**To All Reviewers:** We thank all reviews for your insightful feedback and your appreciation of our $\text{MCR}^2$ formulation. We will incorporate suggestions on minor corrections, references, footnotes, and presentations in the final version.

**Why diverse intra-class representations?** This work aims to introduce a new objective (i.e., $\text{MCR}^2$) for learning representations not only discriminative between classes as with cross-entropy loss, but also *diverse* within class. We believe identifying more discriminative features lead to more reliable classification since the most discriminative feature may not be present in all samples. We rigorously prove that this can be achieved with the proposed $\text{MCR}^2$ loss function. Furthermore, we empirically demonstrate this objective can be used to train deep networks that have good properties in handling label noise (in supervised setting) and achieve SOTA for clustering (in unsupervised setting).

**Robustness to label noise:** The initial motivation of $\text{MCR}^2$ is to promote learning rich discriminative features. It is a nice surprise that so learned deep features are more robust than existing learning objectives including cross entropy and many others shown in Tables 1, 2. Unlike cross entropy that fits labels of individual samples, $\text{MCR}^2$ compresses samples of each class *collectively*. As mentioned in Section 4, given the compelling empirical evidence, a rigorous justification of the robustness is an exciting problem for future work.

Table 1: Comparison with OLE and Large Margin [EKM+18] on learning from noisy labels.

| ResNet18 | Ratio=0.1 | Ratio=0.2 | Ratio=0.3 | Ratio=0.4 | Ratio=0.5 |
|---|---|---|---|---|---|
| OLE | 91.04% | 86.01% | 80.69% | 71.79% | 61.06% |
| [EKM+18] | 90.10% | 87.42% | 83.77% | 78.51% | 72.48% |
| $\text{MCR}^2$ | **91.16%** | **89.70%** | **88.18%** | **86.66%** | **84.30%** |

Table 2: Comparison with Trimmed Loss [SS19] on learning from noisy labels.

| WRN16 | Ratio=0.1 | Ratio=0.3 | Ratio=0.5 | Ratio=0.7 |
|---|---|---|---|---|
| [SS19] | 90.33% | 88.23% | 82.51% | 64.74% |
| $\text{MCR}^2$ | **91.55%** | **88.81%** | **84.25%** | **67.09%** |

**To Reviewer #1:** Please refer to the top of the rebuttal for the motivations of larger intra-class subspace in $\text{MCR}^2$.

**Q1:** *Compare with OLE: " 1). It is not clear why a larger . . . these connections are not crystal clear in the paper. 2). Does the OLE type loss have the same property as Theorem 1? 3). The authors should show more comparison . . . OLE."*

**A1:** 1). We will make these connections more clear in the final version; 2). As mentioned in the paper (line 209-213), OLE loss does *not* have the diversity property of $\text{MCR}^2$ given in Theorem 1; 3). In Table 1, we compare $\text{MCR}^2$ with OLE on the corrupted label task using the same network architecture. $\text{MCR}^2$ achieves significantly better performance.

**Q2:** *Gaussian assumption of data: "I have a concern whether the rate distortion function . . . to be self-contained."*

**A2:** Thank you for your suggestion. As shown in [MDHW07], the rate distortion function can serve as a tight and accurate approximation for a wide range of subspace-like distributions. We will give more details in the final version.

**Q3:** *The paper can be considered as applying existing objective/criterion . . . into learning of deep features.*

**A3:** We disagree. Our $\text{MCR}^2$ objective is new and is different from those in previous works such as OLE. To our best knowledge, $\text{MCR}^2$ is *the first* objective theoretically shown to guarantee both diverse and discriminative properties.

**Q4:** *In fact, a core problem in understanding deep learning . . . on it.*

**A4:** Thanks for your comment. Precisely, we believe identifying a diverse and discriminative representation from the data is an important step to gaining better understandings of the generalizability and robustness of deep learning.

**To Reviewer #2:**

**Q1:** *Relationship with information bottleneck (IB) framework: " In Section 2, the authors seek to . . . Gaussians."*

**A1:** Both $\text{MCR}^2$ and IB are information-theoretic objectives. However, the goal of IB is to find a *minimal* set of most informative representations while $\text{MCR}^2$ aims to capture both diverse and discriminative representations, which is very different. We will better clarify relationships with mutual information-based approaches in our final version.

**Q2:** *Related work on label noise: "The label noise robustness experiments . . . iterative trimmed loss minimization [1]."*

**A2:** In Table 2, we compare $\text{MCR}^2$ with [SS19] using the same network. $\text{MCR}^2$ achieves better performance *without* any noise ratio dependent hyperparameters as required by [SS19]. We will add the comparison in the final version.

**To Reviewer #3:** Please refer to the top of the rebuttal for clarifying the objectives and motivations of our work.

**Q1:** *"My main concern is that, I don't see the benefits . . . lie on a union of subspaces)."*

**A1:** First of all, we do *not* model the original data by subspaces. $\text{MCR}^2$ can guide a deep network to map real data on complicated nonlinear submanifolds to a union of orthogonal subspaces. Secondly, once the subspaces are learned, the nearest subspace classification is computationally *efficient*. Finally, compared with *hidden* representations learned by cross-entropy, the union of discriminative subspaces learned by $\text{MCR}^2$ is geometrically and statistically meaningful.

**Q2:** *"While the theoretical analysis reveals interesting properties . . . the loss, see e.g. [ZF14]."*

**A2:** Our theoretical analysis reveals that the proposed $\text{MCR}^2$ is optimized only when features are the most diverse and discriminative. Our experiments have clearly shown that using $\text{MCR}^2$, deep features learned from real data such as CIFAR10 have the same nice properties that are predicted by our theoretical results. We plan to rigorously justify this phenomenon by studying the interplay of the $\text{MCR}^2$ objective and the choice of network architectures in future work.

**Q3:** *Related work on clustering: "While there is no dedicated related work . . . be included and compared against."*

**A3:** $\text{MCR}^2$ outperforms [HMT+17, JHV19] on both CIFAR10 and CIFAR100 by a large margin. For STL10, [HMT+17] applied pretrained ImageNet models and $\text{MCR}^2$ outperforms [JHV19] when using the same amount of training data.

**Answers to minor comments:** We will add the above comparison and references, and also compared $\text{MCR}^2$ to [EKM+18] on CIAFR10 with label noise (see Table 1). We did not encounter any computation issue when dealing with $\log\det$ and the optimization is stable. The $\boldsymbol{\Pi}$ is defined by the labels and satisfies the simplex constraint (footnote 15).

**To Reviewer #4:** Please refer to the top of the rebuttal for the question regarding the robustness of $\text{MCR}^2$.

**Q1:** *Applying the MCR reduction to the large-scale dataset seems computationally very hard.*

**A1:** The computation only increases *linearly* in the number of classes for the supervised learning setting.