Appendix

A. Experiment Details

A.1. Implementation Details

For CIFAR datasets, we use the model PreAct Resnet18 [19]. For ANIMAL-10N, we use a random initialized model Resnet18 [19]. For Clothing1M, we use an ImageNet pre-trained model Resnet18 [19]. We illustrate our framework in Figure 5. The projection MLP is 3-layer MLP and the prediction MLP is 2-layer MLP as proposed in Sim-siam [9]. We use weak augmentations $\mathcal{A}_w : \mathcal{X} \to \mathcal{X}$ including random resized crop and random horizontal flip for optimizing the cross entropy loss $\mathcal{L}_{\text{ce}}$. Following Sim-Siam [9] [7], we use a strong augmentation $\mathcal{A}_s : \mathcal{X} \to \mathcal{X}$ applied on images twice for optimizing the contrastive regularization term $\widetilde{\mathcal{L}}_{\text{ctr}}$. Specifically, $\{z_i\} = f\big(\mathcal{A}_s(\{x_i\})\big)$ and $\{q_i\} = h\big(f\big(\mathcal{A}_s(\{x_i\})\big)\big)$ for every example x_i, where one strong augmented image is for calculating z and another is for calculating q.

A.2. Algorithm

According to our gradient analysis on two different clean images x_i, x_j with $y_i = y_j$ and a noisy image x_m with $y_m = y_i$, apply the regularization function Eq. (8) can avoid representation learning dominated by the wrong contrastive pair (x_i, x_m). The analysis does not cover the same image with two different augmentations. When applying the strong augmentation twice, each image x has two different augmentations x', x''. The contrastive pair (x', x'') will also dominate the representation learning given the property of Eq. (8). However, focusing on learning similar representations of (x', x'') does not help to form a cluster structure in representation space. As mentioned in [41], learning this self-supervised representations causes representations of data distributed uniformly on the unit hypersphere. Hence, we want the gradient from the pair (x', x'') to be smaller when their representations approach to each other. We use the original contrastive regularization to regularize the pair (x', x''). The pseudocode of the proposed method is given in Algorithm 1.

A.3. Hyperparameters

CIFAR. Our method has two hyperparameters λ and τ. For each noise setting for CIFAR-10, we select the best hyperparameters: λ from $\{50, 130\}$ and τ from $\{0.4, 0.8\}$. For each noise setting for CIFAR-100, we select the best hyperparameters: λ from $\{50, 90\}$ and τ from $\{0.05, 0.7\}$. The batch size is set as 256, and the learning rate is 0.02 using SGD with a momentum of 0.9 and a weight dacay of 0.0005.

ANIMAL-10N & Clothing1M. For ANIMAL-10N, we set $\lambda = 50$, $\tau = 0.8$ and batch size is 256. The learning rate is set as 0.04 with the same SGD optimizer as the CIFAR experiment. For Clothing1M, we set $\lambda = 90$, $\tau = 0.4$ and batch size is 256. The learning rate is set as 0.06 with the same SGD optimizer as above.

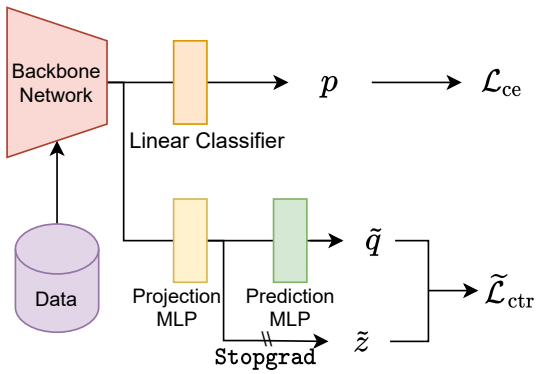

Figure 5. Illustration of our framework.

B. Proofs of Theoretical Results

B.1. Proof for Theorem 1

Theorem. *Representations Z learned by minimizing Eq. (1) maximizes the mutual information $I(Z; X^+)$.*

Proof. We first decompose the mutual information $I(Z; X^+)$:

$$\begin{aligned} I(Z; X^+) =& \mathbb{E}_{Z,X^+} \log \frac{p(Z|X^+)}{p(Z)} \\ =& \mathbb{E}_{X^+}\mathbb{E}_{Z|X^+}[\log p(Z|X^+)] - \mathbb{E}_{Z,X^+}[p(Z)] \\ =& -\mathbb{E}_{X^+}\big[H(Z|X^+)\big] + H(Z). \end{aligned}$$

The first term $\mathbb{E}_{X^+}\big[H(Z|X^+)\big]$ measures the uncertainty of $Z|X^+$, which is minimized when Z can be completely determined by X^+. The second term $H(Z)$ measures the uncertainty of Z itself and it is minimized when outcomes of Z are equally likely.

We next show that Z can be completely determined by X^+ when minimum of Eq. (1) is achieved and uncertainty of Z itself is maintained by an assumption about the framework. By the Cauchy-Schwarz inequality,

$$\begin{aligned} \mathbb{E}_{X,X^+}\big[\mathcal{L}_{\text{ctr}}(X, X^+)\big] \geq& \mathbb{E}_{X,X^+}\big[\|\tilde{q}\|_2 \big\|\tilde{z}^+\big\|_2 \\ &+ \big\|\tilde{q}^+\big\|_2 \|\tilde{z}\|_2\big] = -2. \end{aligned}$$

The equality is attained when $\tilde{q} = \tilde{z}^+$ and $\tilde{q}^+ = \tilde{z}$ for all x, x^+ from the same class. For any three images x_1, x_2, x_3 from the same class, we have:

$$f(x_1) = g(x_3), \quad f(x_2) = g(x_3),$$

where $g = h(f(\cdot))$. We can find $f(x_1) = f(x_2)$ for any images x_1, x_2 from the same class. The result can

Algorithm 1: CTRR Pseudocode in a PyTorch-like style

```
# Training
# f: backbone + projection mlp
# h: prediction mlp
# g: backbone + softmax linear classifier

for x, y in loader:
   bsz = x.size(0)
   x1, x2 = strong_aug(x), strong_aug(x) # strong random augmentation
   x3 = weak_aug(x) # weak random augmentation
   z1, z2 = f(x1), f(x2)
   q1, q2 = h(z1), h(z2)
   p = g(x3)

   # compute representations
   c1 = torch.matmul(q1, z2.t()) # B X B
   c2 = torch.matmul(q2, z1.t()) # B X B

   # compute contrastive loss for each pair
   m1 = torch.zeros(bsz, bsz).fill_diagonal_(1) # identity matrix
   m2 = torch.ones(bsz, bsz).fill_diagonal_(0) # 1-identity matrix
   # - <i,i> + log(1-<i,j>)
   c1 = -c1*m1 + ((1-c1).log()) * m2
   c2 = -c2*m1 + ((1-c2).log()) * m2
   c = torch.cat([c1, c2], dim=0) # 2B X B

   # compute probability threshold
   probs_thred = torch.matmul(p, p.t()).fill_diagonal_(1).detach() # B X B
   mask = (probs_thred >= tau).float()
   probs_thred = probs_thred * mask
   # normalize the threshold
   weight = probs_thred / probs_thred.sum(1, keepdim=True)
   weight = weight.repeat((2, 1)) # 2B X B

   loss_ctr = (contrast_logits * weight).sum(dim=1).mean(0)
```

be easily extended to the general case: $f(X_1) = f(X_2)$ for any $(X_1, Y_1) \sim P(X, Y), (X_2, Y_2) \sim P(X, Y)$ with $Y_1 = Y_2$. Thus Z can be determined by X^+ with the equation $Z = f(X^+)$, which minimizes $\mathbb{E}_{X^+}\big[H(Z|X^+)\big]$.

When $p(Z = c_y|Y = y) = \frac{1}{|\mathcal{Y}|}$, the entropy $H(Z)$ is maximized. With extensive empirical results in Simsiam [9], we assume the collapsed solutions are perfectly avoided by using the SimSiam framework. By this assumption, $c_j \neq c_k$ for any $j \neq k$. The model learns different clusters c_y for different y and representations with different labels have different clusters. Therefore, for a balanced dataset, the outcomes of Z are equally likely and it maximizes the second term $H(Z)$. In summary, the learned representations by Eq. (1) maximizes the mutual information $I(Z; X^+)$. □

B.2. Proof for Theorem 2

Theorem. *Given a distribution $D(X, Y, \tilde{Y})$ that is (ϵ, γ)-Distribution, we have*

$$I(X;Y) - \epsilon \leq I(Z^\star;Y) \leq I(X;Y), \tag{11}$$

$$I(Z^\star;\tilde{Y}) \leq I(X;\tilde{Y}) - \gamma + \epsilon. \tag{12}$$

Proof. The Theorem builds upon the Theorem 5 from [39]. We first provide the proof for the first inequality, which can also be obtained from [39]. Then we provide the proof for the second inequality.

For the first inequality, by adopting Data Processing Inequality in the Markov Chain $Y \leftrightarrow X \rightarrow Z$, we have $I(X;Y) \geq I(Z;Y)$ for any $Z \in \mathcal{Z}$. Then, we have $I(X;Y) \geq I(Z^\star;Y)$. Since $Z^\star = \arg\max_{Z_\theta} I(Z_\theta; X^+)$, and $I(Z_\theta; X^+)$ is maximized at $I(X; X^+)$, then $I(Z^\star; X^+) = I(X; X^+)$ and $I(Z^\star; X^+|Y) = I(X; X^+|Y)$. Meanwhile, use the result $I(Z^\star; X^+; Y) = I(X; X^+; Y)$, which is given by

$$\begin{aligned} I(Z^\star; X^+; Y) &= I(Z^\star; X^+) - I(Z^\star; X^+|Y) \\ &= I(X; X^+) - I(X; X^+|Y) \\ &= I(X; X^+; Y), \end{aligned}$$

we have

$$\begin{aligned} I(Z^\star;Y) =& I(X;X^+;Y) + I(Z^\star;Y|X^+) \\ =& I(X;Y) - I(X;Y|X^+) + I(Z^\star;Y|X^+). \end{aligned} \tag{13}$$

Thus, by Eq. (13) and the Definition 1, we get

$$I(Z^\star;Y) \geq I(X;Y) - I(X;Y|X^+) \geq I(X;Y) - \epsilon \tag{14}$$

Now we present the second inequality $I(Z^\star;\tilde{Y}) \leq I(X;\tilde{Y}) - \gamma + \epsilon$.

Similarly, by Eq. (13), we have

$$I(Z^\star;\tilde{Y}) = I(X;\tilde{Y}) - I(X;\tilde{Y}|X^+) + I(Z^\star;\tilde{Y}|X^+) \tag{15}$$
$$\leq I(X;\tilde{Y}) - \gamma + I(Z^\star;\tilde{Y}|X^+) \tag{16}$$
$$\leq I(X;\tilde{Y}) - \gamma + I(Z^\star;Y|X^+) \tag{17}$$
$$\leq I(X;\tilde{Y}) - \gamma + \epsilon \tag{18}$$

, where the first and the third inequalities are by the definition 1; the second inequality is by the Data Processing Inequality in the Markov Chain $\tilde{Y} \leftarrow Y \leftrightarrow X \rightarrow Z$.

□

B.3. Proof for Lemma 1

Lemma. *Consider a pair of random variables $(X, \tilde{Y})$. Let $\hat{Y}$ be outputs of any classifier based on inputs Z_θ, and $\tilde{e} = \mathbb{1}\{\hat{Y} \neq \tilde{Y}\}$, where $\mathbb{1}\{A\}$ be the indicator function of event A. Then, we have*

$$\mathbb{E}[\tilde{e}] \geq \frac{H(\tilde{Y}) - I(Z_\theta;\tilde{Y}) - H(\tilde{e})}{\log\left(|\tilde{\mathcal{Y}}|\right) - 1}.$$

Proof. If we are given any two of $\{\tilde{e}=1\}, \hat{Y}, \tilde{Y}$, the other one is known. By the properties of conditional entropy, $H(\tilde{Y}, \tilde{e}|\hat{Y}, Z_\theta)$ can be decomposed into the two equivalent forms.

$$\begin{aligned} H(\tilde{Y}, \tilde{e}|\hat{Y}, Z_\theta) &= H(\tilde{Y}|\tilde{e}, \hat{Y}, Z_\theta) + H(\tilde{e}|\hat{Y}, Z_\theta) \\ &= \underbrace{H(\tilde{Y}|\tilde{e}, \hat{Y}, Z_\theta)}_{0} + H(\tilde{Y}|\hat{Y}, Z_\theta) \end{aligned} \tag{19}$$

The first equality can also be decomposed into another form:

$$\begin{aligned} & H(\tilde{Y}, \tilde{e}|\hat{Y}, Z_\theta) \\ =& H(\tilde{Y}|\tilde{e}, \hat{Y}, Z_\theta) + H(\tilde{e}|\hat{Y}, Z_\theta) \\ =& p(\tilde{e}=1) H(\tilde{Y}|\tilde{e}=1, \hat{Y}, Z_\theta) \\ & + p(\tilde{e}=0) \underbrace{H(\tilde{Y}|\tilde{e}=0, \hat{Y}, Z_\theta)}_{0} + H(\tilde{e}|\hat{Y}, Z_\theta) \\ =& p(\tilde{e}=1) H(\tilde{Y}|\tilde{e}=1, \hat{Y}, Z_\theta) + H(\tilde{e}|\hat{Y}, Z_\theta) \end{aligned} \tag{20}$$

Relating Eq. (19) to Eq. (20), we have

$$\begin{aligned} \mathbb{E}[\tilde{e}] &= \frac{H(\tilde{Y}|\hat{Y}, Z_\theta) - H(\tilde{e}|\hat{Y}, Z_\theta)}{H(\tilde{Y}|\tilde{e}=1, \hat{Y}, Z_\theta)} \\ &\geq \frac{H(\tilde{Y}|\hat{Y}, Z_\theta) - H(\tilde{e}|\hat{Y}, Z_\theta)}{\log\left(|\mathcal{Y}| - 1\right)} \\ &\geq \frac{H(\tilde{Y}|\hat{Y}, Z_\theta) - H(\tilde{e})}{\log\left(|\mathcal{Y}| - 1\right)} \\ &= \frac{H(\tilde{Y}) - I(\tilde{Y}; Z_\theta, \hat{Y}) - H(\tilde{e})}{\log\left(|\mathcal{Y}| - 1\right)} \\ &= \frac{H(\tilde{Y}) - I(\tilde{Y}; Z_\theta) - H(\tilde{e})}{\log\left(|\mathcal{Y}| - 1\right)}. \end{aligned}$$

The first inequality is by $H(\tilde{Y}|\tilde{e}=1, \hat{Y}, Z_\theta) \leq \log\left(|\mathcal{Y}| - 1\right)$, where $\tilde{Y}$ can take at most $|\mathcal{Y}| - 1$ values. For the second inequality,

$$\begin{aligned} H(\tilde{e}|\hat{Y}, Z_\theta) &= H(\tilde{e}) - I(\tilde{e}; \hat{Y}, Z_\theta) \\ &\leq H(\tilde{e}). \end{aligned}$$

For the last equality,

$$\begin{aligned} I(\tilde{Y}; Z_\theta, \hat{Y}) =& H(Z_\theta, \hat{Y}) - H(Z_\theta, \hat{Y}|\tilde{Y}) \\ =& H(Z_\theta) + H(\hat{Y}|Z_\theta) \\ & - H(Z_\theta|\tilde{Y}) - H(\hat{Y}|Z_\theta, \tilde{Y}) \\ =& I(Z_\theta, \tilde{Y}) + I(\hat{Y}; \tilde{Y}|Z_\theta) \\ =& I(Z_\theta, \tilde{Y}), \end{aligned}$$

where $I(\hat{Y}; \tilde{Y}|Z_\theta) = 0$ given the Markov Chain $\tilde{Y} \leftarrow Y \leftrightarrow X \rightarrow Z \rightarrow \hat{Y}$:

$$\begin{aligned} I(\hat{Y}; \tilde{Y}|Z_\theta) =& H(\hat{Y}|Z_\theta) - H(\hat{Y}|Z_\theta, \tilde{Y}) \\ =& H(\hat{Y}|Z_\theta) - H(\hat{Y}|Z_\theta) = 0. \end{aligned}$$

□

B.4. Proof for Lemma 2

Lemma. *Let $R(X) = \inf_g \mathbb{E}_{X,Y}[\mathcal{L}(g(X), Y)]$ be the minimum risk over the joint distribution $X \times Y$, where $\mathcal{L}(p, y) = \sum_{i=1}^{\mathcal{Y}} y^{(i)} \log p^{(i)}$ is a CE loss and g is a function mapping from input space to label space. Let $R(Z^\star) = \inf_{g'} \mathbb{E}_{Z^\star,Y}[\mathcal{L}(g'(Z^\star), Y)]$ be the minimum risk over the joint distribution $Z^\star \times Y$ and g' maps from representation space to label space. Then,*

$$R(Z^\star) \leq R(X) + \epsilon.$$

Proof. The lemma is given by the variational form of the conditional entropy $H(Y|Z^\star) = \inf_{g'} \mathbb{E}_{Z^\star,Y}[\mathcal{L}(g'(Z^\star), Y)]$ [11, 22]. According to a property of mutual information,

$$I(A; B) = H(A) - H(A|B),$$

we have $R(Z^\star) = H(Y) - I(Z^\star; Y)$. By the results of Theorem 2,

$$\begin{aligned} R(Z^\star) \leq & H(Y) - I(X;Y) + \epsilon \\ = & H(Y|X) = \inf_g \mathbb{E}_{X,Y}[\mathcal{L}(g(X), Y)]. \end{aligned}$$

□

C. Gradients of Contrastive regularization Functions

For the contrastive regularization function

$$\mathcal{L}'_{\text{ctr}}(x_i, x_j) = -\big(\frac{q_i}{\|q_i\|_2} \cdot \frac{z_j}{\|z_j\|_2} + \frac{q_j}{\|q_j\|_2} \cdot \frac{z_i}{\|z_i\|_2}\big),$$

we only consider the case $\mathbb{1}\{p_i^\top p_j \geq \tau\} = 1$ because $\mathcal{L}'_{\text{ctr}}(x_i, x_j)$ is not calculated in the algorithm when $\mathbb{1}\{p_i^\top p_j \geq \tau\} = 0$. We assume that h is an identity function and x_i, x_j are from the same class for simplicity.

Let $a = \|q_i\|_2$, $b = q_i$, $x = \frac{z_j}{\|z_j\|_2}$ and $c = \frac{b}{a}$. According to the equation $a^2 = b^\top b$, we differentiate both side of the equation and get

$$2a\,\mathrm{d}a = 2b^\top\,\mathrm{d}b. \tag{21}$$

In the meanwhile,

$$\begin{aligned} \partial\big(\frac{b^\top x}{a}\big) &= \frac{\mathrm{d}(b^\top x)a - \mathrm{d}a b^\top x}{a^2} \\ &\overset{(21)}{=} \frac{ax^\top\,\mathrm{d}b}{a^2} - \frac{b^\top\,\mathrm{d}b b^\top x}{a^3} \\ &= \frac{x^\top\,\mathrm{d}b}{a} - \frac{a^2 c^\top x c^\top\,\mathrm{d}b}{a^3} \\ &= \frac{1}{a}\big(x^\top - c^\top x c^\top\big)\,\mathrm{d}b. \end{aligned}$$

Taking a, b, c and x back to the equation, we get the result

$$\frac{\partial \mathcal{L}'_{\text{ctr}}(x_i, x_j)}{\partial q_i} = -\frac{1}{\|q_i\|_2}\big(\frac{q_j}{\|q_j\|_2} - \big(\frac{q_i^\top q_j}{\|q_i\|_2\|q_j\|_2}\big)\frac{q_i}{\|q_i\|_2}\big).$$

Note that $z_i = \texttt{Stopgrad}(q_i)$ because of the identity map h. Let $c_i = 1/\|q_i\|_2^2$ and then we have

$$\left\|\frac{\partial \mathcal{L}'_{\text{ctr}}(x_i, x_j)}{\partial q_i}\right\|_2^2 = c_i(1 - (\tilde{q}_i^\top \tilde{q}_j)^2).$$

Similarly, for the contrastive regularization function

$$\begin{aligned} \widetilde{\mathcal{L}}_{\text{ctr}}(x_i, x_j) = & \Bigg(\log\big(1 - \langle \frac{q_i}{\|q_i\|_2}, \frac{z_j}{\|z_j\|_2}\rangle\big) \\ & + \log\big(1 - \langle \frac{q_j}{\|q_j\|_2}, \frac{z_i}{\|z_i\|_2}\rangle\big)\Bigg), \end{aligned}$$

$$\begin{aligned} \frac{\partial \widetilde{\mathcal{L}}_{\text{ctr}}(x_i, x_j)}{\partial q_i} = & \frac{1}{1 - \tilde{q}_i^\top \tilde{q}_j}\frac{\partial \mathcal{L}'_{\text{ctr}}(x_i, x_j)}{\partial q_i} \\ = & c_i(1 + \tilde{q}_i^\top \tilde{q}_j). \end{aligned}$$