Uniform Noise Distribution and Compact Clusters: Unveil The Key to Self-Supervised Learning's Success in Label Noise

Anonymous authors
Paper under double-blind review

Abstract

Label noise is ubiquitous in real-world datasets, posing significant challenges to machine learning models. While self-supervised learning (SSL) algorithms have empirically demonstrated effectiveness in learning noisy labels, the theoretical understanding of their effectiveness remains underexplored. In this paper, we present a theoretical framework to understand how SSL methods enhance learning with noisy labels, especially for the instance-dependent label noise. We reveal that the uniform and compact cluster structures induced by contrastive SSL play a crucial role in mitigating the adverse effects of label noise. Specifically, we theoretically show that a classifier trained on SSL-learned representations significantly outperforms one trained using traditional supervised learning methods. This results from two key merits of SSL representations over label noise: 1. Uniform Noise Distribution: Label noise becomes uniformly distributed over SSL representations with respect to the true class labels, rather than the noisy ones, leading to an easier learning task. 2. Enhanced Cluster Structure: SSL enhances the formation of well-separated and compact categorical clusters, increasing inter-class distances while tightening intra-class clusters. We further theoretically justify the benefits of training a classifier on such structured representations, demonstrating that it encourages the classifier trained on noisy data to be aligned with the optimal classifier. Extensive experiments validate the robustness of SSL representations in combating label noise, confirming the practical values of our theoretical findings.

1 Introduction

Label noise is ubiquitous in the real world since acquiring accurately annotated datasets is expensive and time-consuming (Patrini et al., 2017; Xiao et al., 2015; Natarajan et al., 2013). Alternatively, a large amount of annotated images can be collected from unreliable sources such as non-expert annotators and image search engines, where label noise is inevitable (Xia et al., 2020a; Li et al., 2017). Recent self-supervised learning (SSL) with contrastive learning paradigms achieved great success in learning meaningful data representations without label information (He et al., 2020; Chen et al., 2020c; Zbontar et al., 2021; Caron et al., 2020).

In SSL, any two augmented examples from the *same image* (referred to as positive pairs) are mapped to nearby locations in the embedding space, whereas two augmented images from *different images* (referred to as negative pairs) are mapped to a distant location (Oord et al., 2018; Purushwalkam & Gupta, 2020; Chen et al., 2020b). Empirical evidence demonstrates that representations learned by SSL can be easily adapted to many downstream tasks such as image classification, objection detection, segmentation, and learning with imbalanced datasets (Grill et al., 2020; Misra & Maaten, 2020; Zhao et al., 2021; Xie et al., 2021; Liu et al., 2021; Yang & Xu, 2020).

Apart from these applications, in this paper, we show how SSL representations enhance learning with label noise. Specifically, we first construct a motivating example of instance-dependent label noise, then we prove that a classifier trained on representations learned by SSL with noisy labels is optimal over clean data distribution. Then we systematically analyze the benefits of representations learned by SSL and find two merits of SSL representations: (1) Uniform Noise Distribution: The label noise **uniformly** spreads over the learned SSL representations, (2) Enhanced Cluster Structure: The learned representations exhibit a

separated and **compact** cluster structure with respect to true labels. It increases the distance of clusters from different classes while reducing the variance of the cluster of the same class.

For point (1), we theoretically show that the label noise is uniformly distributed across the learned representations by SSL in the motivating example, which is easier to address in practice (Cheng et al., 2020; Chen et al., 2021; Zhang et al., 2020). We further extend the relationship between label noise and the representations learned by SSL to a more general case and provide empirical validation. For point (2), we empirically and theoretically justify that representations learned by SSL exhibit a cluster structure based on true labels.

Furthermore, we demonstrate that such a structure encourages the classifier trained on noisy data to align more closely with the optimal classifier learned from clean data. In contrast, representations learned through supervised learning (SL) do not achieve uniform noise distribution or form a good cluster structure. In particular, representations learned through supervised learning still depend on noisy labels and they exhibit clusters with respect to noisy labels instead of true labels.

From the algorithmic perspective, we show that mixup over SSL representations boosts the property of cluster structure discussed in our theory. Specifically, we fix representations learned by SSL and then only train a linear classifier on the frozen representations with different types of label noise methods, which show significant improvements. The main contributions are summarized as follows:

- We provide theoretical analysis to show that representations learned by SSL exhibit a more uniform and discriminative cluster structure under label noise, leading to a better classifier than supervised learning.
- We systematically show that SSL breaks the dependency of label noise and representation, resulting in uniformly distributed label noise over classes. Further, SSL enlarges the distance of clusters from different classes while tightening the cluster of the same class.
- Empirically, we show that mixup based on conventional SSL augmentations satisfies our theoretical motivations and benefits cluster structure learning. Extensive experiments validate our analysis and the effectiveness of SSL in noisy label learning.

2 Related Work

Learning with Noisy Labels. There are different approaches to addressing the issues caused by label noise. Commonly used loss functions such as cross-entropy loss (CE) are not robust to label noise. Therefore, noise-robust loss function methods are proposed to make models fit clean examples but not mislabeled examples (Zhang & Sabuncu, 2018; Wang et al., 2019; Ma et al., 2020; Englesson & Azizpour, 2021). Alternative solutions design selection strategies to improve the confidence of clean examples and filter them from noisy data(Han et al., 2018; Yu et al., 2019; Wei et al., 2020; Song et al., 2019; Yang et al., 2023). Label correction methods correct pseudo labels of noisy samples for computing their loss functions (Zhang et al., 2021; Liu et al., 2020; Yi & Wu, 2019; Zhang et al., 2020). Noise transition matrix methods estimate the underlying label noise distribution and use it to correct the noisy labels and build a robust classifier. Other solutions include using the neural symbolic system to model and reduce the noise Smirnova et al. (2022), filtering noisy labels from sample and parameter levels (Wang et al., 2023), and enhancing the label accuracy with graph fusion (Xu et al., 2022). Cheng et al. (2021) provides a study on distilling the knowledge of SSL representations to supervised learning representations. In contrast, we focus on revealing and investigating how plain SSL representation can successfully address the label noise issue.

Self-supervised Learning. Representations of images learned by SSL have achieved remarkable success. SimCLR (Chen et al., 2020b) requires a large batch size to contain sufficient in-batch negative pairs and domain-specific augmentations such as Gaussian blur, color distortions, and color jittering. However, a large batch size is infeasible. MoCo (He et al., 2020) solves this issue by introducing a memory bank to store representations of data from previous iterations. BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021) propose new frameworks without using negative pairs so they can work with a reasonable batch size. On the other hand, SSL relies on domain-specific image augmentation. That is to assume that image augmentations

such as changing the colors of images should not affect labels of images in downstream tasks (Tsai et al., 2020). DACL (Verma et al., 2021) and I-MIX (Lee et al., 2020) both leverage mixup augmentation (Zhang et al., 2017) as domain-agnostic augmentation and they find that SSL methods with both domain-agnostic and domain-specific augmentations perform better.

Noisy labels with SSL. Applying SSL methods to mitigate label noise has recently been studied. Li et al. (2021) propose the contrastive learning loss in the principle subspace via autoencoder and the mixup in the low dimensional space to learn a robust representation. Ghosh & Lan (2021) empirically validates that the representations pretrained by SSL methods are beneficial for noisy label learning. C2D (Zheltonozhskii et al., 2022) propose a contrast and divide method to empirically address the warm-up obstacle for memorizing the noise. Yi et al. (2022) analyze the memorization issue in cross-entropy issue and propose a contrastive regularization to mitigate it. Xue et al. (2022) shows that contrastive learning boosts robustness by analyzing the singular values of the representation matrix. However, these methods either lack systematic theoretical analysis or rely on extra assumptions of sub-class structure to show how SSL benefits noisy label learning.

Different from these discussed methods, our method provides theoretical analysis that SSL methods benefit noisy label learning by learning a more separated and compact cluster structure. Concretely, we analytically show that SSL can enlarge the cluster distance from different classes and reduce the variance of each cluster. Further, we empirically prove that the built-on mixup can achieve our theoretical motivations and improve noisy label learning.

3 Analysis of SSL representations with noisy labels

We first provide a motivating example to show that SSL can be significantly better than supervised learning, which enables us to explore and investigate the benefits of representations learned by SSL.

We first construct a binary classification problem with two linearly separable clusters, where the samples from clusters are artificially flipped according to a label noise function. We denote y_i by the true label for x_i and assume it is a balanced sample from $\{-1, +1\}$. Then the instance x_i is decided in the following manner:

$$x_i = \begin{cases} e_1\zeta_i + e_2\xi_i, & \text{if } y_i = +1 \\ -e_1\zeta_i - e_2\xi_i, & \text{if } y_i = -1 \end{cases}$$

where $\zeta \sim \mathcal{U}_{[0,4]}$, $\xi \sim \mathcal{U}_{[-1.75,2.25]}$, and $e_1, e_2 \in \mathbb{R}^d$ are two orthogonal unit-norm vectors. We assume $\beta(x, y) = \text{sign}(yx^\top e_2)$ as the instance-dependent label noise function. For each clean example (x_i, y_i), the corresponding noisy example is $(x_i, \tilde{y}_i)$, where $\tilde{y}_i = y_i\beta(x_i, y_i)$. Then we can compute that there are 43.75% mislabeled examples if the noise function $\beta(x, y)$ is adopted.

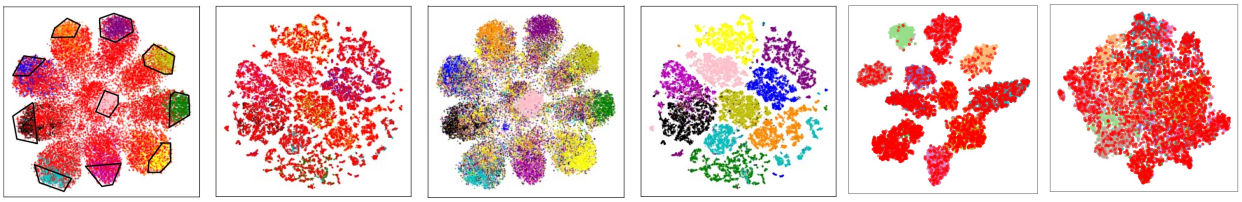
We use a simple linear classifier parameterized by ω and use the gradient descent algorithm to learn the parameter ω over the noisy data $\{x_i, \tilde{y}_i\}_{i=1}^n$, with a logistic loss function. Thus, in conventional supervised learning, we have the loss:

$$\mathcal{L}(\omega) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-\tilde{y}_i\omega^\top x_i\right)\right).$$

In contrast, in the SSL framework, we first learn a linear representation model with parameter $W \in \mathbb{R}^{1\times d}$ from $\{x_i\}_i^n$ in a self-supervised manner. Specifically, we adopt the linear SSL objective function studied in Liu et al. (2021); HaoChen et al. (2021), which tends to pull two positive pairs $(x + \gamma, x + \gamma')$ to nearby locations in the embedding space:

$$W_{\text{SSL}} = \underset{W\in\mathbb{R}^{1\times d}}{\arg\min} -\hat{\mathbb{E}}[(x+\gamma)^\top W^\top W(x+\gamma')] + \frac{1}{2}\left\|W^\top W\right\|_F^2, \tag{1}$$

where $\hat{\mathbb{E}}$ is an empirical expectation over the data, with γ, γ' are independent and identical $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ random variables. Once the optimal representation model W_{SSL} is obtained, we fix W_{SSL} and then learn a linear classifier parameterized by θ on the top of representations with noisy labels $\{(W_{\text{SSL}}x_i, \tilde{y}_i)\}_i$. Analogous to



(a) SL(noisy labels) (b) SSL(noisy labels) (c) SL(true labels) (d) SSL(true labels) (e) SL(noisy labels) (f) SSL(noisy labels)

Figure 1: T-SNE of 60% instance-dependent label noise on CIFAR-10. We train a ResNet34 on the noisy data by supervised learning (SL) and we visualize the representations learned by SL in (a) and (c) with respect to noisy labels and true labels, respectively. Similarly, we also train an SSL ResNet34 and visualize data representations in (b) and (d). We highlight regions with solid polygons that suffer from the label noise in (a), where red points (label noise) represent wrong-labeled data. In (b), the red points are almost uniformly spread over the data representations. Similarly, we also show results on the real-world Clothing1M trained with ResNet50. The similar observation shows that the label noise is not uniformly distributed in SL (e) while almost uniformly distributed in all classes in SSL (f).

supervised learning, we also use the gradient descent with a logistic loss function $\mathcal{L}(\theta)$ to train the classifier.

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-\tilde{y}_i \theta^\top W_{\text{SSL}} x_i\right)\right).$$

The following theorem states the behavior of linear classifiers on input data $\{(x_i, \tilde{y}_i)\}_i$ and representations of inputs data $\{(W_{\text{SSL}} x_i, \tilde{y}_i)\}_i$, respectively.

Theorem 3.1. *Let $\tilde{\omega}, \tilde{\theta}$ be normalized optimal parameters via gradient descent with logistic loss over the data $\{(x_i, \tilde{y}_i)\}_i$ and $\{(W_{\text{SSL}} x_i, \tilde{y}_i)\}_i$, respectively. Then the generalization accuracy in supervised learning is upper bounded by:*

$$\Pr_{(x,y)}[\text{sign}(\tilde{\omega}^\top x) = y] \leq \frac{9}{16} + \frac{2d}{3n}, \tag{2}$$

While the generalization accuracy in SSL is lower bounded by:

$$\Pr_{(x,y)}[\text{sign}(\tilde{\theta}^\top W_{\text{SSL}} x) = y] \geq 1 - 2e^{-n/128}. \tag{3}$$

Remark Theorem 3.1 reveals two interesting facts in the presence of label noise. (1) The prediction accuracy under SSL is guaranteed to be a high value via a provable lower bound. The lower bound could further converge to 1 (perfect prediction without error) when sample size $n \to +\infty$. (2) In contrast, in supervised learning, simply collecting more samples does not guarantee a high accuracy, where the upper bound of the accuracy converges to $9/16$ as $n \to +\infty$.

4 Why SSL Works

To show the benefits of SSL in noisy labels, we begin with analyzing the learned representation W_{SSL} from Eq. (1).

Proposition 4.1. *The optimal solution W_{SSL} in Eq. (1) converges in probability to ke_1 with the constant $k > 0$.*

The solution W_{SSL} is the span of the vector e_1, which is crucial for learning an optimal classifier in Eq.(3). Note that only e_1 determines the true labels of data x. In fact, the injected label noise depends on the non-discriminative feature e_2 but not the discriminative feature e_1. If we orthogonally project data x onto the direction of e_1, the label noise is independent of and is uniformly distributed over the projected data points spanned by e_1. The representation model W_{SSL} exactly maps the data x onto the direction of e_1

orthogonally. Thus, the label noise is uniformly distributed over data representations $W_{\text{SSL}}x$, which makes the label noise easier to address. Specifically:

If the label noise is *uniformly distributed* across the inputs, the classifier trained on data with this label noise can generalize well. Specifically, it can be verified that the optimal Bayes classifier, $h(x) = \text{sign}(e_1^\top x)$, is also the optimal classifier over the clean distribution. On the other hand, the classifier trained with the supervised learning method from Theorem 3.1 is forced to learn spurious correlations between the inputs e_2 and the labels.

Besides, estimating a noise transition matrix is easier when the label noise is uniformly distributed over inputs. In particular, the instance-dependent label noise can be characterized by the noise transition matrix $T(x) \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$, where $T(x)_{ij}$ measures the probability of observing a corrupted label j given the true label i and an instance x. The issue of label noise is then solved by estimating the noise transition matrix $T(x)$ (Xiao et al., 2015; Liu & Tao, 2015; Patrini et al., 2017; Goldberger & Ben-Reuven, 2016). Estimating $T(x)$ for instance-dependent label noise is practically challenging since $T(x)$ can be different for different x and we may need to parametrize n different $T(x)$ from the noisy dataset of size n by a neural network (Cheng et al., 2020; Xia et al., 2020b; Berthon et al., 2021). In contrast, T is the same for all x for *symmetric label noise* (i.e, label noise is uniformly distributed over data) (Patrini et al., 2017) and we only need to estimate a constant noise transition matrix. Therefore, by estimating a single noise transition matrix instead of parameterizing n noise transition matrices by a neural network, the label noise is easier to solve.

5 Generalized Observations in Real World

The benefits of the SSL are not only in the motivating example but can also be observed in real-world datasets. In this section, we empirically justify the benefits by investigating the CIFAR-10 with 60% instance-dependent label noise. We train a ResNet34 (He et al., 2016) over the noisy data through cross-entropy loss, and visualize the representations in Fig 1(a) with respect to noisy labels. The SSL representations with respect to noisy labels are visualized in Fig 1(b), where the representation model ResNet34 is trained with the self-supervised method MoCov2. Fig 1(a) shows that in the representations learned by supervised learning, the label noise and representations are still dependent (the regions are highlighted by solid polygons), whereas the SSL breaks such dependency and makes the label noise uniformly distributed across the data representations. We also get similar observations on the real-world Clothing1M dataset as shown in (e) and (f). The representations of green and yellow classes have less label noise, which indicates that label noise is not uniformly distributed in the supervised learning classifier in Clothing1M.

Besides, we also visualize these representations with respect to their true labels in Fig 1(c-d). We find that representations learned by SSL exhibit an intrinsic cluster structure that is consistent with the true labels (Fig 1(d)). In contrast, Fig 1(c) shows that representations learned by supervised learning do not exhibit a cluster structure with respect to true labels. Thus, it further motivates us to explore how the cluster structure learned by SSL helps noisy label learning.

6 Compact Cluster Structure for Label Noise

In this section, we investigate how the cluster structure can help mitigate the label noise. Concretely, we show that for fixed representations, a good cluster structure encourages the classifier to be aligned to the optimal classifier, resulting in better generalization performance.

For simplicity, we use a two-component Gaussian mixture model to describe the clusters of representations, with each cluster representing one class. We assume that representations from class $+1$ are sampled from $\mathcal{N}(\mu, \Sigma)$ and representations from class -1 are sampled from $\mathcal{N}(-\mu, \Sigma)$, where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$. In this case, the distance between two clusters is controlled by $\|\mu\|$, and the variance of each cluster is controlled by the sum of eigenvalues of Σ, which is equivalent to the trace of Σ.

We connect the cluster structure to $-\widetilde{\nabla\mathcal{L}}(\omega_0)^\top\tilde{\mu}$, which can characterize the performance of the linear classifier, where $\widetilde{\nabla\mathcal{L}}(\omega_0) = \frac{\nabla\mathcal{L}(\omega_0)}{\|\nabla\mathcal{L}(\omega_0)\|}$, $\tilde{\mu} = \frac{\mu}{\|\mu\|}$, and $\nabla\mathcal{L}(\omega_0)$ is the gradient of the logistic loss computed by

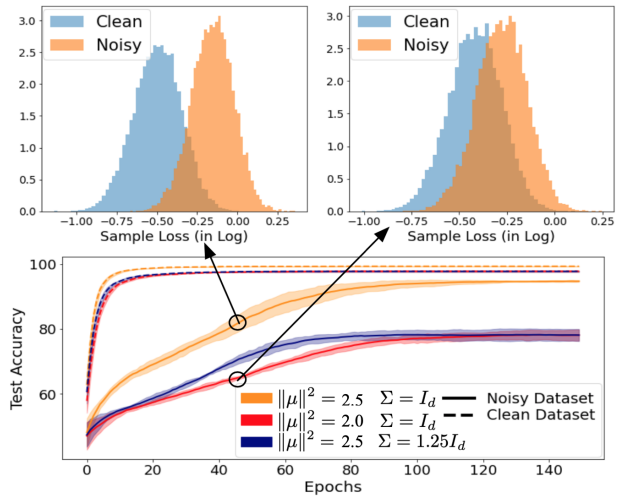


Figure 2: Linear classifiers trained on synthetic datasets with 40% noise level. The dashed line is the performance of classifiers without label noise and the solid line is that with label noise. The histograms are sample loss values at epoch = 50 with respect to whether they are mislabeled.

the linear classifier (initialized by ω_0). The normalized gradient of the loss $-\widetilde{\nabla\mathcal{L}}(\omega_0)$ represents the direction of the steepest descent in the loss function calculated on the noisy data. Given that an optimal classifier obtained from the clean data is $k\mu$ for any scalar $k > 0$, $-\widetilde{\nabla\mathcal{L}}(\omega_0)^\top\tilde{\mu}$ can measure the cosine similarity between the gradient descent direction and the direction where the optimal classifier points. After applying one-step gradient descent on the classifier, the updated classifier is more correlated to the optimal classifier if the cosine similarity is higher. More details can be found in the Appendix. This intuitively explains why $-\widetilde{\nabla\mathcal{L}}(\omega_0)^\top\tilde{\mu}$ can be used to measure the performance of the linear classifier. Now we focus on establishing the relationship between $-\widetilde{\nabla\mathcal{L}}(\omega_0)^\top\tilde{\mu}$ and the cluster structure.

As shown in Section 4, label noise is uniformly distributed over representations. Thus, we define the symmetric label noise function:

$$\beta(x, y) = \begin{cases} -1, & \text{with probability } r \\ +1, & \text{with probability } 1 - r \end{cases}$$

where $0 < r < 1$ controls the noise level. Note that for symmetric label noise, the label noise function $\beta(x, y)$ is independent of the data. The relationship between $-\widetilde{\nabla\mathcal{L}}(\omega_0)^\top\tilde{\mu}$ and the cluster structure is presented in Theorem 6.1.

Theorem 6.1. *If at least half of examples are clean ($r < \frac{1}{2}$)*

$$-\widetilde{\nabla\mathcal{L}}(\omega_0)^\top\tilde{\mu} \geq \sqrt{\frac{\|\mu\|^2}{c\mathrm{Tr}(\Sigma) + \|\mu\|^2}}(1 - 2r) + o(n^{-1/3}), \tag{4}$$

where $\mathrm{Tr}(\Sigma)$ is the trace of Σ and $c > 0$ is a constant.

Theorem 6.1 provides a lower bound for $-\widetilde{\nabla\mathcal{L}}(\omega_0)^\top\tilde{\mu}$. The larger lower bound means the updated classifier is more correlated to the optimal one.

The lower bound can be affected by the noise level r and the following two cluster properties: **1)** the distance between two clusters $\|\mu\|$, **2)** the variance of each cluster $\mathrm{Tr}(\Sigma)$. Without considering any label correction techniques, the noise level r is fixed given a dataset. Therefore, by learning clusters of data representations that are distant from each other (larger $\|\mu\|$) and/or by learning tight representation clusters (smaller $\mathrm{Tr}(\Sigma)$), the classifier generalizes better.

Remark We remark that the spirits of encouraging a good cluster structure are the same for other forms of label noise such as asymmetric label noise, though their expressions of $-\widetilde{\nabla\mathcal{L}}(\omega_0)^\top\tilde{\mu}$ are different. Details are in Appendix C.2.

We empirically justify that linear classifiers get better performance when $-\widetilde{\nabla\mathcal{L}}(\omega_0)^\top\tilde{\mu}$ becomes larger. Fig 2 shows the performances of classifiers trained on data points with different cluster structures given a fixed noise level 40%. Specifically, with the same variance, the classifier trained on clusters with larger distances performs better (orange line v.s. red line). While with the same distance, the classifier trained on tight clusters performs better (orange line v.s. blue line). The two histograms show that the linear classifier with larger $-\widetilde{\nabla\mathcal{L}}(\omega_0)^\top\tilde{\mu}$ (orange) fits clean examples better, compared with the linear classifier (red). It also highlights that representations with better cluster structure help the classifier generalize better on clean data distribution.

7 Benefits of SSL on Cluster Structure

In this section, we rigorously justify that the cluster properties characterized in Theorem 6.1 can be achieved by SSL. In particular, we focus on the SSL objective function Eq. (5) that has been studied in Wang & Isola (2020). Notably, the loss and its variants have been widely adopted in the SSL (Chen et al., 2020b; He et al., 2020; Tsai et al., 2021).

$$\mathcal{L}_{\text{ctr}}(f) = \frac{1}{n}\sum_{i=1}^{n}\underbrace{\mathbb{E}_{\substack{u_i\sim P(u|x_i)\\ u_i^+\sim P(u|x_i)}}\left[\left\|f(u_i)-f(u_i^+)\right\|^2\right]}_{\mathcal{L}_{\text{Align}}(i)} + \lambda\log\left[\frac{1}{n(n-1)}\sum_{i\neq j}\underbrace{\mathbb{E}_{\substack{u_i\sim P(u|x_i)\\ u_j\sim P(u|x_j)}}\left[e^{-\left\|f(u_i)-f(u_j)\right\|^2}\right]}_{\mathcal{L}_{\text{Uniform}}(i,j)}\right], \tag{5}$$

where λ is a hyper-parameter. $\{x_1, x_2, \cdots, x_n\}$ are input instances, $P(u|x)$ denotes the conditional distribution of an augmented instance u gievn x, and $f(\cdot)$ is a representation network that takes u as an input. Intuitively, Eq. (5) minimizes the distance between two representations of different views from the same instance, and the $\mathcal{L}_{\text{Uniform}}$ makes representations uniformly distributed over the embedding space.

To characterize the cluster properties studied in Theorem 6.1, we introduce the notion of δ-cluster closeness.

Definition 7.1 (δ-cluster closeness)**.** *Let S_i be the support where $P(u|x_i) > 0$ for any $u \in S_i$. S_i and S_j are δ-cluster close if $P[u \in S_i \cap S_j|x_i] \geq \delta$ for any two different instances x_i, x_j with $y_i = y_j$.*

The notion of δ-cluster closeness is similar to the cluster assumption in Lafferty & Wasserman (2007); Rigollet (2007); Singh et al. (2008). Definition 7.1 reveals that the instance augmentations should be rich enough so that any two different distorted augmentations from the same class can be overlapped. Following (Verma et al., 2021; Lee et al., 2020), we apply mixup data augmentation on top of the conventional SSL data augmentations (Chen et al., 2020b) in order to have richer instance augmentations, making Definition 7.1 hold with a large δ. With the notion of δ-cluster closeness, we investigate the relationship between Eq. (5) and cluster properties of representations: the distances between any two clusters and the variance of each cluster.

To analyze the cluster properties, we decompose $\mathcal{L}_{\text{ctr}}(f)$ into three components and study their effects individually.

$$\mathcal{L}_{\text{ctr}}(f) = \frac{1}{n}\sum_{m\in\mathcal{Y}}\sum_{i\in J_m}\mathcal{L}_{\text{Align}}(i) + \lambda\log\left[\frac{1}{n(n-1)}\sum_{\substack{m\in\mathcal{Y},n\in\mathcal{Y}\\ m\neq n}}\sum_{\substack{i\in J_m\\ j\in J_n}}\mathcal{L}_{\text{Uniform}}(i,j) + \frac{1}{n(n-1)}\sum_{m\in\mathcal{Y}}\sum_{\substack{i,j\in J_m\\ i\neq j}}\mathcal{L}_{\text{Uniform}}(i,j)\right], \tag{6}$$

The following lemma indicates that a large distance between any two clusters can be achieved by optimizing $\mathcal{L}_{\text{Uniform}}(i,j)$ for some i, j. We denote J_y by the index set corresponding to true class y.

Lemma 7.2. *Let* $\hat{\mu}_i = \sum_{k \in J_i} \frac{f(u_k)}{|J_i|}$, $\hat{\mu}_j = \sum_{k \in J_j} \frac{f(u_k)}{|J_j|}$ *be sample means of cluster* i *and cluster* j *with* $i \neq j$. *Without loss of generality, we assume* $|J_i| = |J_j|$. *Then*

$$\mathbb{E}[\|\hat{\mu}_i - \hat{\mu}_j\|] \geq -\frac{1}{|J_i|} \sum_{k \in J_i} \log \left(\mathcal{L}_{Uniform}(k, g(k)) \right)$$

where the function $g : J_i \to J_j$ *is any bijective function, and the expectation is over the data augmentation.*

Remark Lemma 7.2 indicates that the distance between the cluster i and the cluster j can be lower bounded by $-\log \left(\mathcal{L}_{\text{Uniform}}(k, g(k)) \right)$ for $k \in J_i$. Since $-\mathcal{L}_{\text{Uniform}}(k, g(k))$ measures the distance between $f(u_k)$ from cluster i and $f(u_{g(k)})$ from cluster j, minimizing $\mathcal{L}_{\text{Uniform}}(k, g(k))$ for all $k \in J_i$ increases the distance between the cluster i and the cluster j.

On the other hand, the objective function Eq. (5) also controls the variance of each cluster. The following lemma helps us understand how the SSL objective function Eq. (5) controls the variance of each cluster.

Lemma 7.3. *Let* $\widehat{\Sigma}_y = \frac{1}{|J_y|} \sum_{i \in J_y} (f(u_i) - \hat{\mu}_i)(f(u_i) - \hat{\mu}_i)^\top$ *be the sample covariance matrix. Suppose Definition 7.1 holds. Then for any fixed* $\delta \in (0, 1)$, *we have*

$$\text{Tr}(\mathbb{E}[\widehat{\Sigma}_{\text{y}}]) \leq \frac{2}{\delta |J_y|} \sum_{i \in J_y} \mathcal{L}_{Align}(i),$$

where the expectation is over the data augmentation.

Remark Lemma 7.3 shows that the variance of cluster y is upper bounded by the term $\sum_{i \in J_y} \mathcal{L}_{\text{Align}}(i)$, where the variance is measured by the sum of eigenvalues for the sample covariance matrix computed by representations from the cluster y. In other words, a small variance of cluster y can be achieved by minimizing $\sum_{i \in J_y} \mathcal{L}_{\text{Align}}(i)$.

Lemma 7.2 and Lemma 7.3 have shown the effects of the first two components in Eq. (6). The last component in Eq. (6) serves as a contradiction against the first component. We note that minimizing $\mathcal{L}_{\text{Uniform}}(i, j)$ for i, j from the same cluster undesirably increases the variance of that cluster. This intuition is justified by the following proposition.

Proposition 7.4. *Suppose Definition 7.1 holds with a fixed* δ. *Then*

$$\log[\frac{1}{n(n-1)} \sum_{m \in \mathcal{Y}} \sum_{i,j \in J_m i \neq j} \mathcal{L}_{Uniform}(i, j)] \geq -\alpha \sum_{m \in \mathcal{Y}} \sum_{i \in J_m} \mathcal{L}_{Align}(i),$$

where $\alpha = \frac{2(n/|\mathcal{Y}|-1)}{\delta n(n-1)} > 0$.

Remark Proposition 7.4 indicates that minimizing the third term in Eq. (6) forces the first term to be larger, which makes the variance of clusters to be larger, where the strength is controlled by a factor α. We note that the third term is due to $\mathcal{L}_{\text{Uniform}}$ of the instances from the same class and it cannot be eliminated since the label information is not leveraged. The constraint strength is mitigated when α decreases. It is small when instance augmentations are rich enough (δ is large), which also highlights the importance of data augmentations in learning SSL representations.

8 Experiment

To validate our analysis of the cluster properties of SSL representations, we combine different SSL methods (i.e., MoCo and BYOL) as complementary with label noise methods. Compared to MoCov2, BYOL does not explicitly compute $\mathcal{L}_{\text{Uniform}}$ but implicitly computes it by the momentum update of the network.

Datasets. Following previous state-of-the-arts (Huang et al., 2023; Yang et al., 2023), we evaluate our method on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) with different types of label noise: symmetric

Table 1: Test accuracy on CIFAR-10 and CIFAR-100 datasets with SYM label noise over different noise levels.

Dataset	CIFAR-10					CIFAR-100				
Noise	20%	40%	60%	80%	90%	20%	40%	60%	80%	90%
SSL+CE	92.66±0.05	92.60±0.06	92.53±0.11	92.22±0.04	91.61±0.04	64.16±0.13	63.76±0.72	62.25±0.42	60.56±0.27	57.07±0.96
GCE	93.16±0.18	90.11±0.27	82.35±0.29	74.95±0.51	54.34±0.81	71.71±0.09	67.72±0.19	59.50±0.43	35.80±0.62	14.04±0.97
+MoCo	**95.74**±0.07	**95.67**±0.06	**95.58**±0.04	**95.36**±0.08	**94.68**±0.24	**75.21**±0.03	**74.89**±0.08	**73.36**±0.09	**71.91**±0.31	**68.22**±0.72
+BYOL	**95.55**±0.02	**95.46**±0.05	**95.32**±0.06	**95.11**±0.08	**94.66**±0.16	**73.53**±0.03	**72.04**±0.04	**71.43**±0.09	**69.40**±0.20	**65.94**±0.26
CT	93.66±0.17	92.22±0.16	70.51±0.22	39.75±0.88	27.34±0.98	72.69±0.14	68.81±0.19	61.15±0.28	16.40±0.44	8.22±1.46
+MoCo	**95.43**±0.07	**95.37**±0.08	**95.19**±0.23	**91.97**±0.80	**87.65**±1.65	**73.86**±0.07	**73.37**±0.12	**72.59**±0.41	**67.79**±0.92	**62.69**±2.18
+BYOL	**95.13**±0.02	**94.93**±0.04	**94.71**±0.03	**93.58**±0.55	**87.35**±1.37	**72.19**±0.05	**71.33**±0.18	**69.49**±0.08	**55.55**±3.28	**52.65**±1.22
ELR	93.53±0.10	93.11±0.14	92.22±0.16	85.74±0.52	54.27±1.06	69.64±0.39	65.16±0.30	60.88±0.32	24.92±0.52	10.22±0.76
+MoCo	**95.88**±0.05	**95.81**±0.04	**95.74**±0.03	**95.65**±0.02	**95.60**±0.09	**72.89**±0.39	**72.74**±0.06	**71.74**±0.11	**70.47**±0.19	**66.75**±0.22
+BYOL	**95.55**±0.02	**95.43**±0.03	**95.30**±0.06	**95.11**±0.05	**95.12**±0.10	**72.48**±0.03	**71.73**±0.06	**70.35**±0.10	**68.45**±0.10	**63.70**±0.14
TCL	93.53±0.10	94.73±0.11	93.22±0.13	90.34±0.29	87.51±0.36	76.24±0.26	73.16±0.28	69.45±0.33	62.71±0.41	53.58±0.56
+MoCo	**95.91**±0.03	**95.71**±0.03	**95.82**±0.03	**95.68**±0.02	**95.30**±0.07	**78.81**±0.31	**76.79**±0.16	**72.52**±0.11	**71.87**±0.23	**68.63**±0.22
+BYOL	**95.59**±0.11	**95.60**±0.13	**95.15**±0.08	**95.03**±0.08	**94.95**±0.15	**78.68**±0.08	**76.41**±0.18	**72.25**±0.13	**71.65**±0.22	**66.13**±0.24

Table 2: Test accuracy on CIFAR-10 and CIFAR-100 datasets with ASYM label noise over different noise levels.

Dataset	CIFAR-10					CIFAR-100				
Noise	10%	20%	30%	40%	45%	10%	20%	30%	40%	45%
SSL+CE	92.25±0.06	87.63±0.21	86.29±0.14	83.86±0.22	79.13±0.55	64.01±0.20	63.42±0.11	62.24±0.62	59.06±0.75	53.16±0.54
GCE	93.02±0.10	91.92±0.23	90.85±0.28	89.44±0.44	85.51±0.59	73.52±0.08	70.05±0.31	65.80±0.35	53.49±0.53	44.08±1.22
+MoCo	**95.63**±0.04	**95.37**±0.08	**95.00**±0.31	**93.30**±0.26	**88.31**±0.41	**74.54**±0.04	**73.60**±0.12	**72.63**±0.13	**66.27**±0.24	**56.12**±0.68
+BYOL	**95.50**±0.02	**95.23**±0.14	**94.83**±0.19	**93.56**±0.36	**90.69**±0.16	73.17±0.04	**72.12**±0.09	**70.75**±0.14	**65.09**±0.13	**53.96**±0.50
CT	94.40±0.03	93.32±0.11	90.27±0.15	69.47±0.21	66.08±0.32	73.88±0.04	69.88±0.21	64.64±0.68	55.22±0.71	48.22±1.00
+MoCo	**95.37**±0.07	**95.25**±0.09	**94.33**±0.16	**92.29**±0.32	**86.79**±0.52	**73.48**±0.11	**72.02**±0.26	**69.36**±0.41	**63.30**±0.73	**55.70**±1.58
+BYOL	**95.48**±0.03	**94.14**±0.72	**94.04**±0.24	**90.72**±0.72	**87.33**±1.23	72.01±0.08	**70.46**±0.04	**66.22**±0.37	54.97±0.94	46.62±1.24
ELR	93.90±0.08	93.26±0.10	92.52±0.13	90.93±0.16	**88.49**±0.24	73.89±0.07	73.44±0.20	72.90±0.19	70.62±0.34	65.62±1.31
+MoCo	**95.73**±0.02	**95.69**±0.04	**94.83**±0.12	92.62±1.15	78.92±0.95	**74.87**±0.05	**74.51**±0.10	**73.75**±0.08	**72.26**±0.05	**67.11**±0.28
+BYOL	**95.59**±0.03	**95.49**±0.07	**95.40**±0.03	**94.72**±0.04	86.91±1.73	73.44±0.05	72.95±0.04	71.81±0.05	69.18±0.08	63.27±0.12
TCL	94.10±0.20	93.53±0.17	92.52±0.11	91.14±0.23	90.91±0.36	78.24±0.24	74.77±0.28	73.98±0.22	71.62±0.52	67.38±0.56
+MoCo	**95.18**±0.08	**95.33**±0.09	**95.07**±0.11	**93.85**±0.12	**93.60**±0.09	**80.89**±0.29	**78.54**±0.16	**77.74**±0.11	**73.88**±0.21	**69.35**±0.25
+BYOL	**95.25**±0.12	**95.38**±0.13	**95.10**±0.11	**93.21**±0.09	**93.17**±0.10	**80.28**±0.15	**78.12**±0.08	**77.25**±0.11	**73.65**±0.12	**68.65**±0.11

(SYM), asymmetric (ASYM), and instance-dependent label noise (IDN). We also evaluate our method on two real-world datasets ANIMAL-10N (Song et al., 2019) and Clothing1M (Xiao et al., 2015).

CIFAR-10 and *CIFAR-100* both consist of 50,000 training and 10,000 test color images whose size is 32×32. The difference is that CIFAR-10 has 10 classes containing 5,000 training and 1,000 test images for each class while CIFAR-100 has 100 classes containing 500 training and 100 test images for each class.

ANIMAL-10N is a dataset with animals of confusing appearances crawled from several online search engines including Bing and Google using the predefined labels as the search keyword. It includes 50,000 training and 5,000 test images.

Clothing1M has 1 million training images and 10,000 test images from 14 classes. The noisy images are collected from online shopping websites, and there are many mislabeled samples since the labels are created by the surrounding text provided by the sellers.

Implementations. The implementation of MoCov2 and BYOL includes the backbone and a 2-layer MLP ad the projected head. For CIFAR datasets, the backbone is ResNet34 with a 2-layer MLP as the projection head. The input and output dimensions are 512. For ANIMAL-10N and Clothing1M, the backbone is ResNet50, and the input and output dimensions are 2048.

For the data augmentations used in MoCov2 and BYOL, we use both strong image augmentation from Chen et al. (2020b) and MixUp from Lee et al. (2020). MixUp (Zhang et al., 2017) has a hyper-parameter λ that

Table 3: Test accuracy on CIFAR-10 and CIFAR-100 datasets with IDN label noise over different noise levels.

Dataset	CIFAR-10					CIFAR-100				
Noise	20%	40%	60%	80%	90%	20%	40%	60%	80%	90%
SSL+CE	91.86±0.23	90.79±0.13	89.65±0.23	87.80±0.19	80.01±1.24	65.78±0.11	63.19±0.14	61.47±0.25	58.84±0.49	54.86±0.57
GCE	90.05±0.29	80.35±0.34	66.94±0.51	49.38±0.66	34.49±0.97	69.58±0.16	60.48±0.32	44.63±0.62	28.34±0.84	14.18±1.29
+MoCo	**95.41**±0.07	**95.05**±0.10	**94.35**±0.21	**91.87**±0.33	**87.72**±0.28	**74.19**±0.04	**72.48**±0.09	**70.47**±0.14	**66.67**±0.46	**61.67**±0.48
+BYOL	**95.22**±0.08	**94.80**±0.09	**94.26**±0.25	**92.88**±0.22	**90.33**±0.77	**72.69**±0.06	**70.86**±0.10	**68.24**±0.10	**65.20**±0.20	**60.06**±0.41
CT	91.50±0.35	85.95±0.38	74.09±0.52	30.79±0.82	22.35±0.92	69.81±0.18	62.59±0.31	52.11±0.65	16.10±0.70	7.91±0.57
+MoCo	**95.23**±0.37	**94.68**±0.31	**94.07**±0.58	**83.91**±0.62	**77.87**±2.69	**73.39**±0.22	**72.15**±0.27	**70.14**±0.54	**66.26**±1.08	**58.34**±0.68
+BYOL	**94.97**±0.07	**94.52**±0.18	**94.11**±0.14	**89.09**±0.08	**71.72**±1.47	**71.59**±0.04	**70.15**±0.15	**66.73**±0.72	**57.79**±0.45	**49.88**±1.23
ELR	93.54±0.04	93.20±0.18	92.07±0.20	73.27±0.55	41.39±0.80	70.11±0.32	67.16±0.70	58.11±0.67	21.96±0.74	10.28±1.07
+MoCo	**95.77**±0.07	**95.70**±0.10	**95.65**±0.07	**95.58**±0.04	**91.35**±1.91	**72.74**±0.04	**71.56**±0.12	**69.69**±0.22	**65.94**±0.59	**59.80**±0.84
+BYOL	**95.45**±0.02	**95.25**±0.03	**95.08**±0.04	**95.07**±0.06	**94.91**±0.10	**72.11**±0.18	**70.64**±0.32	**68.72**±0.24	**63.75**±0.50	**57.54**±0.48

controls the strength of interpolation between data points, where we set $\lambda = 1$ for CIFAR datasets and $\lambda = 2$ for ANIMAL-10 and Clothing-1M. Once we train the representation network, we train a linear classifier by different label noise methods on this representation network.

The linear classifier is trained for 100 epochs using SGD, where the learning rate starts from {1, 5, 10, 20, 30} and it is reduced by a factor of 5 after 20, 30 and 40 epochs. For GCE method (Zhang & Sabuncu, 2018), its parameter q is selected from {0.2, 0.4, 0.6, 0.8, 0.9}; for Co-teaching method (Han et al., 2018), the warmup parameters is selected from {5, 8, 10}; for ELR method (Liu et al., 2020), the parameter β is selected from {0.7, 0.9} and the parameter λ is selected from {3, 5, 7}; for TCL method (Huang et al., 2023), two augmentations are used, and the temperature τ of contrastive loss and the α of mixup are 0.25 and 0.1.

Baselines. To show our analysis generally facilitating learning noisy labels, we combine frozen SSL representations with different types of algorithms: robust loss function GCE (Zhang & Sabuncu, 2018), sample selection Co-teaching (Han et al., 2018), label correction ELR (Liu et al., 2020), and TCL (Huang et al., 2023) that models and filter samples simultaneously, and the standard cross-entropy (SSL+CE). GCE designs a loss function to address memorization issues for incorrect labels. Co-teaching selects clean examples to update the neural network. ELR introduces a regularization for pseudo labels. TCL uses the GMM model to model the noisy label distribution and filter wrong labels as out-of-distribution examples.

Main Results. Tables 1-3 show the results for SYM, ASYM, and IDN label noise on CIFAR-10 and CIFAR-100, respectively. Table 6 shows the results on ANIMAL-10N and Clothing1M. Both MoCov2 and BYOL SSL representations can improve efficacy to a wide range of label noise methods: robust loss function methods, sample selection methods, and label correction methods. Results show that training a linear classifier on frozen SSL representations over noisy datasets is significantly better than training a whole neural network over noisy datasets. Asymmetric label noise is to flip labels between semantically-similar classes. For example, cats are inherently more difficult to differentiate from dogs than trucks. To this end, we combine IDN with ASYM to generate more realistic label noise. Specifically, we choose similar images for each class and then we flip their labels to the next class. Results are reported in Table 4.

Following Zhang et al. (2020); Chen et al. (2020a); Lee et al. (2019), semantic label noise is a type of instance-dependent label noise that follows the intuition that hard instances are more likely to be mislabeled, where the hard instances are near the decision boundary of the model. To generate the semantic label noise, we train a VGG-13 (Simonyan & Zisserman, 2014) on training datasets for 30 epochs. Following Chen et al. (2020a), we select instances with the highest mislabeling scores to corrupt. For the first case, we corrupt these instances with random labels. For the second case, we corrupt these instances with predictions of the model VGG-13. We term the former TYPE-1 label noise and the latter TYPE-2 label noise. The results for the two types of label noise are reported in Table 5. Therefore, extensive experiments have demonstrated the effectiveness of applying frozen SSL representations.

Cluster Structure. We evaluate our cluster structure of SSL representations learned by MoCov2 on CIFAR-10. Learning SSL representations do not leverage the label information, so the representations are invariant to label noise, whereas representations learned by supervised learning (SL) are sensitive to label

Table 4: Test accuracy on CIFAR-10 and CIFAR-100 datasets with IDN-ASYM label noise over different noise levels.

Dataset	CIFAR-10					CIFAR-100				
Noise Level	10%	20%	30%	40%	45%	10%	20	30%	40%	45%
GCE	87.02±0.16	77.40±0.29	68.63±0.68	57.85±0.81	54.01±0.92	71.01±0.12	62.42±0.18	52.48±0.33	44.69±0.78	40.02±0.65
+MoCo	**95.20**±0.02	**94.89**±0.05	**93.28**±0.21	**84.26**±0.43	**71.66**±1.10	**73.90**±0.04	**71.52**±0.57	**69.12**±0.21	**61.69**±0.87	**52.17**±2.23
+BYOL	**95.19**±0.05	**94.85**±0.10	**93.66**±0.21	**85.57**±0.37	**70.44**±1.44	**71.58**±0.14	**69.61**±0.13	**67.22**±0.08	**59.33**±0.77	**50.37**±1.57
CT	87.75±0.28	78.37±0.30	69.31±0.55	60.48±0.57	54.62±0.80	71.97±0.09	64.33±0.13	55.61±0.26	47.12±0.32	42.23±0.46
+MoCo	**94.73**±0.45	**93.25**±0.42	**88.92**±0.64	**74.67**±0.91	**61.13**±1.37	**72.77**±0.32	**69.61**±0.49	**66.52**±0.80	**59.37**±1.05	**50.43**±1.43
+BYOL	**94.88**±0.06	**93.55**±0.07	**90.99**±0.49	**74.80**±0.39	**63.67**±1.07	**69.54**±0.03	**67.69**±0.12	**64.44**±0.19	**59.23**±0.32	**51.65**±0.94
ELR	94.08±0.05	93.97±0.08	93.91±0.14	93.79±0.20	77.76±2.24	72.21±0.23	71.96±0.24	71.83±0.41	**70.96**±0.43	**67.33**±0.45
+MoCo	**95.72**±0.05	**95.60**±0.04	**95.46**±0.03	**95.23**±0.09	**95.04**±0.20	**73.90**±0.68	**73.16**±0.57	**72.69**±0.35	70.11±0.64	64.51±1.09
+BYOL	**95.39**±0.02	**95.30**±0.03	**95.20**±0.07	**94.99**±0.14	**85.19**±0.32	70.58±0.12	69.80±0.09	68.60±0.18	66.95±0.35	63.32±0.61

Table 5: Test accuracy on CIFAR-10 datasets semantic label noise over different noise levels.

Dataset	TYPE-1					TYPE-2				
Noise Level	20%	40%	60%	80%	90%	10%	20%	30%	40%	45%
GCE	88.61±0.11	79.01±0.36	68.77±0.43	53.24±0.82	33.66±0.74	85.31±0.08	76.08±0.22	70.44±0.34	64.71±0.63	57.8±0.47
+MoCo	**95.39**±0.04	**94.80**±0.10	**89.94**±0.16	**83.13**±0.95	**71.84**±0.59	**90.92**±0.05	**83.10**±0.12	**76.41**±0.17	**69.60**±0.29	**66.68**±0.55
+BYOL	**94.86**±0.04	**93.69**±0.31	**89.19**±0.36	**83.53**±1.26	**75.77**±1.44	**89.64**±0.27	**83.34**±0.29	**75.65**±0.37	**69.32**±0.15	**66.86**±0.34
CT	91.06±0.33	72.78±0.35	44.30±0.47	25.37±0.18	17.30±0.23	85.52±0.12	76.11±0.12	65.36±0.20	55.54±0.81	48.64±0.38
+MoCo	**95.02**±0.43	**89.84**±0.67	**84.70**±0.96	**73.34**±1.84	**59.22**±3.11	**89.00**±0.45	**83.23**±0.77	**76.55**±0.92	**72.41**±1.53	**66.66**±1.98
+BYOL	**94.58**±0.07	**91.56**±0.42	**86.19**±0.70	**73.27**±0.61	**64.75**±1.95	**91.06**±0.47	**83.10**±0.20	**76.50**±0.18	**68.57**±1.23	**63.82**±1.58
ELR	94.19±0.05	91.75±0.51	82.72±0.43	70.86±0.46	39.05±0.72	86.69±0.02	79.06±0.08	71.02±0.11	62.09±0.27	58.02±0.26
+MoCo	**95.76**±0.01	**94.96**±0.26	**84.30**±0.59	**79.99**±0.45	**63.63**±1.09	**88.27**±0.49	**84.53**±0.27	**78.45**±0.61	**77.57**±0.35	**69.27**±1.70
+BYOL	**95.66**±0.02	**95.35**±0.07	**88.91**±0.31	**77.83**±0.42	**76.98**±0.60	**89.41**±0.87	**85.83**±0.70	**78.30**±0.55	**77.10**±0.62	**71.64**±1.50

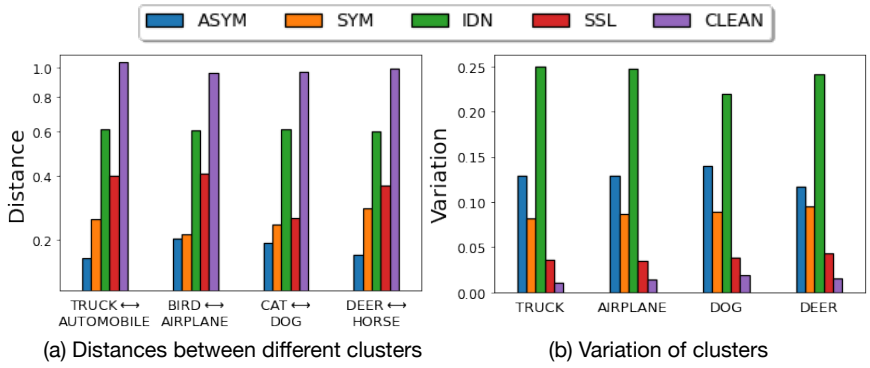


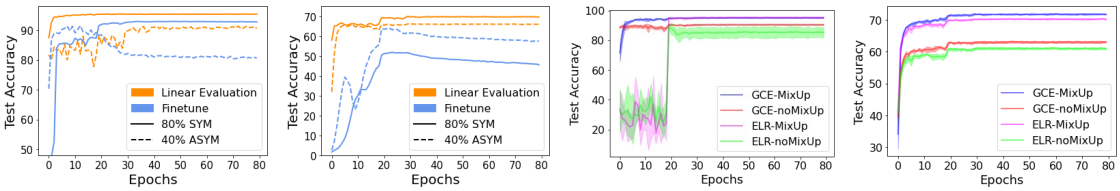
Figure 3: Illustration of cluster structures for CIFAR-10 dataset. Representations are learned in different label noise settings. ASYM (blue): 40% asymmetric noise; SYM (orange): 60% symmetric noise; IDN (green): 60% instance-dependent noise. And we visualize distances between two clusters in (a) and the variance of each cluster in (b). The cluster structure (purple) serves as a baseline that representations are trained by supervised learning without label noise.

noise. We compare the cluster structure of SSL to that of SL in different label noise settings in Figure 3. We find that SSL representations (red) have a better cluster structure than SL representations obtained with different label noise. We note that although the distances between clusters of SL representations (green) learned with IDN are slightly larger than that of SSL representations, the variance of each cluster is 10 times larger. We also highlight the baseline cluster structure with purple, which is trained by the SL method without label noise.

Fine-tuned Performance. Following Chen et al. (2020b), we fine-tune the MoCov2 representation network on CIFAR-10 by GCE algorithm. For noise-free classification tasks, fine-tuning usually outperforms linear

Table 6: Test accuracy on ANIMAL-10N and Clothing1M

Dataset	Animal-10N				Clothing-1M			
Method	GCE	CT	ELR	TCL	GCE	CT	ELR	TCL
Origin	84.58	86.93	86.52	87.90	71.34	71.68	71.89	73.81
+MOCO	**87.35**	**87.66**	**88.51**	**89.51**	**72.61**	**72.41**	**72.71**	**74.31**
+BYOL	**88.42**	**88.36**	**88.68**	**89.43**	**72.90**	**72.63**	**72.98**	**74.38**



CIFAR-10 CIFAR-100 CIFAR-10 CIFAR-100

Figure 4: Left Two: comparison of linear eval and fine-tune on GCE. Right Two: comparison of SSL methods with and without MixUp component enabled on 80% symmetric label noise.

evaluation on various classification datasets (Chen et al., 2020b; Grill et al., 2020). However, Figure 4 illustrates that linear evaluation (orange) performs better than fine-tuning (blue) in the presence of label noise. When label noise exits and the representations are not frozen, fine-tuning degrades the performance of the neural network. We hypothesize that fine-tuning the learned SSL representations destroys the cluster structure regarding true labels, leading to poor performance.

The effects of MixUp augmentation. We study the importance of mixup data augmentations. Our analysis indicates the importance of keeping larger δ in Lemma 7.3 and Proposition 7.4 by data augmentation. Without MixUp, the Definition 7.1 holds with smaller δ. With MixUp enabled, the bound in Lemma 7.3 is tighter and the negative effects in Proposition 7.4 are mitigated. Results in Figure 4 indicate that applying MixUp augmentation significantly improves the performance on noisy datasets.

9 Conclusion

We provide a simple but effective method to address label noise. We first construct a motivating example to theoretically show that the classifier learned on SSL representations generalizes better than that from supervised learning. By further investigating the SSL representations under label noise, we find that: (1) The label noise is uniformly distributed over the data representations. (2) Representations learned by SSL exhibit good cluster properties, which encourages the linear classifier to be aligned with the optimal classifier. From the algorithmic perspective, we show that SSL representations can be applied as a strong complementary to various label noise methods by extensive experiments.

References

Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *International Conference on Machine Learning*, pp. 825–836. PMLR, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. *arXiv preprint arXiv:2012.05458*, 2020a.

Pengfei Chen, Guangyong Chen, Junjie Ye, jingwei zhao, and Pheng-Ann Heng. Noise against noise: stochastic label noise helps combat inherent label noise. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=80FMcTSZ6J0.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020b.

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*, 2020c.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. *arXiv preprint arXiv:2010.02347*, 2020.

Hao Cheng, Zhaowei Zhu, Xing Sun, and Yang Liu. Demystifying how self-supervised features improve training from noisy labels. *arXiv preprint arXiv:2110.09022*, 2021.

Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *arXiv preprint arXiv:2105.04522*, 2021.

Aritra Ghosh and Andrew Lan. Contrastive learning improves model robustness under label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2703–2708, 2021.

Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 2016.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *arXiv preprint arXiv:1804.06872*, 2018.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *arXiv preprint arXiv:2106.04156*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Zhizhong Huang, Junping Zhang, and Hongming Shan. Twin contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11661–11670, 2023.

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

John Lafferty and Larry Wasserman. Statistical analysis of semi-supervised regression. 2007.

Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. I-mix: A domain-agnostic strategy for contrastive representation learning. *arXiv preprint arXiv:2010.08887*, 2020.

Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. In *International Conference on Machine Learning*, pp. 3763–3772. PMLR, 2019.

Junnan Li, Caiming Xiong, and Steven CH Hoi. Learning from noisy data with robust representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9485–9494, 2021.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *arXiv preprint arXiv:2007.00151*, 2020.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International Conference on Machine Learning*, pp. 6543–6553. PMLR, 2020.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26:1196–1204, 2013.

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

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *arXiv preprint arXiv:2007.13916*, 2020.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(7), 2007.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Aarti Singh, Robert Nowak, and Jerry Zhu. Unlabeled data: Now it helps, now it doesn't. *Advances in neural information processing systems*, 21:1513–1520, 2008.

Alisa Smirnova, Jie Yang, Dingqi Yang, and Philippe Cudre-Mauroux. Nessy: A neuro-symbolic system for label noise reduction. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, pp. 5907–5915. PMLR, 2019.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020.

Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Han Zhao, Louis-Philippe Morency, and Ruslan Salakhutdinov. Self-supervised representation learning with relative predictive coding. *arXiv preprint arXiv:2103.11275*, 2021.

Vikas Verma, Thang Luong, Kenji Kawaguchi, Hieu Pham, and Quoc Le. Towards domain-agnostic contrastive learning. In *International Conference on Machine Learning*, pp. 10530–10541. PMLR, 2021.

Haixin Wang, Huiyu Jiang, Jinan Sun, Shikun Zhang, Chong Chen, Xian-Sheng Hua, and Xiao Luo. Dior: Learning to hash with label noise via dual partition and contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 322–330, 2019.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13726–13735, 2020.

Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International Conference on Learning Representations*, 2020a.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33, 2020b.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2691–2699, 2015.

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8392–8401, 2021.

Ning Xu, Yong-Di Wu, Congyu Qiao, Yi Ren, Minxue Zhang, and Xin Geng. Multi-view partial multi-label learning via graph-fusion-based label enhancement. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise. In *International Conference on Machine Learning*, pp. 24851–24871. PMLR, 2022.

Hao Yang, You-Zhi Jin, Zi-Yin Li, Deng-Bao Wang, Xin Geng, and Min-Ling Zhang. Learning from noisy labels via dynamic loss thresholding. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *arXiv preprint arXiv:2006.07529*, 2020.

Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7017–7025, 2019.

Li Yi, Sheng Liu, Qi She, A Ian McLeod, and Boyu Wang. On learning contrastive representations for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16682–16691, 2022.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pp. 7164–7173. PMLR, 2019.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. *arXiv preprint arXiv:2103.07756*, 2021.

Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Zizhao Zhang, Han Zhang, Sercan O Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9294–9303, 2020.

Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10623–10633, 2021.

Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alex M Bronstein, and Or Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1657–1667, 2022.