

# InfoMatch: Entropy Neural Estimation for Semi-Supervised Image Classification

Qi Han, Zhibo Tian, Chengwei Xia and Kun Zhan\*

School of Information Science and Engineering, Lanzhou University  
kzhan@lzu.edu.cn

## Abstract

Semi-supervised image classification, leveraging pseudo supervision and consistency regularization, has demonstrated remarkable success. However, the ongoing challenge lies in fully exploiting the potential of unlabeled data. To address this, we employ information entropy neural estimation to utilize the potential of unlabeled samples. Inspired by contrastive learning, the entropy is estimated by maximizing a lower bound on mutual information across different augmented views. Moreover, we theoretically analyze that the information entropy of the posterior of an image classifier is approximated by maximizing the likelihood function of the softmax predictions. Guided by these insights, we optimize our model from both perspectives to ensure that the predicted probability distribution closely aligns with the ground-truth distribution. Given the theoretical connection to information entropy, we name our method *InfoMatch*. Through extensive experiments, we show its superior performance. The source code is available at <https://github.com/kunzhan/InfoMatch>.

## 1 Introduction

Deep learning image classifiers typically depend on a considerable amount of labeled data, encountering performance limitations as the dataset scale increases. In contrast, semi-supervised learning (SSL) [Lee, 2013; Rasmus *et al.*, 2015] leverages both limited labeled data and a substantial pool of unlabeled data, achieving performance comparable to or even surpassing fully-supervised methods with fewer labels.

Most existing SSL methods utilize two key strategies: pseudo supervision and consistency regularization. FixMatch [Sohn *et al.*, 2020] effectively combines these strategies. It utilizes weak augmented view to generate pseudolabels that serve as guidance for the prediction of strong augmented view. FlexMatch [Zhang *et al.*, 2021], FreeMatch [Wang *et al.*, 2023], and SoftMatch [Chen *et al.*, 2023] primarily focus on improving the quality and quantity

of pseudolabels. These methods center on pseudolabel selection and leverage confidence metrics to assess uncertainty, labeling samples only when confidence surpasses a predefined threshold. Additionally, variants such as MixMatch [Berthelot *et al.*, 2019b] and ReMixMatch [Berthelot *et al.*, 2019a] delve into generating numerous novel samples through diverse data augmentation techniques, intending to introduce noise and transformations for enhanced model robustness.

Our motivation does not focus on pseudolabel selection strategies and data augmentation techniques. Instead, we aim to efficiently exploit the potential of unlabeled data by adapting the entropy neural estimation from unsupervised representation learning [Hjelm *et al.*, 2019; Tian *et al.*, 2023]. The fundamental principle is to maximize the lower bound of mutual information between two augmented views [Nowozin *et al.*, 2016; Belghazi *et al.*, 2018; Hjelm *et al.*, 2019; Tian *et al.*, 2023]. These infomax approaches leverage all available unlabeled data, allowing the model to delve deeper into the inherent structure and patterns of the data, ultimately improving classification accuracy. We generate two strong augmentation views for all unlabeled data, utilizing the mutual information between these views as a lower bound on the dataset entropy. This strategy ensures the model consistently produces accurate output under varying input conditions, accurately capturing the most salient features of the data.

Another motivation of this paper is to improve classifier accuracy by approximating the posterior probability. The probability of a sample belonging to a class is determined by the posterior of the classifier. A neural classifier models its last layer using a softmax function as a parameterized posterior, meaning the classifier’s predictive probability distribution approximates the ground-truth posterior distribution. We investigate the connection between posterior entropy and the likelihood function of its predictions. Our observation suggests that the upper bound of posterior entropy is effectively approximated by maximizing the softmax prediction’s likelihood. This finding is critical for advancing model optimization, ensuring the predicted-probability distribution more accurately reflects the true distribution. We approximate the entropy of the ground-truth posterior. To apply the posterior entropy neural estimator to unlabeled samples, we employ pseudo supervision and weak-to-strong strategies in addition to the supervision loss of labeled data. We generate pseudolabels through a weak augmentation view, guiding the

\*Corresponding Author.

prediction of two strong augmented views. Additionally, we introduce CutMix [Yun *et al.*, 2019] as a strong augmentation view to capture more non-deterministic target features.

In *InfoMatch*, we are motivated by two key objectives: estimating the entropy of the data and estimating the entropy of the ground-truth posterior. We propose a novel entropy-based methodology that integrates a data entropy neural estimator with a posterior entropy neural estimator. This combination prompts the model to thoroughly explore the intrinsic structure defined by entropy, achieved by maximizing the mutual information between augmented views of all unlabeled data. Moreover, we make the crucial observation that the upper bound on posterior entropy is effectively approximated by maximizing the predicted softmax likelihood function. This insightful observation is applied to unlabeled data in our approach through weak-to-strong, pseudo-supervision, and CutMix strategies. By approximating these entropies through gradient descent, our method progressively captures information about unlabeled data and model characteristics. Since our method primarily focuses on information entropy estimations, we aptly name it *InfoMatch*. The extensive experimental results consistently validate the effectiveness of *InfoMatch*, especially in scenarios where labeled data is scarce. Our dual-entropy-based method stands as a robust solution for fully exploiting the potential of unlabeled data, providing valuable insights into both the data structure and the posterior probabilities, thereby enhancing the performance of semi-supervised image classification.

The main contributions are summarized as follows: 1) Our proposed *InfoMatch* effectively exploit the potential of unlabeled data. 2) Leveraging multiple objectives, *InfoMatch* approximates the data entropy and the ground-truth posterior entropy. By utilizing strong-to-strong contrastive, weak-to-strong pseudo supervision, and CutMix strategies, we apply this approach to unlabeled data, enabling more efficient utilization of unlabeled data information. 3) Experiments across various semi-supervised learning benchmarks validate the superior performance of *InfoMatch*.

## 2 Related Works

Semi-supervised learning is a crucial branch in the field of machine learning and computer vision. Its fundamental concept lies in utilizing the data distribution information latent in a substantial quantity of unlabeled samples to enhance learning performance when only a limited number of labeled samples are available. The main strategies behind SSL are pseudo supervision and consistency regularization.

**Consistency regularization.** Based on the smoothness assumption, consistency regularization suggests that minor perturbations applied to unlabeled samples does not result in substantial variations in their predictions. To ensure consistency,  $\Pi$  Model [Rasmus *et al.*, 2015] introduces perturbations through data augmentation and dropout, aiming to maximize the similarity of predictions derived from two forward propagations of identical unlabeled sample. Temporal Ensembling [Samuli and Timo, 2017] employs a time series combination model to minimize the mean square error loss between current and historical predictions, thus streamlin-

ing forward reasoning. MeanTeacher [Tarvainen and Valpola, 2017] converts the exponential moving average of the prediction results into the model weight, penalizing the difference in predictions between student model and teacher model. In addition, Unsupervised Data Augmentation (UDA) [Xie *et al.*, 2020] extends the data augmentation method and employs specific target data augmentation algorithms for specific tasks. In contrast, Pseudo Label [Lee, 2013] works by generating artificial labels for unlabeled data and performing fully-supervised training using both labeled and pseudo-labeled data.

**Holistic methods.** Most SSL methods combine pseudo supervision with consistency regularization to improve performance. MixMatch [Berthelot *et al.*, 2019b] leverages unlabeled data by integrating both consistency regularization and entropy minimization. Based on MixMatch, RemixMatch [Berthelot *et al.*, 2019a] introduces two innovative strategies, distribution alignment and augmentation anchor, to enhance its robustness and accuracy. FixMatch [Sohn *et al.*, 2020] which maintains the use of Augmentation Anchor, simplifies the number of strong augmentations. To improve the accuracy of pseudolabels, a fixed high threshold is established to eliminate unreliable labels. To alleviate class imbalance and address the issue of low utilization of unlabeled data in the early stage, FlexMatch [Zhang *et al.*, 2021] and FreeMatch [Wang *et al.*, 2023] take the model’s learning status and class learning difficulty into consideration to adaptively adjust the confidence threshold. SoftMatch [Chen *et al.*, 2023] uses a truncated Gaussian function to assign weights to samples based on their confidence, addressing the inherent trade-off between the quantity and quality of pseudolabels.

Furthermore, methods such as CoMatch [Li *et al.*, 2021], SimMatch [Zheng *et al.*, 2022], and SimMatchV2 [Zheng *et al.*, 2023] integrate graph-based learning methods into semi-supervised image classification, jointly learning two representations, class probabilities and low-dimensional embeddings, of the training data.

**Mixing augmentation.** The two prevalent mixing methods for images are Mixup [Zhang *et al.*, 2018] and CutMix [Yun *et al.*, 2019]. Mixup generates novel training samples by combining two images and their corresponding labels linearly, aiming to facilitate the model in learning smoother decision boundaries. Conversely, CutMix swaps segments of one image with those of another, updating the corresponding labels simultaneously to encourage the model to comprehend and leverage local information within the image. Interpolation consistency training [Verma *et al.*, 2022], MixMatch [Berthelot *et al.*, 2019b] and RemixMatch [Berthelot *et al.*, 2019a] combine Mixup with consistency regularization to expand the dataset by generating virtual “mixed” training samples, and utilize consistency regularization to ensure its generalization ability on mixed data. While FMixCutMatch [Wei *et al.*, 2021] combines Fourier space-based data cutting and data mixing to augment both labeled and unlabeled samples, thereby enhancing its performance and generalization. Furthermore, [Ghorban *et al.*, 2022] utilizes Mixup and CutMix simultaneously, aiming to blend the semantic information of three images and introduce enhanced perturba-

tions into the data. To further enhance its attention to local information and improve its sensitivity to diverse position cues, we employ CutMix [Yun *et al.*, 2019] in *InfoMatch*.

### 3 Entropy Neural Estimation

We employ an independent and identically distributed (i.i.d.) training dataset denoted as  $\mathcal{X} = \mathcal{X}_l \cup \mathcal{X}_u$ , where  $\mathcal{X}_l$  comprises labeled data pairs  $(\mathbf{x}_i, \mathbf{p}_i), \forall i \in \{1, \dots, n_l\}$ , and  $\mathcal{X}_u$  consists of unlabeled data  $\mathbf{x}_i, \forall i \in \{n_l + 1, \dots, n_l + n_u\}$ . The sizes of labeled and unlabeled datasets are denoted by  $n_l$  and  $n_u$ , respectively, with  $n_l \ll n_u$ , and the total dataset size is  $n = n_l + n_u$ . Here, each  $\mathbf{x}_i$  represents a data point, and  $\mathbf{p}_i$  corresponds to its ground-truth class label, represented using a one-hot encoding.  $\mathcal{C} = \{c_1, \dots, c_k\}$  denotes the class set. If  $\mathbf{x}_i$  belongs to  $c_j$  class, we have  $p(c_j|\mathbf{x}_i) = p_{ij} = 1$ .

We employ an encoder to generate a latent feature  $\mathbf{z}_i = [z_{ij}]$  and define the posterior probability  $\sigma(c_j|\mathbf{x}_i, \theta)$  of  $\mathbf{x}_i$  belonging to the  $j$ -th class  $c_j$  using a softmax function  $\sigma(\cdot)$ :

$$y_{ij} = \sigma(c_j|\mathbf{x}_i, \theta) = \frac{\exp(z_{ij})}{\sum_j \exp(z_{ij})}. \quad (1)$$

$y_{ij}$  denotes the probability of  $\mathbf{x}_i$  belonging to the  $j$ -th class while  $p_{ij} = p(c_j|\mathbf{x}_i) = 1$  representing a one-hot encoding. It means that we model  $\sigma(c_j|\mathbf{x}_i, \theta)$  as an encoder to approximate the truth posterior  $p_{ij} = p(c_j|\mathbf{x}_i) = 1$ .

According to the second term of the information bottleneck theory [Tishby *et al.*, 2000],  $\min I(\mathcal{X}; \mathcal{Y}) - \beta I(\mathcal{Y}; \mathcal{C})$ , the neural network is mainly coding its softmax prediction  $\sigma(c_j|\mathbf{x}_i, \theta)$  to close to the ground-truth posterior  $p(c_j|\mathbf{x}_i)$ .

**Theorem 1.** *For an i.i.d. finite dataset  $(\mathcal{X}, \mathcal{C})$ , the approximation  $H(\mathcal{C}|\mathcal{X}) \simeq -\ln p(\mathcal{C}|\mathcal{X})$  holds.*

*Proof.* Since  $\mathbf{x}_i$  are sampled i.i.d.,  $p(c|\mathbf{x}_i)$  corresponds point-to-point with  $\mathbf{x}_i$  and is regarded as an i.i.d. set, the posterior is  $p(\mathcal{C}|\mathcal{X}) = \prod_{i=1}^n p(c|\mathbf{x}_i)$ . Then the expectation with respect to  $p(\mathbf{x})$  is approximated by a finite sum over  $\mathcal{X}$ , so that

$$\begin{aligned} H(\mathcal{C}|\mathcal{X}) &= -\mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{p(c|\mathbf{x})} \ln p(c|\mathbf{x}) \\ &\simeq -\sum_{i=1}^n \ln p(c|\mathbf{x}_i) = -\ln \prod_{i=1}^n p(c|\mathbf{x}_i) \\ &= -\ln p(\mathcal{C}|\mathcal{X}). \end{aligned}$$

Thus, the equation  $H(\mathcal{C}|\mathcal{X}) \simeq -\ln p(\mathcal{C}|\mathcal{X})$  holds.  $\square$

#### 3.1 Posterior Entropy Neural Estimation

Suppose that the ground-truth posterior is being generated from the unknown distribution  $p(c|\mathbf{x}_i)$  that we aim to model. We approximate  $p(c|\mathbf{x}_i)$  using the parametric probability distribution  $\sigma(c|\mathbf{x}, \theta)$ ,

$$\text{div}_{\text{kl}}(p||\sigma) \simeq \sum_{i=1}^n \left\{ -\ln \sigma(c|\mathbf{x}_i, \theta) + \ln p(c|\mathbf{x}_i) \right\} \quad (2)$$

$$= -\left\{ \ln \sigma(\mathcal{C}|\mathcal{X}, \theta) - \ln p(\mathcal{C}|\mathcal{X}) \right\} \geq 0 \quad (3)$$

$$H(\mathcal{C}|\mathcal{X}) \leq -\ln \sigma(\mathcal{C}|\mathcal{X}, \theta). \quad (4)$$

where  $\text{div}_{\text{kl}}$  denotes the Kullback-Leibler divergence. The left-hand side of Eq. (4) is independent of  $\theta$ , and the right-hand term is the negative log likelihood function for  $\theta$  under

the posterior distribution evaluated using the training dataset. Thus we see that minimizing Eq. (3) is equivalent to maximizing the likelihood function, *i.e.*, minimizing this upper bound of the posterior entropy. For an i.i.d. finite dataset  $\mathcal{X}$ , when approximating  $p(c|\mathbf{x})$  through a parametric distribution  $\sigma(c|\mathbf{x}, \theta)$ , the following Lemma 1 holds:

**Lemma 1.** *Maximizing the likelihood function is equivalent to minimizing the upper bound on the posterior entropy.*

#### 3.2 Data Entropy Neural Estimation

We approach  $H(\mathcal{X})$  by maximizing the mutual information of pairwise augmentation views of  $\mathcal{X}$  [Oord *et al.*, 2018; Belghazi *et al.*, 2018; Ma *et al.*, 2023]. Then, we have

$$H(\mathcal{X}) = I(\mathcal{X}; \mathcal{X}) \geq I(\mathcal{X}^{(1)}; \mathcal{X}^{(2)}). \quad (5)$$

Then, the mutual information between the two augmented views  $\mathcal{X}^{(1)}$  and  $\mathcal{X}^{(2)}$  is given by

$$I(\mathcal{X}^{(1)}; \mathcal{X}^{(2)}) = \text{div}_{\text{kl}}(p(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})||p(\mathbf{z}^{(1)})p(\mathbf{z}^{(2)})) \quad (6)$$

where  $\mathbf{z}^{(1)}$  and  $\mathbf{z}^{(2)}$  are the potential features corresponding to the two augmented views respectively.

Following contrastive learning [Oord *et al.*, 2018; Belghazi *et al.*, 2018; Tian *et al.*, 2023], the maximization of the mutual information between two views turns into a lower bound maximization problem. Then, the following Lemma 2 holds,

**Lemma 2.** *Maximizing the mutual information between two augmentation views is equivalent to maximizing the lower bound of the entropy.*

#### 3.3 Entropy Estimation for SSL

We assume that both  $\mathcal{X}_l$  and  $\mathcal{X}_u$  share the same distribution as  $\mathcal{X}$ , and implement an encoder to optimize it from two bounds.

Given the modeling of labeled data point  $(\mathbf{x}, \mathbf{p})$  by the parameter  $\theta$  to produce  $\mathbf{y} = \sigma(c_j|\mathbf{x}_i, \theta)$ , we encode all labeled data  $\mathcal{X}_l$ , and the corresponding labeled data set is  $\mathcal{C}_l$ . During the training process, we strive to align the predictive coding of *InfoMatch* with the ground-truth coding, which essentially means modeling  $y_{ij}$  to align closely with  $p_{ij}$ . In other words, we aim to approximate the entropy of the ground-truth posterior  $-\ln p(\mathcal{C}|\mathcal{X})$  with its negative log likelihood  $-\ln \sigma(\mathcal{C}|\mathcal{X}, \theta)$ . Referring to Lemma 1, the likelihood is

$$\sigma(\mathcal{C}_l|\mathcal{X}_l, \theta) = \prod_{i=1}^{n_l} \prod_{j=1}^k \sigma(c_j|\mathbf{x}_i, \theta)^{p_{ij}} = \prod_{i=1}^{n_l} \prod_{j=1}^k y_{ij}^{p_{ij}}, \quad (7)$$

where  $p_{ij}$  is an element of  $n_l \times k$  matrix of ground-truth matrix  $P = [p_{ij}]$ . The loss is defined by taking the negative logarithm of the likelihood, resulting in the cross-entropy loss:

$$\ell_{\text{upper}}^l = -\ln \sigma(\mathcal{C}_l|\mathcal{X}_l, \theta) = -\sum_{i=1}^{n_l} \sum_{j=1}^k p_{ij} \ln y_{ij}. \quad (8)$$

Referring to Lemma 2, we employ the strategy of contrastive learning by maximizing the lower bound of entropy for  $\mathcal{X}_u$ . We choose Jensen-Shannon divergence over

Kullback-Leibler divergence, following the methodology in [Nowozin *et al.*, 2016]. The lower bound is

$$\begin{aligned} & \text{div}_{\text{js}}(p(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) \| p(\mathbf{z}^{(1)})p(\mathbf{z}^{(2)})) \\ & \geq \mathbb{E}_{p(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})} \log(d(\mathbf{z}^{(1)}, \mathbf{z}^{(2)} | \theta)) \\ & + \mathbb{E}_{p(\mathbf{x}^{(1)})p(\mathbf{x}^{(2)})} \log(1 - d(\mathbf{z}^{(1)}, \mathbf{z}^{(2)} | \theta)) = -\mathcal{L}_{\text{lower}} \end{aligned} \quad (9)$$

where  $\text{div}_{\text{js}}$  represents the Jensen-Shannon divergence.

Similar to [Hjelm *et al.*, 2019; Tian *et al.*, 2023], we employ the view-wise contrastive loss, *i.e.*,  $\mathcal{L}_{\text{lower}}$ ,

$$\begin{aligned} \mathcal{L}_{\text{lower}} = & -\frac{1}{|\mathcal{P}|} \sum_{(i,i) \in \mathcal{P}} \log d(\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} | \theta) \\ & - \frac{1}{|\mathcal{N}|} \sum_{(i,j) \in \mathcal{N}} \log(1 - d(\mathbf{z}_i^{(1)}, \mathbf{z}_j^{(2)} | \theta)) \end{aligned} \quad (10)$$

where  $d(\cdot, \cdot)$  is the similarity score of pairwise logits, and  $\mathcal{P}$  and  $\mathcal{N}$  denote the sets of positive and negative pairs, respectively, *i.e.*,  $\mathbf{z}_i^{(1)}$  and  $\mathbf{z}_i^{(2)}$  are belong to positive pairs  $(i, i) \in \mathcal{P}$  while  $\mathbf{z}_i^{(1)}$  and  $\mathbf{z}_j^{(2)}$  are negative pairs  $(i, j) \in \mathcal{N}, \forall i \neq j$ .

## 4 InfoMatch

Utilizing a neural network  $\sigma(c_j | \mathbf{x}_i, \theta)$  to approximate the information entropy for semi-supervised image classification, we call our method *InfoMatch*. For each unlabeled data point  $\mathbf{x} \in \mathcal{X}_u$ , we conduct augmentation processing, creating one weak view, *i.e.*, performing random flip, and two strong augmented views [Cubuk *et al.*, 2020] represented by  $\mathbf{x}'$ ,  $\mathbf{x}^{(1)}$ , and  $\mathbf{x}^{(2)}$ , respectively.

### 4.1 Minimize Upper Bound

In *InfoMatch*, neural networks aim to approximate the upper bound of posterior entropy by maximizing the likelihood.

According to Eq. (8), minimizing the cross entropy between the predicted-probability distribution and the ground-truth distribution amounts to minimizing the likelihood function. By employing optimization algorithms like gradient descent, *InfoMatch* is gradually steered away from the incorrect distribution towards the accurate one.

In practice, when dealing with labeled data, standard supervised learning which relies on cross-entropy loss is employed, and the corresponding loss function is designated as  $\mathcal{L}_{\text{upper}}^l$ .

To incorporate unlabeled data into training, we utilize pseudo supervision and weak-to-strong strategies. Let the one-hot vector  $\hat{\mathbf{p}}_i = [\hat{p}_{ij}]$  represent the pseudolabel corresponding to the weak-view prediction  $\mathbf{y}'_i$ , the pseudo-supervised loss function from weak to strong is given by:

$$\mathcal{L}_{\text{upper}}^u = -\frac{1}{2} \sum_{i=1}^{n_u} \sum_{j=1}^k m_{ij} \hat{p}_{ij} (\ln y_{ij}^{(1)} + \ln y_{ij}^{(2)}) \quad (11)$$

where  $\mathbf{y}_i^{(1)}$  and  $\mathbf{y}_i^{(2)}$  correspond to the predictions for  $\mathbf{x}_i^{(1)}$  and  $\mathbf{x}_i^{(2)}$ , respectively, and  $M = [m_{ij}]$  is utilized to mask pseudolabels with confidence levels below a threshold.

Additionally, we use CutMix to generate a new strong augmentation view on the top of weak augmentation, aiming to

create more challenging samples and compel *InfoMatch* to fully extract meaningful features. To enhance computational efficiency, we shuffle the dataset in batches and randomly select an image, denoted as  $\mathbf{x}'_r$ , at the corresponding position of  $\mathbf{x}'_i$ . Subsequently, these images are used to generate CutMix images in a one-to-one manner. Furthermore, we evaluate the quantity of semantic information retained from the original images within the CutMix images, according to the size of the region. Consequently, the CutMix image and its corresponding pseudolabel is expressed by:

$$\mathbf{x}_i^c = \mathbf{b}_\eta \odot \mathbf{x}'_i + (\mathbf{1} - \mathbf{b}_\eta) \odot \mathbf{x}'_r \quad (12)$$

$$\hat{\mathbf{p}}_{ij}^c = \eta m_{ij} \hat{p}_{ij} + (1 - \eta) m_{rj} \hat{p}_{rj} \quad (13)$$

where  $\mathbf{b}_\eta$  represents a random binary mask indicating where to exclude and incorporate information from two images,  $\eta$  denotes the area proportion of  $\mathbf{x}'_i$  in the mixed image, obtained by averaging the values of the binary matrix  $\mathbf{b}_\eta$ , and  $\odot$  performs element-wise multiplication. The loss for the CutMix image is given by

$$\mathcal{L}_{\text{upper}}^c = -\sum_{i=1}^{n_u} \sum_{j=1}^k \hat{p}_{ij}^c \ln y_{ij}^c \quad (14)$$

where the prediction of  $\mathbf{x}_i^c$  is  $\mathbf{y}_i^c$ . By combining aforementioned losses, we obtain the upper bound of posterior entropy,

$$\mathcal{L}_{\text{upper}} = \mathcal{L}_{\text{upper}}^l + \mathcal{L}_{\text{upper}}^u + \mathcal{L}_{\text{upper}}^c. \quad (15)$$

### 4.2 Maximize Lower Bound

*InfoMatch* effectively captures the mutual information between pairs of augmentation views derived from the original dataset. By Eq. (10), minimizing the contrast loss is synonymous with maximizing the mutual information between pairwise augmentation views, thereby compelling *InfoMatch* to delve into the inherent structure dictated by  $H(\mathcal{X})$ . After multiple random augmentations, the distribution of augmentation views will gradually align with the distribution of the raw dataset, *i.e.*, through multiple trainings with random augmentations of diverse samples, the neural network progressively approaches the information entropy of the original dataset.

Inspired by BYOL [Grill *et al.*, 2020], we compute the contrastive loss by utilizing the positive pairs only, guaranteeing that samples of the same class are situated adjacent to each other in the embedding space. Furthermore, latent features often serve as covert structures or patterns derived from the data, offering a deeper understanding of the relationships within the data than merely predicting probability distributions. Consequently, we utilize the mutual information between the latent features corresponding to the two augmented views to approximate the lower bound of entropy.

The similarity  $d(\cdot, \cdot)$  of positive logits is defined by the Gaussian function [Tian *et al.*, 2023],

$$d(\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} | \theta) = \exp\left(-\|\mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)}\|_2^2\right), \quad (16)$$

and the similarity defined in Eq. (16) implies that if any two logits are identical, the similarity is one; otherwise, it tends to zero as their distance increases significantly.

---

**Algorithm 1** The *InfoMatch* algorithm.

---

**Input:**  $\mathcal{X} = \mathcal{X}_l \cup \mathcal{X}_u$ , batch size  $n_l^b$  and  $n_u^b$ , parameter  $\lambda$ .

**Output:** Optimal model parameters  $\theta^*$ .

```

1: Initialization:  $epoch = 0, epoch_{\max}$ , and  $\theta$ .
2: while  $epoch \leq epoch_{\max}$  do
3:   for mini-batch samples in  $\mathcal{X}$  do
4:     Obtain augmented images  $\mathbf{x}'$ ,  $\mathbf{x}^{(1)}$ ,  $\mathbf{x}^{(2)}$  and  $\mathbf{x}^c$ ;
5:     Feed samples into model for forward propagation;
6:     Calculate  $\mathcal{L}_{\text{upper}}^l$  by Eq. (8);
7:     Obtain the mask  $M$  for filtering pseudolabels;
8:     Calculate  $\mathcal{L}_{\text{upper}}^u$  and  $\mathcal{L}_{\text{upper}}^c$  by Eqs. (11) and (14);
9:     Calculate  $\mathcal{L}_{\text{upper}} = \mathcal{L}_{\text{upper}}^l + \mathcal{L}_{\text{upper}}^u + \mathcal{L}_{\text{upper}}^c$ 
10:    Calculate  $\mathcal{L}_{\text{lower}}$  by Eq. (17);
11:    Calculate  $\mathcal{L} = \mathcal{L}_{\text{upper}} + \lambda \mathcal{L}_{\text{lower}}$ 
12:    Update parameter  $\theta$  by minimizing  $\mathcal{L}$ ;
13:     $epoch = epoch + 1$ ;
14:   end for
15: end while

```

---

Substituting Eq. (16) into Eq. (10), we obtain

$$\mathcal{L}_{\text{lower}} = \frac{1}{n_u} \sum_{i=1}^{n_u} \|z_i^{(1)} - z_i^{(2)}\|_2^2. \quad (17)$$

### 4.3 InfoMatch Algorithm

We introduce an innovative semi-supervised image classification algorithm, called *InfoMatch*, which treats the classification task as an entropy approximation problem. For labeled data, we utilize the cross entropy loss function as an upper bound for posterior entropy. *InfoMatch* mainly focuses on effectively leveraging vast amounts of unlabeled data.

In *InfoMatch* algorithm, we initiate with pseudo supervision and weak-to-strong strategies. This allows us to transform the entropy upper bound into pseudo-supervised cross entropy loss, serving as a supervisory signal for unlabeled data. Furthermore, to enrich the diversity of our dataset, we introduce a new strong augmentation method, CutMix, that enhances its generalization capabilities by formulating corresponding loss. Subsequently, we consider the potential feature contrastive loss between the two augmented views as a lower bound for dataset entropy. By maximizing this lower bound and minimizing the upper bound, we gradually optimize *InfoMatch*, even when limited labeled data is available.

The overall objective function incorporates the three cross-entropy losses of Eq. (15) and contrastive loss of Eq. (17),

$$\mathcal{L} = \mathcal{L}_{\text{upper}} + \lambda \mathcal{L}_{\text{lower}} \quad (18)$$

where  $\lambda$  serves as a non-negative hyperparameter, regulating the relative influence of the upper and lower bounds. During the experiment, we employ mini-batches for training, utilizing both labeled and unlabeled samples with batch sizes of  $n_l^b$  and  $n_u^b$ , respectively. Algorithm 1 offers a comprehensive overview of the learning process.

## 5 Experimental Results

### 5.1 Experimental Setup

We evaluate *InfoMatch* on well-known benchmark datasets, including CIFAR-10/100 [Krizhevsky and Hinton, 2009],

SVHN [Netzer *et al.*, 2011], STL-10 [Coates *et al.*, 2011], and ImageNet [Deng *et al.*, 2009]. Additionally, we conduct experiments using varying amounts of labeled data.

To ensure a fair comparison, we follow the experimental setup as in FixMatch [Sohn *et al.*, 2020] and FreeMatch [Wang *et al.*, 2023]. Specifically, we employ standard stochastic gradient descent algorithm with cosine learning rate decay as the optimizer across all datasets, with an initial learning rate of 0.03 and a momentum of 0.9. For all experiments, we set the total number of iterations to  $2^{20}$ . *InfoMatch* performance is then evaluated using the EMA with a parameter of 0.999. Additionally, for ImageNet, we maintain a batch size of 128 for both labeled and unlabeled samples, *i.e.*,  $n_l^b = n_u^b = 128$ , and utilize the ResNet-50 architecture [He *et al.*, 2016]. While for other datasets, we adjust the batch sizes to  $n_l^b = 64$  and  $n_u^b = 448$ , and employ the Wide ResNet variants, such as Wide ResNet-28-2 [Zagoruyko and Komodakis, 2016] and Wide ResNet-28-8 [Zhou *et al.*, 2020].

To evaluate the impact of the two distinct terms of the loss Eq. (18) in *InfoMatch*, we introduce a non-negative hyperparameter  $\lambda$ . Then, we ensure an equitable representation of both real and pseudolabels in the likelihood function and maintain equilibrium between RandAugment [Cubuk *et al.*, 2020] and CutMix [Yun *et al.*, 2019]. Subsequently, we adjust the parameter  $\lambda$  that regulates the entropy bounds to 0.002.

In addition, we employ distinct threshold selection strategies to eliminate the impact of pseudolabel errors on *InfoMatch*. For SVHN, we adopt a fixed threshold of 0.95. Conversely, for other datasets, we utilize the self-adaptive thresholding method proposed by FreeMatch, which measures the overall learning progress by utilizing the expectation of the highest confidence across all unlabeled data in the current batch as the global threshold and assess the class-specific learning status by calculating the average prediction probabilities corresponding to each class.

Finally, to guarantee a precise and unbiased evaluation, we conduct multiple training sessions for each model using various random seeds and calculate the mean and standard deviation of the optimal accuracy achieved.

### 5.2 Main Results

We compare *InfoMatch* with full-supervised learning method and a range of representative semi-supervised learning methods, including pseudolabel-based methods such as FlexMatch [Zhang *et al.*, 2021] and FreeMatch [Wang *et al.*, 2023], as well as graph-based methods like SimMatch [Zheng *et al.*, 2022] and SimMatchv2 [Zheng *et al.*, 2023]. The Top-1 error rates for CIFAR-10/100, SVHN, and STL-10 under various labeled data sizes are presented in Tabs. 1 and 3.

Based on these results, it is evident that *InfoMatch* exhibits superior performance across all benchmarks, enhancing performance by an average of 1.49%, 0.55%, 0.13%, and 2.12% on the four datasets. In particular, *InfoMatch* impressively lower the mean Top-1 error rate from 8.07% to 4.39% (−3.68%) in the CIFAR-10 dataset, which contains just 10 labeled data. While in the STL-10 dataset with 40 labeled data, the Top-1 error rate is reduces from 13.74% to 9.86% (−3.88%). Additionally, on both CIFAR-10 and SVHN datasets, our approach achieves remarkable performance im-

Dataset	CIFAR-10				CIFAR-100
# Label	10	40	250	4000	10000
II Model [Rasmus <i>et al.</i> , 2015]	79.18 $\pm$ 1.11	74.34 $\pm$ 1.76	46.24 $\pm$ 1.29	13.13 $\pm$ 0.59	36.65 $\pm$ 0.00
Pseudo Label [Lee, 2013]	80.21 $\pm$ 0.55	74.61 $\pm$ 0.26	46.49 $\pm$ 2.20	15.08 $\pm$ 0.19	36.55 $\pm$ 0.24
VAT [Miyato <i>et al.</i> , 2018]	79.81 $\pm$ 1.17	74.66 $\pm$ 2.12	41.03 $\pm$ 1.79	10.51 $\pm$ 0.12	32.14 $\pm$ 0.19
MeanTeacher [Tarvainen and Valpola, 2017]	76.37 $\pm$ 0.44	70.09 $\pm$ 1.60	37.46 $\pm$ 3.30	8.10 $\pm$ 0.21	31.75 $\pm$ 0.23
MixMatch [Berthelot <i>et al.</i> , 2019b]	65.76 $\pm$ 7.06	36.19 $\pm$ 6.48	13.63 $\pm$ 0.59	6.66 $\pm$ 0.26	27.78 $\pm$ 0.29
ReMixMatch [Berthelot <i>et al.</i> , 2019a]	20.77 $\pm$ 7.48	9.88 $\pm$ 1.03	6.30 $\pm$ 0.05	4.84 $\pm$ 0.01	<u>20.02</u> $\pm$ 0.27
UDA [Xie <i>et al.</i> , 2020]	34.53 $\pm$ 10.69	10.62 $\pm$ 3.75	5.16 $\pm$ 0.06	4.29 $\pm$ 0.07	22.49 $\pm$ 0.23
FixMatch [Sohn <i>et al.</i> , 2020]	24.79 $\pm$ 7.65	7.47 $\pm$ 0.28	4.86 $\pm$ 0.05	4.21 $\pm$ 0.08	22.20 $\pm$ 0.12
Dash [Xu <i>et al.</i> , 2021]	27.28 $\pm$ 14.09	8.93 $\pm$ 3.11	5.16 $\pm$ 0.23	4.36 $\pm$ 0.11	21.88 $\pm$ 0.07
MPL [Pham <i>et al.</i> , 2021]	23.55 $\pm$ 6.01	6.62 $\pm$ 0.91	5.76 $\pm$ 0.24	4.55 $\pm$ 0.04	21.74 $\pm$ 0.09
FlexMatch [Zhang <i>et al.</i> , 2021]	13.85 $\pm$ 12.04	4.97 $\pm$ 0.06	4.98 $\pm$ 0.09	4.19 $\pm$ 0.01	21.90 $\pm$ 0.15
FreeMatch [Wang <i>et al.</i> , 2023]	<u>8.07</u> $\pm$ 4.24	<u>4.90</u> $\pm$ 0.04	4.88 $\pm$ 0.18	4.10 $\pm$ 0.02	21.68 $\pm$ 0.03
CoMatch [Li <i>et al.</i> , 2021]	-	6.91 $\pm$ 1.39	4.91 $\pm$ 0.33	4.27 $\pm$ 0.12	22.11 $\pm$ 0.22
SimMatch [Zheng <i>et al.</i> , 2022]	-	5.60 $\pm$ 1.37	<u>4.84</u> $\pm$ 0.39	<u>3.96</u> $\pm$ 0.01	20.58 $\pm$ 0.11
SimMatchV2 [Zheng <i>et al.</i> , 2023]	-	<u>4.90</u> $\pm$ 0.16	5.04 $\pm$ 0.09	4.33 $\pm$ 0.16	21.37 $\pm$ 0.20
<b>InfoMatch (Ours)</b>	<b>4.39</b> $\pm$ 0.22	<b>4.22</b> $\pm$ 0.14	<b>4.01</b> $\pm$ 0.07	<b>3.29</b> $\pm$ 0.08	<b>19.47</b> $\pm$ 0.56
Fully-Supervised	4.62 $\pm$ 0.05				19.30 $\pm$ 0.09

Table 1: Top-1 error rates (%) on CIFAR-10/100 datasets. **Bold** indicates the best result, while underline indicates the second-best result.

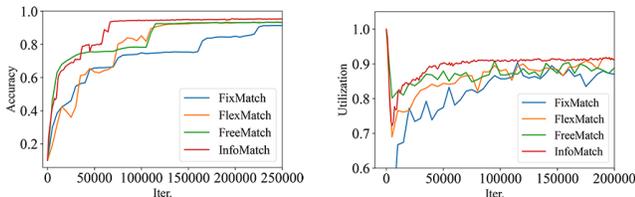


Figure 1: Compare *InfoMatch* with FixMatch, FlexMatch, and FreeMatch on the CIFAR-10 dataset with 40 labeled data in terms of Top-1 accuracy and utilization of unlabeled data. (a) Top-1 accuracy, and (b) Utilization of unlabeled data.

provements, surpassing not only other semi-supervised baselines but also outperforming fully-supervised learning methods across all benchmarks. Furthermore, comparing to other methods, *InfoMatch* exhibits a smaller standard deviation across multiple experiments with varying seeds, thereby highlighting the superior stability and robustness of *InfoMatch*.

It is worth noting that *InfoMatch* significantly outperforms other methods when the number of labeled data is extremely limited. Especially in the CIFAR-10 dataset, even when only one labeled data is available for each class, the average Top-1 error rate reaches 4.39%, which is lower than that of fully-supervised learning methods ( $-0.23\%$ ).

We assess *InfoMatch* on ImageNet to show its efficacy. Following the settings of FreeMatch, we select 100 labeled samples per class. Table 2 illustrates a comparison of the average Top-1 and Top-5 errors across different models. With identical parameters, *InfoMatch* has a significant improvement of 4.49% in Top-1 accuracy and 2.86% in Top-5 accuracy over FreeMatch.

Method	Top-1	Top-5
FixMatch [Sohn <i>et al.</i> , 2020]	43.66	21.80
FlexMatch [Zhang <i>et al.</i> , 2021]	41.85	19.48
FreeMatch [Wang <i>et al.</i> , 2023]	40.57	18.77
<b>InfoMatch (Ours)</b>	<b>36.21</b>	<b>15.91</b>

Table 2: Error rates (%) on ImageNet with 100 labels per class.

### 5.3 Ablation Study

*InfoMatch* primarily explores how to utilize of unlabeled data, where we use the weak-to-strong supervision based on two strong augmentation methods, RandAugment [Cubuk *et al.*, 2020] and CutMix [Yun *et al.*, 2019], represented by  $\mathcal{L}_{\text{upper}}^u$  and  $\mathcal{L}_{\text{upper}}^c$  respectively, as well as the contrastive loss  $\mathcal{L}_{\text{lower}}$  between two strong augmented views. To evaluate the impact of these losses, we conduct in-depth ablation studies on the CIFAR-10 dataset with 40 labels.

To determine the effectiveness of each loss term, we conduct experiments by excluding some components. As shown in Fig. 2, Results indicate that eliminating certain losses lead to different levels of decline in model performance. This suggests that each component is important and complementary.

Comparing to RandAugment [Cubuk *et al.*, 2020], CutMix [Yun *et al.*, 2019] employs a stronger augmentation by combining two images, addressing the limitations of low mask-based augmentation ratio. This approach enables *InfoMatch* to more effectively comprehend and capture the essential features, thus enhancing the generalization performance of *InfoMatch*. However, utilizing CutMix exclusively may distort the original features, potentially leading *InfoMatch* to learning incorrect or distorted features during training, ultimately resulting in performance deterioration.

Dataset	SVHN			STL-10	
# Label	40	250	1000	40	1000
II Model [Rasmus et al., 2015]	67.48 $\pm$ 0.95	13.30 $\pm$ 1.12	7.16 $\pm$ 0.11	74.31 $\pm$ 0.85	32.78 $\pm$ 0.40
Pseudo Label [Lee, 2013]	64.61 $\pm$ 5.6	15.59 $\pm$ 0.95	9.40 $\pm$ 0.32	74.68 $\pm$ 0.99	32.64 $\pm$ 0.71
VAT [Miyato et al., 2018]	74.75 $\pm$ 3.38	4.33 $\pm$ 0.12	4.11 $\pm$ 0.20	74.74 $\pm$ 0.38	37.95 $\pm$ 1.12
MeanTeacher [Tarvainen and Valpola, 2017]	36.09 $\pm$ 3.98	3.45 $\pm$ 0.03	3.27 $\pm$ 0.05	71.72 $\pm$ 1.45	33.90 $\pm$ 1.37
MixMatch [Berthelot et al., 2019b]	30.60 $\pm$ 8.39	4.56 $\pm$ 0.32	3.69 $\pm$ 0.37	54.93 $\pm$ 0.96	21.70 $\pm$ 0.68
ReMixMatch [Berthelot et al., 2019a]	24.04 $\pm$ 9.13	6.36 $\pm$ 0.22	5.16 $\pm$ 0.31	32.12 $\pm$ 6.24	6.74 $\pm$ 0.14
UDA [Xie et al., 2020]	5.12 $\pm$ 4.27	<u>1.92</u> $\pm$ 0.05	<u>1.89</u> $\pm$ 0.01	37.42 $\pm$ 8.44	6.64 $\pm$ 0.17
FixMatch [Sohn et al., 2020]	3.81 $\pm$ 1.18	2.02 $\pm$ 0.02	1.96 $\pm$ 0.03	35.97 $\pm$ 4.14	6.25 $\pm$ 0.33
Dash [Xu et al., 2021]	2.19 $\pm$ 0.18	2.04 $\pm$ 0.02	1.97 $\pm$ 0.01	34.52 $\pm$ 4.30	6.39 $\pm$ 0.56
MPL [Pham et al., 2021]	9.33 $\pm$ 8.02	2.29 $\pm$ 0.04	2.28 $\pm$ 0.02	35.76 $\pm$ 4.83	6.66 $\pm$ 0.00
FlexMatch [Zhang et al., 2021]	8.19 $\pm$ 3.20	6.59 $\pm$ 2.29	6.72 $\pm$ 0.30	29.15 $\pm$ 4.16	5.77 $\pm$ 0.18
FreeMatch [Wang et al., 2023]	<u>1.97</u> $\pm$ 0.02	1.97 $\pm$ 0.01	1.96 $\pm$ 0.03	15.56 $\pm$ 0.55	<u>5.63</u> $\pm$ 0.15
CoMatch [Li et al., 2021]	8.20 $\pm$ 5.32	2.16 $\pm$ 0.04	2.01 $\pm$ 0.04	<u>13.74</u> $\pm$ 4.20	5.71 $\pm$ 0.08
SimMatch [Zheng et al., 2022]	7.60 $\pm$ 2.11	2.48 $\pm$ 0.61	2.05 $\pm$ 0.05	16.98 $\pm$ 4.24	5.74 $\pm$ 0.31
SimMatchV2 [Zheng et al., 2023]	7.92 $\pm$ 2.80	2.92 $\pm$ 0.81	2.85 $\pm$ 0.91	15.85 $\pm$ 2.62	5.65 $\pm$ 0.26
<b>InfoMatch (Ours)</b>	<b>1.84</b> $\pm$ 0.07	<b>1.79</b> $\pm$ 0.01	<b>1.75</b> $\pm$ 0.03	<b>9.86</b> $\pm$ 1.13	<b>5.27</b> $\pm$ 0.09
Fully-Supervised		2.13 $\pm$ 0.01		-	

Table 3: Top-1 error rates (%) on SVHN and STL-10 datasets. **Bold** indicates the best result, while underline indicates the second-best result.

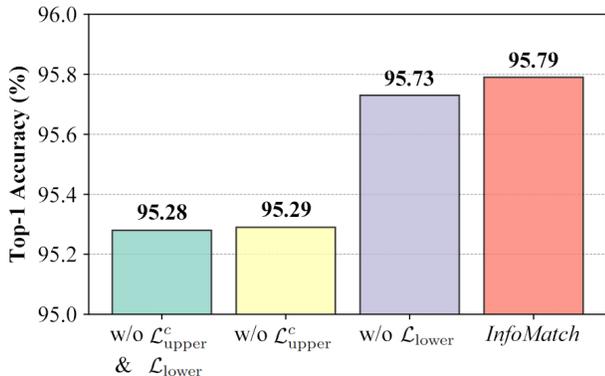


Figure 2: Ablation studies on the Cifar-10 dataset with 40 labels.

By minimizing  $\mathcal{L}_{lower}$ , we optimize *InfoMatch* from a different perspective. *InfoMatch* maximizes the mutual information between features in different augmented views, enabling *InfoMatch* to learn feature representations that express the same information from different perspectives. Consequently, the performance of *InfoMatch* is further improved.

#### 5.4 Qualitative Analysis

To further evaluate the performance of *InfoMatch* in classification tasks, we conduct a comparative analysis with FixMatch, FlexMatch, and FreeMatch in terms of Top-1 accuracy and utilization of unlabeled data. This analysis is conducted using the CIFAR-10 dataset with 40 labeled data.

Initially, a noteworthy advantage of *InfoMatch* is its excellent convergence speed, which is illustrated in Fig. 2(a). Within 100 epochs, *InfoMatch* achieves a 94.70% accuracy rate. By the time it reaches 200 epochs, our accuracy even exceeds the highest accuracy achieved by FreeMatch, reach-

ing 95.49%. Additionally, as shown in Fig. 2(b), in contrast to other methods, the utilization rate of unlabeled data in *InfoMatch* quickly rises to and stabilizes above 90% after a brief decline, further encouraging *InfoMatch* to incorporate more unlabeled samples into training during the early stages.

## 6 Conclusion

We present an extension of the entropy neural estimation method to the semi-supervised domain, enabling to measure the uncertainty of classification models from an entropy perspective. We introduce an upper bound on the entropy of the ground-truth posterior, indicating that maximizing the likelihood function of *InfoMatch* corresponds to minimizing the upper bound on the posterior entropy. Similarly, we establish a lower bound on entropy for a given dataset, demonstrating that maximizing the mutual information between different augmented views of the data aligns with maximizing the lower bound of entropy. Drawing upon these two theories, we introduce *InfoMatch* to optimize it from two perspectives. To leverage unlabeled data more effectively, we employ strategies like weak-to-strong and pseudo supervision, and introduce CutMix as a new strong augmentation view. Using optimization methods such as gradient descent, *InfoMatch* ensures that its predicted probability distribution gradually aligns with the ground-truth distribution, even in the presence of input perturbations. We introduce *InfoMatch* that seamlessly integrates pseudo supervision, consistency regularization, and mixing strategies to improve its generalization capabilities. The results from various SSL benchmark tests demonstrate that *InfoMatch* achieves good performance, especially in scenarios with limited labels, surpassing even fully-supervised learning methods. The experimental results confirm the effectiveness of *InfoMatch*.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62176108, Natural Science Foundation of Qinghai Province of China under No. 2022-ZJ-929, Fundamental Research Funds for the Central Universities under No. lzujbky-2022-ct06, and Supercomputing Center of Lanzhou University. The corresponding author of the paper is Kun Zhan.

## References

- [Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *ICML*, pages 531–540, 2018.
- [Berthelot *et al.*, 2019a] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *ICLR*, 2019.
- [Berthelot *et al.*, 2019b] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. MixMatch: A holistic approach to semi-supervised learning. In *NeurIPS*, volume 32, 2019.
- [Chen *et al.*, 2023] H Chen, R Tao, Yue Fan, Y Wang, M Savvides, J Wang, B Raj, X Xie, and Bernt Schiele. Softmatch: Addressing the quantity-quality tradeoff in semi-supervised learning. In *ICLR*, 2023.
- [Coates *et al.*, 2011] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTAT*, pages 215–223, 2011.
- [Cubuk *et al.*, 2020] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR workshop*, pages 702–703, 2020.
- [Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.
- [Ghorban *et al.*, 2022] Farzin Ghorban, Nesreen Hasan, Jörg Velten, and Anton Kummert. Individual weighting of unlabeled data points via confidence-awareness in semi-supervised learning. In *ISCAS*, pages 1215–1219, 2022.
- [Grill *et al.*, 2020] Jean-Bastien Grill, Florian Strub, Altché, et al. Bootstrap your own latent—a new approach to self-supervised learning. In *NeurIPS*, volume 33, pages 21271–21284, 2020.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [Hjelm *et al.*, 2019] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.
- [Krizhevsky and Hinton, 2009] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.
- [Lee, 2013] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, volume 3, 2013.
- [Li *et al.*, 2021] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *CVPR*, pages 9475–9484, 2021.
- [Ma *et al.*, 2023] Yixuan Ma, Xiaolin Zhang, Peng Zhang, and Kun Zhan. Entropy neural estimation for graph contrastive learning. In *ACM Multimedia*, volume 31, pages 435–443, 2023.
- [Miyato *et al.*, 2018] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *TPAMI*, 41(8):1979–1993, 2018.
- [Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop*, 2011.
- [Nowozin *et al.*, 2016] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka.  $f$ -GAN: Training generative neural samplers using variational divergence minimization. In *NeurIPS*, pages 271–279, 2016.
- [Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [Pham *et al.*, 2021] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *CVPR*, pages 11557–11568, 2021.
- [Rasmus *et al.*, 2015] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NeurIPS*, volume 28, 2015.
- [Samuli and Timo, 2017] Laine Samuli and Aila Timo. Temporal ensembling for semi-supervised learning. In *ICLR*, volume 4, 2017.
- [Sohn *et al.*, 2020] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, volume 33, pages 596–608, 2020.
- [Taruainen and Valpola, 2017] Antti Taruainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, volume 30, 2017.
- [Tian *et al.*, 2023] Zhibo Tian, Xiaolin Zhang, Peng Zhang, and Kun Zhan. Improving semi-supervised semantic segmentation with dual-level Siamese structure network. In *ACM Multimedia*, volume 31, pages 4200–4208, 2023.

- [Tishby *et al.*, 2000] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication*, pages 368–377, 2000.
- [Verma *et al.*, 2022] Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145:90–106, 2022.
- [Wang *et al.*, 2023] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. In *ICLR*, 2023.
- [Wei *et al.*, 2021] Xiang Wei, Xiaotao Wei, Xiangyuan Kong, Siyang Lu, Weiwei Xing, and Wei Lu. FMixCut-Match for semi-supervised deep learning. *Neural Networks*, 133:166–176, 2021.
- [Xie *et al.*, 2020] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *NeurIPS*, volume 33, pages 6256–6268, 2020.
- [Xu *et al.*, 2021] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yufeng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *ICML*, pages 11525–11536, 2021.
- [Yun *et al.*, 2019] Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019.
- [Zagoruyko and Komodakis, 2016] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.
- [Zhang *et al.*, 2018] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. MixUp: Beyond empirical risk minimization. In *ICLR*, 2018.
- [Zhang *et al.*, 2021] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *NeurIPS*, volume 34, pages 18408–18419, 2021.
- [Zheng *et al.*, 2022] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *CVPR*, pages 14471–14481, 2022.
- [Zheng *et al.*, 2023] Mingkai Zheng, Shan You, Lang Huang, Chen Luo, Fei Wang, Chen Qian, and Chang Xu. Simmatchv2: Semi-supervised learning with graph consistency. In *ICCV*, pages 16432–16442, 2023.
- [Zhou *et al.*, 2020] Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Time-consistent self-supervision for semi-supervised learning. In *ICML*, pages 11523–11533, 2020.