SelfMatch: Combining Contrastive Self-Supervision and Consistency for Semi-Supervised Learning

Byoungjip Kim, Jinho Choo, Yeong-Dae Kwon, Seongho Joe, Seungjai Min, Youngjune Gwon
Samsung SDS
{bjip.kim, jinho12.choo, y.d.kwon, drizzle.cho,
seungjai.min, gyj.gwon}@samsung.com

Abstract

This paper introduces SelfMatch, a semi-supervised learning method that combines the power of contrastive self-supervised learning and consistency regularization. SelfMatch consists of two stages: (1) self-supervised pre-training based on contrastive learning and (2) semi-supervised fine-tuning based on augmentation consistency regularization. We empirically demonstrate that SelfMatch achieves the state-of-the-art results on standard benchmark datasets such as CIFAR-10 and SVHN. For example, for CIFAR-10 with 40 labeled examples, SelfMatch achieves 93.19% accuracy that outperforms the strong previous methods such as MixMatch (52.46%), UDA (70.95%), ReMixMatch (80.9%), and FixMatch (86.19%). We note that SelfMatch can close the gap between supervised learning (95.87%) and semi-supervised learning (93.19%) by using only a few labels for each class.

1 Introduction

Deep neural networks have shown that they can achieve human-level performance in many tasks [13][14]. However, such high performance is usually achieved in a supervised learning setting that exploits a very large number of labeled training examples. Since labeling a large number of examples requires considerable time and cost, label efficient learning algorithms are in high demand.

Semi-supervised learning [4] is one of the attractive approaches that addresses the label inefficiency problem. Semi-supervised learning enables a deep neural network to learn with small labeled data by leveraging large unlabeled data [20][31][19][34][2][35][3][29].

Recently, self-supervised learning [17] is attracting high attention as a means of unsupervised representation learning. It learns representations by performing a pretext task in which labels can be automatically generated [10][37][25][11]. Also, contrastive self-supervised learning [16][26][1][5][32][15][33] learns representations by contrasting different views of examples in a task-agnostic way. It is shown that self-supervised learning provides representations good for downstream tasks.
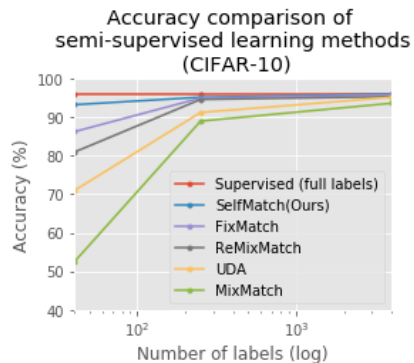

Figure 1: Accuracy comparison of semi-supervised learning methods. SelfMatch can close the gap between supervised and semi-supervised learning by using only a few labels.

In this paper, we introduce SelfMatch, a semi-supervised learning method that combines the power of contrastive self-supervised learning and consistency regularization. SelfMatch consists of two

stages: (1) self-supervised pre-training based on contrastive learning and (2) semi-supervised fine-tuning based on augmentation consistency regularization. We adopt SimCLR [5] for self-supervised pre-training, and FixMatch [29] for semi-supervised fine-tuning. SelfMatch achieves new state-of-the-art results on standard benchmarks such as CIFAR-10 [18] and SVHN [24] (see Figure 1). We conjecture that this is because the two stages are complementary in using unlabeled data. Contrastive self-supervised pre-training learns representations by maximizing the mutual information between different views of the data. Meanwhile, consistency regularization further enhances the learned representations by minimizing the cross entropy between class predictions of augmented data (see Appendix E). By combining two closely related but complementary stages, SelfMatch improves from other methods such as S4L [36] that also uses self-supervised pre-training (see Appendix A).

Our contributions can be summarized as follows.

- We introduce SelfMatch, a semi-supervised learning method that consists of two stages: (1) self-supervised pre-training based on contrastive learning and (2) semi-supervised fine-tuning based on augmentation consistency regularization. (see Figure 2).
- We empirically demonstrate that SelfMatch achieves the state-of-the-art results on standard benchmarks such as CIFAR-10 and SVHN (see Table 1). Especially, we show that SelfMatch can close the gap between supervised learning and semi-supervised learning by using only a few labels (see Figure 1).

2 Related Work

2.1 Self-supervised representation learning

Self-supervised learning [17] is a class of methods to unsupervised representation learning. Specifically, it aims to learn representations that can be used for downstream tasks via pretext tasks in which the training labels are automatically generated. Such pretext tasks include context prediction [10], image colorization [37], Jigsaw puzzle solving [25], and rotation degree classification (RotNet) [11].

Contrastive self-supervised learning is a task-agnostic approach to unsupervised representation learning. It aims to learn representations by contrasting two views from the same or different samples. Its objective functions are based on InfoMax principle [33]. This approach includes Deep InfoMax (DIM) [16], CPC [26], Augmented Multiscale DIM [1], SimCLR [5], CMC [32], MoCo [15], etc.

2.2 Semi-supervised learning

Semi-supervised learning enables a deep neural network to learn with small labeled data by leveraging large unlabeled data. To leverage unlabeled data, diverse methods like consistency regularization [27][31][19], entropy minimization [12], and pseudo-labeling [20] have been proposed.

In pseudo-labeling [20], a model is first trained with a small number of labeled examples, and then it assigns pseudo-labels with high confidence on unlabeled examples. And then, the pseudo-labeled examples are used to further train the model in a supervised manner. However, the vanilla pseudo-labeling method suffers from the over-fitting and noisy labels.

Consistency regularization has been introduced by Π-model [27], and is further developed by many following works such as Mean Teacher [31][19]. In the basic consistency regularization, an input image is transformed by a stochastic transformation function and the training objective is to minimize the distance between the model predictions of randomly transformed images by using L_2 distance.

Very recently, advanced consistency regularization methods have been introduced. These methods enhance basic consistency regularization methods with pseudo-labeling and improved data augmentation, and provide very high accuracy that is comparable to supervised learning with full labels. They include ICT [34], MixMatch [2], UDA [35], ReMixMatch [3], and FixMatch [29].

3 Method

An overview of SelfMatch is shown in Figure 2. SelfMatch consists of two stages: (1) self-supervised pre-training based on contrastive learning and (2) semi-supervised fine-tuning based on augmentation consistency regularization. The model $p_{model}(y|x)$ consists of an encoder $f(\cdot)$ and a head $c(\cdot)$.
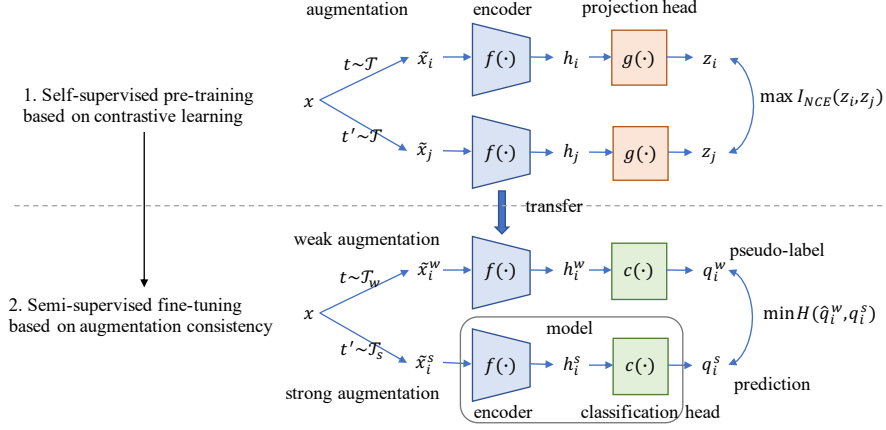
Figure 2: SelfMatch overview.

3.1 Self-supervised pre-training based on contrastive learning

To achieve the high accuracy with only a few labels, SelfMatch pre-trains encoder $f(\cdot)$ by leveraging unlabeled data. For this unsupervised pre-training, SelfMatch adopts SimCLR [5], one of the most promising self-supervised learning methods. SimCLR learns representations by encouraging two views $\tilde{x}_i$ and $\tilde{x}_j$ from the same image x to be similar, and two views $\tilde{x}_i$ and $\tilde{x}_k$ $(k \neq i)$ from different images to be dissimilar.

As shown in Figure 2, SimCLR consists of four components: data augmentation $\mathcal{T}(\cdot)$, base encoder $f(\cdot)$, projection head $g(\cdot)$, and contrastive loss $\mathcal{L}_c$. For data augmentation, SelfMatch uses random crop and color distortion. For the base encoder, we use ResNet-34 [13]. For the projection head, we use a MLP consisting of two layers with Dropout and ReLU activation. Finally, the contrastive loss is formulated as follows: $\mathcal{L}_c = -\log \frac{\exp(\mathrm{sim}(z_i,z_j)/\tau)}{\sum_{k=1}^{2B} \mathbb{1}(k \neq i)\exp(\mathrm{sim}(z_i,z_k)/\tau)}$, where $\mathrm{sim}(\cdot)$ is a similarity measure function, τ is a temperature parameter scaling the similarity, $\mathbb{1}(k \neq i)$ is an indicator function evaluating to 1 iff $k \neq i$, and B is a batch size. We set hyperparameters to $\tau = 0.5$ and $B = 512$.

3.2 Semi-supervised fine-tuning based on augmentation consistency regularization

Unlike the related methods such as S4L [36] (using semi-supervised fine-tuning based on VAT [23]) and SimCLRv2 [6] (using supervised fine-tuning with few labels), SelfMatch exploits semi-supervised fine-tuning based on augmentation consistency. For augmentation consistency, we adopt FixMatch [29]. It encourages the consistent prediction between weakly and strongly augmented examples ($\tilde{x}_i^w$ and $\tilde{x}_i^s$). More concretely, FixMatch uses a model output q_i^w of a weakly augmented input as the pseudo-label for a strongly augmented input $\tilde{x}_i^s$ (see Figure 2). For weak augmentation, we use random crop and random horizontal flip. For strong augmentation, SelfMatch adopts RandAugment (RA) [8], an effective automated augmentation method (see Appendix B). For the classification head $c(\cdot)$, SelfMatch uses a MLP consisting of two layers with Dropout [30] and ReLU activation.

The loss function of FixMatch consists of supervised and unsupervised loss: $\mathcal{L}_{semi} = \mathcal{L}_s + \lambda_u \mathcal{L}_u$. The supervised loss $\mathcal{L}_s$ is formulated as follows: $\mathcal{L}_s = \frac{1}{B} \sum_{i=1}^{B} H(p_i, q_i^w)$, where B is a batch size, p_i is a one-hot encoded label of sample x_i, and q_i^w is a model output $p_{model}(y|\tilde{x}_i^w)$ of a weakly augmented input $\tilde{x}_i^w$. For the experiments, we set B to 64.

The unsupervised loss $\mathcal{L}_u$ is formulated as follow: $\mathcal{L}_u = \frac{1}{\mu B} \sum_{i=1}^{\mu B} \mathbb{1}(\max(q_i^w) \geq c)H(\hat{q}_i^w, q_i^s)$, where $\mathbb{1}(\max(q_i^w) \geq c)$ is an indicator function, c is a confidence threshold, B is a batch size, and μ is the ratio of labeled and unlabeled samples in a batch. Here, $\hat{q}_i^w$ is $\arg\max(q_i^w)$ and q_i^s is a model output of a strongly augmented input $\tilde{x}_i^s$. We set hyperparameters to $c = 0.95$, $\mu = 7$, and $\lambda_u = 1$.

For training, SelfMatch uses a standard SGD optimizer with momentum $\beta = 0.9$ and exploits cosine learning rate decay [22] with the initial learning rate $\eta = 0.03$. Also, it utilizes Exponential Moving Average (EMA) [31] with weight decay 0.999 for stable training and inference.

3

Table 1: Comparison of accuracy for CIFAR-10 and SVHN.

	CIFAR-10			SVHN		
Method	40 labels	250 labels	4000 labels	40 labels	250 labels	1000 labels
Supervised		**95.87**			**97.41**	
Pseudo-Label	-	50.22±0.43	83.91±0.28	-	79.79±1.09	90.06±0.61
Π-Model	-	45.74±3.87	85.99±0.38	-	81.04±1.92	92.46±0.36
Mean Teacher	-	67.68±2.30	90.81±0.19	-	96.43±0.11	96.58±0.07
MixMatch	52.46±11.50	88.95±0.86	93.58±0.10	57.45±14.53	96.02±0.23	96.5±0.28
UDA	70.95±5.93	91.18±1.08	95.12±0.18	47.37±20.51	94.31±2.76	97.54±0.24
ReMixMatch	80.90±9.64	94.56±0.05	95.28±0.13	**96.66**±0.20	97.08±0.48	97.35±0.08
FixMatch(RA)	86.19±3.37	94.93±0.65	95.74±0.05	96.04±2.17	**97.52**±0.38	**97.72**±0.11
SelfMatch	**93.19**±1.08	**95.13**±0.26	**95.94**±0.08	**96.58**±1.02	97.37±0.43	97.49±0.07

Table 2: Effect of self-supervised pre-training on semi-supervised learning (CIFAR-10).

Method	40	250	4000	**Method**	40	250	4000
MixMatch	52.46	88.95	93.58	FixMatch	86.19	94.93	95.74
SimCLR + MixMatch	63.42 (+10.96)	92.58 (+3.63)	94.96 (+1.38)	SimCLR + FixMatch	93.19 (+7.00)	95.13 (+0.2)	95.94 (+0.2)

4 Experiments

We evaluate the performance of SelfMatch on standard classification benchmark datasets such as CIFAR-10 [18] and SVHN [24]. To see the label efficiency, we perform evaluation with varying amount of labeled data. For CIFAR-10, we use 50,000 unlabeled training images and 10,000 testing images. For SVHN, we use 73,357 unlabeled training images and 26,032 test images. We present SelfMatch results using the average value from three evaluation runs with different random seeds.

CIFAR-10 Table 1 shows the results for CIFAR-10. SelfMatch achieves the state-of-the-art for all number of labeled examples. More important, at 40 labeled examples (4 labels for each class), SelfMatch achieves 93.19% accuracy that outperforms the strong previous methods including MixMatch [2] (52.46%), UDA [35] (70.95%), ReMixMatch [3] (80.9%), and FixMatch [29] (86.19%). It is remarkable that the gap between the supervised learning (95.87%) and the semi-supervised learning (93.19%) can be narrowed down to a few percent with only 4 labels per class (see also Figure 1). To further demonstrate the effectiveness of SelfMatch, a comparison of learning curves of different methods is presented in Appendix D.

SVHN Table 1 shows the results for SVHN. SelfMatch achieves very high accuracy that are within the margin of error of other state-of-the-art results.

Ablation study Table 2 shows the effect of self-supervised pre-training for two representative semi-supervised learning methods, MixMatch [2] and FixMatch [29]. At 40 labeled examples, self-supervised pre-training improves MixMatch by 10.96%, and FixMatch by 7.00%. Further ablation study can be found in Appendix C.

5 Conclusion

In this paper, we introduce SelfMatch, a semi-supervised learning method that consists of two stages: (1) self-supervised pre-training based on contrastive learning and (2) semi-supervised fine-tuning based on augmentation consistency regularization. We empirically demonstrate that SelfMatch can close the gap between supervised learning and semi-supervised learning using only a few labels.

References

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[3] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *International Conference on Learning Representations (ICLR)*, 2020.

[4] Olivier Chapelle, Bernhard Schölkopf, Alexander Zien, et al. Semi-supervised learning, vol. 2. *Cambridge: MIT Press.*, 2006.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[7] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[10] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

[11] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[12] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2005.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollr, and Ross Girshick. Mask rcnn. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[16] 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 *International Conference on Learning Representations (ICLR)*, 2019.

[17] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[19] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2016.

[20] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.

[21] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[22] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

[23] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.

[24] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*. 2011.

[25] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[27] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[28] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[29] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958, 2014.

[31] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[32] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, 2020.

[33] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[34] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2019.

[35] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. In *International Conference on Learning Representations (ICLR)*, 2020.

[36] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[37] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, 2016.

A Comparison with related method

A comparison with a closely related method is shown in Table 3. Very recently, Zhai et al. presented S4L [36], a semi-supervised learning method that consists of three stages: (1) self-supervised pre-training based on rotation prediction (RotNet [11]), (2) semi-supervised fine-tuning based on adversarial consistency (VAT [23]), and (3) supervised fine-tuning with a few labeled data. SelfMatch is different from S4L. First, SelfMatch proposes to use contrastive self-supervised pre-training (e.g., SimCLR [5], MoCo [15], etc). Since contrastive self-supervised learning is task-agnostic, it provides better representations for downstream tasks. Second, SelfMatch proposes to use semi-supervised fine-tuning based on augmentation consistency (e.g., FixMatch [29], ReMixMatch [3], etc.). Since contastive self-supervised learning and augmentation consistency regularization are commonly based on data augmentation, SelfMatch is architecturally simpler and has more rooms to be optimized.

Table 3: A comparison with a related method.

Phase	S4L [36]	SelfMatch (Ours)
1. Pre-training	Task-specific self-supervised (RotNet [11])	Task-agnostic self-supervised (SimCLR [5])
2. Fine-tuning A	Semi-supervised (adversarial consistency, VAT [23])	Semi-supervised (augmentation consistency, FixMatch [29])
3. Fine-tuning B	Supervised with few labels	-

B Transformations in RandAugment

SelfMatch uses RandAugment [8] for strong augmentation in the semi-supervised fine-tuning stage. Table 4 shows the list of transformations used in RandAugment. At each strong data augmentation process, two transformations with random magnitude are randomly selected and serially performed.

Table 4: List of transformations used in RandAugment [8].

Transformation	Description
Autocontrast	Maximizing the image contrast.
Brightness	Adjusting the brightness of the image. (0: black image, 1: original image)
Color	Adjusting the color balance of the image. (0: black & while image, 1: original image)
Contrast	Adjusting the contrast of the image. (0: a gray image, 1: original image)
Equalize	Equalizing the image histogram.
Identity	The original image.
Posterize	Reducing each pixel to [4, 8] bits.
Rotate	Rotating the image by [-30, 30] degrees.
Sharpness	Adjusting the sharpness of the image. (0: blurred image, 1: original image)
Shear-x	Shearing the image along the horizontal axis with rate [-0.3, 0.3].
Shear-y	Shearing the image along the vertical axis with rate [-0.3, 0.3].
Solarize	Inverting all pixels above a threshold value of [0, 1].
Translate-x	Translating the image horizontally by ([-0.3, 0.3] x image width) pixels.
Translate-y	Translating the image vertically by ([-0.3, 0.3] x image height) pixels.

C Further ablation study

Table 5 shows further ablation study. More specifically, it shows baseline accuracy results of self-supervised pre-training for CIFAR-10 [18] with 4000 labels. Usually, in unsupervised representation learning, the performance of learned representation is measured by using a linear classification head and full labeled data [5]. As shown in Table 5, when following such an evaluation protocol for CIFAR-10, SimCLR [5] provides about 92.27% accuracy. When using a MLP head consisting of two layers with Dropout [30] and ReLU activation, SimCLR provides about 92.69% accuracy (+0.42%). In semi-supervised learning setting with 4000 labels, it decreases to 91.36% (-1.33%). This can be considered as a baseline method that exploits self-supervised pre-training. More specifically, the baseline method consists of (1) self-supervised pre-training and (2) supervised fine-turning with 4000 labels. Unlike this, using semi-supervised fine-truing based on augmentation consistency (FixMatch [29]) in the second stage, SelfMatch improves the baseline method up to 95.94% in a large margin (+4.58%).

Table 5: Baseline accuracy results of self-supervised pre-training (CIFAR-10).

Method	Encoder (ResNet-34)	Head	Labels	Accuracy (%)
Supervised	random init.	Linear (512×10)	50000	95.87
Supervised	random init.	Linear (512×10)	4000	79.74
SimCLR	frozen	Linear (512×10)	50000	92.27
SimCLR	frozen	MLP (2 layers)	50000	92.69
SimCLR	frozen	Linear (512×10)	4000	90.99
SimCLR	frozen	MLP (2 layers)	4000	**91.36**
SelfMatch	fine-tuned with FixMatch	MLP (2 layers)	4000	**95.94** (+4.58)

D Comparison of learning curves

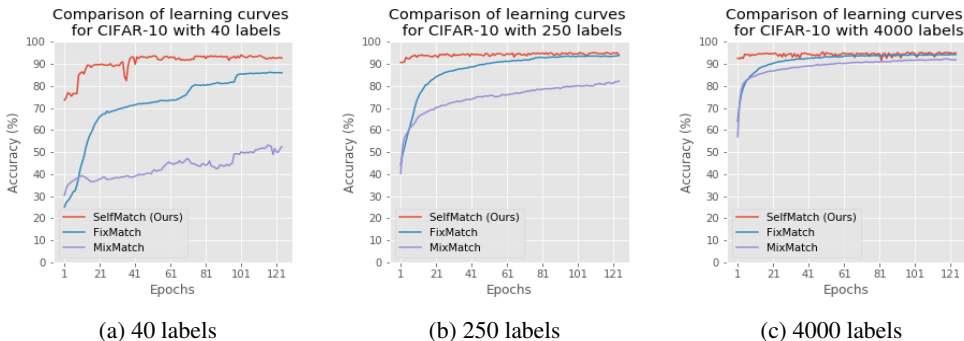(a) 40 labels   (b) 250 labels   (c) 4000 labels

Figure 3: A comparison of learning curves for CIFAR-10.

To demonstrate the effectiveness of SelfMatch in more detail, we present a comparison of learning curves for CIFAR-10 in Figure 3. Each figure shows a comparison of learning curves of three most advanced methods including SelfMatch (ours), FixMatch [29] and MixMatch [2] for varying number of labels (i.e., 40, 250, 4000 labels). Note that the learning curves of SelfMatch only present the second stage of SelfMatch, that is the semi-supervised fine-tuning stage after the self-supervised pre-training stage.

As shown in Figure 3, compared to other methods, SelfMatch not only achieves higher accuracy, but also reaches such high accuracy more rapidly. The accuracy gain is more prominent in the case where only a few labels are available (see Figure 3a). These results are also summarized in Table 1. More importantly, SelfMatch reaches such high accuracy more rapidly. For 250 and 4000 labels, SelfMatch reaches an accuracy of greater than 90% through just one epoch. In the case of 40 labels, SelfMatch achieves an accuracy of 90% through around 10 epochs, while FixMatch and MixMatch do not reach such high accuracy even though iterating through a very large number of training epochs.

E   Discussion

Relation between contrastive self-supervision and consistency regularization   In this paper, we introduce SelfMatch, a semi-supervised learning method that combines the power of contrastive self-supervised learning and consistency regularization. To further motivate the reason why we combine these two approaches, we summarize the relation between the two approaches in Table 6. Note that the relation is based on SimCLR [5] and FixMatch [29] adopted by SelfMatch. If other methods are considered for each approach, the comparison would be slightly different.

SelfMatch consists of two stages: (1) self-supervised pre-training based on contrastive learning and (2) semi-supervised fine-tuning based on consistency regularization. We empirically found that SelfMatch not only achieves higher accuracy, but also reaches such high accuracy more rapidly. As mentioned in Section 1, we conjecture that this is because the two stages are complementary in using unlabeled data. Contrastive self-supervised learning methods such as SimCLR [5] use loss functions based on the InfoMax principle [21], and the mutual information is usually estimated by InfoNCE [26]. Also, Michael et al. [33] have shown that InfoNCE losses are related with triplet losses [28]. We have a similar perspective. In the pre-training stage of SelfMatch, contrastive self-supervised learning learns representations by maximizing a lower bound (e.g., InfoNCE) of the mutual information between different views of the data. This InfoNCE loss can be interpreted as encouraging two objectives: (1) maximizing the similarity between different views from the same sample and (2) minimizing the similarity between different views from different samples. The former objective can be considered as using intra-sample statistics, and the later objective can be understood as leveraging inter-sample statistics. Meanwhile, in the fine-tuning stage of SelfMatch, consistency regularization [29] further enhances the learned representations by minimizing the cross entropy (or distance) between class predictions of different views from the same sample. This objective can be considered as exploiting intra-sample statistics.

Furthermore, two approaches are slightly different in exploiting intra-sample statistics. The target entities to optimize are normalized representations in contrastive self-supervised learning, while they are class probability distributions in consistency regularization. Also, consistency regularization of FixMatch exploits asymmetric augmentation (i.e., weak and strong augmentation of the same sample).

Table 6: Relation between contrastive self-supervision and consistency regularization.

	Contrastive self-supervision	**Consistency regularization**
Method	SimCLR [5]	FixMatch [29]
Loss function	$-\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2B} \mathbb{1}(k \neq i)\exp(\mathrm{sim}(z_i, z_k)/\tau)}$	$\frac{1}{\mu B} \sum_{i=1}^{\mu B} \mathbb{1}(\max(q_i^w) \geq c) H(\hat{q}_i^w, q_i^s)$
	Maximizing the similarity between different views form the same sample (z_i and positive sample z_j) Minimizing the similarity between different views from different samples (z_i and negative sample z_k)	Minimizing the cross entropy (or distance) between predictions of different views of the same sample ($\hat{q}_i^w$ and q_i^s)
Negative samples	Yes	No
Target entity	Representations (z_i)	Class probability distributions (q_i)
Transformation	Symmetric (two weak augmentations)	Asymmetric (weak and strong augmentations)

Large-scale experiment   In this paper, we imperially demonstrate the effectiveness of SelfMatch by using standard benchmark datasets such as CIFAR-10 [18] and SVHN [24]. However, these datasets are rather small. Therefore, we are doing experiments by using larger datasets such as STL-10 [7] and ImageNet [9], and will add the results in an extended version of this paper.