Noise Is Also Useful: Negative Correlation-Steered Latent Contrastive Learning

Jiexi Yan[1], Lei Luo[2], Chenghao Xu[1], Cheng Deng[1]*, Heng Huang[2]
[1]School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2]Department of Electrical and Computer Engineering, University of Pittsburgh, PA 15260, USA
{jxyan1995, luoleipitt, shydyl1456113691, chdeng.xd,henghuanghh}@gmail.com

Abstract

*How to effectively handle label noise has been one of the most practical but challenging tasks in Deep Neural Networks (DNNs). Recent popular methods for training DNNs with noisy labels mainly focus on directly filtering out samples with low confidence or repeatedly mining valuable information from low-confident samples. However, they cannot guarantee the robust generalization of models due to the ignorance of useful information hidden in noisy data. To address this issue, we propose a new effective method named as LaCoL (**La**tent **Co**ntrastive **L**earning) to leverage the negative correlations from the noisy data. Specifically, in label space, we exploit the weakly-augmented data to filter samples and adopt classification loss on strong augmentations of the selected sample set, which can preserve the training diversity. While in metric space, we utilize weakly-supervised contrastive learning to excavate these negative correlations hidden in noisy data. Moreover, a cross-space similarity consistency regularization is provided to constrain the gap between label space and metric space. Extensive experiments have validated the superiority of our approach over existing state-of-the-art methods.*

1. Introduction

In the past ten years, Deep Neural Networks (DNNs) have achieved impressive performance and revolutionized a wide variety of computer vision applications, such as image recognition [13,14], semantic segmentation [31,36], object detection [10,27], and cross-modal retrieval [30,34,48]. The remarkable success in training DNNs has been benefiting from the collection of large-scale datasets with high-quality human annotations (*e.g.*, ImageNet [8] and MS-COCO [28]).

However, it is expensive and time-consuming to obtain high-quality annotations for large-scale data in most real-world scenarios. To overcome this limitation, online key search engine [4, 26] or crowdsourcing [52] methods are proposed to efficiently and cheaply gain the desired training datasets with low-quality labels, in which noisy labels are likely to be introduced consequently. Although DNNs have high model capacities, they often overfit the noisy labels due to the memorization effect, resulting in poor classification and generalization performance [53]. Therefore, developing an effective method to improve the robustness of DNNs against noisy labels is of great practical importance.

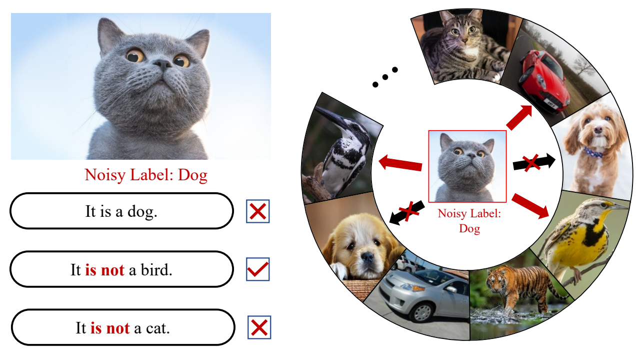

Figure 1. An illustration of negative information and pairwise negative correlation. Negative information (Left) provides a new complementary label to the training sample. If the assigned complementary label is wrong, it will attenuate the performance severely. Pairwise negative correlation (Right) randomly selects K negative samples that are not in the same category as the anchor image, *e.g.*, the anchor image has the negative correlation to the image labeled "car". Instead of using noisy positive correlation (black arrows), these pairwise negative correlations (red arrows) can be well captured in metric space to improve the robustness of DNNs.

Early robust learning methods primarily model noisy labels with the noise transition matrix [11, 15, 40] and use it refine losses. However, it is difficult to correctly estimate the noise transition matrix, as it heavily relies on either prior knowledge or a subset of high-quality labeled data. Considering the memorization effect to noisy labels, recent methods attempt to select high-confident samples as clean data and filter out others through a human-defined rule. For example, the small-loss trick widely-used in many

*C.D. is corresponding author.

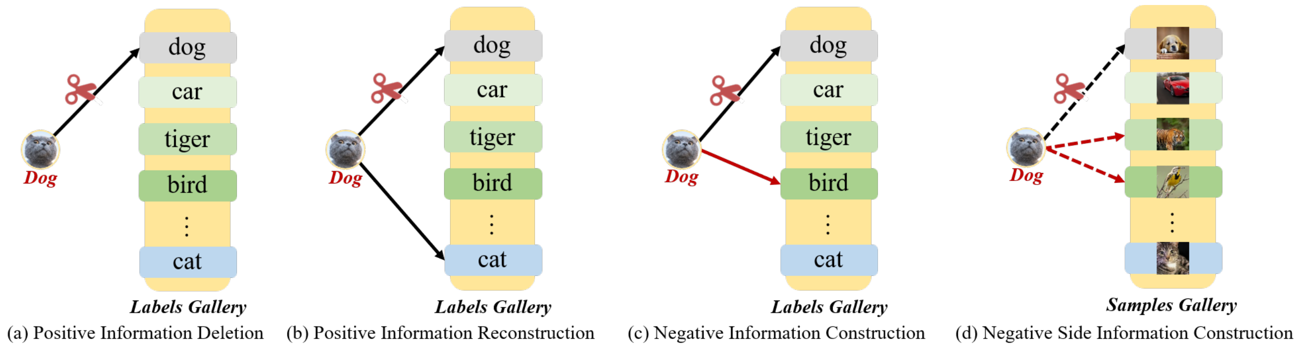

Figure 2. Conceptual illustration of different methods that leverage noisy data. The black solid line means "belong to", the red solid line means "is not", the black dotted line denotes a positive similarity relationship, the red dotted line denotes a negative similarity relationship, and the red scissors mean "filter out this relationship". (a) Sample selection methods adopt a sample selection strategy to find clean data and filter out noisy data. (b) Relabeling methods give new pseudo-labels for noisy data using the model's predictions. (c) Negative learning methods randomly assign a complementary label for each noisy sample. (d) LaCoL randomly selects K negative samples for each noisy sample.

methods, such as Co-teaching [12], Co-teaching+ [51] and JoCoR [43], selects a proportion of small-loss samples as clean ones. However, these methods cannot fully exploit the hidden information in the filtered samples, which degenerates the robustness of DNNs. To further take advantage of noisy training data, a series of methods, represented by semi-supervised learning based approaches (*e.g.*, DivideMix [22], and ELR+ [29]), relabel noisy samples using the model's predictions. Whereas the semi-supervised learning strategy increases computation cost, and relabeling noisy samples according to the model's predictions could cause confirmation bias, where the prediction error accumulates and harms performance.

Different from relabeling based methods, the recently proposed negative learning [18, 19] can effectively capture the underlying negative information of each noisy sample, which uses complementary labels to replace the original noisy labels and train DNNs by virtue of the learned negative information more effectively. For example, as shown in the left of Figure 1, the image of a cat is assigned to the wrong label "dog". Negative learning will randomly give it a new complementary label other than "dog", *e.g.*, "bird". Although the negative learning provides the "right" information (*e.g.*, the image of a cat shown in Figure 1 is not "bird") with a high probability in this manner, selecting a true label as a complementary label (*e.g.*, the image of a cat shown in Figure 1 is not "cat") is inevitable, which will severely degenerate the performance of the model. Meanwhile, this influence will be exacerbated due to the strong discrimination ability of cross-entropy (CE)-like loss used in these negative learning methods.

In this paper, we propose a simple yet effective method, named LaCoL (**La**tent **Co**ntrastive **L**earning), to improve the robustness and generalization of DNNs through excavating the implicit negative correlations in noisy data. Specifically, for each anchor sample, we randomly select K other samples, that are not in the same category with the anchor image, as negative samples, and use these negative pairs to construct negative correlations (as shown in Figure 2, compared with the existing approaches, pairwise negative correlation can make full use of noisy data and has higher confidence, which is beneficial for enhancing the robustness of DNNs). To better capture these negative correlations, we exploit the weakly-supervised contrastive learning method in the latent metric space, which is robust against wrongly assigned negative samples. Considering that incorporating different augmentation strategies during training can improve the generalization of models [9, 16, 35], we adopt weak (*e.g.*, using only crop-and-flip) and strong (*e.g.*, using RandAugment [7]) augmentations. Specifically, given the anchor sample with weak augmentation, its strong augmentation is the positive point and the negative points are derived by exploiting negative correlations in our method. Meanwhile, inspired by the alternate sample selection in Co-teaching [12], we use weak augmentations to select high-confident samples, and then apply strong augmentations to the back-propagation in label space, which can keep the divergence of sample selection procedure. Furthermore, we provide a cross-space similarity consistency regularization to narrow the gap between label space and metric space, which makes the learned negative correlation in metric space more powerful to improve the performance of the classification task in label space. The main contributions of this work are summarized as follows:

1) We propose a latent contrastive learning method exploiting the useful negative correlation hidden in noisy data, which can improve the robustness and generalization of traditional DNNs.

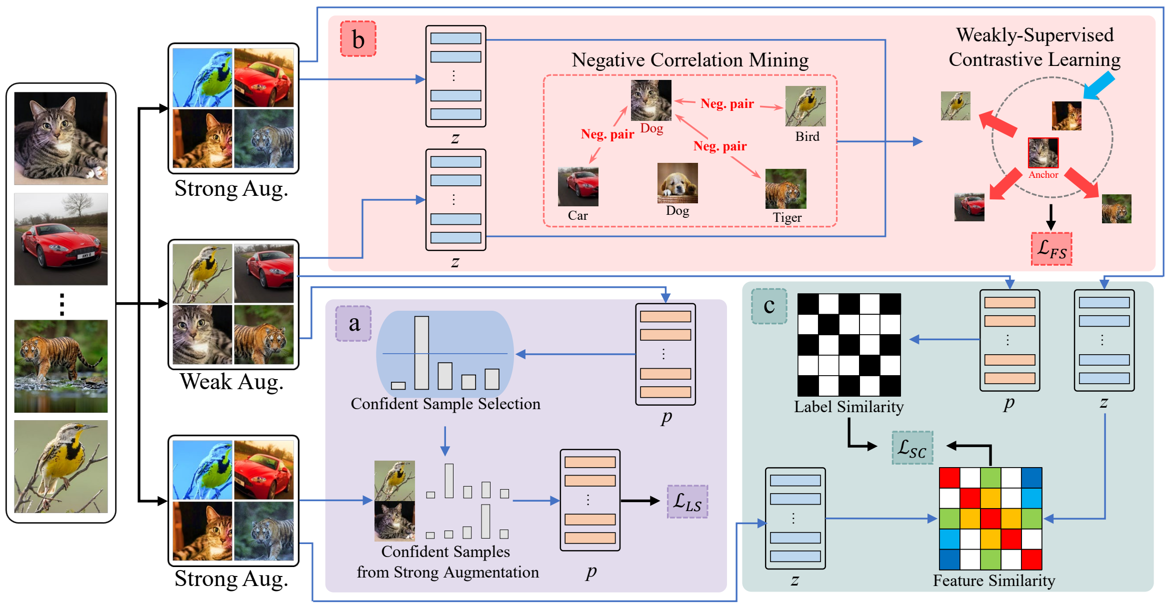

Figure 3. The framework of the proposed LaCoL. Given the training data with noisy labels, their weakly-augmented images are used to filter samples, and then the strong augmentations of the selected high-confident images are applied to optimize classification loss (Part a). Meanwhile, the weakly-augmented and strongly-augmented images are both projected in metric space to capture the implicit negative correlation guided by a weakly-supervised contrastive learning loss (Part b). Furthermore, the similarity matrices in label space and metric space are calculated to train both classification head and embedding head such that images with similar classification probabilities have similar embeddings (Part c).

2) To keep the divergence during sample selection in each iteration, we use weak augmentations to calculate confidence for selection, and then apply strong augmentations of high-confident data to train DNNs.

3) To make latent contrastive learning in metric space better guide classification tasks in label space, we present a cross-space similarity consistency regularization to constrain the gap between label space and metric space.

4) Extensive experiments demonstrate that LaCoL significantly outperforms state-of-the-art methods on both synthetic and real-world noisy datasets.

2. Related Work

Learning with Noisy Labels. Recently, learning with noisy data has been well studied and achieved great advances [1, 25, 38, 44, 45, 49]. Considering that the memorization effect of noisy labels in DNNs usually results in inferior model performance, existing state-of-the-art methods primarily adopt a sample selection strategy, which selects high-confident samples for subsequent training. For example, Co-teaching [12] trains two networks and feeds the small-loss samples of each network to its peer for parameter updating. The small-loss inputs have high confidence to be clean because that DNNs fit the underlying clean distribution before overfitting to noisy labels.

Straightly throwing away low-confident samples means that we ignore the underlying information implied in them, which makes DNNs insufficiently trained. To alleviate this phenomenon, some methods perform label correction using predictions from the network [29, 33]. The recent semi-supervised techniques such as MixUp exhibit good robustness to label noise. Inspired by this, MixMatch [2] leverages low-confident samples as unlabeled data in a semi-supervised learning paradigm. The recently proposed DivideMix [22] effectively combines label correction and sample selection with the MixUp data augmentation under a co-training framework. However, the usage of semi-supervised learning increases the computation cost, and the error of learned pseudo-label will be accumulated and degenerate the model performance [25].

Contrastive Learning. Self-supervised learning [3, 50, 54] has attracted much attention in unsupervised representation learning, due to its ability to directly leverage unlabeled data for model pre-training. Recently, contrastive learning and its variants [5, 6, 17, 37] develop rapidly and are widely adopted in many practical applications [24, 25, 48] to learn informative representations from unlabeled data. In contrastive learning, two different augmented images are randomly generated for each input image. Then the feature embeddings from the same source image are pulled together while the feature embeddings from different source images are pushed apart through the designed contrastive loss. For example, SimCLR [5] calculates the pairwise similarity of images from the same batch, whereas MoCo [6] maintains a queue of feature embeddings from the EMA model. Considering the existence of false-negative samples, some methods such as supervised contrastive learning [17] are proposed to select more informative negative samples.

3. Methodology

In this section, we will first illustrate the existing problem on learning with noisy labels. After that, we formulate the framework of the proposed LaCoL, which contains three parts: (a) classification task in label space, (b) weakly-supervised contrastive learning in metric space, and (c) cross-space similarity consistency regularization shown in Figure 3.

3.1. Overview

Label Noise Problem. Considering the influence of memorization effect on DNNs, most of the latest robust learning methods try to filter noisy samples and mine extra information from noisy data for training DNNs. As shown in Figure 2, a classical method represented by Co-teaching [12] just filters out low-confident samples according to the value of output probability. However, the filtering of noise data makes the model training insufficient. To address this issue, relabeling based methods give pseudo-labels to low-confident samples, while negative learning methods assign a new complementary label to each sample. Whereas both methods suffer from confirmation bias, *i.e.*, the error of derived implicit supervised information will accumulate and harm performance. Meanwhile, relabeling procedure increases computation cost. Therefore, it is critical to propose an efficient and effective method to enhance the robustness of models through mining more useful and reliable information hidden in noisy data.

To this end, we propose a new latent contrastive learning (LaCoL) framework for combating noisy labels. Different from most existing robust learning methods, LaCoL jointly learns the encoder $g(\cdot)$, the classification head $h(\cdot)$ shown in Figure 4, and the embedding head $f(\cdot)$ with two different data augmentations (*i.e.*, weak augmentation $\mathcal{A}_w(\cdot)$ and strong augmentation $\mathcal{A}_s(\cdot)$). As shown in Figure 3, given the noisy training data $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, where $\mathbf{y}_i \in \{0,1\}^C$ is the one-hot label vector corresponding to $\mathbf{x}_i$ over C classes, we perform one weak augmentation $\mathcal{A}_w(\mathbf{x}_i)$ using only crop-and-flip, and two strong augmentations $\mathcal{A}_s(\mathbf{x}_i)$ and $\mathcal{A}'_s(\mathbf{x}_i)$ using RandAugment [7] for each sample $\mathbf{x}_i$. And then LaCoL jointly optimizes three losses: (1) a supervised classification loss on strong augmentations of selected high-confidence data in label space $\mathcal{L}_{LS}$, (2)a latent contrastive learning loss that is weakly supervised by learned pairwise negative correlation in metric space $\mathcal{L}_{MS}$, and (3) a cross-space similarity consistency loss $\mathcal{L}_{SC}$.

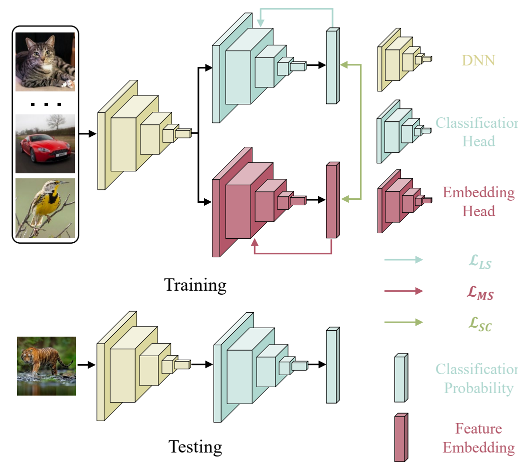

Figure 4. The pipeline of the proposed LaCoL. In the training procedure, we jointly learn the encoder $g(\cdot)$, the classification head $h(\cdot)$ and the embedding head $f(\cdot)$ optimized by three loss functions: (1) a supervised classification loss in label space $\mathcal{L}_{LS}$, a latent contrastive learning loss in metric space $\mathcal{L}_{MS}$, and (3) a cross-space similarity consistency loss $\mathcal{L}_{SC}$. For testing, only the encoder $g(\cdot)$ and the classification head $h(\cdot)$ with classification loss $\mathcal{L}_{LS}$ are used.

3.2. Latent Contrastive Learning

Pairwise Negative Correlation. For a sample with the noisy label, it has a non-true label and the remaining labels thus contain its real label. When negative learning assigns a complementary label from the remaining labels, its real label would be treated as the negative information. This error will be accumulated progressively and degenerate the performance. To alleviate this problem, we randomly select several negative samples rather than a single complementary sample shown in Figure 1. This sample-wise negative correlation is richer in diversity than class-wise negative information, which makes it robust against wrong assignment.

For each sample $\mathbf{x}_i$, we randomly construct a negative set with K samples as follows,

$$\mathcal{N}_i = \{\mathbf{x}_j\}_{j=1}^K, \forall \mathbf{y}_j \neq \mathbf{y}_i. \tag{1}$$

The derived negative pairs $\{(\mathbf{x}_i, \mathbf{x}_j)\}_{j=1}^K$ make up the sample-wise negative correlation, which is more diverse and informative to guide the classification task with noisy data.

Weakly-supervised Contrastive Learning. After achieving the pairwise negative correlation, we need to select a suitable objective function to capture it. Most existing methods still process auxiliary information extracted from noisy data constrained by classification loss (*e.g.*, Cross-

Entropy loss) in label space. Due to the strong discrimination ability of classification loss, it is sensitive to the error of predicted pseudo-label or assigned complementary label. Meanwhile, class-wise classification loss cannot well apply sample-wise negative correlation. To better capture negative correlation in a robust manner, we propose latent contrastive learning (LaCoL) that is weakly supervised by learned negative correlation in metric space. Considering that the mined negative correlation only contains the negative similarity relationship, we introduce the different augmentation of the anchor sample as the self-supervised positive similarity relationship.

We conduct latent contrastive learning for all training data. That being said, our method is weakly supervised by the negative correlations, thus less wrong negative samples will be involved during training, which can improve the robustness of models. Specially, for clean data, our method can be regarded as self-supervised contrastive learning without the wrong negative samples, which is informative to guide the classification task.

In LaCoL, we project weakly-augmented input $\mathcal{A}_w(\mathbf{x}_i)$ and strongly-augmented $\mathcal{A}_s(\mathbf{x}_i)$ into metric space, and derive the feature embedding $\tilde{\mathbf{z}}_i = f \circ h\left(\mathcal{A}_w(\mathbf{x}_i)\right)$ and $\hat{\mathbf{z}}_i = f \circ h\left(\mathcal{A}_s(\mathbf{x}_i)\right)$, respectively. Within the self-supervised positive similarity relationship and weakly-supervised negative similarity relationship, the latent contrastive learning loss in metric space is defined as:

$$\mathcal{L}_{MS} = \sum_{i=1}^{N} -\log \frac{\exp\left(\langle \tilde{\mathbf{z}}_i, \hat{\mathbf{z}}_i \rangle / \tau\right)}{\sum_{j \in \mathcal{N}_i} \exp\left(\left(\langle \tilde{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle + \langle \tilde{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle\right) / \tau\right)}, \quad (2)$$

where $< \cdot, \cdot >$ denotes the cosine similarity, and τ is the temperature parameter.

3.3. Divergence Preserving

As mentioned above, we adopt two different augmentation strategies for feature embedding in metric space. And then, we further take full advantage of weak augmentation and strong augmentation, and extend these augmentation strategies to train classification head, since incorporating different augmentation strategies during training can improve generalization and robustness. To preserve the diversity during training, some methods represented by Coteaching [12] train two networks alternately, *i.e.* one select high-confident samples with smaller loss, the other optimize the loss function using selected samples. However, training two networks simultaneously increase computation cost. Therefore, we utilize two different augmentation policies mentioned above to achieve the diversity preserving with the single network.

We first use weakly-augmented data $\mathcal{W} = \{\mathcal{A}_w(\mathbf{x}_i)\}_{i=1}^{N}$ to select high-confident samples which can facilitate the model training. Following [1], we treat samples whose predictions are consistent with given labels as high-confident samples. The high-confident sample set $\bar{\mathcal{D}}$ can be derived as follows:

$$\begin{aligned} &\bar{\mathcal{D}} = \left\{(\mathbf{x}_i, \mathbf{y}_i) | \mathbf{y}_i = \bar{\mathbf{y}}_i, i = 1, \cdots, N\right\}, \\ &\bar{\mathbf{y}}_i = \text{Sharpen}\left(h \circ g\left(\mathcal{A}_w(\mathbf{x}_i)\right)\right), \end{aligned} \quad (3)$$

where $\tilde{\mathbf{p}}_i = h \circ g\left(\mathcal{A}_w(\mathbf{x}_i)\right)$ denotes the classification probability in label space, and Sharpen operation sets the maximum term 1 and the others 0. After obtaining the high-confident data set, we adopt the strong augmentation of high-confident samples $\bar{\mathcal{S}} = \{\mathcal{A}_s(\mathbf{x}_i) | (\mathbf{x}_i, \mathbf{y}_i) \in \bar{\mathcal{D}}\}$ for training in label space with a weighted classification loss:

$$\begin{aligned} \mathcal{L}_{LS} &= \sum_{\mathcal{A}_s(\mathbf{x}_i) \in \bar{\mathcal{S}}} -\mu_k \mathbf{y}_i^{\top} \log\left(\hat{\mathbf{p}}_i\right), \\ \mu_k &= \frac{\epsilon_k}{\sum_{j=1}^{C} \epsilon_i}, \end{aligned} \quad (4)$$

where $\hat{\mathbf{p}}_i = h \circ g\left(\mathcal{A}_s(\mathbf{x}_i)\right)$, ϵ_k is the number of high-confident samples belonging to the k-th class, and k denotes that the sample $\mathbf{x}_i$ belongs to the k-th class.

3.4. Cross-Space Similarity Consistency

During training, the classification loss and latent contrastive loss are optimized in different spaces, which leads to a semantic gap between the learned feature embedding in metric space and the derived classification probability in label space. To make the latent contrastive learning better guide the classification task, we propose a cross-space similarity consistency regularization since it can guarantee the classification probabilities and feature embeddings to guide each other.

The representations in metric space should have the same similarity relationship as the classification results in label space. To ensure this cross-space similarity consistency, we minimize the cross-entropy between the similarity matrices in label space and metric space.

Given the weakly-augmented data $\{\mathcal{A}_w(\mathbf{x}_i)\}_{i=1}^{N}$ and their output probability $\{\tilde{\mathbf{p}}_i\}_{i=1}^{N}$, the similarity matrix in label space can be constructed as follows:

$$w_{ij}^{l} = \begin{cases} 1 & \text{if } i = j \\ \langle \tilde{\mathbf{p}}_i, \tilde{\mathbf{p}}_j \rangle & \text{if } i \neq j \text{ and } \langle \tilde{\mathbf{p}}_i, \tilde{\mathbf{p}}_i \rangle \geq \rho \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where ρ is a threshold.

To obtain the similarity matrix in metric space, we conduct two strong augmentations $\{\mathcal{A}_s(\mathbf{x}_i)\}_{i=1}^{N}$ and $\{\mathcal{A}'_s(\mathbf{x}_i)\}_{i=1}^{N}$. Their feature embedding can be represented as $\{\hat{\mathbf{z}}_i\}_{i=1}^{N}$ and $\{\hat{\mathbf{z}}'_i\}_{i=1}^{N}$ in label space. Then we build the similarity matrix in label space as follows:

$$w_{ij}^{m} = \begin{cases} \exp(\langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}'_i \rangle / \tau) & \text{if } i = j \\ \exp(\langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle / \tau) & \text{if } i \neq j \end{cases} \quad (6)$$

Algorithm 1 Latent Constrastive Learning Algorithm

Input:

Training dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ with label noise;

Hyper-parameter τ, ρ, λ_{MS}, λ_{SC}, and K;

The number of epochs E.

Pre-train and initialize the model parameters θ.
for $epoch = 1, 2, \cdots, E$ **do**
  Conduct weak and strong augmentations $\{\mathcal{A}_w(\mathbf{x}_i), \mathcal{A}_s(\mathbf{x}_i), \mathcal{A}'_s(\mathbf{x}_i)\}$ for each sample following the description in subsection 3.2;
  Filter the training data according to the predictions of weak augmentation $\{\mathcal{A}_w(\mathbf{x}_i)\}_{i=1}^N$, and derive the strongly-augmented data with high confidence $\bar{\mathcal{S}}$;
  Calculate the classification loss Eq.(4) in label space using the obtained $\bar{\mathcal{S}}$;
  Randomly select K negative samples for each input sample to construct the pairwise negative correlations acccording to Eq.(1);
  Calculate the latent contrastive learning loss Eq.(2) weakly-supervised by learned negative correlations;
  Construct similarity matrices in both label space and metric space described in Eqs.(5) and (6);
  Calculate the cross-space similarity consistency loss Eq.(7) with two similarity matrices;
  Optimize the parameters θ using the joint loss Eq.(8);
end for
return θ.

Based on two similarity matrices, the cross-space similarity consistency loss is defined as:

$$\mathcal{L}_{SC} = \frac{1}{N}\sum_{i=1}^{N} \ell_{ce}(\check{\mathbf{w}}_i^l, \check{\mathbf{w}}_i^m), \tag{7}$$

where $\ell_{ce}(\cdot,\cdot)$ denotes the cross-entropy loss function, and $\check{\mathbf{w}}_i^l$ and $\check{\mathbf{w}}_i^m$ are the normalized similarity vectors between i-th sample and other samples in label space and metric space, respectively. Due to the similarity consistency regularization, the negative information captured in metric space can better boost the classification task in label space.

To this end, our overall training objective function is:

$$\mathcal{L} = \mathcal{L}_{LS} + \lambda_{MS}\mathcal{L}_{MS} + \lambda_{SC}\mathcal{L}_{SC}, \tag{8}$$

where λ_{MS} and λ_{SC} are two scalar hyper-parameters.

Note that the proposed LaCoL can capture more reliable and discriminative information (*i.e.*, pairwise negative correlation) from noisy data by applying weakly-supervised contrastive learning in metric space. Guided by informative feature embedding in metric space, DNNs can also achieve appealing performance with label noise.

4. Experiments

4.1. Experimental Settings

Datasets. To evaluate the performance of the proposed LaCoL, we conduct the experiments on two benchmarks CIFAR-10 and CIFAR-100 [20] with different levels of symmetric, asymmetric, and instance-dependent label noise (abbreviated as instance label noise), and a large-scale real-world dataset Clothing1M [46]. CIFAR-10 and CIFAR-100 are both composed of 50k training images and 10k test images of size 32×32. Following previous works [12, 25, 29, 45], symmetric noise is generated by uniformly flipping labels for a percentage of the training dataset to all possible labels. Asymmetric noise is class-dependant, where labels are only changed to similar classes. And, instance noise is generated by image features. More details about the synthetic label noise are given in the *supplementary material*. Clothing1M consists of 1 million training images collected from online shopping websites with noisy labels generated from surrounding texts. Its noise level is estimated at 38.5%, and some pairs of classes are often confused with each other (*e.g.*, Hoodie and Windbreaker).

Baselines. To evaluate the performance on CIFAR-10 and CIFAR-100, we compare our method against standard CE, along with recent state-of-the-art approaches including Co-teaching [12], Co-teaching+ [51], JoCoR [43], CDR [45], APL [32] and JNPL [19]. To perform evaluation on Clothing-1M, besides the above methods, other state-of-the-art methods like Joint-Optim [41], MLNT [23], DMI [47], and PENCIL [50] are also compared.

Evaluation Metrics. Following the standard protocol [12, 22], we use the test accuracy, *i.e.*, test accuracy = (# of correct predictions) / (# of test dataset) to measure the performance. Higher test accuracy implies that the method is more robust to the label noise.

Implementation details. We implement our method in PyTorch [39]. Same as the previous works [18, 19], we use ResNet-34 [14] for CIFAR-10 and CIFAR-100 [20]. We adopt SGD with 0.9 momentum as the optimizer and train the network for 200 epochs. The initial learning rate is set as 0.1 and decayed with a factor of 10 at the 100th and 150th epoch respectively, and weight decay set $1e-4$. For Clothing-1M [46], we follow the setting of [41] with ResNet-50 [14] pre-trained on IamgeNet [21]. We train the network for 6 epochs and use SGD with 0.9 as the optimizer with a weight decay of $1e-3$. The initial learning rate is $5e-3$ and is decayed by a factor of 10 at the 3rd and 4th epoch, respectively. For the hyper-parameters, we fix $\tau = 0.2$, $\rho = 0.8$, $\lambda_{MS} = 1$ and $\lambda_{SC} = 0.5$.

4.2. Experimental Results

Comparison on synthetic datasets. We first evaluate the performance of our proposed method on two synthetic

Datasets	Model	Methods	Symmetric			Asymmetric	
			20%	40%	60%	20%	40%
CIFAR-10	ResNet-18	Standard (Cross-Entropy loss)	78.37	51.15	32.51	79.31	50.55
	ResNet-18	Co-teaching [12]	91.01	87.36	84.22	92.58	72.10
	ResNet-18	Co-teaching+ [51]	91.66	88.08	82.18	90.47	70.58
	ResNet-34	JoCoR [43]	91.84	88.15	59.20	91.19	83.61
	ResNet-34	NFL+RCE [32]	90.50	85.16	70.77	89.66	78.30
	ResNet-34	NCE+RCE [32]	90.36	84.57	74.09	90.13	78.48
	ResNet-34	JNPL [19]	93.53	91.89	88.45	93.45	90.72
	ResNet-34	Ours	**94.12**	**92.33**	**88.72**	**93.76**	**91.07**
CIFAR-100	ResNet34	Standard (Cross-Entropy loss)	57.32	45.64	24.30	62.12	44.55
		Co-teaching [12]	69.56	62.81	51.12	67.46	52.86
		Co-teaching+ [51]	68.18	60.15	48.97	70.13	52.40
		JoCoR [43]	71.75	63.96	37.84	65.05	45.14
		NFL+RCE [32]	58.70	42.76	24.77	56.45	37.52
		NCE+RCE [32]	57.41	43.75	25.87	56.84	36.40
		JNPL [19]	70.94	68.11	61.26	69.95	59.51
		Ours	**71.24**	**68.59**	**61.93**	**70.84**	**59.73**

Table 1. Comparison with state-of-the-art methods on CIFAR-10 and CIFAR-100 with symmetric and asymmetric label noise from different levels. We show the test accuracy (%). Bold indicates best performance.

Dataset	CIFAR-10		CIFAR-100	
Methods/Noise	Instance - 20%	Instance - 40%	Instance - 20%	Instance - 40%
Standard (Cross-Entropy loss)	85.10	77.00	52.19	42.26
Co-teaching [12]	86.54	80.98	57.24	45.69
Joint-Optim [41]	89.69	82.62	65.15	55.57
DMI [47]	89.14	84.78	58.05	47.36
CDR [45]	90.41	83.07	67.33	55.94
Ours	**92.47**	**86.76**	**70.16**	**63.87**

Table 2. Comparison with state-of-the-art methods on CIFAR-10 and CIFAR-100 with instance-dependent label noise from different levels. We show the test accuracy (%). Bold indicates best performance.

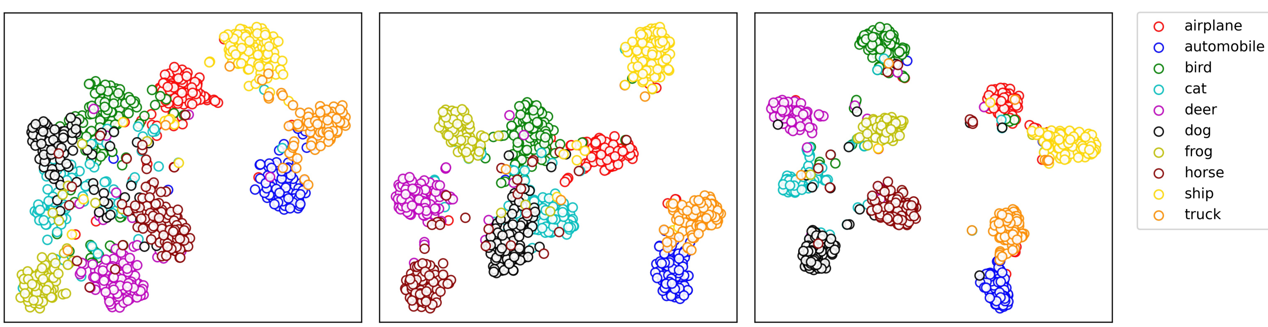

Figure 5. t-SNE visualization results on CIFAR-10 with 40% symmetric label noise. (a) Left: Baseline (Standard CE loss + normal sample selection); (b) Middle: Baseline + Diversity Preserving Strategy; (c) Right: Our LaCoL. It is clear that the learned representations in middle and right images are more discriminative than the left image.

datasets CIFAR-10 and CIFAR-100 [20] with different levels of symmetric, asymmetric, and instance label noise. Results are presented in Tables 1 and 2, which show that our proposed method can consistently outperform all other

Method	Model	Test Accuracy (%)
Standard (CE loss)		69.21
Joint-Optim [41]		72.16
MLNT [23]		73.47
DMI [47]	ResNet-50	72.46
PENCIL [50]		73.49
JNPL [19]		74.15
Ours		**74.68**

Table 3. Compassion with state-of-the-art methods on Clothing-1M. Results of baseline methods are taken from the original papers. Bold indicates best performance.

baselines in all cases. These empirical results support our proposal that the proposed LaCoL can effectively extract informative and robust representations to guide the classification task, which helps improve the robustness and generalization of DNNs training with label noise.

Comparison on real-world dataset. To verify the effectiveness of the proposed method, we also perform experiments on a real-world dataset Clothing-1M [46] with compared methods such as CE, Joint-Optim [41], MLNT [23], DMI [47], PENCIL [50] and JNPL [19]. The overall results are reported in Table 3, from which we can easily observe that the proposed LaCoL can outperform all baselines. This demonstrates that, through applying latent contrastive learning that is weakly supervised by pairwise negative correlation, our method is more effective to handle such real-world noise problems.

	Symmetric	Asymmetric
w/o diversity preserving	67.24	58.22
w/o $\mathcal{L}_{MS}$	59.12	47.71
w/o $\mathcal{L}_{SC}$	66.91	56.49
Ours	**68.59**	**59.73**

Table 4. Effect of the proposed components. We show the test accuracy (%) on CIFAR-100 with 40% label noise.

4.3. Ablation Study

Performance contributions of different components in the proposed method. In Table 4, we study the effect of three components from the proposed methods including the diversity preserving strategy, the latent contrastive loss $\mathcal{L}_{MS}$ and the cross-space similarity consistency loss $\mathcal{L}_{SC}$. The results show that all the components improve the model's performance, especially that $\mathcal{L}_{MS}$ is most crucial to the model's performance. We also show the t-SNE [42] visualization of the feature embeddings on CIFAR-10 with 40% symmetric label noise in Figure 5. It is clear that the learned representations in middle and right images are more discriminative than the left image, which demonstrates that the components of the proposed method can all improve the performance of DNNs training with label noise.

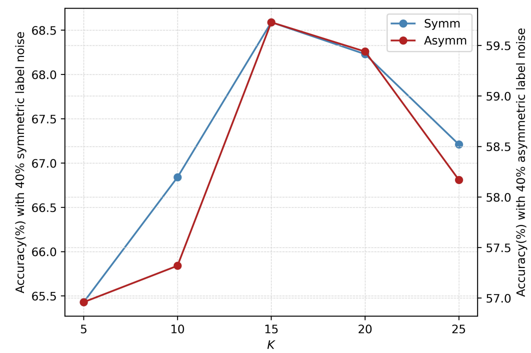

Figure 6. Classification performance on CIFAR-100 with 40% symmetric and asymmetric label noise in comparison with different number of negative samples K.

Impact of the different number of negative samples. During training, we randomly assign K negative samples for each training sample. The number of negative samples K is a critical parameter for the proposed method. We analyze the impact of different values of K to the model's performance, and the corresponding results are shown in Figure 6. The value of K ranges from 5, 10, 15, 20 to 25. We can see that the best value of K is 15.

5. Conclusion

In this paper, we propose a new latent contrastive learning (LaCoL) method for learning with noisy labels. We excavate the underlying negative correlation in noisy data and capture it with a weakly-supervised contrastive learning loss in metric space. Meanwhile, we exploit weakly-augmented data to select samples and optimized classification loss on strong augmentations of the selected sample set. Furthermore, we provide a cross-space similarity consistency regularization to make the learned feature embedding more informative to guide the classification task. Extensive experiments show that our method achieves the state-of-the-art performance on multiple noisy datasets.

6. Acknowledgment

Our work was supported in part by the National Key R&D Program of China under Grant 2017YFE0104100; in part by the National Natural Science Foundation of China under Grant 62132016, Grant 62171343, Grant 62071361, and Grant 62102293; in part by Key Research and Development Program of Shaanxi under Grant 2021ZDLGY01-03; and in part by the Fundamental Research Funds for the Central Universities ZDRC2102.

References

[1] Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *arXiv preprint arXiv:2106.15853*, 2021. 3, 5

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 3

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149, 2018. 3

[4] Youngchul Cha and Junghoo Cho. Social-network analysis using topic models. In *SIGIR*, pages 565–574, 2012. 1

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 3, 4

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3, 4

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, pages 702–703, 2020. 2, 4

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1

[9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2

[10] Ross Girshick. Fast r-cnn. In *CVPR*, pages 1440–1448, 2015. 1

[11] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 2017. 1

[12] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018. 2, 3, 4, 5, 6, 7

[13] Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *CVPR*, pages 5927–5935, 2017. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 6

[15] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *NIPS*, 31:10456–10465, 2018. 1

[16] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 2

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 33, 2020. 3, 4

[18] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *ICCV*, pages 101–110, 2019. 2, 6

[19] Youngdong Kim, Juseung Yun, Hyounguk Shon, and Junmo Kim. Joint negative and positive learning for noisy labels. In *CVPR*, pages 9442–9451, 2021. 2, 6, 7, 8

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6, 7

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 6

[22] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020. 2, 3, 6

[23] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5051–5059, 2019. 6, 8

[24] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9475–9484, 2021. 3

[25] Junnan Li, Caiming Xiong, and Steven CH Hoi. Learning from noisy data with robust representation learning. In *ICCV*, pages 9485–9494, 2021. 3, 6

[26] 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. 1

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 1

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[29] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *NeurIPS*, 33, 2020. 2, 3, 6

[30] Yang Liu, Qingchao Chen, and Samuel Albanie. Adaptive cross-modal prototypes for cross-domain visual-language retrieval. In *CVPR*, pages 14954–14964, 2021. 1

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1

[32] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *ICML*, pages 6543–6553. PMLR, 2020. 6, 7

[33] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *ICML*, pages 3355–3364. PMLR, 2018. 3

[34] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*, pages 9826–9836, 2021. 1

[35] Kento Nishi, Yi Ding, Alex Rich, and Tobias Hollerer. Augmentation strategies for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8022–8031, 2021. 2

[36] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *CVPR*, pages 1520–1528, 2015. 1

[37] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[38] Diego Ortego, Eric Arazo, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6606–6615, 2021. 3

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32:8026–8037, 2019. 6

[40] 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 *CVPR*, pages 1944–1952, 2017. 1

[41] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, pages 5552–5560, 2018. 6, 7, 8

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[43] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, pages 13726–13735, 2020. 2, 6, 7

[44] Songhua Wu, Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Nannan Wang, Haifeng Liu, and Gang Niu. Class2simi: A noise reduction perspective on learning with noisy labels. In *International Conference on Machine Learning*, pages 11285–11295. PMLR, 2021. 3

[45] 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 *ICLR*, 2020. 3, 6, 7

[46] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pages 2691–2699, 2015. 6, 8

[47] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*, pages 6222–6233, 2019. 6, 7, 8

[48] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *CVPR*, pages 11562–11572, 2021. 1, 3

[49] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-src: A contrastive approach for combating noisy labels. In *CVPR*, pages 5192–5201, 2021. 3

[50] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *CVPR*, pages 7017–7025, 2019. 3, 6, 8

[51] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, pages 7164–7173. PMLR, 2019. 2, 6, 7

[52] Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *ECCV*, pages 68–83, 2018. 1

[53] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 1

[54] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by autoencoding transformations rather than data. In *CVPR*, pages 2547–2555, 2019. 3