Robust Contrastive Learning against Noisy Views

Ching-Yao Chuang[†]   R Devon Hjelm[‡]   Xin Wang[‡]   Vibhav Vineet[‡]

Neel Joshi[‡]   Antonio Torralba[†]   Stefanie Jegelka[†]   Yale Song[‡]

[†]MIT CSAIL   [‡]Microsoft Research

https://github.com/chingyaoc/RINCE

Abstract

Contrastive learning relies on an assumption that positive pairs contain related views, e.g., patches of an image or co-occurring multimodal signals of a video, that share certain underlying information about an instance. But what if this assumption is violated? The literature suggests that contrastive learning produces suboptimal representations in the presence of noisy views, e.g., false positive pairs with no apparent shared information. In this work, we propose a new contrastive loss function that is robust against noisy views. We provide rigorous theoretical justifications by showing connections to robust symmetric losses for noisy binary classification and by establishing a new contrastive bound for mutual information maximization based on the Wasserstein distance measure. The proposed loss is completely modality-agnostic and a simple drop-in replacement for the InfoNCE loss, which makes it easy to apply to existing contrastive frameworks. We show that our approach provides consistent improvements over the state-of-the-art on image, video, and graph contrastive learning benchmarks that exhibit a variety of real-world noise patterns.

1. Introduction

Contrastive learning [1, 2, 3] has become one of the most prominent self-supervised approaches to learn representations of high-dimensional signals, producing impressive results with image [4, 5, 6, 7, 8, 9], text [10, 11, 12, 13], audio [14, 15, 16], and video [17, 18, 19]. The central idea is to learn representations that capture the underlying information shared between different "views" of data [3, 20]. For images, the views are typically constructed by applying common data augmentation techniques, such as jittering, cropping, resizing and rotation [6], and for video the views are often chosen as adjacent frames [21] or co-occurring multimodal signals, such as video and the corresponding optical flow [22], audio [19] and transcribed speech [17].

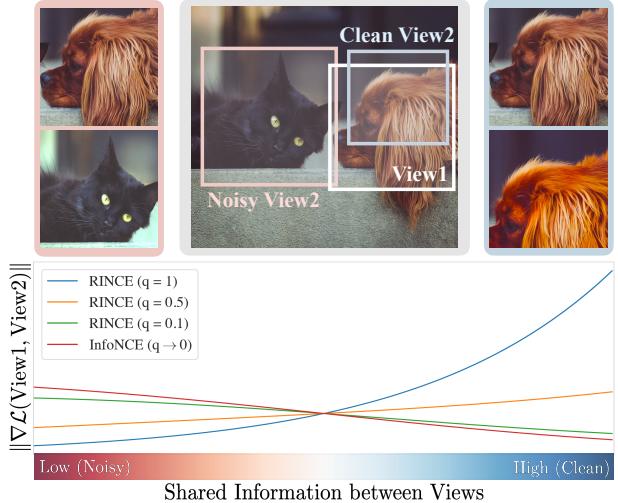Designing the *right* contrasting views has shown to be

Figure 1. **Noisy views can deteriorate contrastive learning.** We propose a new contrastive loss function (RINCE) that rescales the sample importance in the gradient space based on an estimated noise level. With a simple turn of a knob ($q \in (0, 1]$), we can up-weight or downweight sample pairs with low shared information.

a key ingredient of contrastive learning [6, 23]. This often requires domain knowledge, intuition, trial-and-error (and luck!). What would happen if the views are wrongly chosen and do not provide meaningful shared information? Prior work has reported deteriorating effects of such *noisy* views in contrastive learning under various scenarios, e.g., unrelated image patches due to extreme augmentation [20], irrelevant video-audio pairs due to overdubbing [24], and misaligned video-caption pairs [17]. The major issue with noisy views is that representations of different views are forced to align with each other even if there is no meaningful shared information. This often leads to suboptimal representations that merely capture spurious correlations [25] or make them collapse to a trivial solution [26]. Worse yet, when we attempt to learn from large-scale unlabeled data – i.e., the scenario where self-supervised learning is particularly expected to shine – the issue is only aggravated

because of the increased noise in the real-world data [27], hindering the ultimate success of contrastive learning.

Consequently, a few attempts have been made to design contrastive approaches that are noise-tolerant. For example, Morgado et al. [24] optimize a soft instance discrimination loss to weaken the impact of noisy views. Miech et al. [17] address the misalignment between video and captions by aligning multiple neighboring segments of a video. However, existing approaches are often tied to specific modalities or make assumptions that may not hold for general scenarios, e.g., MIL-NCE [17] is not designed to address the issues of irrelevant audio-visual signals.

In this work, we develop a principled approach to make contrastive learning robust against noisy views. We start by making connections between contrastive learning and the classical noisy binary classification in supervised learning [28, 29]. This allows us to explore the wealth of literature on learning with noisy labels [30, 31, 32]. In particular, we focus on a family of robust loss functions that has the *symmetric* property [29], which provides strong theoretical guarantees against noisy labels in binary classification. We then show a functional form of contrastive learning that can satisfy the symmetry condition if given a proper symmetric loss function, motivating the design of new contrastive loss functions that provide similar theoretical guarantees.

This leads us to propose **R**obust **I**nfo**NCE** (RINCE), a contrastive loss function that satisfies the symmetry condition. RINCE can be understood as a generalized form of the contrastive objective that is robust against noisy views. Intuitively, its symmetric property provides an implicit means to reweight sample importance in the gradient space *without requiring* an explicit form of noise estimator. It also provides a simple "knob" (a real-valued scalar $q \in (0, 1]$) that controls the behavior of the loss function balancing the exploration-exploitation trade-off (i.e., from being conservative to playing adventures on potentially noisy samples).

We also provide a theoretical analysis of the proposed RINCE objective and show that it extends the analyses by Ghosh et al. [29] to the self-supervised contrastive learning regime. Furthermore, we relate the proposed loss function to dependency measurement. Analogous to InfoNCE loss, which is a lower bound of mutual information between two views [3], we show that RINCE is a lower bound of Wasserstein Dependency Measure (WDM) [33] even in the noisy setting. By replacing the KL divergence in the mutual information estimator with the Wasserstein distance, WDM is able to capture the geometry of the representation space via the equipped metric space and provides robustness against noisy views better than the KL divergence, both in theory and practice. In particular, the features learned with RINCE achieve better class-wise separation, which is proved to be crucial to improve generalization [34].

Despite its rigorous theoretical background, implement-ing RINCE requires only a few lines code and can be a simple drop-in replacement for the InfoNCE loss to make contrastive learning robust against noisy views. Since InfoNCE sets the basis for many modern contrastive methods such as SimCLR [6] and MoCo-v1/v2/v3 [5, 7, 35], our construction can be easily applied to many existing frameworks.

Finally, we provide strong empirical evidence demonstrating the robustness of RINCE against noisy views under various scenarios with different modalities and noise types. We show that RINCE improves over the state-of-the-art in image [36, 37], video [27, 38] and graph [39] self-supervised learning benchmarks, demonstrating its generalizability across multiple modalities. We also show that RINCE exhibits strong robustness against different types of noise such augmentation noise [20, 40], label noise [28, 41], and noisy audio-visual correspondence [24]. The improvement is consistently observed across different dataset scales and training epochs, demonstrating the scalability and computational efficiency. In short, our main contributions are:

- We propose RINCE, a new contrastive learning objective that is robust against noisy views of data;

- We provide a theoretical analysis to relate the proposed loss to symmetric losses and dependency measurement;

- We demonstrate our approach on real-world scenarios of image, video, and graph contrastive learning.

2. Related Work

Contrastive Learning   Contrastive approaches have become prominent in unsupervised representation learning [1, 2, 42]: InfoNCE [3] and its variants [5, 6, 8, 9] achieve state-of-the-art across different modalities [10, 14, 17, 18, 19]. Modern approaches improve upon InfoNCE from different directions. One line of work focuses on modifying training mechanisms, e.g., appending projection head [6], momentum encoder with dynamic dictionary update [5, 7], siamese networks with stop gradient trick [8, 43], and online cluster assignment [44]. Another line of work refines the loss function itself to make it more effective, e.g., upweight hard negatives [12, 45], correct false negatives [11], and alleviate feature suppression [46]. Along this second line of work, we propose a new contrastive loss function robust against noisy views. Some of prior work in this direction [17, 24] was demonstrated on limited modalities only; we demonstrate its generality on image [37], video [27, 38], and graph [40, 47, 48] contrastive learning scenarios. Our approach is orthogonal to the first line of work; our loss function can easily be applied to some of existing training mechanisms such as SimCLR [6] and MoCo-v1/v2/v3 [5, 7, 35].

Robust Loss against Noisy Labels   Learning with noisy labels has been actively explored in recent years [28, 29, 31, 32, 49, 50, 51, 52, 53, 54, 55]. One line of work

attempts to develop robust loss functions that are noise-tolerant [29, 30, 56, 57]. Ghosh et al. [29] prove that symmetric loss functions are robust against noisy labels, e.g., Mean Absolute Error (MAE) [30], while commonly used Cross Entropy (CE) loss is not. Based on this idea, Zhang and Sabuncu [56] propose the generalized cross entropy loss to combine MAE and CE loss functions. A similar idea is adopted in [57] by combining the reversed cross entropy loss with CE loss. In the next section, we relate noisy views to noisy labels by interpreting contrastive learning as binary classification, and developed a robust symmetric contrastive loss that enjoys the similar theoretical guarantees.

3. Prelim: From Noisy Labels to Noisy Views

We start by connecting two seemingly different but related frameworks: supervised binary classification with noisy labels and self-supervised contrastive learning with noisy views. We then introduce a family of symmetric loss functions that is noise-tolerant and show how we can transform contrastive objectives to a symmetric form.

3.1. Symmetric Losses for Noisy Labels

Denoting the input space by $\mathcal{X}$ and the binary output space by $\mathcal{Y} = \{-1, 1\}$, let $\mathcal{S} = \{x_i, y_i\}_{i=1}^m$ be the unobserved clean dataset that is drawn i.i.d. from the data distribution $\mathcal{D}$. In the noisy setting, the learner obtains a noisy dataset $\mathcal{S}_\eta = \{x_i, \hat{y}_i\}_{i=1}^m$, where $\hat{y}_i = y_i$ with probability $1 - \eta_{x_i}$ and $\hat{y}_i = -y_i$ with probability η_{x_i}. Note that the noise rate η_x is data point-dependent. For a classifier $f \in \mathcal{F} : \mathcal{X} \to \mathbb{R}$, the expected risk under the noise-free scenario is $R_\ell(f) = \mathbb{E}_\mathcal{D}[\ell(f(x), y)]$ where $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is a binary classification loss function. When the noise exists, the learner minimizes the noisy expected risk $R_\ell^\eta(f) = \mathbb{E}_{\mathcal{D}_\eta}[\ell(f(x), \hat{y})]$.

Ghosh et al. [29] show that *symmetric* loss functions are robust against noisy labels in binary classification. In particular, a loss function ℓ is symmetric if it sums to a constant:

$$\ell(s, 1) + \ell(s, -1) = c, \quad \forall s \in \mathbb{R}, \tag{1}$$

where s is the prediction score from f. Note that the symmetry condition should also hold with the gradients w.r.t. s. They show that if the noise rate is $\eta_x \leq \eta_{\max} < 0.5, \forall x \in \mathcal{X}$ and if the loss is symmetric and non-negative, the minimizer of the noisy risk $f_\eta^* = \arg\inf_{f \in \mathcal{F}} R^\eta(f)$ approximately minimizes the clean risk:

$$R(f_\eta^*) \leq \epsilon/(1 - 2\eta_{\max}),$$

where $\epsilon = \inf_{f \in \mathcal{F}} R(f)$ is the optimal clean risk. This implies that the noisy risk under symmetric loss is a good surrogate of the clean risk. In Appendix A.2, we further relax the non-negative constraint on the loss with a corollary.[1]

---

[1] This is important for our proposed RINCE loss that involves an exponential function $\ell(s, y) = -ye^s$, which can produce negative values.

3.2. Towards Symmetric Contrastive Objectives

The results above suggest that we can achieve robustness against noisy views if a contrastive objective can be expressed in a form that satisfies the symmetry condition in the binary classification framework. To this end, we first relate contrastive learning to binary classification, and then express it in a form where symmetry can be achieved.

Contrastive learning as binary classification. Given two views X and V, we can interpret contrastive learning as noisy binary classification operating over pairs of samples (x, v) with a label 1 if it is sampled from the joint distribution, $(x, v) \sim P_{XV}$, and -1 if it comes from the product of marginals, $(x, v') \sim P_X P_V$. In the presence of noisy views, some negative pairs $(x, v') \sim P_X P_V$ could be mislabeled as positive, introducing noisy labels.

To see this more concretely, let us consider the InfoNCE loss [3], one of the most widely adopted contrastive objectives [4, 6, 11, 58]. It minimizes the following loss function:

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{s}) = -\log \frac{e^{s^+}}{e^{s^+} + \sum_{i=1}^K e^{s_i^-}}$$

$$:= -\log \frac{e^{f(x)^T g(v)/t}}{e^{f(x)^T g(v)/t} + \sum_{i=1}^K e^{f(x)^T g(v_i)/t}}, \tag{2}$$

where $\mathbf{s} = \{s^+, \{s_i^-\}_{i=1}^K\}$, s^+ and s_i^- are the scores of related (positive) and unrelated (negative) pairs and t is the temperature parameter introduced to avoid gradient saturation. The expectation of the loss is taken over $(x, v) \sim P_{XV}$ and K independent samples $v_i \sim P_V$, where P_{XV} denotes the joint distribution over pairs of views such as transformations of the same image or co-occurring multimodal signals. Although InfoNCE has a functional form of the $(K + 1)$-way softmax cross entropy loss, the model ultimately learns to classify whether a pair (x, v) is positive or negative by maximizing/minimizing the positive score s^+/negative scores s_i^-. Therefore, InfoNCE under noisy views can be seen as binary classification with noisy labels. We acknowledge that similar interpretations have been made in prior works under different contexts [4, 59, 60].

Symmetric form of contrastive learning. Now we turn to a functional form of contrastive learning that can achieve the symmetric property. Assume that we have a noise-tolerant loss function ℓ that satisfies the symmetry condition of equation 1. We say a contrastive learning objective is symmetric if it accepts the following form

$$\mathcal{L}(\mathbf{s}) = \underbrace{\ell(s^+, 1)}_{\text{Positive Pair}} + \lambda \underbrace{\sum_{i=1}^K \ell(s_i^-, -1)}_{K \text{ Negative Pairs}} \tag{3}$$

which consists of a collection of $(K + 1)$ binary classification losses; $\lambda > 0$ is a density weighting term controlling

```
1  # pos: exponent for positive example
2  # neg: sum of exponents for negative examples
3  # q, lam: hyperparameters of RINCE
4
5  info_nce_loss = -log(pos/(pos+neg))
6  rince_loss = -pos**q/q + (lam*(pos+neg))**q/q
```

Figure 2. **Pseudocode for RINCE.** The implementation only requires a small modification to the InfoNCE code.

the ratio between classes 1 (positive pairs) and -1 (negative pairs). Reducing λ places more weight on the positive score s^+, while setting λ to zero recovers the negative-pair-free contrastive loss such as BYOL [8].

Contrastive objectives that satisfy the symmetric form enjoy strong theoretical guarantees against noisy labels as described in Ghosh et al. [29], as long as we plug in the right contrastive loss function ℓ that satisfies the symmetry condition. Unfortunately, the InfoNCE loss [3] does not satisfy the symmetry condition in the gradients w.r.t. $s^{+/-}$ (we provide the full derivations in Appendix A.5). This motivates us to develop a new contrastive loss function that satisfies the symmetry condition, described next.

4. Robust InfoNCE Loss

Based on the idea of robust symmetric classification loss, we present the following Robust InfoNCE (RINCE) loss:

$$\mathcal{L}_{\text{RINCE}}^{\lambda,q}(\mathbf{s}) = \frac{-e^{q \cdot s^+}}{q} + \frac{(\lambda \cdot (e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}))^q}{q},$$

where $q, \lambda \in (0,1]$. Figure 2 shows the pseudo-code of RINCE: it is simple to implement. When $q = 1$, RINCE becomes a contrastive loss that fully satisfies the symmetry property in the form of equation 3 with $\ell(s, y) = -ye^s$:

$$\mathcal{L}_{\text{RINCE}}^{\lambda,q=1}(\mathbf{s}) = -(1-\lambda)e^{s^+} + \lambda \sum_{i=1}^{K} e^{s_i^-}.$$

Notice that the exponential loss $-ye^s$ satisfies the symmetric condition defined in equation 1 with $c = 0$. Therefore, when $q \to 1$, we achieve robustness against noisy views in the same manner as binary classification with noisy labels.

In the limit of $q \to 0$, RINCE becomes asymptotically equivalent to InfoNCE, as the following lemma describes:

Lemma 1. *For any $\lambda > 0$, it holds that*

$$\lim_{q \to 0} \mathcal{L}_{\text{RINCE}}^{\lambda,q}(\mathbf{s}) = \mathcal{L}_{\text{InfoNCE}}(\mathbf{s}) + \log(\lambda);$$

$$\lim_{q \to 0} \frac{\partial}{\partial \mathbf{s}} \mathcal{L}_{\text{RINCE}}^{\lambda,q}(\mathbf{s}) = \frac{\partial}{\partial \mathbf{s}} \mathcal{L}_{\text{InfoNCE}}(\mathbf{s}).$$

We defer the proofs to Appendix A. Note that the convergence also holds for the derivatives: optimizing RINCE in the limit of $q \to 0$ is mathematically equivalent to optimizing InfoNCE. Therefore, by controlling $q \in (0,1]$ we smoothly interpolate between the InfoNCE loss ($q \to 0$) and the RINCE loss in its fully symmetric form ($q \to 1$).
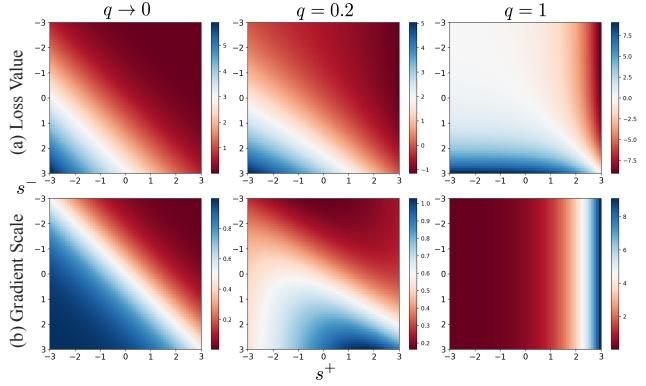

Figure 3. **Loss Visualization.** We visualize the (a) loss value and the (b) gradient scale with respect to the positive score s^+ for different q while setting $\lambda = 0.5$. The gradient scale of InfoNCE ($q \to 0$) is larger when the positive score is smaller (hard positive pair). In contrast, for fully symmetric RINCE ($q = 1$), the gradient is larger when positive score is large (easy positive pair).

4.1. Intuition behind RINCE

We now analyze the behavior of RINCE through the lens of exploration-exploitation trade-off. In particular, we reveal an implicit easy/hard positive mining scheme by inspecting the gradients of RINCE under different q values, and show that we achieve stronger robustness (more exploitation) with larger q at the cost of potentially useful clean hard positive samples (less exploration).

To simplify the analysis, we consider InfoNCE and RINCE with a single negative pair ($K = 1$):

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{s}) = -\log(e^{s^+}/(e^{s^+} + e^{s^-}));$$

$$\mathcal{L}_{\text{RINCE}}^{\lambda,q}(\mathbf{s}) = \frac{-e^{q \cdot s^+}}{q} + \frac{(\lambda \cdot (e^{s^+} + e^{s^-}))^q}{q}.$$

We visualize the loss and the scale of the gradients with respect to positive scores s^+ in Figure 3. Although the loss values are different for each q, they follow the same principle: The loss achieves its minimum when the positive score s^+ is maximized and the negative score s^- is minimized.

The interesting bit lies in the gradients. The InfoNCE loss ($q \to 0$) places more emphasis on *hard* positive pairs, i.e., the pairs with *low* positive scores s^+ (the left-most part in the plot). In contrast, the fully symmetric RINCE loss ($q = 1$) places more weights on *easy* positive pairs (the right-most part). Note that both $q \to 0$ and $q \to 1$ naturally perform hard negative mining; both their derivatives put exponentially more weights on hard negative pairs.

This reveals an implicit trade-off between exploration (convergence) and exploitation (robustness). When $q \to 0$, the loss performs hard positive mining, providing faster convergence in the noise-free setting. But in the presence of noise, exploration is harmful; it wrongly puts higher weights to false positive pairs because noisy samples tend

4

to induce larger losses [24, 55, 61, 62], and this could hinder convergence. In contrast, when $q \to 1$, we perform easy positive mining. This provides robustness especially against false positives; but this is done at the cost of exploration with clean hard positives. An important aspect here is that RINCE does not require an explicit form of noise estimator: the scores s^+ and s^-, and the relationship between the two (which is what the loss function measures) act as noise estimates. In practice, we set $q \in [0.1, 0.5]$ to strike the balance between exploration and exploitation.

4.2. Theoretical Underpinnings

Next, we provide an information-theoretic explanation on what makes RINCE robust against noisy views. In particular, we show that RINCE is a contrastive lower-bound of mutual information (MI) expressed in Wasserstein dependency measure (WDM) [33], which provides superior robustness against sample noise compared to the Kullback–Leibler (KL) divergence thanks to strong geometric properties of the Wasserstein metric. We further show that, even in the presence of noise, RINCE is a lower bound of clean WDM, indicating its robustness against noisy views.

Limitations of KL divergence in MI estimation. Without loss of generality, let $f = g$ and consider $f = f' \circ \phi$, where ϕ is a representation encoder and f' is a projection head [6]. Also, let $P^\phi = \phi_\# P$ be the pushforward measure of P with respect to ϕ. It has been shown [20, 63] that InfoNCE is a variational lower-bound of MI in the representation space expressed with KL-divergence:

$$-\mathbb{E}\left[\mathcal{L}_{\text{InfoNCE}}(\mathbf{s})\right] + \log(K) \leq I(\phi(X), \phi(V))$$
$$= D_{\text{KL}}(P^\phi_{XV}, P^\phi_X P^\phi_V).$$

Intuitively, maximizing MI can be interpreted as maximizing the discrepancy between positive and negative pairs. However, prior works [33, 64] have identified theoretical limitations of maximizing MI using the KL divergence: Because KL divergence is not a metric, it is sensitive to small differences in data samples regardless of the geometry of the underlying data distributions. Therefore, the encoder ϕ can capture limited information shared between X and V as long as the differences are sufficient to maximize the KL divergence. Note that this can be especially detrimental in the presence of noisy views, as the learner can quickly settle on spurious correlations in false positive pairs due to the absence of the actual shared information.

RINCE is a lower bound of WDM. We now establish RINCE as a lower bound of WDM [33], which is proposed as a replacement for the KL divergence in MI estimation.

WDM is based on the Wasserstein distance, a distance metric between probability distributions defined via an optimal transport cost. Letting μ and $\nu \in \text{Prob}(\mathbb{R}^d \times \mathbb{R}^d)$

be two probability measures, we define the Wasserstein-1 distance with a Euclidean cost function as

$$\mathcal{W}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{\substack{(X,V) \\ (X',V')}^{\sim \pi}} \left[\left\| X - X' \right\| + \left\| V - V' \right\| \right]$$

where $\Pi(\mu, \nu)$ denotes the set of measure couplings whose marginals are μ and ν, respectively. By virtue of symmetry when $q = 1$, if $\lambda > 1/(K+1)$, the Kantorovich-Rubinstein duality [65] implies that (full theorem in Appendix A.3):

$$-\mathbb{E}\left[\mathcal{L}_{\text{RINCE}}^{\lambda, q=1}(\mathbf{s})\right] \leq L \cdot I_{\mathcal{W}}(\phi(X), \phi(V))$$
$$:= L \cdot \mathcal{W}(P^\phi_{XV}, P^\phi_X P^\phi_V), \quad (4)$$

where $I_{\mathcal{W}}(\phi(X), \phi(V))$ is the WDM defined in [33] and L is a constant that depends on t, λ, and the Lispchitz constant of the projection head f. Note that we are not aware of any work that showed it is possible to establish a similar bound with WDM for the InfoNCE loss.

This provides another explanation of what makes RINCE robust against noisy views. Unlike InfoNCE which maximizes the KL divergence, optimizing RINCE is equivalent to maximizing the WDM with a Lipschitz function. Equipped with a proper metric, this allows RINCE to measure the divergence between two distributions P^ϕ_{XV} and $P^\phi_X P^\phi_V$ without being overly sensitive to individual sample noise, as long as the noise does not alter the geometry of the distributions. This also allows the encoder ϕ to learn more complete representations, as maximizing the Wasserstein distance requires the encoder to not only model the density ratio between the two distributions but also the optimal cost of transporting one distribution to another.

RINCE is still a lower bound of WDM even with noise. Finally, we show that RINCE still maximizes the noise-less WDM under additive noise, corroborating the robustness of RINCE. Let's consider a simple mixture noise model:

$$P^\eta_{XV} = (1 - \eta) P_{XV} + \eta P_X P_V,$$

where η is the noise rate and the noisy joint distribution P^η_{XV} is a weighted sum between the noise-less positive distribution P_{XV} and negative distribution $P_X P_V$. Note that the marginals of P^η_{XV} are still P_X and P_V by construction. The intuition behind mixture noise model is that when we draw positive pairs from P^η_{XV}, we obtain false positives from $P_X P_V$ with probability η. Via the symmetry of the contrastive loss, we can extend bound (4) as follows (proof in Appendix A.4):

$$-\mathbb{E}_{P^\eta_{XV}}\left[\mathcal{L}_{\text{RINCE}}^{\lambda, q=1}(\mathbf{s})\right] \leq (1 - \eta) \cdot L \cdot I_{\mathcal{W}}(\phi(X), \phi(V)).$$

Comparing to the bound (4), the right hand side is rewieghted with $(1 - \eta)$. This implies that minimizing RINCE with noisy views still maximizes a lower bound of
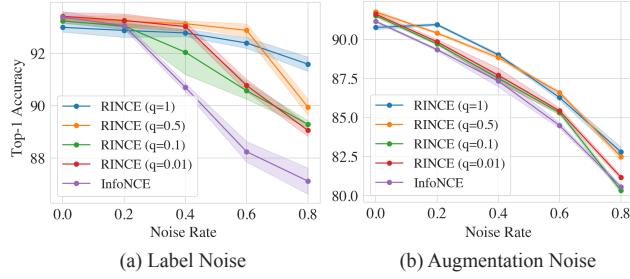
(a) Label Noise   (b) Augmentation Noise

Figure 4. **Noisy CIFAR-10.** We show the top1 accuracy of RINCE with different values of q across different noise rate η. Large q ($q = 0.5, 1$) leads to better robustness, while smaller q ($q = 0.01$) performs similar to InfoNCE ($q \to 0$).
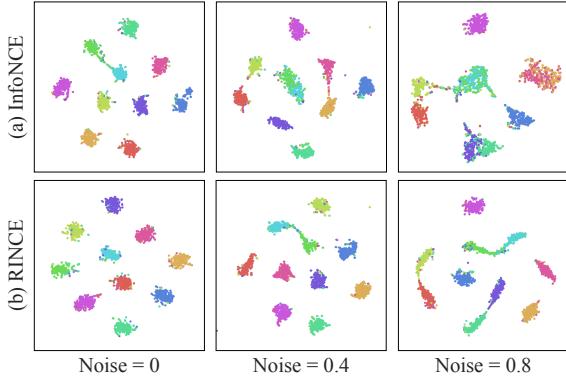

Noise = 0   Noise = 0.4   Noise = 0.8

Figure 5. **t-SNE Visualization on CIFAR-10 with label noise.** Colors indicate classes. RINCE leads to better class-wise separation than the InfoNCE loss in both noise-less and noisy cases.

noise-less WDM. Despite the simplicity of the analysis, it intuitively relates dependency measures and the noisy views with interpretable bounds. It would be an interesting future direction to extend the analysis to more complicated noise models, e.g., $P_{XV}^{\eta} = (1 - \eta)P_{XV} + \eta Q_{XV}$, where Q is an unknown perturbation on positive distribution.

5. Experiments

We evaluate RINCE on various contrastive learning scenarios involving images (CIFAR-10 [36], ImageNet [37]), videos (ACAV100M [27], Kinetics400 [38]) and graphs (TUDataset [39]). Empirically, we find that RINCE is insensitive to the choice of λ; we simply set $\lambda = 0.01$ for all vision experiments and $\lambda = 0.025$ for graph experiments.

5.1. Noisy CIFAR-10

We begin with controlled experiments on CIFAR-10 to verify the robustness of RINCE against synthetic noise by controlling the noise rate η. We consider two noise types:

Label noise: We start with the case of supervised contrastive learning [41] where positive pairs are different images of the same label. This allows us to control noise in the traditional sense, i.e., learning with noisy labels. Similar to

Method	Δ to SimCLR [6]	Top 1	Top 5
Supervised [67]	N/A	76.5	-
SimSiam [43]	No negative pairs	71.3	-
BYOL [8]	No negative pairs	74.3	91.6
Barlow Twins [9]	Redundancy reduction	73.2	91.0
SwAV [44]	Cluster discrimination	75.3	-
SimCLR [6]	None	69.3	89.0
+RINCE (Ours)	Symmetry controller q	**70.0**	**89.8**
MoCo [5]	Momentum encoder	60.6	-
MoCov2 [7]	Momentum encoder	71.1	90.1
MoCov3 [35]	Momentum encoder	73.8	-
+RINCE (Ours)	Symmetry controller q	**74.2**	**91.8**

Table 1. **Linear Evaluation on ImageNet.** All the methods use ResNet-50 [67] as backbone architecture with 24M parameters. Note that RINCE subsumes InfoNCE when $q \to 0$.

[56], we flip the true labels to semantically related ones, e.g., CAT $\leftrightarrow$ DOG with probability $\eta/2$. This is commonly referred to as class-dependent noise [55, 56, 57].

Augmentation noise: We consider the self-supervised learning scenario and vary the crop size during data augmentation similar to [20], i.e., after applying all the transformations as in SimCLR [6], images are further cropped into $1/5$ of their original size with probability η. This effectively controls the noise rate as cropped patches will most likely be too small to contain any shared information.

Figure 4 shows the results of SimCLR trained with InfoNCE and RINCE with different choices of q and λ. When the augmentation noise is present, e.g., $\eta = 0.4$, the accuracy of InfoNCE drops from 91.14% to 87.33%. In contrast, the robustness of RINCE is enhanced by increasing q, achieving 89.01% when $q = 1.0$,. InfoNCE also fails to address label noise and suffers from significant performance drop (93.38% $\to$ 87.11% when $\eta = 0.8$). In comparison, RINCE retains the performance even when the noise rate is large (91.59% for $q = 1.0$). In both cases, reducing the value of q makes the performance of RINCE closer to InfoNCE, verifying our analysis in Lemma 1.

Figure 5 shows t-SNE visualization [66] of representations learned with InfoNCE and RINCE ($q = 1.0$) under different label noise. As the noise rate increases, representations of different classes start to tangle up for InfoNCE, while RINCE still achieves decent class-wise separation.

5.2. Image Contrastive Learning

We verify our approach on the well-established ImageNet benchmark [37]. We adopt the same training protocol and hyperparameter settings of SimCLR [6] and MoCov3 [35] and simply replace the InfoNCE with our RINCE

Method	Backbone	Finetune Input Size	HMDB	UCF
3D-RotNet [68]	R3D-18	16×112^2	33.7	62.9
ClipOrder [69]	R3D-18	16×112^2	30.9	72.4
DPC [70]	R3D-18	25×128^2	35.7	75.7
CBT [71]	S3D	16×112^2	44.6	79.5
AVTS [72]	MC3-18	25×224^2	56.9	85.8
SeLaVi [73]	R(2+1)D-18	32×112^2	47.1	83.1
XDC [74]	R(2+1)D-18	32×224^2	52.6	86.8
Robust-xID [24]	R(2+1)D-18	32×224^2	55.0	85.6
Cross-AVID [19]	R(2+1)D-18	32×224^2	59.9	86.9
AVID+CMA [19]	R(2+1)D-18	32×224^2	60.8	87.5
InfoNCE (Ours)	R(2+1)D-18	32×224^2	57.8	88.6
RINCE (Ours)	R(2+1)D-18	32×224^2	**61.6**	**88.8**
GDT [23]	R(2+1)D-18	30×112^2	62.3*	90.9*

Based on an advanced hierarchical data augmentation during pretraining.

Table 2. **Kinetics400-pretrained performance on UCF101 and HMDB51** (top-1 accuracy). Ours use the same data augmentation approach as Cross-AVID and AVID+CMA, while GDT uses an advanced hierarchical sampling process.

loss ($q = 0.1$ and $q = 0.6$, respectively) as shown in Figure 2. Table 1 shows that RINCE improves InfoNCE (SimCLR and MoCov3) by a non-trivial margin. We also include results from the SOTA baselines, where they improve SimCLR by introducing dynamic dictionary plus momentum encoder (MoCo-v1/v2/v3 [5, 7, 35]), removing negative pairs plus the stop-gradient trick (SimSiam [43], BYOL [8]), or online cluster assignment (SwAV [44]). In comparison, our work is orthogonal to the recent developments, and the existing tricks can be applied along with RINCE.

Figure 6 shows the positive pairs from SimCLR augmentations and the corresponding positive scores $s^+ = f(x)^T g(v)$ output by trained RINCE model. Examples with lower positive scores contain pairs that is less informative to each other, while semantically meaningful pairs often have higher scores. This implies that positive scores are good noise detectors, and down-weighting the samples with lower positive score brings robustness during training, verifying our analysis in section 4.1.

5.3. Video Contrastive Learning

We examine our approach in the audio-visual learning scenario using two video datasets: Kinetics400 [38] and ACAV100M [27]. Here, we find that simple *q-warmup* improves the stability of RINCE, i.e., q starts at 0.01 and linearly increases to 0.4 until the last epoch. We apply this to all RINCE models in this section. As we show below, RINCE outperforms SOTA noise-robust contrastive methods [19, 24] on Kinetics400, while also providing scalability and computational efficiency compared to InfoNCE.

Kinetics400 For fair comparison to SOTA, we follow the same experimental protocol and hyperparameter settings
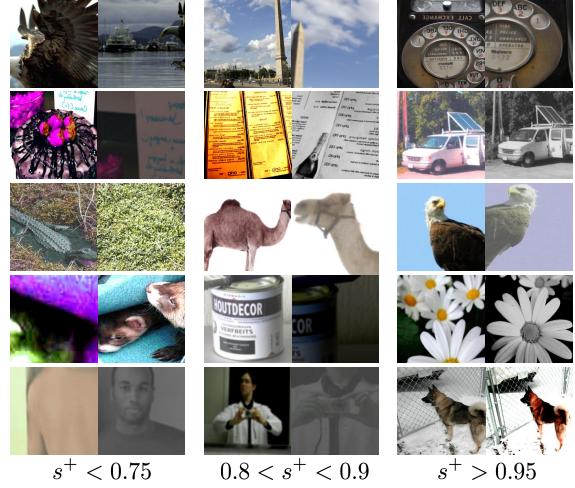

Figure 6. **Positive pairs and their scores.** The positive scores $s^+ \in [-1, 1]$ are output by the trained RINCE model (temperature $= 1$). Pairs that have lower scores are visually noisy, while informative pairs often have higher scores.

of [19] and simply replace their loss functions with InfoNCE and RINCE loss as shown in Figure 2. We use the same network architecture, i.e., 18-layer R(2+1)D video encoder [75], 9-layer VGG-like audio encoder, and 3-layer MLP projection head producing 128-dim embeddings. We use the ADAM optimizer [76] for 400 epochs with 4,096 batch size, 1e-4 learning rate and 1e-5 weight decay. The pretrained encoders are finetuned on UCF-101 [77] and HMDB-51 [78] with clips composed of 32 frames of size 224×224. We defer the full experimental details to Appendix B.

Table 2 shows that RINCE outperforms most of the baseline approaches, including Robust-xID [24] and AVID+CMA [19] which are recent InfoNCE-based SOTA methods proposed to address the noisy view issues in audio-visual contrastive learning. Considering the only change required is the simple replacement of the InfoNCE with our RINCE loss, the results clearly show the effectiveness of our approach. The simplicity means we can easily apply RINCE to a variety of InfoNCE-based approaches, such as GDT [23] that uses advanced data augmentation mechanisms to achieve SOTA results.

ACAV100M We conduct an in-depth analysis of RINCE on ACAV100M [27], a recent large-scale video dataset for self-supervised learning. Compared to Kinetics400 which is limited to human actions, ACAV100M contains videos "in-the-wild" exhibiting a wide variety of audio-visual patterns. The unconstrained nature of the dataset makes it a good benchmark to investigate the robustness of RINCE to various types of real-world noise, e.g., background music, overdubbed audio, studio narrations, etc.

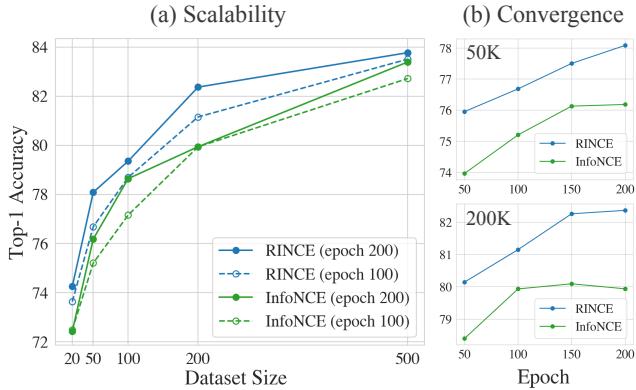We focus on evaluating the (a) scalability and (b) conver-

Figure 7. **RINCE outperforms InfoNCE with fewer epochs across different scales** The results are based on ACAV100M-pretrained models transferred to UCF-101.

gence rate of RINCE, thereby answering the question: *Will it retrain its edge over InfoNCE (a) even in the large-scale regime and (b) with a longer training time?* We follow the same experimental setup as described above, but reduce the batch size to 512 and report the results only on the first split of UCF-101 to make our experiments tractable.

Figure 7 (a) shows the top-1 accuracy of RINCE and InfoNCE across different data scales and training epochs. RINCE outperforms InfoNCE by a large margin at every data scale. In terms of the convergence rate, RINCE is comparable to or even outperforms fully-trained (200 epochs) InfoNCE models with only 100 or fewer epochs. Figure 7 (b) gives a closer look at the convergence at 50K and 200K scales. Interestingly, InfoNCE saturates and even degenerates after epoch 150, while RINCE keeps improving. This verifies our analysis in section 4.1: InfoNCE can overfit noisy samples due to its exploration property, while RINCE downweights them and continue to obtain the learning signal from clean ones, achieving robustness against noise.

5.4. Graph Contrastive Learning

To see whether the modality-agnostic nature of RINCE applies beyond image and video data, we examine our approach on TUDataset [39], a popular benchmark suite for graph inference on molecules (BZR, NCI1), bioinformatics (PROTEINS), and social network (RDT-B, IMDB-B). Unlike vision datasets, data augmentation for graphs requires careful engineering with domain knowledge, limiting the applicability of InfoNCE-type contrastive objectives.

For fair comparison, we follow the protocol of [40] and train graph isomorphism networks [79] with four types of data augmentation: node dropout, edge perturbation, attribute masking, and subgraph sampling. We train models using ADAM [76] for 20 epochs with a learning rate 0.01 and report mean and standard deviation over 5 independent trials. We set $q = 0.1$ for all the experiments in this section.

Table 3 shows that RINCE outperforms three SOTA

Methods	RDT-B	NCI1	PROTEINS	DD
node2vec [80]	-	54.9±1.6	57.5±3.6	-
sub2vec [81]	71.5±0.4	52.8±1.5	53.0±5.6	-
graph2vec [82]	75.8±1.0	73.2±1.8	73.3±2.1	-
InfoGraph [47]	82.5±1.4	76.2±1.1	74.4±0.3	72.9±1.8
GraphCL [40]	89.5±0.8	77.9±0.4	74.4±0.5	78.6±0.4
JOAO [83]	85.3±1.4	78.1±0.5	74.6±0.4	77.3±0.5
JOAOv2 [83]	86.4±1.5	78.4±0.5	74.1±1.1	77.4±1.2
InfoNCE* (Ours)	89.9±0.4	78.2±0.8	74.4±0.5	78.6±0.8
RINCE (Ours)	**90.9±0.6**	**78.6±0.4**	**74.7±0.8**	**78.7±0.4**

*GraphCL [40] but uses the same data augmentation as RINCE.

Table 3. **Self-supervised representation learning on TUDataset**: The baseline results are excerpted from the published papers.
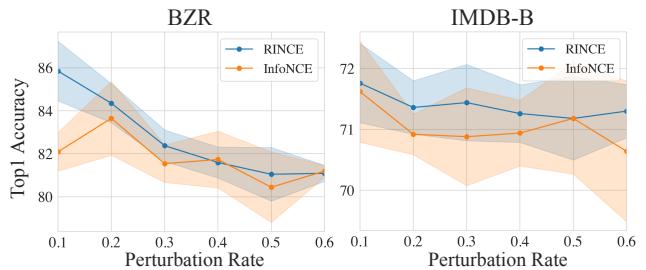

Figure 8. **Performance v.s. Perturbation Rate:** We increase the perturbation rate of node dropping, edge perturbation, and attribute masking from 10% to 60%. RINCE outperforms InfoNCE in terms of accuracy and variance when perturbation enhances.

InfoNCE-based contrastive methods, GraphCL and JOAO /JOAOv2, setting the new records on all four datasets. GraphCL applies different augmentations for different datasets, while JOAO/JOAOv2 require solving bi-level optimization to choose optimal augmentation per dataset. In contrast, we apply the same augmentation across all four datasets and achieve competitive performance, demonstrating its generality and robustness. In Figure 8, we control perturbation rate by applying three augmentation types (node dropout, edge perturbation, attribute masking) to different % of nodes/edges. We show results on two datasets most sensitive to augmentation. Again, RINCE consistently outperform InfoNCE and has relatively smaller variances when the noise rate increases.

6. Conclusion

We presented *Robust InfoNCE (RINCE)* as a simple drop-in replacement for the InfoNCE loss in contrastive learning. Despite its simplicity, it comes with strong theoretical justifications and guarantees against noisy views. Empirically, we provided extensive results across image, video, and graph contrastive learning scenarios demonstrating its robustness against a variety of realistic noise patterns.

References

[1] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546. 1, 2

[2] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742. 1, 2

[3] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018. 1, 2, 3, 4

[4] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 776–794. 1, 3

[5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738. 1, 2, 6, 7

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607. 1, 2, 3, 5, 6, 17

[7] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," in *2020 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'20)*. IEEE, 2020. 1, 2, 6, 7

[8] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *Advances in neural information processing systems*, 2020. 1, 2, 4, 6, 7

[9] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021. 1, 2, 6

[10] L. Logeswaran and H. Lee, "An efficient framework for learning sentence representations," in *International Conference on Learning Representations*, 2018. 1, 2

[11] C.-Y. Chuang, J. Robinson, L. Yen-Chen, A. Torralba, and S. Jegelka, "Debiased contrastive learning," *Advances in neural information processing systems*, 2020. 1, 2, 3, 17

[12] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," in *International Conference on Learning Representations*, 2021. 1, 2

[13] J. M. Giorgi, O. Nitski, G. D. Bader, and B. Wang, "Declutr: Deep contrastive learning for unsupervised textual representations," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2020. 1

[14] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in neural information processing systems*, 2020. 1, 2

[15] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3875–3879. 1

[16] L. Wang and A. v. d. Oord, "Multi-format contrastive learning of audio representations," *arXiv preprint arXiv:2103.06508*, 2021. 1

[17] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9879–9889. 1, 2

[18] S. Ma, Z. Zeng, D. McDuff, and Y. Song, "Active contrastive learning of audio-visual video representations," in *International Conference on Learning Representations*, 2020. 1, 2

[19] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 475–12 486. 1, 2, 7, 17, 18

[20] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" in *Advances in neural information processing systems*, 2020. 1, 2, 5, 6

[21] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *2018 IEEE international conference on robotics and automation*, 2018. 1

[22] T. Han, W. Xie, and A. Zisserman, "Self-supervised co-training for video representation learning," *Advances in neural information processing systems*, 2020. 1

[23] M. Patrick, Y. M. Asano, P. Kuznetsova, R. Fong, J. F. Henriques, G. Zweig, and A. Vedaldi, "On compositions of transformations in contrastive self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1, 7

[24] P. Morgado, I. Misra, and N. Vasconcelos, "Robust audio-visual instance discrimination," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 934–12 945. 1, 2, 5, 7

[25] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019. 1

[26] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," *arXiv preprint arXiv:2110.09348*, 2021. 1

[27] S. Lee, J. Chung, Y. Yu, G. Kim, T. Breuel, G. Chechik, and Y. Song, "Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 274–10 284. 2, 6, 7

[28] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," *Advances in neural information processing systems*, vol. 26, pp. 1196–1204, 2013. 2

[29] A. Ghosh, N. Manwani, and P. Sastry, "Making risk minimization tolerant to label noise," *Neurocomputing*, vol. 160, pp. 93–107, 2015. 2, 3, 4, 14

[30] A. Ghosh, H. Kumar, and P. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017. 2, 3

[31] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[32] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie, "Learning from noisy large-scale datasets with minimal supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 2

[33] S. Ozair, C. Lynch, Y. Bengio, A. v. d. Oord, S. Levine, and P. Sermanet, "Wasserstein dependency measure for representation learning," *arXiv preprint arXiv:1903.11780*, 2019. 2, 5, 15

[34] C.-Y. Chuang, Y. Mroueh, K. Greenewald, A. Torralba, and S. Jegelka, "Measuring generalization with optimal transport," in *Advances in Neural Information Processing Systems*, 2021. 2

[35] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," *arXiv preprint arXiv:2104.02057*, 2021. 2, 6, 7, 17

[36] A. Krizhevsky *et al.*, "Learning multiple layers of features from tiny images." Citeseer, 2009. 2, 6

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. 2, 6

[38] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017. 2, 6, 7

[39] C. Morris, N. M. Kriege, F. Bause, K. Kersting, P. Mutzel, and M. Neumann, "Tudataset: A collection of benchmark datasets for learning with graphs," *arXiv preprint arXiv:2007.08663*, 2020. 2, 6, 8

[40] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5812–5823, 2020. 2, 8, 18

[41] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, 2020. 2, 6

[42] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018. 2

[43] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 6, 7

[44] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 6, 7

[45] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[46] J. Robinson, L. Sun, K. Yu, K. Batmanghelich, S. Jegelka, and S. Sra, "Can contrastive learning avoid shortcut solutions?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[47] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," *arXiv preprint arXiv:1809.10341*, 2018. 2, 8

[48] K. Hassani and A. H. Khasahmadi, "Contrastive multi-view representation learning on graphs," in *International Conference on Machine Learning*. PMLR, 2020, pp. 4116–4126. 2

[49] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015. 2

[50] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2691–2699. 2

[51] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 447–461, 2015. 2

[52] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. 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*, 2017, pp. 1944–1952. 2

[53] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *International Conference on Machine Learning*, 2018. 2

[54] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4334–4343. 2

[55] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Advances in neural information processing systems*, 2018. 2, 5, 6

[56] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, 2018. 3, 6

[57] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 322–330. 3, 6

[58] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," *arXiv preprint arXiv:1906.00910*, 2019. 3

[59] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304. 3

[60] Z. Wu, Y. Xiong, S. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance-level discrimination," *arXiv preprint arXiv:1805.01978*, 2018. 3

[61] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 233–242. 5

[62] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama, "How does disagreement help generalization against label corruption?" in *International Conference on Machine Learning*. PMLR, 2019, pp. 7164–7173. 5

[63] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5171–5180. 5

[64] D. McAllester and K. Stratos, "Formal limitations on the measurement of mutual information," in *International Conference on Artificial Intelligence and Statistics*, 2020. 5

[65] F. H. N. H. L. Hörmander, N. S. B. Totaro, and A. V. M. Waldschmidt, "Grundlehren der mathematischen wissenschaften 332." Springer, 2006. 5

[66] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008. 6

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 6

[68] L. Jing and Y. Tian, "Self-supervised spatiotemporal feature learning by video geometric transformations," *arXiv preprint arXiv:1811.11387*, vol. 2, no. 7, p. 8, 2018. 7

[69] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang, "Self-supervised spatiotemporal learning via video clip order prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 334–10 343. 7

[70] T. Han, W. Xie, and A. Zisserman, "Video representation learning by dense predictive coding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0. 7

[71] C. Sun, F. Baradel, K. Murphy, and C. Schmid, "Learning video representations using contrastive bidirectional transformer," *arXiv preprint arXiv:1906.05743*, 2019. 7

[72] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 7

[73] Y. M. Asano, M. Patrick, C. Rupprecht, and A. Vedaldi, "Labelling unlabelled videos from scratch with multi-modal self-supervision," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 7

[74] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 7

[75] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459. 7

[76] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015. 7, 8

[77] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012. 7

[78] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563. 7

[79] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *International Conference on Learning Representations*, 2018. 8

[80] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864. 8

[81] B. Adhikari, Y. Zhang, N. Ramakrishnan, and B. A. Prakash, "Sub2vec: Feature learning for subgraphs," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2018, pp. 170–182. 8

[82] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal, "graph2vec: Learning distributed representations of graphs," *arXiv preprint arXiv:1707.05005*, 2017. 8

[83] Y. You, T. Chen, Y. Shen, and Z. Wang, "Graph contrastive learning automated," *arXiv preprint arXiv:2106.07594*, 2021. 8

[84] W. Falcon and K. Cho, "A framework for contrastive self-supervised learning and designing a new approach," *arXiv preprint arXiv:2009.00104*, 2020. 17

A. Theory and Proofs

A.1. Proof of Lemma 1: From RINCE to InfoNCE

We show that RINCE becomes asymptotically equivalent to InfoNCE when $q \to 0$. In particular, we prove the convergence of RINCE and its derivative in the limit of $q \to 0$.

Proof. We first prove the convergence in the function space with the L'Hôpital's rule:

$$\lim_{q \to 0} \mathcal{L}_{\text{RINCE}}^{\lambda,q}(\mathbf{s})$$

$$= \lim_{q \to 0} \frac{-e^{q \cdot s^+}}{q} + \frac{\left(\lambda \cdot (e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})\right)^q}{q}$$

$$= \lim_{q \to 0} \frac{1 - e^{q \cdot s^+}}{q} + \frac{-1 + \left(\lambda \cdot (e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})\right)^q}{q}$$

$$= \lim_{q \to 0} \frac{1 - e^{q \cdot s^+}}{q} + \lim_{q \to 0} \frac{-1 + \left(\lambda \cdot (e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})\right)^q}{q}$$

$$= -\log(e^{s^+}) + \log\left(\lambda(e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})\right) \qquad \text{(L'Hôpital's rule)}$$

$$= -\log \frac{e^{s^+}}{\lambda\left(e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}\right)}$$

$$= \mathcal{L}_{\text{InfoNCE}}(\mathbf{s}) + \log(\lambda).$$

To prove the convergence in its derivative, we analyze the derivative with respect to the positive score s^+ and the negative score s_i^-. We begin with RINCE:

$$\text{(positive score)} \qquad \lim_{q \to 0} \frac{\partial}{\partial s^+} \mathcal{L}_{\text{RINCE}}^{\lambda,q}(\mathbf{s})$$

$$= \lim_{q \to 0} \frac{\partial}{\partial s^+} \frac{-e^{q \cdot s^+}}{q} + \frac{\partial}{\partial s^+} \frac{\left(\lambda \cdot (e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})\right)^q}{q}$$

$$= \lim_{q \to 0} -e^{q \cdot s^+} + (\lambda \cdot (e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}))^{q-1} \cdot \lambda \cdot e^{s^+}$$

$$= -1 + \frac{e^{s^+}}{e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}};$$

$$\text{(negative score)} \qquad \lim_{q \to 0} \frac{\partial}{\partial s_i^-} \mathcal{L}_{\text{RINCE}}^{\lambda,q}(\mathbf{s})$$

$$= \lim_{q \to 0} \frac{\partial}{\partial s_i^-} \frac{\left(\lambda \cdot (e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})\right)^q}{q}$$

$$= \lim_{q \to 0} (\lambda \cdot (e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}))^{q-1} \cdot \lambda \cdot e^{s_i^-}$$

$$= \frac{e^{s_i^-}}{e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}}.$$

We can see that the derivatives match the ones of InfoNCE

(positive score)
$$\frac{\partial}{\partial s^+} \mathcal{L}_{\text{InfoNCE}}(\mathbf{s})$$

$$= \frac{\partial}{\partial s^+} - \log \frac{e^{s^+}}{e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}}$$

$$= -\frac{e^{s^+}}{e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}} \cdot \frac{(e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}) \cdot e^{s^+} - e^{2s^+}}{(e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})^2}$$

$$= -1 + \frac{e^{s^+}}{e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}};$$

(negative score)
$$\frac{\partial}{\partial s_i^-} \mathcal{L}_{\text{InfoNCE}}(\mathbf{s})$$

$$= -\frac{e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}}{e^{s_i^-}} \cdot \frac{-e^{s^+} \cdot e^{s_i^-}}{(e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})^2}$$

$$= \frac{e^{s_i^-}}{e^{s^+} + \sum_{i=1}^{K} e^{s_i^-}}.$$

□

A.2. Noisy Risk Bound for Exponential Loss

We justify the robustness of RINCE when $q = 1$ by extending Ghosh et al. [29]'s theorem to the exponential loss. The proof technique can be applied to other bounded symmetric classification losses.

Corollary 2. *Consider the setting of Ghosh et al. [29] and the exponential loss function $\mathcal{L}(s, y) = -ye^s$. Let $f_\eta^* = \arg\inf_{f \in \mathcal{F}} R_\mathcal{L}^\eta(f)$ be the minimizer of the noisy risk and $\epsilon = \inf_{f \in \mathcal{F}} R_\mathcal{L}(f)$ be the optimal risk. If $\eta_x \le \eta_{\max} < 0.5$ for all $x \in \mathcal{X}$. If the prediction score is bounded by $s_{\max}$, we have $R(f_\eta^*) \le (\epsilon + 2\eta_{\max} e^{s_{\max}})/(1 - 2\eta_{\max})$.*

Proof. Consider a binary classification loss with the following form:

$$\tilde{\mathcal{L}}_x(f(x), y) = B + \mathcal{L}_x(f(x), y) = B - y \cdot e^{f(x)} \ge 0,$$

where the prediction score $f(x)$ is bounded by $s_{\max} = \log(B)$. Note that the boundedness assumption holds for general representation learning on hypersphere, where the prediction score is the inner product between normalized feature vectors. Importantly, the loss satisfies

$$\tilde{\mathcal{L}}(f(x), 1) + \tilde{\mathcal{L}}(f(x), -1) = 2B.$$

By construction, the optimal risk takes the following value:

$$\inf_{f \in \mathcal{F}} R_{\tilde{\mathcal{L}}}(f) = \inf_{f \in \mathcal{F}} \mathbb{E}_{x \sim \mu}[\tilde{\mathcal{L}}(f(x), y_x)] = \epsilon + B := \tilde{\epsilon},$$

and $f^* = \arg\inf_{f \in \mathcal{F}} R_{\tilde{\mathcal{L}}}(f)$. Note that f^* is also a minimizer w.r.t. the original loss $\mathcal{L}$ ($f^* = \arg\inf_{f \in \mathcal{F}} R_\mathcal{L}(f)$), as an additive constant will not change the optimum solutions. Expanding the noisy risk gives

$$R_{\tilde{\mathcal{L}}}^\eta(f) = \mathbb{E}_{(x,y) \sim \mu}[(1 - \eta_x)\tilde{\mathcal{L}}(f(x), y_x) + \eta_x \tilde{\mathcal{L}}(f(x), -y_x)]$$
$$= \mathbb{E}_{x \sim \mu}[(1 - \eta_x)\tilde{\mathcal{L}}(f(x), y_x) + \eta_x(2B - \tilde{\mathcal{L}}(f(x), y_x))] \quad \text{(Symmetry)}$$
$$= \mathbb{E}_{x \sim \mu}[(1 - 2\eta_x)\tilde{\mathcal{L}}(f(x), y_x)] + 2B\mathbb{E}_{x \sim \mu}[\eta_x].$$

Let $f_\eta^* = \arg\inf R_\mathcal{L}^\eta(f_\eta^*) = \arg\inf R_{\tilde{\mathcal{L}}}^\eta(f_\eta^*)$, we have

$$R_{\tilde{\mathcal{L}}}^\eta(f^*) - R_{\tilde{\mathcal{L}}}^\eta(f_\eta^*) = \mathbb{E}_{x \sim \mu}[(1 - 2\eta_x)(\tilde{\mathcal{L}}(f^*(x), y_x) - \tilde{\mathcal{L}}(f_\eta^*(x), y_x))] \ge 0$$

14

since f_η^* is the minimizer of $R_{\tilde{\mathcal{L}}}^\eta$, which implies that

$$E_{x\sim\mu}[(1-2\eta_x)\tilde{\mathcal{L}}(f_\eta^*(x), y_x)] \leq E_{x\sim\mu}[(1-2\eta_x)\tilde{\mathcal{L}}(f^*(x), y_x)] \leq \tilde{\epsilon},$$

since $0 < 1 - 2\eta_x \leq 1$ by assumption. Let $\eta_{max} = \sup_{x\in\mathcal{X}}\eta_x$, we have

$$(1-2\eta_{\max})E_{x\sim\mu}[\tilde{\mathcal{L}}(f_\eta^*(x), y_x)] \leq \epsilon,$$

since the loss is non-negative, which implies

$$R_{\tilde{\mathcal{L}}}(f_\eta^*) \leq \frac{\tilde{\epsilon}}{1-2\eta_{\max}}.$$

Finally, we recover the original exponential loss without the additive term B. Plugging the form we have

$$B + R_{\mathcal{L}}(f_\eta^*) \leq \frac{\epsilon + B}{1 - 2\eta_{max}},$$

which implies

$$R_{\mathcal{L}}(f_\eta^*) \leq \frac{\epsilon + B}{1 - 2\eta_{max}} - B = \frac{\epsilon + 2B\eta_{max}}{1 - 2\eta_{max}}.$$

For exponential loss, setting B to $e^{s_{\max}}$ completes the proof. $\qquad\square$

For instance, when the noise level is 40%, we have $R_{\mathcal{L}}(f_\eta^*) \leq 5\epsilon + 4B$. Note that the prediction score is bounded by $1/t$ in our case as the representations are projected onto the unit hypersphere.

A.3. Lower bound of Wasserstein Distance

We now establish RINCE as a lower bound of WDM [33]. WDM is based on the Wasserstein distance, a distance metric between probability distributions defined via an optimal transport cost. Letting μ and $\nu \in \text{Prob}(\mathbb{R}^d \times \mathbb{R}^d)$ be two probability measures, we define the Wasserstein-1 distance with a Euclidean cost function as

$$\mathcal{W}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{\substack{(X,V)\\(X',V')}\sim\pi}\left[\|X - X'\| + \|V - V'\|\right],$$

where $\Pi(\mu, \nu)$ denotes the set couplings whose marginals are μ and ν, respectively. We are now ready to state our theorem.

Theorem 3. *If $\lambda K > 1 - \lambda$ and f projects the representation to a unit hypersphere, we have*

$$-\mathbb{E}\left[\mathcal{L}_{\text{RINCE}}^{\lambda,q=1}(s)\right] \leq \frac{\text{Lip}(f)\cdot(1-\lambda)\cdot e^{1/t}}{t}\mathcal{W}_1(P_{XV}^\phi, P_X^\phi P_V^\phi).$$

Proof. By the additivity of expectation, we can bound the negative symmetric loss as follows

$$-\mathbb{E}\left[\mathcal{L}_{\text{RINCE}}^{\lambda,q=1}(\mathbf{s})\right]$$

$$= \mathbb{E}_{\substack{x\sim P_X\\v\sim P_{V|X=x}\\v_i\sim P_V}}[(1-\lambda)e^{f(\phi(x))^T f(\phi(v)/t} - \lambda\sum_{i=1}^K e^{f(\phi(x))^T f(\phi(v_i)/t}]$$

$$= \mathbb{E}_{(x,v)\sim P_{XV}}\left[(1-\lambda)e^{f(\phi(x))^T f(\phi(v)/t}\right] - \mathbb{E}_{\substack{x\sim P_X\\v_i\sim P_V}}\left[\lambda\sum_{i=1}^K e^{f(\phi(x))^T f(\phi(v_i)/t}\right]$$

$$= \mathbb{E}_{(x,v)\sim P_{XV}}\left[(1-\lambda)e^{f(\phi(x))^T f(\phi(v)/t}\right] - \mathbb{E}_{\substack{x\sim P_X\\v_i\sim P_V}}\left[\lambda\sum_{i=1}^K e^{f(\phi(x))^T f(\phi(v_i)/t}\right]$$

$$= \mathbb{E}_{(x,v)\sim P_{XV}}\left[(1-\lambda)e^{f(\phi(x))^T f(\phi(v)/t}\right] - \lambda K\mathbb{E}_{\substack{x\sim P_X\\v\sim P_V}}\left[e^{f(\phi(x))^T f(\phi(v)/t}\right]$$

$$\leq (1-\lambda)\cdot(\mathbb{E}_{(x,v)\sim P_{XV}}\left[e^{f(\phi(x))^T f(\phi(v)/t}\right] - \mathbb{E}_{\substack{x\sim P_X\\v\sim P_V}}\left[e^{f(\phi(x))^T f(\phi(v)/t}\right]),$$

15

where the last equality follows by $\lambda K > 1 - \lambda$. Note that for $\frac{-1}{t} \leq s \leq \frac{1}{t}$, which implies $|\nabla_s e^s| \leq e^{1/t}$. Therefore, by the mean value theorem, we have

$$|e^{f(\phi(x))^T f(\phi(v))/t} - e^{f(\phi(x'))^T f(\phi(v'))/t}|$$

$$\leq \frac{e^{1/t}}{t}|\langle f(\phi(x)), f(\phi(v))\rangle - \langle f(\phi(x')), f(\phi(v'))\rangle| \qquad \text{(Mean Value Theorem)}$$

$$= \frac{e^{1/t}}{t}|\langle f(\phi(x)) - f(\phi(x')), f(\phi(v))\rangle + \langle f(\phi(x')), f(\phi(v) - f(\phi(v')))\rangle|$$

$$\leq \frac{e^{1/t}}{t}(|\langle f(\phi(x)) - f(\phi(x')), f(\phi(v))\rangle| + |\langle f(\phi(x')), f(\phi(v) - f(\phi(v')))\rangle|)$$

$$\leq \frac{e^{1/t}}{t}(\|f(\phi(x)) - f(\phi(x'))\|\|f(\phi(v))\| + \|f(\phi(v) - f(\phi(v')))\|\|f(\phi(x'))\|) \qquad \text{(Cauchy–Schwarz Ineq.)}$$

$$= \frac{e^{1/t}}{t}(\|f(\phi(x)) - f(\phi(x'))\| + \|f(\phi(v) - f(\phi(v')))\|) \qquad (f(\phi(x)) \text{ is unit norm})$$

$$\leq \frac{\text{Lip}(f) \cdot e^{1/t}}{t}(\|\phi(x) - \phi(x')\| + \|\phi(v) - \phi(v')\|)$$

$$= \frac{\text{Lip}(f) \cdot e^{1/t}}{t}d((\phi(x), \phi(v)), (\phi(x'), \phi(v'))).$$

We can see that the Lipschitz constant of $\exp(f(\cdot, \cdot))$ with respect to the metric d is bounded by $\frac{\text{Lip}(f) \cdot e^{1/t}}{t}$. Therefore, by Kantorovich-Rubinstein duality, we have

$$-\mathbb{E}\left[\mathcal{L}_{\text{RINCE}}^{\lambda, q=1}(\mathbf{s})\right]$$

$$\leq (1 - \lambda) \cdot (\mathbb{E}_{(x,v) \sim P_{XV}}[e^{f(\phi(x))^T f(\phi(v)/t}]$$

$$- \mathbb{E}_{\substack{x \sim P_X \\ v \sim P_V}}[e^{f(\phi(x))^T f(\phi(v)/t}]),$$

$$\leq \frac{\text{Lip}(f) \cdot (1 - \lambda) \cdot e^{1/t}}{t}\mathcal{W}_1(\phi_\# P_{XV}, \phi_\# P_X \cdot \phi_\# P_V)$$

$$\square$$

A.4. Noisy Wasserstein Dependency Measure

The result is a simple combination of Corollary 2 and Theorem 3. If $\lambda \geq \frac{\eta K - \eta + 1}{\eta K - \eta + 1 + K}$, by the assumption of additive noisy models and the symmetry of loss, we have

$$-\mathbb{E}\left[\mathcal{L}_{\text{RINCE}}^{\lambda, q=1}(\mathbf{s})\right]$$

$$= \mathbb{E}_{\substack{(x,v) \sim P_{XV}^\eta \\ v_i \sim P_V}}[(1 - \lambda)e^{f(\phi(x))^T f(\phi(v)/t} - \lambda \sum_{i=1}^K e^{f(\phi(x))^T f(\phi(v_i)/t}]$$

$$= \mathbb{E}_{(x,v) \sim P_{XV}^\eta}\left[(1 - \lambda)e^{f(\phi(x))^T f(\phi(v)/t}\right] - \mathbb{E}_{\substack{x \sim P_X \\ v_i \sim P_V}}\left[\lambda \sum_{i=1}^K e^{f(\phi(x))^T f(\phi(v_i)/t}\right]$$

$$= (1 - \lambda)(1 - \eta)\mathbb{E}_{(x,v) \sim P_{XV}}\left[e^{f(\phi(x))^T f(\phi(v)/t}\right] - K \cdot (\lambda - \eta + \eta\lambda)\mathbb{E}_{\substack{x \sim P_X \\ v \sim P_V}}\left[e^{f(\phi(x))^T f(\phi(v)/t}\right] \qquad \text{(symmetry)}$$

$$\leq (1 - \lambda)(1 - \eta) \cdot (\mathbb{E}_{(x,v) \sim P_{XV}}\left[e^{f(\phi(x))^T f(\phi(v)/t}\right] - \mathbb{E}_{\substack{x \sim P_X \\ v_i \sim P_V}}\left[e^{f(\phi(x))^T f(\phi(v_i)/t}\right]) \qquad (\lambda \geq \frac{\eta K - \eta + 1}{\eta K - \eta + 1 + K})$$

$$\leq (1 - \eta) \cdot \frac{\text{Lip}(f) \cdot (1 - \lambda) \cdot e^{1/t}}{t}\mathcal{W}_1(\phi_\# P_{XV}, \phi_\# P_X \cdot \phi_\# P_V)$$

$$= (1 - \eta) \cdot L \cdot I_{\mathcal{W}}(\phi(X), \phi(V)).$$

A.5. InfoNCE is not symmetric

Note that by taking the derivative with respect to the prediction score s, the definition is equivalent to $\frac{\partial \mathcal{L}(s,1)}{\partial s} + \frac{\partial \mathcal{L}(s,-1)}{\partial s} = 0 \ \forall s \in \mathbb{R}$.

$$\frac{\partial \mathcal{L}_{\text{InfoNCE}}(\mathbf{s})}{\partial s^+} = \frac{-1}{\mathcal{L}_{\text{InfoNCE}}(\mathbf{s})} \cdot \frac{e^{s^+} \cdot \sum_{i=1}^{K} e^{s_i^-}}{(e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})^2}$$

$$\frac{\partial \mathcal{L}_{\text{InfoNCE}}(\mathbf{s})}{\partial s_i^-} = \frac{-1}{\mathcal{L}_{\text{InfoNCE}}(\mathbf{s})} \cdot \frac{e^{s^+}(1 - e^{s_i^-}) + \sum_{i=1}^{K} e^{s_i^-}}{(e^{s^+} + \sum_{i=1}^{K} e^{s_i^-})^2}.$$

Within a batch of data, the gradients with respect to s^+ and s^- are entangled and do not sum to a constant, which fail to meet the symmetry condition.

B. Experiment Details

B.1. CIFAR-10

We follow the experiment setup in [11], where the SimCLR [6] models are trained with Adam optimizer for 500 epochs with learning rate 0.001 and weight decay 1e−6. The encoder is ResNet-50 and the dimension of the latent vector is 128. The temperature is set to $t = 0.5$. The models are then evaluated by training a linear classifier for 100 epochs with learning rate 0.001 and weight decay 1e−6. We use the PyTorch code in Figure 9 to generate the data augmentation noise.

```python
def get_train_transform(noise_rate):
    train_transform = transforms.Compose([
        transforms.RandomResizedCrop(32),
        transforms.RandomApply([transforms.RandomResizedCrop(32, scale=(0.2, 0.2))], p=noise_rate
    ),
        transforms.RandomHorizontalFlip(p=0.5),
        transforms.RandomApply([transforms.ColorJitter(0.4, 0.4, 0.4, 0.1)], p=0.8),
        transforms.RandomGrayscale(p=0.2),
        transforms.GaussianBlur(kernel_size=int(0.1*32)),
        transforms.ToTensor(),
        transforms.Normalize([0.4914, 0.4822, 0.4465], [0.2023, 0.1994, 0.2010])])
    return train_transform
```

Figure 9. PyTorch code for CIFAR-10 data augmentation noise.

B.2. ImageNet

SimCLR We adopt the SimCLR implementation[2] from PyTorch Lightning [84]. In addition, we spot a bug and fix the implementation of negative masking of PyTorch Lightning according to Figure 10 and achieve 68.9 top-1 accuracy on ImageNet (the one reported in the PyTorch Lightning's website is 68.4). To implement RINCE, we only modify the lines that calculates loss according to Figure 2.

Mocov3 We adopt the official code[3] from Mocov3 [35]. To implement RINCE, we only modify the lines that calculates loss in `moco/builder.py` according to Figure 2.

B.3. Kinetics-400

We adopt the official implementation[4] from [19]. Similarly, we only modify the loss function in the `criterions` directory. In particular, we use the SimCLR style implementation for both InfoNCE and RINCE loss. We also adopt the same hyperparameters described in the git repository for training. We set the learning rate to $1e − 3$ to finetune the models on downstream classification tasks such as UCF101 and HMDB51 with the provided evaluation code.

---

[2]https://github.com/PyTorchLightning/lightning-bolts/tree/master/pl_bolts/models/self_supervised/simclr
[3]https://github.com/facebookresearch/moco-v3
[4]https://github.com/facebookresearch/AVID-CMA

```
1   def compute_neg_mask(self):
2       total_images = self.num_nodes * self.gpus * self.batch_size * self.num_pos
3       world_size = self.num_nodes * self.gpus
4       batch_size = self.batch_size * self.num_pos
5       orig_images = self.batch_size
6       rank = int(os.environ["LOCAL_RANK"])
7
8       neg_mask = torch.zeros(batch_size, total_images)
9       all_indices = np.arange(total_images)
10      pos_members = orig_images * world_size * np.arange(self.num_pos)
11      for anchor in np.arange(self.num_pos):
12          for img_idx in range(orig_images):
13              delete_inds = orig_images * rank + img_idx + pos_members
14              neg_inds = torch.tensor(np.delete(all_indices, delete_inds)).long()
15              neg_mask[anchor * orig_images + img_idx, neg_inds] = 1
16      neg_mask = neg_mask.cuda(non_blocking=True)
17
18      return neg_mask
19
20  def nt_xent_loss(self, out_1, out_2, temperature):
21      if torch.distributed.is_available() and torch.distributed.is_initialized():
22          out_1_dist = SyncFunction.apply(out_1)
23          out_2_dist = SyncFunction.apply(out_2)
24      else:
25          out_1_dist = out_1
26          out_2_dist = out_2
27
28      out = torch.cat([out_1, out_2], dim=0)
29      out_dist = torch.cat([out_1_dist, out_2_dist], dim=0)
30
31      similarity = torch.exp(torch.mm(out, out_dist.t()) / temperature)
32
33      ################################ original code ################################
34      # # from each row, subtract e^(1/temp) to remove similarity measure for x1.x1        #
35      # neg = similarity.sum(dim=-1)                                                        #
36      # row_sub = Tensor(neg.shape).fill_(math.e ** (1 / temperature)).to(neg.device)       #
37      # neg = torch.clamp(neg - row_sub, min=eps)  # clamp for numerical stability          #
38      ################################################################################
39
40      neg_mask = self.compute_neg_mask()
41      neg = torch.sum(similarity * neg_mask, 1)
42
43      pos = torch.exp(torch.sum(out_1 * out_2, dim=-1) / temperature)
44      pos = torch.cat([pos, pos], dim=0)
45
46      loss = -(torch.mean(torch.log(pos / (pos + neg))))
47
48      return loss
```

Figure 10. PyTorch Lightening implementation of SimCLR. The original implementation of negative masking (commented out) is problematic because it subtracts $e^{1/t}$ to remove similarity measure for pairs that consist of the same images. However, subtracting a constant does not alter the gradient with respect to the model parameters. In particular, there are still gradients backpropagating through the false positive pairs. We fix it by directly filtering out those false pairs with a negative mask.

B.4. ACAV100M

We again modify the official implementation of [19] for the ACAV100M experiments, where we modify the data loader to adopt it to ACAV100M. Different from Kinetics-400 experiments, the input size is set to 8×224^2 during the finetuning process for computational efficiency. We again use the exact same set of hyperparameters from [19] for both training and testing.

B.5. TU-Dataset

We adopt the official implementation[5] from [40]. To implement RINCE, we only modify the loss in gsimclr.py file.

---

[5]https://github.com/Shen-Lab/GraphCL/tree/master/unsupervised_TU

C. Additional Results

C.1. Exact Number of CIFAR-10 and ACAV100M Experiments

We first provide the exact numbers for CIFAR-10 and ACAV100M experiments.

η	InfoNCE	$q = 0.01$	$q = 0.1$	$q = 0.5$	$q = 1.0$
0.0	93.4±0.2	93.4±0.2	93.2±0.1	93.3±0.1	93.0±0.2
0.2	93.1±0.1	93.3±0.3	93.0±0.1	93.2±0.2	92.9±0.3
0.4	90.7±0.2	93.0±0.2	92.0±0.9	93.1±0.1	92.8±0.1
0.6	88.2±0.4	90.8±0.2	90.6±0.3	92.9±0.2	92.4±0.2
0.8	87.1±0.5	89.1±0.2	89.3±0.1	89.9±0.3	91.6±0.3
1.0	87.1±1.0	88.7±0.1	89.3±0.4	89.3±0.6	88.2±0.3

Table 4. CIFAR-10 Label Noise

η	InfoNCE	$q = 0.01$	$q = 0.1$	$q = 0.5$	$q = 1.0$
0.0	91.1±0.1	91.6±0.1	91.5±0.1	91.8±0.2	90.7±0.1
0.2	89.3±0.1	89.8±0.2	89.7±0.1	90.4±0.1	90.9±0.1
0.4	87.3±0.4	87.7±0.5	87.5±0.2	88.8±0.1	89.0±0.1
0.6	84.5±0.2	85.4±0.2	85.3±0.2	86.6±0.1	86.3±0.2
0.8	80.6±0.1	81.2±0.2	80.3±0.2	82.5±0.2	82.8±0.3
1.0	71.0±0.5	71.2±0.6	71.8±0.4	71.5±0.3	72.7±0.2

Table 5. CIFAR-10 Augmentation Noise

model	20K	50K	100K	200K	500K
InfoNCE (100 epoch)	72.482	75.205	77.161	79.937	82.717
InfoNCE (150 epoch)	72.429	76.13	78.8	80.095	83.082
InfoNCE (200 epoch)	72.429	76.183	78.641	79.94	83.388
RINCE (100 epoch)	73.635	76.685	78.694	81.153	83.505
RINCE (150 epoch)	74.632	77.505	79.064	82.263	83.399
RINCE (200 epoch)	74.253	78.086	79.355	82.368	83.769

Table 6. Top1 accuracy on UCF101 of models trained on ACAV100M.

C.2. Positive Scores and Views, Continue

We extend our analysis of Figure 6 to InfoNCE baseline and discuss the impact of implicit weighting. We can see that the positive scores in both InfoNCE and RINCE models are correlated to the noisiness of positive pairs.

We then study the distribution of positive scores and compare the positive scores output by InfoNCE and RINCE on noisy views. As Figure 12 (a) shows, the positive scores of clean pairs output by RINCE is slightly higher, making the density of RINCE around score 1.0 larger than InfoNCE. Figure 12 (b) gives a closer look on scores versus noisy views. We can see that InfoNCE tends to output higher scores for noisy views than RINCE, corroborating our analysis: InfoNCE tends to maximize the positive score of hard (noisy) pairs. This inherently makes the positive scores of clean pairs lower for InfoNCE, explaining the discrepancy between InfoNCE and RINCE in (a).

C.3. Ablation Study on λ

Finally, we provide an ablation study on how λ affect the performance of RINCE with CIFAR-10 augmentation noise experiments. We can see that in both clean and noise setting, RINCE is not sensitive to the choice of λ as long as it is not too large. Therefore, we simply set $\lambda = 0.01$ for all vision experiments and $\lambda = 0.025$ for graph experiments.
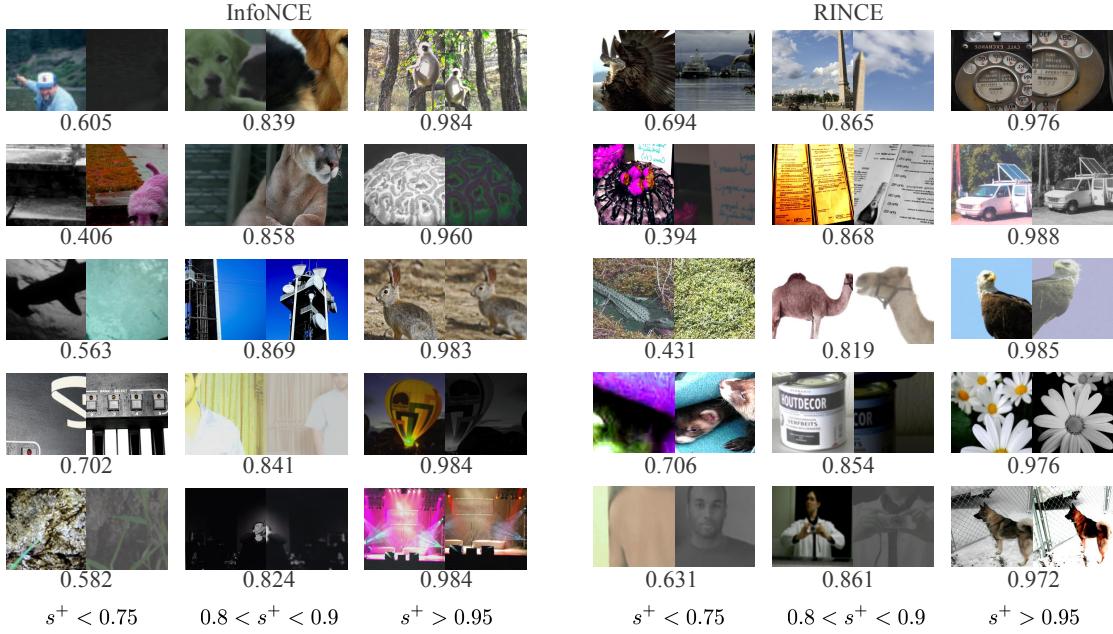
Figure 11. **Positive pairs and their scores.** The corresponding positive scores are shown below the image pairs. The positive scores $s^+ \in [-1, 1]$ are output by the trained InfoNCE and RINCE model (temperature = 1). Pairs that have lower scores are visually noisy, while informative pairs often have higher scores.

(a) Distribution of Positive Scores
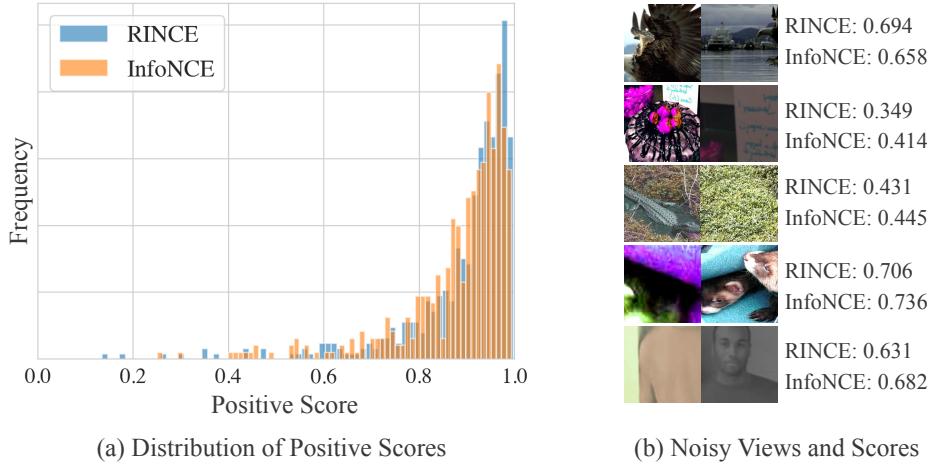
(b) Noisy Views and Scores

Figure 12. **Comparison between RINCE and InfoNCE.** (a) Distribution of Positive Scores for RINCE and InfoNCE; (b) InfoNCE outputs higher scores for noisy pairs.

Noise Rate	0.0	0.4
RINCE ($\lambda = 0.01$)	91.54	89.65
RINCE ($\lambda = 0.05$)	**91.81**	**89.81**
RINCE ($\lambda = 0.1$)	91.32	89.9
RINCE ($\lambda = 0.2$)	90.55	89.69
RINCE ($\lambda = 0.4$)	90.89	89.39

Table 7. CIFAR-10 Augmentation Noise