$.

Fig. 4: Ablation study of the Second-level Interaction: (a)Error rate of the varying number of neighbors. (b)Error rate of the varying threshold of aggregated pseudo-labeling. (c)Error rate of the varying number of neighbors with $\tau_1 = 0.6$.

of the number of neighbors K with the aggregation threshold $\tau_1 = 0.6$. We observe that performance decreases as the number of neighbors increases with a low confidence threshold. The number of neighbors interacts with the aggregation threshold to ensure the reliability of aggregated pseudo-labeling.

5 Conclusion

Our paper introduces a novel dual-interaction method for SSL that regulates diverse augmented representations with consistent class predictions and different class predictions with coherent feature representations. Leveraging this new perspective, we present a new SSL technique named DualMatch. DualMatch could learn more data-efficient representation and provide more robust pseudo-labels than the previous single-interaction-based SSL methods. Experimental results on both standard semi-supervised settings and more challenging class-imbalanced semi-supervised settings clearly demonstrate that DualMatch can achieve significant performance improvement.

Acknowledgements This work was supported in part by National Natural Science Foundation of China, Grant Number: 62206108 and Natural Science Foundation of Jilin Province, Grant Number: 20220101107JC.

Ethical Statement

The purpose of this research paper is to explore a picture classification task under semi-supervised learning. In our study, we strictly adhere to ethical practice standards.

A number of ethical considerations were taken into account in conducting this study:

- We ensure that no private data with others is directly involved in the study.
- We ensure that no indirect leakage of researcher or participant privacy occurs or privacy can be inferred in the course of the study.
- The data were collected from publicly available datasets. The data was analyzed using open source models. We ensure that the data is reliable and public, and that our analysis methods are widely accepted by the open source community and do not contain any bias or undue influence.

We also considered potential ethical issues that may arise in the course of the study. Semi-supervised learning has been widely used in various real-world scenarios, and this study explores the potential feature of data in semi-supervised scenarios, which is uninterpretable, and uses this feature to improve the performance and robustness of the model. With the development of deep learning, the potential features provided by the encoder may be interpreted, which may lead to the leakage of data privacy when improperly handled in the application of realistic scenarios.

We can conclude that our study is based on the open source community's code, models and public datasets. At this stage it does not cause privacy issues such as personal data leakage. Moreover, our study is currently not applied in real-world scenarios and there is no conflict of interest.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2020)
2. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. International Conference on Learning Representations (2019)
3. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems **32** (2019)
4. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning. 2006. Cambridge, Massachusettes: The MIT Press View Article **2** (2006)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607 (2020)
6. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems **33** (2020)
7. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
9. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33** (2020)
10. Kim, J., Hur, Y., Park, S., Yang, E., Hwang, S.J., Shin, J.: Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. Advances in neural information processing systems **33** (2020)
11. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. International Conference on Learning Representations (2016)
12. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521** (2015)
13. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. p. 896 (2013)
14. Lee, D., Kim, S., Kim, I., Cheon, Y., Cho, M., Han, W.S.: Contrastive regularization for semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3911–3920 (2022)
15. Li, J., Xiong, C., Hoi, S.C.: Comatch: Semi-supervised learning with contrastive graph regularization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9475–9484 (2021)
16. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European conference on computer vision (ECCV). pp. 181–196 (2018)
17. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence **41** (2018)

18. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic evaluation of deep semi-supervised learning algorithms. Advances in neural information processing systems **31** (2018)
19. Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T.: Semi-supervised learning with ladder networks. Advances in neural information processing systems **28** (2015)
20. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. Advances in neural information processing systems **29** (2016)
21. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33** (2020)
22. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15** (2014)
23. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: International conference on machine learning. pp. 1139–1147 (2013)
24. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
25. Wei, C., Sohn, K., Mellina, C., Yuille, A., Yang, F.: Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10857–10866 (2021)
26. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems **33** (2020)
27. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: British Machine Vision Conference 2016 (2016)
28. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: Beyond empirical risk minimization. International Conference on Learning Representations (2018)