ConMatch: Semi-Supervised Learning with Confidence-Guided Consistency Regularization

Jiwon Kim[1,2†*], Youngjo Min[1*], Daehwan Kim[3*], Gyuseong Lee[1], Junyoung Seo[1], Kwangrok Ryoo[1], and Seungryong Kim[1]

[1] Korea University
[2] NAVER AI Lab
[3] Samsung Electro-Mechanics

Abstract. We present a novel semi-supervised learning framework that intelligently leverages the consistency regularization between the model's predictions from two strongly-augmented views of an image, weighted by a confidence of pseudo-label, dubbed ConMatch. While the latest semi-supervised learning methods use weakly- and strongly-augmented views of an image to define a directional consistency loss, how to define such direction for the consistency regularization between two strongly-augmented views remains unexplored. To account for this, we present novel confidence measures for pseudo-labels from strongly-augmented views by means of weakly-augmented view as an anchor in non-parametric and parametric approaches. Especially, in parametric approach, we present, for the first time, to learn the confidence of pseudo-label within the networks, which is learned with backbone model in an end-to-end manner. In addition, we also present a stage-wise training to boost the convergence of training. When incorporated in existing semi-supervised learners, ConMatch consistently boosts the performance. We conduct experiments to demonstrate the effectiveness of our ConMatch over the latest methods and provide extensive ablation studies. Code has been made publicly available at https://github.com/JiwonCocoder/ConMatch

1 Introduction

Semi-supervised learning has emerged as an attractive solution to mitigate the reliance on large labeled data, which is often laborious to obtain, and intelligently leverage a large amount of unlabeled data, to the point of being deployed in many computer vision applications, especially image classification [38, 51, 53]. Generally, this task have adopted pseudo-labeling [1,19,29,38,44,49,59] or consistency regularization [17, 23, 28, 34, 46, 51]. Some methods [4, 5, 40, 45, 50, 52, 56] proposed to integrate both approaches in a unified framework, which is often called holistic approach. As one of pioneering works, FixMatch [45] first generates a pseudo-label from the model's prediction on the weakly-augmented instance and then encourages the prediction from the strongly-augmented instance to follow

† Work done in Korea University
* Equal contribution

the pseudo-label. Their success inspired many variants that use, e.g., curriculum learning [52,56].

On the other hand, concurrent to the race for better semi-supervised learning methods [45,52,56], substantial progress has been made in self-supervised representation learning, especially with contrastive learning [3,6,8,10,20,22], aiming at learning a task-agnostic feature representation without any supervision, which can be well transferred to the downstream tasks. Formally, they encourage the features extracted from two differently-augmented images to be pulled against each other, which injects some invariance or robustness into the models. Not surprisingly, semi-supervised learning frameworks can definitely benefit from self-supervised representation learning [24, 32, 33] in that good representation from the feature encoder yields better performance with semi-supervised learning, and thus, some methods [24,32] attempt to combine the aforementioned two paradigms to boost the performance by achieving the better feature encoder.

Extending techniques presented in existing self-supervised representation learning [3, 6, 8, 10, 20, 22], which only focus on learning feature encoder, to further consider the model's prediction itself would be an appealing solution to effectively combine the two paradigms, which allows for boosting not only feature encoder but also classifier. However, compared to feature representation learning [3,6,8,10,20,22], the consistency between the model's predictions from two different augmentations should be defined by considering which direction is better to achieve not only invariance but also high accuracy in image classification. Without this, simply pulling the model's predictions as done in [3,6,8,10,20,22] may hinder the classifier output, thereby decreasing the accuracy.

In this paper, we present a novel framework for semi-supervised learning, dubbed ConMatch, that intelligently leverages the confidence-guided consistency regularization between the model's predictions from two strongly-augmented images. Built upon conventional frameworks [45, 56], we consider two strongly-augmented images and one weakly-augmented image, and define the consistency between the model's predictions from two strongly-augmented images, while still using an unsupervised loss between the model's predictions from one of the strongly-augmented images and the weakly-augmented image, as done in [45,56]. Since defining the direction of consistency regularization between two strongly-augmented images is of prime importance, rather than selecting in a deterministic manner, we present a probabilistic technique by measuring the confidence of pseudo-labels from each strongly-augmented image, and weighting the consistency loss with this confidence. To measure the confidence of pseudo-labels, we present two techniques, including non-parametric and parametric approaches. With this confidence-guided consistency regularization, our framework dramatically boosts the performance of existing semi-supervised learners [45, 56]. In addition, we also present a stage-wise training scheme to boost the convergence of training. Our framework is a plug-and-play module, and thus various semi-supervised learners [4, 24, 32, 33, 45, 50, 52, 56] can benefit from our framework. We briefly summarize our method with other highly relevant works in semi-supervised learning in Table 1. Experimental results and ablation studies show

Table 1. Comparison of our ConMatch to other relevant works which have a form of consistency regularization combining pseudo-labeling [5,24,32,33, 45,52,56]

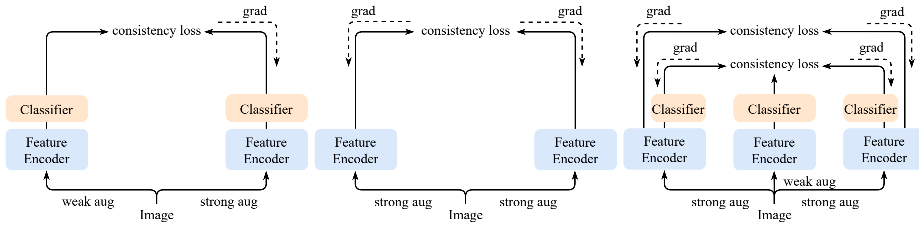
	MixMatch [5]	FixMatch [45]	FlexMatch [56]	Dash [52]	SelfMatch [24]	CoMatch [32]	LESS [33]	ConMatch (Ours)
Using pseudo-labeling	✗	✓	✓	✓	✓	✓	✓	✓
Using two strong branches	✗	✗	✗	✗	✓	✓	✗	✓
Learning confidence measure	✗	✗	✗	✗	✗	✗	✗	✓
Using stage-wise training	✗	✗	✗	✗	✓	✗	✗	✓

that the proposed framework not only boosts the convergence but also achieves the state-of-the-art performance on most standard benchmarks [12,27,35].

2 Related Works

Semi-supervised Learning. Semi-supervised learning has been an effective paradigm for leveraging an abundance of unlabeled data along with limited labeled data. For this task, various methods such as pseudo-labeling [19, 29] and consistency regularization [28,42,46] have been proposed. In pseudo-labeling [29], a model uses unlabeled samples with high confidence as training targets, which reduces the density of data points at the decision boundary [19,41]. Consistency regularization has been first introduced by π-model [42], which is further improved by numerous following works [17, 23, 28, 34, 46, 51]. In the consistency regularization, the model should minimize the distance between the model's predictions when fed perturbed versions of the input [23, 28, 34, 37, 46, 49, 51] or the model [23, 28, 37, 46, 51, 57]. Very recently, advanced consistency regularization methods [4,45,50] have been introduced by combining with pseudo-labeling. These methods show high accuracy, comparable to supervised learning in a fully-labeled setting, e.g., ICT [48], MixMatch [5], UDA [50], ReMixMatch [4], and FixMatch [45]. The aforementioned methods can be highly boosted by simultaneously considering the techniques proposed in recent self-supervised representation learning methods [3, 6, 8, 20, 22].

Self-supervised Representation Learning. Self-supervised representation learning has recently attracted much attention [3, 6, 8, 16, 18, 20, 22, 36, 58] due to its competitive performance. Specifically, contrastive learning [3, 6, 8, 20, 22] becomes a dominant framework. It formally maximizes the agreement between different augmented views of the same image [16,18,36,58]. Most previous methods benefit from a large amount of negative pairs to preclude constant outputs and avoid a collapse problem [8]. An alternative to approximate the loss is to use cluster-based approach by discriminating between groups of images with similar features [6]. Some methods [10, 20] mitigated to use negative samples by using a momentum encoder [20] and a stop-gradient technique [10]. The aforementioned methods applied the consistency loss at the feature-level, unlike recent semi-supervised learning methods [45, 56] that consider the consistency loss in the logit-level, which may not be optimal to be incorporated with



(a) Semi-sup. (b) Self-sup. (c) Semi-sup. with self-sup.

Fig. 1. Conceptual illustration of existing methods that leverage unlabeled data: (a) semi-supervised learning- the model uses the model's prediction itself to produce a pseudo-label for unlabeled data [4, 5, 17, 28, 29, 34, 42, 45, 46, 48, 50], (b) self-supervised representation learning- the model is learned to generate the same feature embedding for two augmented views from unlabeled data [3, 6, 8, 10, 20, 20, 22], and (c) semi-supervised learning with self-supervision representation learning- the model simultaneously learns a feature representation with self-supervised representation loss, while learning all the networks with semi-supervised learning [24, 32, 33].

semi-supervised learners. Formulating the consistency loss in the logit-level as self-supervision is challenging because a direction between two augmented views should be determined. Without this, simply pulling the model's predictions as done in [3,6,8,10,20,22] may hinder the classifier output, thereby decreasing the accuracy.

Self-supervision in Semi-supervised Learning. Many recent state-of-the-art semi-supervised learning methods adopt the self-supervised representation learning methods [9,55] to jointly learn good feature representation. Self-supervised pre-training, followed by supervised fine-tuning, has shown strong performance on semi-supervised learning settings. Specifically, SelfMatch [24] adopted SimCLR [8] for self-supervised pre-training and FixMatch [45] for semi-supervised fine-tuning. However, it may learn sub-optimal representation for the image classification task due to the task-agnostic learning. On the other hand, some methods [30, 32] unify pseudo-labeling and self-supervised learning. [31] alternates between self- and semi-supervised learning. There lacks a study to effectively use self-supervision, rather than simply adopting this.

Confidence Estimation in Semi-supervised Learning. In semi-supervised learning, a confidence-based strategy has been widely used along with pseudo labeling so that the unlabeled data are used only when the predictions are sufficiently confident. Such confidence in pseudo-labeling has been often measured by the peak values of the predicted probability distribution [40, 45, 50, 52, 56]. Although the selection of unlabeled samples with high confidence predictions moves decision boundaries to low density regions [7], many of these selected predictions are incorrect due to the poor calibration of neural networks [21], which has the discrepancy between the confidence level of a network's individual predictions and its overall accuracy and leads to noisy training and poor generalization [14,15]. However, there was no study how to learn the confidence of pseudo-labels, which is the topic of this paper.

3 Methodology

3.1 Preliminaries

Let us define a batch of *labeled* instances as $\mathcal{X} = \{(x_b, y_b)\}_{b=1}^{B}$, where x_b is an instance and y_b is a label representing one of Y labels. In addition, let us define a batch of *unlabeled* instances as $\mathcal{U} = \{u_b\}_{b=1}^{\mu B}$, where μ is a hyper-parameter that determines the size of $\mathcal{U}$ relative to $\mathcal{X}$. The objective of semi-supervised learning is to use both $\mathcal{X}$ and $\mathcal{U}$ to train a model with parameters θ taking an instance $r \in \mathcal{X} \cup \mathcal{U}$ as input and outputting a distribution over class labels y such that $p_{\text{model}}(y|r;\theta)$. The model generally consists of an feature encoder $f(\cdot)$ and a classifier $g(\cdot)$, and thus, $p_{\text{model}}(y|r;\theta) = g(f(r))$.

For semi-supervised learning, most state-of-the-art methods are based on consistency regularization approaches [2,28,42] that rely on the assumption that the model should generate similar predictions when perturbed versions of the same instance are fed, e.g., using data augmentation [34], or model perturbation [28,46]. These methods formally extract a pseudo-label from one branch, filtered by confidence, and use this as a target for another branch. For instance, FixMatch [45] utilizes two types of augmentations such as *weak* and *strong*, denoted by $\alpha(\cdot)$ and $\mathcal{A}(\cdot)$, and a pseudo-label from weakly-augmented version of an image is used as a target for strongly-augmented version of the same image. This loss function is formally defined such that

$$\mathcal{L}_{\text{un}} = c(r)\mathcal{H}\left(q(r), p_{\text{model}}\left(y|\mathcal{A}(r);\theta\right)\right), \tag{1}$$

where $c(r)$ denotes a confidence of $q(r)$, and $q(r)$ denotes a pseudo-label generated from $p_{\text{model}}(y|\alpha(r);\theta)$, which can be either an one-hot label [40,45,52,56] or a sharpened one [4,5,50], and $\mathcal{H}(\cdot,\cdot)$ is often defined as a cross-entropy loss. In this framework, measuring confidence $c(r)$ is of prime importance, but conventional methods simply measure this, e.g., by the peak value of the softmax predictions [40,45,50,52,56].

On the other hands, semi-supervised learning framework can definitely benefit from existing self-supervised representation learning [24,32,33] in that good representation from the feature encoder $f(\cdot)$ yields better performance with semi-supervised learner. In this light, some methods attempted to combine semi-supervised learning and self-supervised representation learning to achieve the better feature encoder [24,32]. Concurrent to the race for better semi-supervised learning methods, substantial progress has been made in self-supervised representation learning, especially with contrastive learning [3, 6, 8, 10, 20, 22]. The loss function for this task can also be defined as a consistency regularization loss, similar to [40,45,50,52,56] but in the feature-level, such that

$$\mathcal{L}_{\text{self}} = \mathcal{D}(F_i(r), F_j(r)), \tag{2}$$

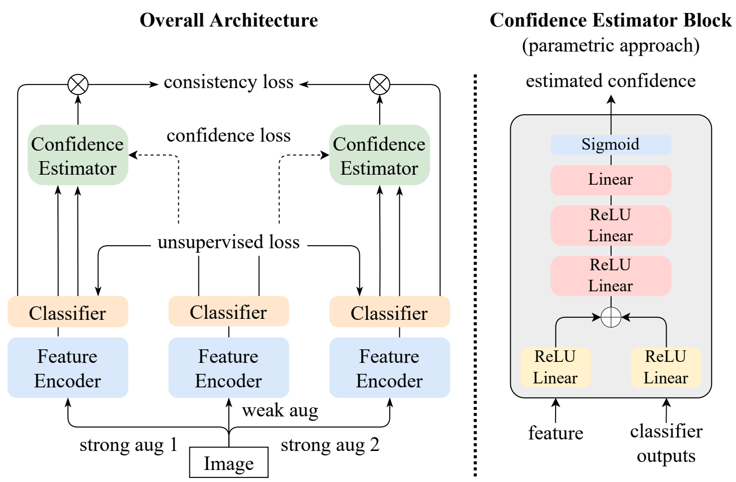


Fig. 2. Network configuration of ConMatch. A semi-supervised learning framework built upon consistency loss with an additional strong branch to leverage confidence loss between two strong branches. In the parametric approach, the confidence estimator block takes a concatenated heterogeneous feature as an input and produces the estimated confidence of pseudo label.

where $F_i(r) = f(\mathcal{A}_i(r))$ and $F_j(r) = f(\mathcal{A}_j(r))$ extracted from images with two different strongly-augmented images $\mathcal{A}_i(\cdot)$ and $\mathcal{A}_j(\cdot)$, respectively. $\mathcal{D}(\cdot,\cdot)$ can be defined as contrastive loss [22] or negative cosine similarity [10]. Even though this loss helps to boost learning the feature encoder $f(\cdot)$, the mechanism that simply pulls the features $F_i(r)$ and $F_j(r)$ may not be optimal to boost a semi-supervised learner and break the latent feature space, without considering a direction representing which branch is better.

3.2 Formulation

To combine the semi- and self-supervised learning paradigm in a boosting fashion, unlike [24, 32, 33], we present to effectively exploit a self-supervision between two *strong* branches tailored for boosting semi-supervised learning, called ConMatch. Unlike existing self-supervised representation learning methods, e.g., SimSiam [10], we formulate the consistency regularization loss at class logit-level[4], as done in semi-supervised learning methods [45, 56], and estimate the confidences of each pseudo-label from two *strongly*-augmented images, $\mathcal{A}_i(r)$ and $\mathcal{A}_j(r)$ for r, and use them to consider the probability of each direction between them. Since measuring such confidences is notoriously challenging, we present novel confidence estimators by using the output from *weak*-augmented image $\alpha(r)$ as an anchor in non-parametric and parametric approaches.

An overview of our ConMatch is illustrated in Fig. 2. Specifically, there exist two branches for *strongly*-augmented images (called strong branches) and one

[4] In the paper, class logit means the output of the network, i.e., $p_{\text{model}}(y|r;\theta)$ for r.

branch for *weakly*-augmented image (called weak branch). Similar to existing semi-supervised representation learning methods [45,52,56], we attempt to apply the consistency loss between a pair of each strong branch and weak branch. But, tailored to semi-supervised learning, we present a confidence-guided consistency regularization loss $\mathcal{L}_{\text{ccr}}$ between two strong branches such that

$$\mathcal{L}_{\text{ccr}} = c_i(r)\mathcal{H}(q_i(r), p_{\text{model}}(y|\mathcal{A}_j(r);\theta)) + c_j(r)\mathcal{H}(q_j(r), p_{\text{model}}(y|\mathcal{A}_i(r);\theta)), \quad (3)$$

where $q_i(r)$ and $q_j(r)$ denote the pseudo-labels generated from $p_{\text{model}}(y|\mathcal{A}_i(r);\theta)$ and $p_{\text{model}}(y|\mathcal{A}_j(r);\theta)$, respectively. $c_i(r)$ and $c_j(r)$ denote estimated confidences of $q_i(r)$ and $q_j(r)$. Our proposed loss function is different from conventional self-supervised representation learning loss $\mathcal{L}_{\text{self}}$ in that the consistency is applied in the logit-level (not feature-level) similar to [45,56], and adjusted by the estimated confidence. However, unlike [40, 45, 50, 52, 56], we can learn the better feature representation by considering two strongly-augmented views, while improving semi-supervised learning performance at the same time. It should be noted that this simple loss function can be incorporated with any semi-supervised learners, e.g., FixMatch [45] or FlexMatch [56].

To measure the confidences $c_i(r)$ and $c_j(r)$, we present two kinds of confidence estimators, based on non-parametric and parametric approaches. In the following, we explain how to measure these confidences in detail.

3.3 Measuring Confidence: Non-parametric Approach

Existing semi-supervised learning methods [29, 42, 45] have selected unlabeled samples with high confidence as training targets (i.e., pseudo-labels) in a straightforward way; which can be viewed as a form of entropy minimization [19]. It has been well known that it is non-trivial to set an appropriate threshold for such handcrafted confidence estimation, and thus, confidence-based strategies commonly suffer from a dilemma between pseudo-label exploration and accuracy depending on the threshold [1,33].

In our framework, estimating the confidence of pseudo-labels from strong branches may suffer from similar limitations if the conventional handcrafted methods [29, 42, 45] are simply used. To overcome this, we present a novel way to measure the confidences, $c_i(r)$ and $c_j(r)$, based on the similarity between outputs of *strongly*-augmented images and *weakly*-augmented images. Based on the hypothesis that the similarity between the logits or probabilities from strongly-augmented images and weakly-augmented images can be directly used as a confidence estimator, we present to measure confidence of each strong branch loss by the cross-entropy loss value itself between strongly-augmented and weakly-augmented images. Specifically, we measure such a confidence with the following:

$$s_i(r) = \frac{1}{\mathcal{H}(p_{\text{model}}(y|\alpha(r);\theta), p_{\text{model}}(y|\mathcal{A}_i(r);\theta))}, \quad (4)$$

where the smaller $\mathcal{H}(p_{\text{model}}(y|\alpha(r);\theta), p_{\text{model}}(y|\mathcal{A}_i(r);\theta))$, the higher $s_i(r)$ is. $s_j(r)$ can be similarly defined with $\alpha(r)$ and $\mathcal{A}_j(r)$. Finally, $c_i(r)$ is computed such that $c_i(r) = s_i(r)/(s_i(r) + s_j(r))$, and $c_j(r)$ is similarly computed.

Algorithm 1: ConMatch-P (Parametric Approach)

Notation: strong augmentation $\mathcal{A}$, weak augmentation α, model $p_{\text{model}}(\cdot;\theta)$ consisting of feature encoder f and classifier g, confidence estimator $h(\cdot;\theta_{\text{conf}})$, pseudo label q, leranable confidence c
Input: $\mathcal{X} = \{(x_b, y_b) : b \in (1,\dots,B)\}$, $\mathcal{U} = \{u_b : b \in (1,\dots,\mu B)\}$
for $b = 1$ to B **do**
  $F(\alpha(x_b)), L(\alpha(x_b)) = f(\alpha(x_b)), g(f(\alpha(x_b)))$
  $c(\alpha(x_b)) = h(F(\alpha(x_b)), L(\alpha(x_b)); \theta_{\text{conf}})$
  if $y_b == \text{argmax}_y p_{\text{model}}(y|\alpha(x_b);\theta)$ **then**
    $c_{\text{GT}}(\alpha(x_b)) = 1$
  else
    $c_{\text{GT}}(\alpha(x_b)) = 0$
  end if
end for
$\mathcal{L}_{\text{sup}} = \sum_{b=1}^{B} \mathcal{H}(y_b, p_{\text{model}}(y|\alpha(x_b);\theta))$
$\mathcal{L}_{\text{conf}-\text{sup}} = \mathcal{H}(c_{\text{GT}}(\alpha(x_b)), h(F(\alpha(x_b)), L(\alpha(x_b)); \theta_{\text{conf}}))$
for $b = 1$ to μB **do**
  $(F_i, F_j), (L_i, L_j) = f(\mathcal{A}_i(u_b), \mathcal{A}_j(u_b)), g(f(\mathcal{A}_i(u_b), \mathcal{A}_j(u_b)))$
  $c_i, c_j = h(F_i, L_i; \theta_{\text{conf}}), h(F_j, L_j; \theta_{\text{conf}})$
  Generate pseudo labels for differently augmented versions $\alpha, \mathcal{A}_i, \mathcal{A}_j$
end for
Calculate $\mathcal{L}_{\text{un}}$ using c, q from $\alpha(u_b)$, and $p_{\text{model}}(y|\mathcal{A}(u_b);\theta)$ via Eq. 2
Calculate $\mathcal{L}_{\text{ccr}}$ using c, q from $\mathcal{A}(u_b)$ and $p_{\text{model}}(y|\mathcal{A}(u_b);\theta)$ via Eq. 3
Calculate $\mathcal{L}_{\text{conf}}$ using c from $\mathcal{A}(u_b)$ and $p_{\text{model}}(y|\alpha(u_b);\theta)$ via Eq. 6
Update θ by minimizing $L_{\text{sup}}, L_{\text{un}}$ and L_{ccr}
Update θ_{conf} by minimizing $L_{\text{conf}-\text{sup}}$ and L_{conf}
Return: Model parameters $\{\theta, \theta_{\text{conf}}\}$

In this case, the total loss for the non-parametric approach is as follows:

$$\mathcal{L}_{\text{total}}^{\text{np}} = \lambda_{\text{sup}}\mathcal{L}_{\text{sup}} + \lambda_{\text{un}}\mathcal{L}_{\text{un}} + \lambda_{\text{ccr}}\mathcal{L}_{\text{ccr}}, \tag{5}$$

where λ_{sup}, λ_{un}, and λ_{ccr} are weights for $\mathcal{L}_{\text{sup}}$, $\mathcal{L}_{\text{un}}$, and $\mathcal{L}_{\text{ccr}}$, respectively. Note that for weakly-augmented labeled images $\alpha(x_b)$ with labels y_b, a simple classification loss $\mathcal{L}_{\text{sup}}$ is applied as $\mathcal{H}(y_b, p_{\text{model}}(y|\alpha(x_b);\theta))$, as done in [45].

3.4 Measuring Confidence: Parametric Approach

Even though the above confidence estimator with non-parametric approach yields comparable performance to some extent (which will be discussed in experiments), it solely depends on each image, and thus it may be sensitive to outliers or errors without any modules to learn a prior from the dataset. To overcome this, we present an additional parametric approach for confidence estimation. Motivated by stereo confidence estimation [11, 39, 43, 47], obtaining a confidence measure from the networks by extracting the confidence features from input and predicting the confidence with a classifier, we also introduce a learnable confidence measure for pseudo-labels. Unlike existing methods that simply use the model output as confidence [40, 45, 50, 52, 56], such learned confidence can intelligently select a subset of pseudo-labels that are less noisy, which helps the network to converge significantly faster and achieve improved performance by utilizing the false negative samples excluded from training by high threshold at early training iterations.

Specifically, we define an additional network for learnable confidence estimation such that $c(r) = h(F(r), L(r); \theta_{\mathrm{conf}})$, where $h(\cdot)$ is a confidence estimator with model parameters θ_{conf}, $F(r)$ is a feature, and $L(r)$ is a logit from an instance r, as shown in Fig. 2. For the network architecture, the concatenation of feature $F(r)$ and logit $L(r)$ transformed by individual non-linear projection heads is used, based on the intuition that a direct concatenation of two their heterogeneous confidence features does not provide an optimal performance [25], followed by the final classifier for confidence estimation. The detailed network architecture is described in the supplementary material.

The confidence estimator is learned with the following loss function:

$$\mathcal{L}_{\mathrm{conf}} = c_i(r)\mathcal{H}(p_{\mathrm{model}}(y|\alpha(r); \theta_{\mathrm{freeze}}), p_{\mathrm{model}}(y|\mathcal{A}_i(r); \theta_{\mathrm{freeze}})) + \log(1/c_i(r)), \tag{6}$$

where θ_{freeze} is a freezed network parameter with a stop gradient. The intuition behind is that during the confidence network training, we just want to make the network learn the confidence itself, rather than collapsing to trivial solution to learn the feature encoder simultaneously. In addition, we also use the supervised loss for confidence estimator $\mathcal{L}_{\mathrm{conf-sup}} = \mathcal{H}(c_{\mathrm{GT}}, h(F(\alpha(x_b)), L(\alpha(x_b)); \theta_{\mathrm{conf}}))$; $c_{\mathrm{GT}} = 1$ if y_b is equal to $\mathrm{argmax}_y\, p_{\mathrm{model}}(y|\alpha(x_b); \theta)$, and $c_{\mathrm{GT}} = 0$ otherwise.

The total loss for the parametric case can be written as

$$\mathcal{L}_{\mathrm{total}}^{\mathrm{param}} = \lambda_{\mathrm{sup}}\mathcal{L}_{\mathrm{sup}} + \lambda_{\mathrm{un}}\mathcal{L}_{\mathrm{un}} + \lambda_{\mathrm{conf}}\mathcal{L}_{\mathrm{conf}} + \lambda_{\mathrm{conf-sup}}\mathcal{L}_{\mathrm{conf-sup}} + \lambda_{\mathrm{ccr}}\mathcal{L}_{\mathrm{ccr}} \tag{7}$$

where λ_{conf} and $\lambda_{\mathrm{conf-sup}}$ are the weights for $\mathcal{L}_{\mathrm{conf}}$ and $\mathcal{L}_{\mathrm{conf-sup}}$, respectively. We explain an algorithm for ConMatch of parametric approach in Alg. 1.

3.5 Stage-Wise Training

Even though our framework can be trained in an end-to-end manner, we further propose a stage-wise training strategy to boost the convergence of training. This stage-wise training consists of three stages, 1) pre-training for the feature encoder, 2) pre-training for the confidence estimator (for parametric approach only), and 3) fine-tuning for both feature encoder and confidence estimator (for parametric approach only). Specifically, we first warm up the feature encoder by solely using the standard semi-supervised loss functions with $\mathcal{L}_{\mathrm{sup}}$ and $\mathcal{L}_{\mathrm{un}}$. We then train the confidence estimator based on the outputs of the pre-trained feature encoder in the parametric approach. As mentioned in [26], this kind of simple technique highly boosts the convergence to discriminate between confident and unconfident outputs from the networks. Finally, we fine-tune all the networks with the proposed confidence-guided self-supervised loss $\mathcal{L}_{\mathrm{ccr}}$. We empirically demonstrate the effectiveness of the stage-wise training by achieving state-of-the-art results on standard benchmark datasets [12, 27, 35].

4 Experiments

4.1 Experimental Settings

In experiments, we extensively evaluate the performance of our ConMatch on various standard datasets [12, 27, 35] with various label fraction settings in comparison to state-of-the-art algorithms, such as UDA [50], FixMatch [45], FlexMatch [56], SelfMatch [24], LESS [33] and Dash [52]. Our proposed methods have two variants; ConMatch-NP (non-parametric approach), and ConMatch-P (parametric approach) integrated to FlexMatch [56], which is the state-of-the-art semi-supervised learner, even though it can be easily integrated to others [32, 45, 49].

Datasets. We consider four standard benchmarks, including CIFAR-10/100 [27], SVHN [35], and STL-10 [12]. CIFAR-10 [27] contains 50,000 training images and 10,000 test images, which have resolution 32×32 with ten classes. Similar to CIFAR-10, CIFAR-100 [27] has the same number of training/test images and image size, but it differently classifies as 100 fine-grained classes. SVHN [35] consists of 73,257 training images with 26,032 test images, having also 32×32 resolution images, belonging to ten different classes of numeric digits. STL-10 [12] contains 5,000 labeled images with size of 96×96 from 10 classes and 100,000 unlabeled images with size of 96×96.

Evaluation Metrics. For quantitative evaluation, we compute the mean and standard deviation of error rates, when trained on 3 different folds for labeled data, based on the standard evaluation protocol of selecting a subset of the training data while keeping the remainder unlabeled. In addition, as in [32, 56], we evaluate the quality of pseudo labels by training curves of precision, recall, and F1 values.

4.2 Implementation Details

For a fair comparison, we generally follow the same hyperparameters with FixMatch [45]. Specifically, we use Wide ResNet (WRN) [54] as a feature encoder for the experiments, especially WRN-28-2 for CIFAR-10 [27] and SVHN [35], WRN-28-8 for CIFAR-100 [27], and WRN-37-2 for STL-10 [12]. We use a batch size of labeled data $B = 64$, the ratio of unlabeled data $\mu = 7$, and SGD optimizer with a learning rate starting from 0.03, The detailed hyperparameter settings are described in the supplementary material. For a weakly-augmented sample, we use a crop-and-flip, and for a strongly-augmented sample, we use RandAugmnet [13].

4.3 Comparison to State-Of-The-Art Methods

On standard semi-supervised learning benchmarks, we evaluate the performance of our frameworks, ConMatch-P and ConMatch-NP, compared to various state-of-the-art methods, as shown in Table 2 and Table 3. We observe that the performance difference between ConMatch-NP and ConMatch-P is not large, except in

Table 2. Comparison on error rates on CIFAR-10 [27] and CIFAR-100 [27] benchmarks on 3 different folds.

Methods	CIFAR-10 40	CIFAR-10 250	CIFAR-10 4,000	CIFAR-100 400	CIFAR-100 2,500
UDA [50]	29.05±5.93	8.82±1.08	4.88±0.18	59.28±0.88	33.13±0.22
FixMatch (RA) [45]	13.81±3.37	5.07±0.65	4.26±0.06	48.85±1.75	28.29±0.11
FlexMatch [56]	4.97±0.06	4.98±0.09	4.19±0.01	39.94±1.62	26.49±0.20
SelfMatch [24]	6.81±1.08	4.87±0.26	4.06±0.08	-	-
CoMatch [32]	6.91±8.47	4.91±0.33	-	-	-
LESS [33]	6.80±1.10	4.90±0.80	-	48.70±12.40	-
Dash (RA) [52]	13.22±3.75	**4.56±0.13**	4.08±0.06	44.76±0.96	27.18±0.21
ConMatch-NP	4.89±0.07	5.00±0.37	4.36±0.42	44.90±1.34	26.91±1.35
ConMatch-P	**4.43±0.13**	4.70±0.25	**3.92±0.08**	**38.89±2.18**	**25.39±0.20**

Table 3. Comparison on error rates on SVHN [35] and STL-10 [12] benchamarks on 3 different folds.

Method	SVHN 40	SVHN 250	STL-10 1,000
UDA [50]	52.63±20.51	5.69±2.76	7.66±0.56
FixMatch (RA) [45]	**3.96±2.17**	2.48±0.38	7.98±1.50
FlexMatch [56]	4.97±0.06	4.98±0.09	5.77±0.18
SelfMatch [24]	3.42±1.02	2.63±0.43	-
CoMatch [32]	6.91±8.47	4.91±0.33	20.20±0.38
Dash (RA) [52]	**3.03±1.59**	2.17±0.10	7.26±0.40
ConMatch-NP	6.20±3.44	5.80±0.74	6.02±0.08
ConMatch-P	3.14±0.57	3.13±0.72	**5.26±0.04**

the label-scare setting. This may be explained by the fact that non-parametric method highly depends on baseline performance since it does not consider other samples which can be modeled as a prior. We show our superiority on most benchmarks with extensive label setting, but we mainly focus the label-scare setting, since it corresponds to the central goal of semi-supervised learning, reducing the need for labeled data. We achieves 4.43% and 38.89% error rate for CIFAR-10 and CIFAR-100 settings [27] with only 4 labels per class respectively. Compared to the results of SelfMatch [24] and CoMatch [32], closely related to ours, adopting self-supervised methods, we can prove the competitiveness of our method by achieving 2.38% and 2.48% improvements at CIFAR-10 with 40 labels. On the other datasets, CIFAR-100 [35] and STL-10 [12], we record the lowest error rate of 38.89% and 25.39% with 400 and 2500 labels setting, and also slightly better than baseline [56] by recording 5.26% in STL-10 dataset.

4.4 Ablation Study

Effects of Different Baseline. We first evaluate our ConMatch with two baselines, FixMatch [45] and FlexMatch [56], in both parametric (ConMatch-P) and non-parametric (ConMatch-NP) approaches as shown in Table 4. ConMatch-P w/ [45] boosts the performance significantly on CIFAR-10 with 40 labels from

Table 4. Ablation study of different semi-supervised baselines. We evaluate non-parametric (ConMatch-NP) and parametric (ConMatch-P) approaches with different baselines, Fixmatch [45] and FlexMatch [56].

Methods	CIFAR-10		CIFAR-100
	40	250	400
FixMatch [45]	13.81	5.07	48.85
ConMatch-NP w/ [45]	<u>6.83</u>	4.73	<u>48.73</u>
ConMatch-P w/ [45]	**5.13**	4.64	**48.00**
FlexMatch [56]	4.97	4.98	<u>39.94</u>
ConMatch-NP w/ [56]	<u>4.84</u>	<u>4.74</u>	44.90
ConMatch-P w/ [56]	**4.68**	**4.70**	**38.89**

Table 5. Ablation study of training schemes. `E` means end-to-end training and `S` means stage-wise training.

Methods	Status	CIFAR-10		CIFAR-100
		40	250	400
ConMatch w/ FixMatch [45]	`E`	4.85	4.77	47.81
	`S`	5.13	4.60	48.00
ConMatch w/ FlexMatch [56]	`E`	4.68	4.70	57.16
	`S`	4.43	4.70	38.89

13.81% to 5.13%, achieving the state-of-the-art result. The performance gains of ConMatch-P w/ [45] is relatively higher than one w/ [56] on most setting since [45] does not adaptively adjust the threshold depending on the difficulty level of samples. Note that the thresholds of FixMatch [45] and FlexMatch [56] are used only for $\mathcal{L}_{\mathrm{un}}$.

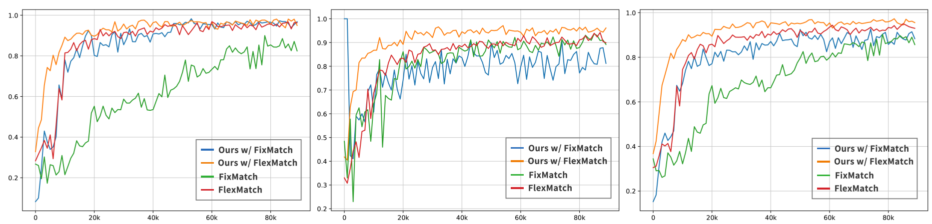
Effectiveness of Confidence Measure. In Table 4, we evaluate two confidence measures in non-parametric and parametric approach. In extremely label-scare setting, such as CIFAR-10 with 4 labels per class, the non-parametric approach achieves relatively lower performance, 1.70% and 0.16%, in both FixMatch and FlexMatch baseline, while the parametric approach (ConMatch-P w/ [56]) reaches the state-of-the-art performance. But, as the number of labels increases, the gap between non-parametric and parametric approach decreases, indicating that a certain number of labeled samples are required to measure the confidence without the confidence estimator.

Effectiveness of Stage-Wise Training. In Table 5, we report the performance difference between end-to-end training and stage-wise training. We can observe that ConMatch-P has obtained meaningful enhancements in both training schemes, but stage-wise training shows more larger gap between baseline.

Architecture. Here we analyze the key components of ConMatch, the confidence estimator and guided consistency regularization as shown in Table 6. For the fair comparison, we construct three branches on FixMatch [45] as baseline **(I)**, one branch for a weakly-augmented sample and two branches for strongly-augmented samples. **(II)** uses logit-level self-supervised loss, but not weighted

Table 6. Ablation study of our component on CIFAR-10 [27] with 40 labels.

	Three branches	Logit-level self-sup.	Confidence net. input: logits	Confidence net. input: features	Error rate
(I)	✓	✗	✗	✗	18.11
(II)	✓	✓	✗	✗	77.50
(III)	✓	✓	✓	✗	7.05
(IV)	✓	✓	✓	✓	5.13



(a) Precision (b) Recall (c) F1-score

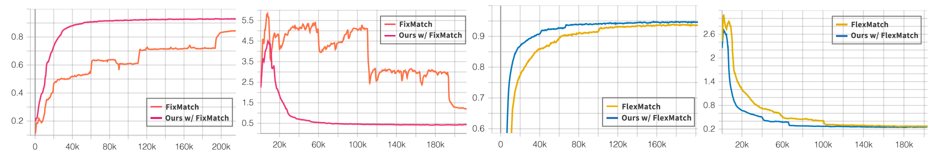
Fig. 3. Plots of evolution of pseudo-labeling between ours and baselines [45, 56] as training progresses on CIFAR-10 [27] with 40 labels: in terms of (a) Precision, (b) Recall, and (c) F1-Score.

by confidence, i.e., $\mathcal{L}_{\text{ccr}}$ with $c_i(r), c_j(r) = 1/2$. **(III)** and **(IV)** weight confidences of strongly-augmented instances to logit-level self-supervised loss. **(III)** only takes logits as an input of confidence estimator while both logits and features are fed into **(IV)**. The result of this ablation study shows that logit-level self-supervised loss without confidence guidance causes network collapse. The collapse occurred in **(II)** is one of the reasons why other semi-supervised methods [32] could not use self-supervision at logit-level and should use negative pairs. **(III)** and **(IV)** show a significant performance improvement compared to **(I)** without such collapse.

Evaluating Confidence Estimation. To evaluate the effectiveness of our confidence estimator, we measure precision, recall, and F1-score of ConMatch and FixMatch [45] as evolving the training iterations on CIFAR-10 [27] with 40 labels as shown in Fig. 3. The confident sample is defined as an unlabeled sample having max probability over than threshold in the baseline and confidence measures over than 0.5 in ConMatch. The quality of the confident sample is important to determine precisely to prevent the confirmation bias problem, significantly degrading the performances. The three classification metric, precision, recall and F1-score, are effective to evaluate the quality of the confidence. By Fig. 3, we can observe that ConMatch, starting from the scratch for the fair comparison, shows higher values in all metric compared to the baseline.

4.5 Analysis

Convergence Speed. One of the advantages of our ConMatch is its superior convergence speed. Based on the results as shown in Fig. 4 (b) and (d), the loss of



(a) (b) (c) (d)

Fig. 4. Convergence analysis of baselines [45,56] and ConMatch: A comparison of top-1-accuracy and loss between FixMatch [45] and ConMatch w/ [45] are shown at (a) and (b). A comparison between FlexMatch [56] and ConMatch w/ [56] is shown at (c) and (d). Evaluations are done every 200K iterations on CIFAR-10 with 40 labels.

ConMatch decreases much faster and smoother than corresponding baseline [45], demonstrating our superior convergence speed. Furthermore, the result of the accuracy in Fig. 4 (a) also proves that the global optimum is quickly reached. We also prove our effectiveness of our method by comparing the another baseline, FlexMatch [56]. The convergence speed gap is relatively smaller than FixMatch since it dynamically adjust class-wise thresholds at each time step, leading to the stable training, but ConMatch achieves fast convergence at all time step from the early phase where the predictions of the model are still unstable. It is manifest that the introduction of ConMatch successfully encourages the model to proactively improve the overall learning effect.

5 Conclusion

In this paper, we have proposed a novel semi-supervised learning framework built upon conventional consistency regularization frameworks with an additional strong branch to define the proposed confidence-guided consistency loss between two strong branches. To account for the direction of such consistency loss, we present confidence measures in non-parametric and parametric approaches. Also, we also presented a stage-wise training to boost the convergence of training. Our experiments have shown that our framework boosts the performance of base semi-supervised learners, and is clearly state-of-the-art on several benchmarks.

Acknowledgements. This research was supported by the MSIT, Korea (IITP-2022-2020-0-01819, ICT Creative Consilience program), and National Research Foundation of Korea (NRF-2021R1C1C1006897).

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: IJCNN (2020)
2. Bachman, P., Alsharif, O., Precup, D.: Learning with pseudo-ensembles. In: NeurIPS (2014)
3. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. In: NeurIPS (2019)
4. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv:1911.09785 (2019)
5. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. In: NeurIPS (2019)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020)
7. Chapelle, O., Zien, A.: Semi-supervised classification by low density separation. In: AISTATS workshops (2005)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
9. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: NeurIPS (2020)
10. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (2021)
11. Choi, H., Lee, H., Kim, S., Kim, S., Kim, S., Sohn, K., Min, D.: Adaptive confidence thresholding for monocular depth estimation. In: ICCV (2021)
12. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: AISTATS (2011)
13. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: CVPR workshops (2020)
14. Dawid, A.P.: The well-calibrated bayesian. JASA (1982)
15. DeGroot, M.H., Fienberg, S.E.: The comparison and evaluation of forecasters. Journal of the Royal Statistical Society: Series D (The Statistician) (1983)
16. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: ICML (2014)
17. French, G., Mackiewicz, M., Fisher, M.: Self-ensembling for visual domain adaptation. arXiv:1706.05208 (2017)
18. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv:1803.07728 (2018)
19. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: NeurIPS (2004)
20. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: NeurIPS (2020)
21. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML (2017)
22. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
23. Ke, Z., Wang, D., Yan, Q., Ren, J., Lau, R.W.: Dual student: Breaking the limits of the teacher in semi-supervised learning. In: ICCV (2019)

24. Kim, B., Choo, J., Kwon, Y.D., Joe, S., Min, S., Gwon, Y.: Selfmatch: Combining contrastive self-supervision and consistency for semi-supervised learning. arXiv:2101.06480 (2021)
25. Kim, S., Min, D., Kim, S., Sohn, K.: Unified confidence estimation networks for robust stereo matching. TIP (2018)
26. Kim, S., Min, D., Kim, S., Sohn, K.: Adversarial confidence estimation networks for robust stereo matching. T-ITS (2020)
27. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
28. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv:1610.02242 (2016)
29. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML workshops (2013)
30. Lee, D., Kim, S., Kim, I., Cheon, Y., Cho, M., Han, W.S.: Contrastive regularization for semi-supervised learning. arXiv:2201.06247 (2022)
31. Lerner, B., Shiran, G., Weinshall, D.: Boosting the performance of semi-supervised learning with unsupervised clustering. arXiv:2012.00504 (2020)
32. Li, J., Xiong, C., Hoi, S.C.: Comatch: Semi-supervised learning with contrastive graph regularization. In: ICCV (2021)
33. Lucas, T., Weinzaepfel, P., Rogez, G.: Barely-supervised learning: Semi-supervised learning with very few labeled images. arXiv:2112.12004 (2021)
34. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. TPAMI (2018)
35. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)
36. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016)
37. Park, S., Park, J., Shin, S.J., Moon, I.C.: Adversarial dropout for supervised and semi-supervised learning. In: AAAI (2018)
38. Pham, H., Dai, Z., Xie, Q., Le, Q.V.: Meta pseudo labels. In: CVPR (2021)
39. Poggi, M., Mattoccia, S.: Learning from scratch a confidence measure. In: BMVC (2016)
40. Rizve, M.N., Duarte, K., Rawat, Y.S., Shah, M.: In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. arXiv:2101.06329 (2021)
41. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Mutual exclusivity loss for semi-supervised deep learning. In: ICIP (2016)
42. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: NeurIPS (2016)
43. Seki, A., Pollefeys, M.: Patch based confidence prediction for dense disparity map. In: BMVC (2016)
44. Shi, W., Gong, Y., Ding, C., Tao, Z.M., Zheng, N.: Transductive semi-supervised deep learning using min-max features. In: ECCV (2018)
45. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: NeurIPS (2020)
46. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS (2017)
47. Tosi, F., Poggi, M., Benincasa, A., Mattoccia, S.: Beyond local reasoning for stereo confidence estimation with deep learning. In: ECCV (2018)

48. Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. arXiv:1903.03825 (2019)
49. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. In: NeurIPS (2020)
50. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. In: NeurIPS (2020)
51. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: CVPR (2020)
52. Xu, Y., Shang, L., Ye, J., Qian, Q., Li, Y.F., Sun, B., Li, H., Jin, R.: Dash: Semi-supervised learning with dynamic thresholding. In: ICML (2021)
53. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. arXiv:1905.00546 (2019)
54. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv:1605.07146 (2016)
55. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4l: Self-supervised semi-supervised learning. In: ICCV (2019)
56. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In: NeurIPS (2021)
57. Zhang, L., Qi, G.J.: Wcp: Worst-case perturbations for semi-supervised deep learning. In: CVPR (2020)
58. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
59. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.: Rethinking pre-training and self-training. In: NeurIPS (2020)