

# RegMixMatch: Optimizing Mixup Utilization in Semi-Supervised Learning

Haorong Han<sup>1,2</sup>, Jidong Yuan<sup>1,2\*</sup>, Chixuan Wei<sup>3</sup>, Zhongyang Yu<sup>1,2</sup>

<sup>1</sup>Key Laboratory of Big Data and Artificial Intelligence in Transportation, Ministry of Education, China

<sup>2</sup>School of Computer Science and Technology, Beijing Jiaotong University

<sup>3</sup>School of Data Science and Intelligent Media, Communication University of China

{harohan, yuanjd, zhongyangyu}@bjtu.edu.cn, chxwei@cuc.edu.cn

## Abstract

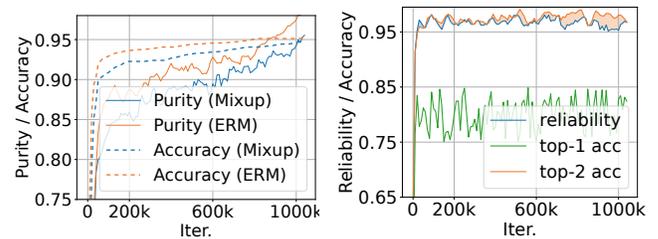
Consistency regularization and pseudo-labeling have significantly advanced semi-supervised learning (SSL). Prior works have effectively employed Mixup for consistency regularization in SSL. However, our findings indicate that applying Mixup for consistency regularization may degrade SSL performance by compromising the purity of artificial labels. Moreover, most pseudo-labeling based methods utilize thresholding strategy to exclude low-confidence data, aiming to mitigate confirmation bias; however, this approach limits the utility of unlabeled samples. To address these challenges, we propose RegMixMatch, a novel framework that optimizes the use of Mixup with both high- and low-confidence samples in SSL. First, we introduce semi-supervised Reg-Mixup, which effectively addresses reduced artificial labels purity by using both mixed samples and clean samples for training. Second, we develop a class-aware Mixup technique that integrates information from the top-2 predicted classes into low-confidence samples and their artificial labels, reducing the confirmation bias associated with these samples and enhancing their effective utilization. Experimental results demonstrate that RegMixMatch achieves state-of-the-art performance across various SSL benchmarks.

**Code** — <https://github.com/hhrd9/regmixmatch>

**Extended version** — <https://arxiv.org/abs/2412.10741>

## Introduction

Semi-supervised learning (SSL) aims to leverage a small amount of labeled data to enable the model to extract useful information from a large volume of unlabeled data during training. FixMatch (Sohn et al. 2020) and its variants (Zhang et al. 2021; Xu et al. 2021; Wang et al. 2023; Huang et al. 2023) have demonstrated competitive results in SSL by retaining pseudo-labels (Lee et al. 2013) for high-confidence samples using a threshold strategy combined with consistency regularization. However, these methods heavily rely on data augmentation techniques for consistency regularization, and their pseudo-labeling approach is constrained to high-confidence samples alone. This raises two critical questions: 1) *Are existing data augmentation techniques optimal for SSL?* 2) *Are the low-confidence samples filtered by*



(a) Purity and accuracy.

(b) Reliability and top- $n$  accuracy.

Figure 1: Motivating example of RegMixMatch. (a) Mixup’s confidence-reducing behavior undermines SSL. (b) The top-2 accuracy rate across all unlabeled data surpasses the reliability of pseudo labels.

*the threshold truly unusable?* At the core of these questions is the challenge of effectively utilizing both high- and low-confidence samples. In this paper, we aim to address these questions by proposing solutions that optimize the use of the Mixup technique (Zhang et al. 2018) in SSL.

Mixup, which generates augmented samples by interpolating between randomly shuffled data, has been introduced into SSL (Berthelot et al. 2019b,a) to enhance model generalization, yielding significant results. However, recent studies (Wen et al. 2021; Pinto et al. 2022) have revealed that Mixup can induce confidence-reducing behavior because it inherently predicts interpolated labels for every input. We argue that this confidence reduction is problematic in SSL because creating effective artificial labels for unlabeled samples is crucial for training. Methods like sharpening (Berthelot et al. 2019a) and pseudo-labeling (Sohn et al. 2020), which harden softmax prediction probabilities, have been shown to be effective. This suggests that *SSL expects to achieve purer artificial labels through low-entropy prediction probabilities, thereby better supporting model training*. However, Mixup increases the entropy of prediction probabilities, adding noise to artificial labels and consequently affecting SSL classification performance due to the lack of high-purity labels. As illustrated in Figure 1(a), the purity (the proportion of artificial labels where the highest predicted probability exceeds a threshold, reflecting the model’s prediction confidence) and test accuracy are compared when FixMatch is trained using Empirical Risk Min-

\*Corresponding Author

imization (ERM) (Vapnik 1991) and Mixup methods with pseudo labels. The results indicate that Mixup reduces prediction confidence, decreasing the number of high-purity artificial labels and leading to lower classification performance than ERM, contrasting its success in supervised learning (Zhang et al. 2018). RegMixup (Pinto et al. 2022) addresses this by using Mixup as a regularizer, training with both unmixed and mixed samples to mitigate the high-entropy issue. Inspired by this, and to counter the reduction in artificial label purity caused by Mixup’s high-entropy behavior in SSL, we propose Semi-supervised RegMixup (SRM). SRM combines SSL’s unique pseudo-labeling technique with weak-to-strong consistency regularization, applying RegMixup within SSL. However, pseudo-labeling alone only allows SRM to effectively utilize high-confidence samples. In the following section, we will demonstrate how to effectively utilize low-confidence data.

Pseudo-labeling typically employs a fixed or dynamic threshold (Zhang et al. 2021; Xu et al. 2021; Wang et al. 2023; Wei et al. 2023) to filter out low-confidence samples, retaining pseudo labels only for those with confidence above the threshold. While this strategy fully exploits high-confidence samples, it leaves samples with confidence below the threshold underutilized, as the supervisory information from these samples is generally noisier. However, unlike previous approaches that focus exclusively on leveraging high-confidence samples, we argue that low-confidence samples also hold significant potential. In Figure 1(b), we present the reliability of pseudo labels (i.e., the accuracy of predictions for unlabeled data that surpass the threshold) during FixMatch training, alongside the top-1 and top-2 accuracy rates of the model’s predictions for all unlabeled data. Notably, while the top-1 accuracy is relatively low, the top-2 accuracy exceeds the reliability of pseudo labels. This indicates that by developing methods that incorporate information from the top-2 classes into the samples and their artificial labels, we could enhance the utilization of a broader range of samples, thereby generating more reliable artificial labels even for those with low initial confidence.

To this end, we introduce a heuristic method called Class-Aware Mixup (CAM), designed to effectively utilize low-confidence samples by generating mixed samples for model training. Since artificial labels for low-confidence samples often contain significant noise due to uncertain predictions, mixing these samples with high-confidence samples of the same predicted class helps to reduce noise in the artificial labels, thereby improving their quality for these challenging samples. Our approach, dubbed **RegMixMatch**, integrates SRM and CAM, both of which leverage Mixup to exploit high- and low-confidence samples, respectively. To validate the effectiveness of this approach, we compare the training time and classification performance of RegMixMatch with previous Mixup-based SSL methods in our experiments, demonstrating significant improvements. The key contributions of our work can be summarized as follows:

- We discover and verify that Mixup reduces the purity of artificial labels in SSL, ultimately leading to a decline in model performance.

- We propose SRM, which applies RegMixup to SSL by combining pseudo-labeling and consistency regularization to address the aforementioned issue.
- We introduce CAM for low-confidence samples, aimed at mitigating confirmation bias and fully exploiting the potential of low-confidence data.
- We conduct extensive experiments to validate the performance of RegMixMatch, achieving state-of-the-art results in most scenarios.

## Related Work

Since the advent of deep learning, SSL has experienced rapid advancement. In this paper, we present an overview of SSL from two key perspectives: consistency regularization and pseudo-labeling. Additionally, we provide a brief outline of data augmentation techniques based on mixing strategies.

**Consistency Regularization.** Consistency regularization (Bachman, Alsharif, and Precup 2014) strengthens SSL by enforcing the principle that a classifier should maintain consistent class predictions for an unlabeled example, even after the application of data augmentation. This principle has led to the development of various SSL algorithms based on different augmentation techniques. Traditional SSL methods typically employ softmax or sharpened softmax outputs to supervise perturbed data, thereby achieving consistency regularization. For instance, Temporal Ensembling (Laine and Aila 2017) ensures model consistency by periodically updating the output mean and minimizing the difference in predictions for the same input across different training epochs. Mean Teacher (Tarvainen and Valpola 2017) enhances learning targets by generating a superior teacher model through the exponential moving average of model weights. VAT (Miyato et al. 2018) introduces adversarial perturbations to input data, increasing the model’s robustness. More recently, MixMatch (Berthelot et al. 2019b) and ReMixMatch (Berthelot et al. 2019a) have incorporated the Mixup technique to create augmented samples, thereby boosting generalization performance. ReMixMatch, along with UDA (Xie et al. 2020), also integrates advanced image augmentation techniques such as AutoAugment (Cubuk et al. 2019) and RandAugment (Cubuk et al. 2020), which generate heavily distorted yet semantically intact images, further enhancing classification performance.

**Pseudo-Labeling.** Pseudo-labeling (Lee et al. 2013) has been demonstrated to offer superior supervision for augmented samples in SSL. To address the issue of confirmation bias inherent in pseudo-labeling, confidence-based thresholding techniques have been developed to enhance the reliability of pseudo labels. FixMatch (Sohn et al. 2020) employs a fixed threshold to generate pseudo labels from high-confidence predictions on weakly augmented images, ensuring consistency with their strongly augmented counterparts. FlexMatch (Zhang et al. 2021) and FreeMatch (Wang et al. 2023) introduce class-specific thresholds that adapt based on each class’s learning progress. Dash (Xu et al. 2021) dynamically adjusts the threshold during training to refine the

filtering of pseudo labels. MPL (Pham et al. 2021) leverages a teacher network to produce adaptive pseudo labels for guiding a student network on unlabeled data. SoftMatch (Chen et al. 2023) tackles the trade-off between pseudo-label quality and quantity through a weighting function. Building on pseudo-labeling, FlatMatch (Huang et al. 2023) minimizes a cross-sharpness measure to ensure consistent learning across labeled and unlabeled datasets, while SequenceMatch (Nguyen 2024) introduces consistency constraints between augmented sample pairs, reducing discrepancies in the model’s predicted distributions across different augmented views. Although pseudo-labeling enhances supervision and overall performance in these methods, its inherent limitations restrict the utilization to primarily high-confidence samples.

In this paper, we employ the SRM, specifically designed for SSL, to achieve consistency regularization by addressing the issue of reduced artificial label purity through the retention of a portion of clean samples for training. Additionally, we explore methods to utilize low-confidence samples through CAM. The differences from previous Mixup-based approaches are detailed in the appendix.

**Mixing-Based Data Augmentation.** Mixup (Zhang et al. 2018) generates augmented data by performing a linear interpolation between samples and their corresponding labels, aiming to achieve smoother decision boundaries. Unlike pixel-level Mixup, Manifold Mixup (Verma et al. 2019) applies interpolation within the model’s hidden layers, allowing regularization at different representation levels. CutMix (Yun et al. 2019) combines the principles of Mixup and Cutout (DeVries and Taylor 2017) by cutting a region from one image and pasting it onto another, thereby mixing samples while preserving the local structure of the images. SaliencyMix (Uddin et al. 2021) further improves upon CutMix by detecting salient regions, ensuring that the cropped regions express label-relevant features, thus enhancing performance. ResizeMix (Qin et al. 2020) simplifies the process by resizing the source image to a small patch and pasting it onto another image, effectively retaining essential image information. For a more detailed comparison of different image mixing methods, we recommend reviewing OpenMixup (Li et al. 2023). In this work, ResizeMix is chosen as the primary mixing method, with results based on Mixup presented in the appendix for comparison.

### Preliminary: From Mixup To RegMixup

Given a training dataset  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ , ERM (Vapnik 1991) uses the empirical data distribution  $P_\delta(x, y)$  from the training set to approximate the true data distribution  $P(x, y)$ :

$$P_\delta(x, y) = \frac{1}{n} \sum_{i=1}^n \delta(x = x_i, y = y_i), \quad (1)$$

where  $\delta(x = x_i, y = y_i)$  denotes the Dirac mass centered at  $(x_i, y_i)$ . The loss function is then minimized over the data distribution  $P_\delta$ . In contrast to ERM, Vicinal Risk Minimization (VRM) (Chapelle et al. 2000) employs strategies to generate new data distributions in the vicinity of each sample,

thereby estimating a richer distribution that provides a more informed risk assessment within the neighborhood around each sample. Based on VRM, Mixup (Zhang et al. 2018) achieves a convex combination of samples to approximate the true data distribution, enhancing the model’s generalization performance. The vicinal distribution of Mixup can be represented as

$$P_v(x, y) = \frac{1}{n} \sum_{i=1}^n \delta(x = \bar{x}_i, y = \bar{y}_i), \quad (2)$$

where  $\bar{x}_i = \lambda x_i + (1 - \lambda)x_j$  and  $\bar{y}_i = \lambda y_i + (1 - \lambda)y_j$ . For each  $(x_i, y_i)$  feature-target vector, the corresponding  $(x_j, y_j)$  is drawn at random from the training data.  $\lambda$  follows a Beta distribution with parameters  $(\alpha, \alpha)$ . The larger the  $\alpha$ , the closer  $\lambda$  is to 0.5, resulting in stronger interpolation.

Research on out-of-distribution (OOD) detection by RegMixup (Pinto et al. 2022) reveals that Mixup can induce confidence-reducing behavior, impairing the model’s ability to effectively differentiate between in-distribution and out-of-distribution samples. To counteract this issue, they propose that by using Mixup to generate mixed samples while retaining clean samples for training, the high-entropy behavior can be mitigated, thereby enhancing the model’s robustness to OOD samples. In this context, the approximate data distribution  $P_v(x, y)$  can be constructed as

$$P_v(x, y) = \frac{1}{n} \sum_{i=1}^n [\delta(x = x_i, y = y_i) + \delta(x = \bar{x}_i, y = \bar{y}_i)]. \quad (3)$$

Based on this distribution, the following loss function is minimized:

$$H(y_i, p_\theta(y | x_i)) + H(\bar{y}_i, p_\theta(y | \bar{x}_i)), \quad (4)$$

where  $H(\cdot, \cdot)$  denotes the standard cross-entropy loss, and  $p_\theta$  represents the softmax output of a neural network parameterized by  $\theta$ . Later, we will introduce the proposed SRM, which integrates RegMixup into SSL by utilizing pseudo-labeling to address the reduction in artificial label purity caused by Mixup.

### RegMixMatch

Given a batch of labeled data containing  $B$  samples  $\mathcal{X} = \{(x_b^l, y_b) : b \in (1, \dots, B)\}$  and a batch of unlabeled data containing  $\mu B$  samples  $\mathcal{U} = \{(x_b^u) : b \in (1, \dots, \mu B)\}$ ,  $\mu$  is a hyperparameter that determines the relative sizes of  $\mathcal{X}$  and  $\mathcal{U}$ . The supervised loss for the labeled samples in SSL methods is typically constructed as the standard cross-entropy loss between the model’s predictions and the true labels:

$$\mathcal{L}_s = \frac{1}{B} \sum_{b=1}^B H(y_b, p_m(y | x_b^l)), \quad (5)$$

where  $p_m(y | x)$  is the predicted class distribution produced by the model for input  $x$ .

Despite the variety of data augmentation methods, they all follow consistency regularization when utilizing unlabeled data: the expectation is that the model outputs similar results for differently perturbed versions of the unlabeled data.

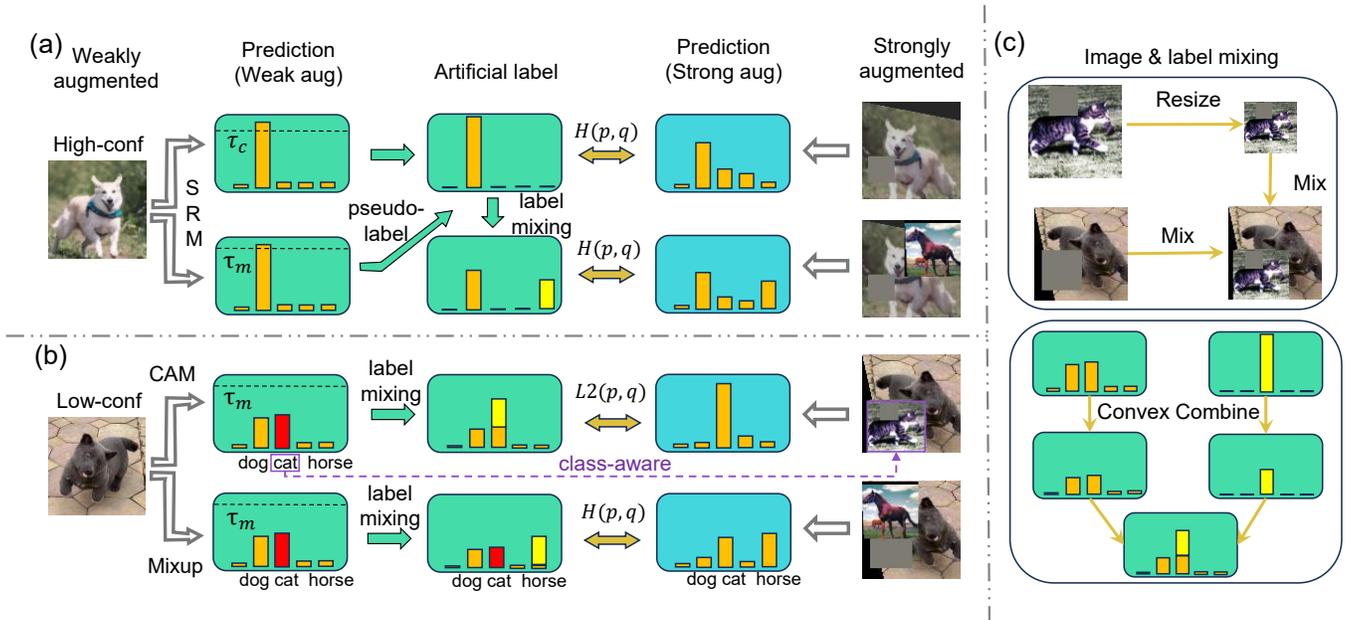


Figure 2: Overview of RegMixMatch. (a) shows the main idea of SRM. A weakly augmented image is fed into the model to obtain predictions. If the prediction confidence surpasses the threshold  $\tau_c$ , a pseudo label is used to compute consistency loss against the model’s prediction on the corresponding strongly augmented view. If the confidence also exceeds the threshold  $\tau_m$ , its strongly augmented view and pseudo label are mixed with those of another high-confidence sample to implement Mixup. (b) illustrates why CAM works. When a low-confidence sample is mispredicted, CAM converts noise (red part) from artificial label into useful information, while MixUp suffers from it. (c) demonstrates how ResizeMix generates mixed images and labels.

Based on this idea, the consistency loss is typically constructed as

$$\frac{1}{\mu B} \sum_{b=1}^{\mu B} \ell_c(p_m(y|\alpha(x_b^u)), p_m(y|\alpha(x_b^u))), \quad (6)$$

where  $\ell_c$  is the consistency loss, which can be cross-entropy or  $\ell_2$  loss.  $\alpha(x)$  represents a random augmentation of input  $x$ , meaning the two terms in Equation 6 can have different values due to the different random augmentations applied.

### From RegMixup to Semi-supervised RegMixup

The overall architecture of SRM is depicted in Figure 2(a). In supervised learning, RegMixup employs cross-entropy loss on both clean and mixed samples. However, in SSL, a substantial portion of the data remains unlabeled. By leveraging pseudo-labeling, we retain pseudo labels for unlabeled samples that exceed a specific confidence threshold. In this section, we will illustrate how the two terms in Equation 4 are adapted for SSL.

For an unlabeled image  $x_b^u$ , we first apply a weak augmentation  $\alpha(x_b^u)$  and compute the model’s predicted class distribution  $q_b = p_m(y|\alpha(x_b^u))$ . The class with the highest probability, denoted as  $\hat{q}_b = \arg \max(q_b)$ , is then assigned as the pseudo label for the corresponding strongly augmented view  $\mathcal{A}(x_b^u)$ . Consequently, we transform the RegMixup loss function for clean samples into the following form:

$$\mathcal{L}_u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}(\max(q_b) \geq \tau_c) H(\hat{q}_b, p_m(y|\mathcal{A}(x_b^u))), \quad (7)$$

where  $\tau_c$  is the threshold above which a pseudo-label is retained, and  $\max(q_b)$  represent the maximum predicted probability of the weak augmentation  $\alpha(x_b^u)$ .

To implement Mixup, we also apply thresholding to retain pseudo labels only for high-confidence unlabeled images. Specifically, for a weakly augmented view  $\alpha(x_i^u)$  of an unlabeled image within a data batch, we retain the pseudo label  $\hat{q}_i$  only if its confidence exceeds the threshold  $\tau_m$ . High-confidence images from this batch are then grouped into a set denoted as  $\mathcal{H} = \{x_b^u | \max(q_b) > \tau_m\}$ . For the loss of mixed data, as shown in Figure 2(a), given a strongly augmented view  $\mathcal{A}(x_i^u)$  and its pseudo label  $\hat{q}_i$  from the set  $\mathcal{H}$ , we randomly select  $\mathcal{A}(x_j^u)$  and  $\hat{q}_j$  from the same set  $\mathcal{H}$ , generating mixed images and labels using ResizeMix technique for the calculation of  $H(\cdot, \cdot)$ . Thus, we transform the loss function of RegMixup for mixed samples into

$$\mathcal{L}_m = \frac{1}{|\mathcal{H}|} \sum_{i=1}^{|\mathcal{H}|} H(\hat{q}_i \oplus \hat{q}_j, p_m(y|\mathcal{A}(x_i^u) \oplus \mathcal{A}(x_j^u))), \quad (8)$$

where the symbol  $\oplus$  denotes the image or label mixing operation, and  $|\cdot|$  represents the number of elements in the set.

Note that SRM employs two different thresholds,  $\tau_c$  and  $\tau_m$ . It has been observed that when  $\tau_m$  is close to or lower than  $\tau_c$ , the model is prone to significant confirmation bias, particularly during the early training stages. To mitigate this,  $\tau_m$  is set to a higher value. Additionally, as described in the preliminary,  $\alpha$  is used to control the mixing intensity; a higher  $\alpha$  results in a stronger mixing intensity. Our parameter  $\alpha_h$ , which controls the mixing intensity of high-

confidence samples, can be set to a higher value than previous Mixup-based SSL methods due to the retention of clean samples, allowing for better generalization performance.

In summary, we select high-confidence samples as clean samples using a high threshold  $\tau_c$  and retain their pseudo labels to compute the consistency loss. Concurrently, we set an even higher threshold  $\tau_m$  to filter and retain the high-confidence samples and their pseudo labels for implementing Mixup. Consequently, we can transform Equation 4 into the following form to achieve the transition from RegMixup to SRM:

$$\mathcal{L}_{rm} = \mathcal{L}_u + \mathcal{L}_m. \quad (9)$$

### Class-Aware Mixup

Pseudo-labeling enables our SRM to effectively leverage high-confidence samples. However, low-confidence samples are often discarded due to the uncertainty in their predictions. Our proposed CAM method facilitates the utilization of these low-confidence samples while maintaining the quality of the artificial labels.

Figure 2(b) provides a conceptual overview of our CAM. Mixup randomly selects a target image, which can belong to any class, to blend with the source image. For low-confidence unlabeled samples, where the predicted class often diverges from the true class, randomly selecting a target image can lead to confirmation bias. This is because the mixed label may incorporate noisy features from the predicted class, while the mixed image does not reflect these features. However, by limiting the samples mixed with low-confidence ones to high-confidence samples sharing the same predicted classes (i.e., class-aware), the noise can be reduced due to the inclusion of predicted class features. Additionally, Figure 1(b) shows that the top-2 accuracy is high, indicating a strong likelihood that the correct class is among the top-2 predicted classes in the mixed labels, making the inclusion of predicted class information sufficient.

We denote the complement of  $\mathcal{H}$  as  $\mathcal{H}^c$ . Given a strongly augmented view  $\mathcal{A}(x_i^u)$  of a low-confidence data in  $\mathcal{H}^c$  and its softmax output  $q_i$ , we use the strongly augmented view  $\mathcal{A}(x_j^u)$  and its pseudo label  $\hat{q}_j$  of the class-aware sample ( $\hat{q}_j = \arg \max(q_i)$ ) from the high-confidence set  $\mathcal{H}$ , generating mixed images and labels using ResizeMix technique for the calculation of  $\ell_2$  loss (as shown in Equation 10).

$$\mathcal{L}_{cm} = \frac{1}{|\mathcal{H}^c|} \sum_{i=1}^{|\mathcal{H}^c|} \|q_i \oplus \hat{q}_j - p_m(y|\mathcal{A}(x_i^u) \oplus \mathcal{A}(x_j^u))\|_2^2. \quad (10)$$

To reduce the negative impact of uncertain predictions, we use softmax outputs as artificial labels for low-confidence data and apply  $\ell_2$  loss instead of cross-entropy loss. Importantly, the parameter  $\alpha_l$  in CAM, which controls the mixing intensity for low-confidence samples, is set higher than  $\alpha_h$ . This choice helps to minimize the risk of the model overfitting to specific noisy samples, a point that will be further explored in the sensitivity analysis. Although the top- $n$  ( $n \geq 3$ ) accuracy of the model’s predictions would be higher, our experiments show that mixing more than two samples does not yield additional improvements. Ultimately,

---

### Algorithm 1: RegMixMatch Algorithm

---

- 1: **Input:** Labeled batch  $\mathcal{X} = \{(x_b^l, y_b) : b \in (1, \dots, B)\}$ , unlabeled batch  $\mathcal{U} = \{x_b^u : b \in (1, \dots, \mu B)\}$ , confidence threshold  $\tau_c$  and  $\tau_m$ , mixing intensities  $\alpha_h$  and  $\alpha_l$  for high- and low-confidence data
  - 2: Calculate  $\mathcal{L}_s$  using Equation 5  $\triangleright$  Loss for labeled data
  - 3: **for**  $b = 1$  to  $\mu B$  **do**
  - 4:  $q_b = p_m(y|\alpha(x_b^u); \theta)$   $\triangleright$  Prediction after weak augmentation
  - 5:  $\hat{q}_b = \arg \max(q_b)$   $\triangleright$  Pseudo labels for unlabeled data
  - 6: **end for**
  - 7: Calculate  $\mathcal{L}_u$  using Equation 7  $\triangleright$  Loss for clean unlabeled data
  - 8: Construct high-confidence set  $\mathcal{H} = \{x_b^u | \max(q_b) > \tau_m\}$
  - 9: Construct low-confidence set  $\mathcal{H}^c = \mathcal{U} \setminus \mathcal{H}$
  - 10: Calculate  $\mathcal{L}_m$  using Equation 8  $\triangleright$  Loss for mixed high-confidence unlabeled data
  - 11:  $\mathcal{L}_{rm} = \mathcal{L}_m + \mathcal{L}_u$   $\triangleright$  Using Mixup as a regularizer
  - 12: Calculate  $\mathcal{L}_{cm}$  using Equation 10  $\triangleright$  Loss for mixed low-confidence unlabeled data
  - 13: **return**  $\mathcal{L}_s + \mathcal{L}_{rm} + \mathcal{L}_{cm}$
- 

CAM enables the incorporation of information from the top-2 classes into the samples and their artificial labels, thereby facilitating the utilization of low-confidence samples.

Overall, our proposed SRM employs pseudo-labeling and consistency regularization for high-confidence data, applying Mixup to generate mixed samples while preserving unmixed samples for training in SSL. SRM enhances model generalization by leveraging Mixup, while simultaneously addressing the issue of degraded artificial label purity due to Mixup’s high-entropy behavior. For low-confidence samples, our CAM method mixes them with samples from specific classes, effectively utilizing these challenging samples while mitigating confirmation bias. The full algorithm of RegMixMatch is shown in Algorithm 1. The overall loss of RegMixMatch is formulated as

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_{rm} + \mathcal{L}_{cm}. \quad (11)$$

## Experiments

In this section, we present an extensive experimental evaluation of the proposed RegMixMatch method. We assess its performance across a variety of widely-used SSL datasets, including CIFAR-10/100 (Krizhevsky, Hinton et al. 2009), SVHN (Netzer et al. 2011), STL-10 (Coates, Ng, and Lee 2011), and ImageNet (Deng et al. 2009), under different labeled data conditions. The experimental results are benchmarked against 13 established SSL algorithms, including VAT (Miyato et al. 2018), Mean Teacher (Tarvainen and Valpola 2017), MixMatch (Berthelot et al. 2019b), ReMixMatch (Berthelot et al. 2019a), UDA (Xie et al. 2020), FixMatch (Sohn et al. 2020), Dash (Xu et al. 2021), MPL (Pham et al. 2021), FlexMatch (Zhang et al. 2021), SoftMatch (Chen et al. 2023), FreeMatch (Wang et al. 2023), SequenceMatch (Nguyen 2024), and FlatMatch (Huang et al. 2023).

Dataset	CIFAR10				CIFAR100			SVHN			STL10	
# Label	10	40	250	4000	400	2500	10000	40	250	1000	40	1000
VAT (Miyato et al. 2018)	79.81	74.66	41.03	10.51	85.20	46.84	32.14	74.75	4.33	4.11	74.74	37.95
Mean Teacher (Tarvainen and Valpola 2017)	76.37	70.09	37.46	8.10	81.11	45.17	31.75	36.09	3.45	3.27	71.72	33.90
MixMatch (Berthelot et al. 2019b)	65.76	36.19	13.63	6.66	67.59	39.76	27.78	30.60	4.56	3.69	54.93	21.70
ReMixMatch (Berthelot et al. 2019a)	20.77	9.88	6.30	4.84	42.75	26.03	20.02	24.04	6.36	5.16	32.12	6.74
UDA (Xie et al. 2020)	34.53	10.62	5.16	4.29	46.39	27.73	22.49	5.12	1.92	1.89	37.42	6.64
FixMatch (Sohn et al. 2020)	24.79	7.47	4.86	4.21	46.42	28.03	22.20	3.81	2.02	1.96	35.97	6.25
Dash (Xu et al. 2021)	27.28	8.93	5.16	4.36	44.82	27.15	21.88	2.19	2.04	1.97	34.52	6.39
MPL (Pham et al. 2021)	23.55	6.62	5.76	4.55	46.26	27.71	21.74	9.33	2.29	2.28	35.76	6.66
FlexMatch (Zhang et al. 2021)	13.85	4.97	4.98	4.19	39.94	26.49	21.90	8.19	6.59	6.72	29.15	5.77
SoftMatch (Chen et al. 2023)	-	4.91	4.82	4.04	37.10	26.66	22.03	2.33	-	2.01	21.42	5.73
FreeMatch (Wang et al. 2023)	<u>8.07</u>	4.90	4.88	4.10	37.98	26.47	21.68	1.97	1.97	1.96	15.56	5.63
SequenceMatch (Nguyen 2024)	-	4.80	4.75	4.15	37.86	25.99	20.10	1.96	1.89	1.79	15.45	5.56
FlatMatch (Huang et al. 2023)	15.23	5.58	4.22	<u>3.61</u>	38.76	<u>25.38</u>	<b>19.01</b>	2.46	<b>1.43</b>	<b>1.41</b>	16.20	4.82
RegMixMatch	<b>4.35</b>	<b>4.24</b>	<b>4.21</b>	<b>3.38</b>	<b>35.27</b>	<b>23.78</b>	<u>19.41</u>	<b>1.81</b>	1.77	<u>1.79</u>	<b>11.74</b>	<b>4.66</b>
Fully-Supervised	4.62				19.30			2.13			-	

Table 1: Error rates on CIFAR10/100, SVHN, and STL10 datasets. The fully-supervised results of STL10 are unavailable since we do not have label information for its unlabeled data. The best results are highlighted with **Bold** and the second-best results are highlighted with underline. We ran each task three times, and the results with standard deviation are presented in the appendix.

To underscore the rationale behind leveraging Mixup and the effectiveness of RegMixMatch, we compare its training time and classification performance with prior Mixup-based SSL methods (Berthelot et al. 2019a) and the latest state-of-the-art SSL approaches (Wang et al. 2023; Huang et al. 2023). Additionally, we perform comprehensive ablation studies and hyperparameter analysis to validate the design choices behind RegMixMatch. We also present the results of RegMixMatch on pre-trained backbones in the appendix.

For a fair comparison, and in line with previous SSL studies, we use Wide ResNet-28-2 (Zagoruyko and Komodakis 2016) for CIFAR-10 and SVHN, Wide ResNet-28-8 for CIFAR-100, and ResNet-37-2 (He et al. 2016) for STL-10. All hyperparameter settings are kept consistent with those used in prior work, as detailed in the appendix. For the implementation of RegMixMatch, the threshold  $\tau_c$  for the consistency loss is set in accordance with the FreeMatch approach, as adopted in the state-of-the-art FlatMatch method. Additionally, we provide experimental results based on the settings of FixMatch in the appendix. The values of  $\tau_m$ ,  $\alpha_h$ , and  $\alpha_l$  are set to 0.999, 1.0, and 16.0, respectively.

## Main Results

Table 1 compares the performance of RegMixMatch with existing methods across various datasets. In 12 commonly used SSL scenarios, RegMixMatch achieves state-of-the-art results in 9 cases, demonstrating significant improvement. In the remaining 3 cases, it secures the second-best performance. Notably, previous methods either exhibit minimal improvement or perform poorly in specific scenarios, such as certain datasets or particular labeled data quantities. Our experimental results highlight the comprehensiveness of RegMixMatch. Specifically, RegMixMatch achieves an error rate of only 4.35% on CIFAR-10 with 10 labels and 11.74% on STL-10 with 40 labels, surpassing the second-best results by 3.72% and 3.71%, respectively.

Table 2 presents the results of training RegMixMatch on the ImageNet dataset using MAE pre-trained ViT-B (He et al. 2022). RegMixMatch notably outperforms other methods, particularly achieving a 3.66% improvement on ImageNet with 10,000 labels. More comprehensive results and hyperparameter settings are detailed in the appendix.

# Label	FixMatch	FlexMatch	FreeMatch	RegMixMatch
1w	46.39	45.79	45.31	<b>41.65</b>
10w	28.47	27.83	27.43	<b>26.34</b>

Table 2: Error rates on ImageNet.

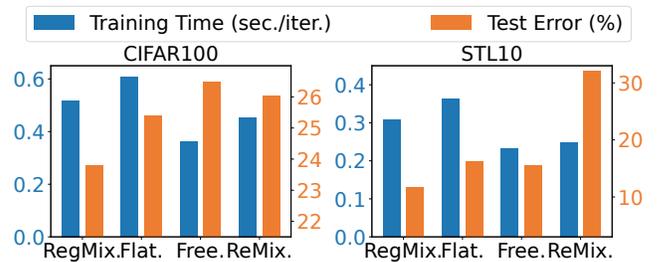


Figure 3: Efficiency analysis of RegMixMatch.

geNet with 10,000 labels. More comprehensive results and hyperparameter settings are detailed in the appendix.

## Efficiency Study

For algorithms handling classification tasks, both runtime and classification performance are critical factors. In this section, we compare RegMixMatch with several SSL algorithms in terms of both metrics. Specifically, we select FlatMatch (the previous state-of-the-art method), FreeMatch (benchmark), and ReMixMatch (Mixup-based SSL method) for comparison. The experiments are conducted on 2 24GB RTX 3090 GPU. Figure 3 presents a comparison of the 4 algorithms under two settings: CIFAR-100 with 2500 labels and STL-10 with 40 labels. The results show that RegMixMatch not only significantly improves classification performance compared to FlatMatch but also requires less training time, highlighting the efficiency of RegMixMatch. Moreover, compared to ReMixMatch and FreeMatch, RegMixMatch achieves substantial performance gains with only a minimal increase in runtime, demonstrating the effectiveness of leveraging Mixup.

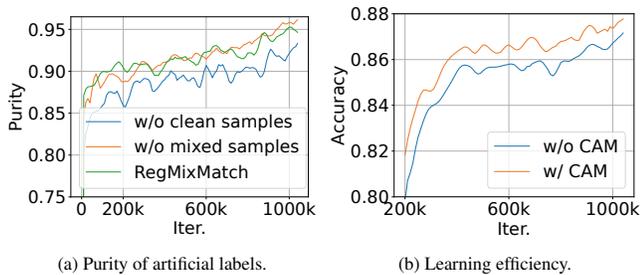


Figure 4: Ablation study of RegMixMatch. (a) demonstrates that RegMixMatch alleviates the reduced purity caused by Mixup. (b) depicts CAM improves learning efficiency.

### Ablation Study

In addition to utilizing low-confidence samples, a key distinction of RegMixMatch from other Mixup-based SSL methods is its retention of clean samples within SRM during training. To verify the rationale behind this design, we compare RegMixMatch with two scenarios: one where only clean samples are retained during training (w/o mixed samples) and another where only mixed samples are used (w/o clean samples). To further validate the rationale and effectiveness of CAM in leveraging low-confidence data, we also present results for two additional configurations: removing CAM entirely (w/o CAM), and replacing CAM with Mixup (CAM  $\rightarrow$  Mixup). The results for these ablation studies on STL-10 with 40 labels are shown in Table 3.

Table 3 highlights several key findings: 1) Training with only clean samples or only mixed samples yields suboptimal results compared to using both. RegMixMatch leverages SRM not only to enhance generalization performance but also to mitigate the negative effects of high-entropy behavior in SSL, as shown in Figure 4(a). 2) CAM improves both classification accuracy and learning efficiency, as evidenced by the higher accuracy at the same iterations in Figure 4(b). This suggests that CAM facilitates the safer utilization of low-confidence samples, enriching the training process with more informative data. 3) Replacing CAM with Mixup results in performance degradation.

### Sensitivity Analysis

RegMixMatch incorporates three key hyperparameters: the threshold  $\tau_m$ , which retains high-confidence samples for SRM; the parameter  $\alpha_h$ , which controls the mixing intensity for high-confidence samples; and the parameter  $\alpha_l$ , which

Ablation	Avg. Error Rate (%)
<b>RegMixMatch</b>	<b>11.74</b>
w/o mixed samples	15.56 (+3.82)
w/o clean samples	15.78 (+4.04)
w/o CAM	12.30 (+0.56)
CAM $\rightarrow$ Mixup	12.12 (+0.38)

Table 3: Ablation results of RegMixMatch.

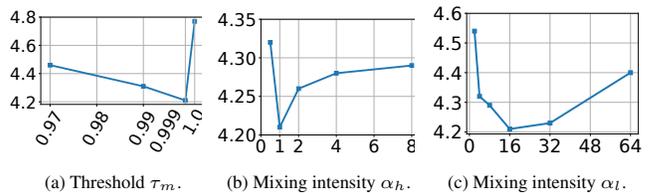


Figure 5: Parameter sensitivity analysis of RegMixMatch.

controls the mixing intensity for low-confidence samples. To assess the specific impact of these hyperparameters on model performance, we conduct experiments on CIFAR-10 with 250 labels.

**Confidence Threshold.** As illustrated in Figure 5(a), the optimal value for  $\tau_m$  is 0.999, which is higher than the threshold  $\tau_c$  (0.95). During model training, high-confidence clean samples are used to compute the consistency loss based on pseudo labels, validating the model’s predictions on these samples. If  $\tau_m$  is set to a lower value (e.g., 0.95), it would further validate the model’s predictions on these samples, leading to overconfidence and exacerbating confirmation bias. Thus,  $\tau_m = 0.999$  is chosen to balance the quality and quantity of artificial labels for mixed samples.

**Mixing Intensity.** In addition to retaining clean samples, RegMixMatch distinguishes itself from previous Mixup-based SSL methods by generating more robust mixed samples to enhance generalization. The parameter  $\alpha$  typically set to a value less than 1.0 in prior work. However, as shown in Figures 5(b) and 5(c), the optimal values for  $\alpha_h$  and  $\alpha_l$  in RegMixMatch are 1.0 and 16.0, respectively. A higher  $\alpha$  indicates stronger augmentation, improving generalization performance. It is noteworthy that  $\alpha_l \gg \alpha_h$ . As observed with Mixup, a greater mixing intensity enhances the model’s robustness to noise. Since the supervisory information for low-confidence samples is more prone to noise, a higher  $\alpha_l$  generates more intermediate-state samples, increasing the quantity and variability of the training data, and thereby reducing the risk of overfitting to specific noisy samples.

### Conclusion

In this paper, we first demonstrate that Mixup’s high-entropy behavior degrades the purity of artificial labels in SSL. To address this issue, we introduce SRM, a framework that integrates pseudo-labeling with consistency regularization through the application of RegMixup in SSL. We show that SRM effectively mitigates the reduced artificial labels purity, enhancing Mixup’s utility for high-confidence samples. Additionally, we investigate strategies for leveraging low-confidence samples. Specifically, we propose CAM, a method that combines low-confidence samples with specific class samples to enhance the quality and utility of artificial labels. Extensive experiments validate the efficiency and effectiveness of RegMixMatch. Our approach represents an initial exploration into the potential of low-confidence samples, with further in-depth studies recommended as a direction for future research.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62402031) and the Beijing Nova Program (No. 20240484620).

## References

- Bachman, P.; Alsharif, O.; and Precup, D. 2014. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27.
- Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; and Raffel, C. 2019a. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*.
- Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019b. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.
- Chapelle, O.; Weston, J.; Bottou, L.; and Vapnik, V. 2000. Vicinal risk minimization. *Advances in neural information processing systems*, 13.
- Chen, H.; Tao, R.; Fan, Y.; Wang, Y.; Wang, J.; Schiele, B.; Xie, X.; Raj, B.; and Savvides, M. 2023. SoftMatch: Addressing the quantity-quality tradeoff in semi-supervised learning. In *The eleventh international conference on learning representations*.
- Coates, A.; Ng, A.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 215–223.
- Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 113–123.
- Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.
- He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Huang, Z.; Shen, L.; Yu, J.; Han, B.; and Liu, T. 2023. Flatmatch: Bridging labeled data and unlabeled data with cross-sharpness for semi-supervised learning. *Advances in neural information processing systems*, 36: 18474–18494.
- Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.
- Laine, S.; and Aila, T. 2017. Temporal ensembling for semi-supervised Learning. In *International conference on learning representations*.
- Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896. Atlanta.
- Li, S.; Wang, Z.; Liu, Z.; Wu, D.; and Li, S. Z. 2023. Openmixup: Open mixup toolbox and benchmark for visual representation learning. *arXiv preprint arXiv:2209.04851*.
- Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1979–1993.
- Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y.; et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, 4. Granada.
- Nguyen, K.-B. 2024. SequenceMatch: Revisiting the design of weak-strong augmentations for semi-supervised learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 96–106.
- Pham, H.; Dai, Z.; Xie, Q.; and Le, Q. V. 2021. Meta pseudo labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11557–11568.
- Pinto, F.; Yang, H.; Lim, S. N.; Torr, P.; and Dokania, P. 2022. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. *Advances in neural information processing systems*, 35: 14608–14622.
- Qin, J.; Fang, J.; Zhang, Q.; Liu, W.; Wang, X.; and Wang, X. 2020. Resizemix: Mixing data with preserved object information and true labels. *arXiv preprint arXiv:2012.11101*.
- Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.
- Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.
- Uddin, A. F. M. S.; Monira, M. S.; Shin, W.; Chung, T.; and Bae, S.-H. 2021. SaliencyMix: A saliency guided data augmentation strategy for better regularization. In *International conference on learning representations*.
- Vapnik, V. 1991. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4.
- Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, 6438–6447. PMLR.

- Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; Schiele, B.; and Xie, X. 2023. FreeMatch: Self-adaptive thresholding for semi-supervised learning. In *The eleventh international conference on learning representations*.
- Wei, C.; Wang, Z.; Yuan, J.; Li, C.; and Chen, S. 2023. Time-frequency based multi-task learning for semi-supervised time series classification. *Information sciences*, 619: 762–780.
- Wen, Y.; Jerfel, G.; Muller, R.; Dusenberry, M. W.; Snoek, J.; Lakshminarayanan, B.; and Tran, D. 2021. Combining ensembles and data augmentation can harm your calibration. In *International conference on learning representations*.
- Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Un-supervised data augmentation for consistency training. *Advances in neural information processing systems*, 33: 6256–6268.
- Xu, Y.; Shang, L.; Ye, J.; Qian, Q.; Li, Y.-F.; Sun, B.; Li, H.; and Jin, R. 2021. Dash: Semi-supervised learning with dynamic thresholding. In *International conference on machine learning*, 11525–11536. PMLR.
- Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.
- Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.
- Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in neural information processing systems*, 34: 18408–18419.
- Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond empirical risk minimization. In *International conference on learning representations*.