# Distance Metric Learning with Joint Representation Diversification

**Xu Chu** [1 2] **Yang Lin** [1 2] **Yasha Wang** [2 3] **Xiting Wang** [4] **Hailong Yu** [1 2] **Xin Gao** [1 2] **Qi Tong** [2 5]

## Abstract

Distance metric learning (DML) is to learn a representation space equipped with a metric, such that similar examples are closer than dissimilar examples concerning the metric. The recent success of DNNs motivates many DML losses that encourage the intra-class compactness and inter-class separability. The trade-off between inter-class compactness and inter-class separability shapes the DML representation space by determining how much information of the original inputs to retain. In this paper, we propose a Distance Metric Learning with Joint Representation Diversification (JRD) that allows a better balancing point between intra-class compactness and inter-class separability. Specifically, we propose a Joint Representation Similarity regularizer that captures different abstract levels of invariant features and diversifies the joint distributions of representations across multiple layers. Experiments on three deep DML benchmark datasets demonstrate the effectiveness of the proposed approach.

## 1. Introduction

Distance metric learning (DML) is a class of approaches that learns a mapping from original high-dimensional feature space to a compact representation space where the metric directly corresponds to a measure of semantic similarity. With proper training, the learned mapping could generalize to unseen classes so that DML has been found especially useful for visual open-set classification tasks such as zero-shot learning (Weston et al., 2011; Frome et al., 2013), retrieval (Zhou et al., 2004; Yang et al., 2018), and face recognition (Chopra et al., 2005; Schroff et al., 2015).

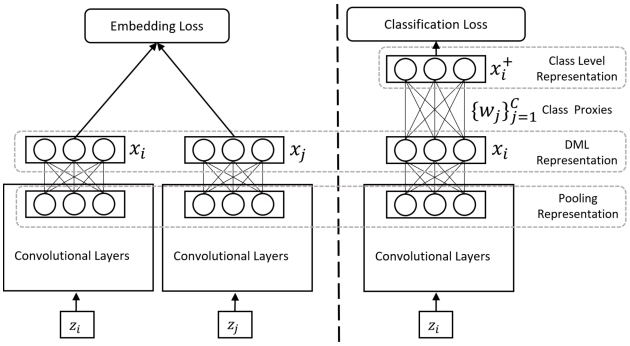


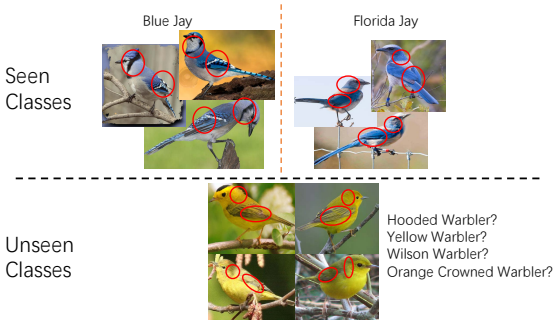

*Figure 1.* The architectures of embedding loss methods and classification loss methods. The embedding losses measure the intra-class compactness by computing intra-class distance $d(x_i, x_j)$ between DML representations $x_i$ and $x_j$. The classification losses measure the intra-class compactness by computing distance $d(x_i, w_{y_i})$ between DML representation $x_i$ and the correct class proxy $w_{y_i}$. The explicit penalization on intra-class distances of DML representations in embedding losses imposes stronger restrictions on intra-class compactness than classification losses.

In the conventional DML methods (Xing et al., 2003; Weinberger et al., 2006; Davis et al., 2007), examples are represented by hand-crafted features in $\mathbb{R}^D$. For similar examples represented by $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^D$, one commonly used mapping is parameterized by a projection matrix $\mathbf{P} \in \mathbb{R}^{d \times D}$ such that $||\mathbf{P}\mathbf{x}_j - \mathbf{P}\mathbf{x}_j||_2$ is minimized under certain regularization constraints (Xie, 2015). Recently, the success of deep neural networks (DNNs) (Krizhevsky et al., 2012) motivates more data-driven deep DML methods that learn representations by DNNs from raw data, leading to a substantial boost in classification performance of visual tasks (Schroff et al., 2015). In general, there are two goals of deep DML optimization objectives: 1) the inner-class compactness and 2) the inter-class separability. The trade-off between inter-class compactness and the inter-class separability shapes the DML representation space by determining how much information of the original inputs to retain.

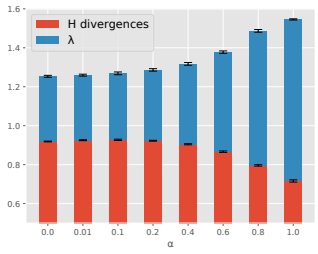
There are mainly two categories of deep DML loss functions: 1) the embedding losses (Chopra et al., 2005; Schroff

[1]School of Electronics Engineering and Computer Science, Peking University, Beijing, China [2]Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing, China [3]National Engineering Research Center of Software Engineering, Peking University, Beijing, China [4]Microsoft Research Asia, Beijing, China [5]School of Software and Microelectronics, Peking University, Beijing, China. Correspondence to: Yasha Wang <wangyasha@pku.edu.cn>.

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Online, PMLR 119, 2020. 




(a)




(b)

*Figure 2.* Hidden limitation of explicit penalization on intra-class distances. (a) The information discriminating seen classes may not suffice to discriminate unseen classes. (b) $\mathcal{H}$-divergence and $\lambda$ (with $95\%$ confidence bands) w.r.t. magnitude $\alpha$ of penalization on intra-class distances of representations for nearest neighbor classifiers. *Best viewed in color.*

et al., 2015; Sohn, 2016; Oh Song et al., 2016; Wang et al., 2019a), and 2) the classification losses (Liu et al., 2016; Movshovitz-Attias et al., 2017; Qian et al., 2019). The architectures are shown in Figure 1. A major difference between these two categories of losses is how they measure the intra-class compactness. The embedding losses measure the intra-class compactness by computing intra-class distance $\mathbf{d}(\mathbf{x}_i, \mathbf{x}_j)$ between DML representations $\mathbf{x}_i$ and $\mathbf{x}_j$. While the classification losses take advantage of the hierarchical representations of DNNs by learning class proxies $\{\mathbf{w}_j\}_{j=1}^{C}$ for all seen classes, and measure the intra-class compactness by computing distance $\mathbf{d}(\mathbf{x}_i, \mathbf{w}_{y_i})$ between DML representation $\mathbf{x}_i$ and the correct class proxy $\mathbf{w}_{y_i}$. The explicit penalization on intra-class distances of DML representations in embedding losses imposes stronger restrictions on intra-class compactness than classification losses.

The stronger restrictions might result in an overemphasizing on intra-class compactness such that the DNNs filter out information that contributes to discriminating examples of unseen classes. The generalization error bound on DML representations of unseen classes in a hypothesis space, can be explicitly estimated by the sum of three components (Ben-David et al., 2010): a) the error on representations of seen classes, b) the $\mathcal{H}$-divergence between representations of seen classes and unseen classes and c) the error of an ideal joint hypothesis $\lambda$ on the representation of seen and unseen classes. A preliminary empirical investigation in Figure 2 demonstrates that additional explicit penalization on intra-class distances of representations for AMSoftmax loss (Wang et al., 2018) (a representative classification loss), would inflate the error of an ideal joint hypothesis acutely, implying explicit penalization on intra-class distances of representations is error-prone for the high tendency of overemphasizing intra-class compactness. The above observation also motivates a better balancing point between intra-class compactness and inter-class separability by putting more emphasis on inter-class separability, which allows more flexible information retaining.

In the context of representation learning, there is a distinctive advantage of DNNs: deep architectures can lead to progressively more abstract features at higher layers (Bengio et al., 2013). Each layer captures a different abstraction level of concepts and is invariant to different level of changes in the input. Hence encouraging the separability of less abstract concepts constructing the DML representation, as well as encouraging the separability of more abstract concepts constructed by the DML representation, are beneficial to enhance the inter-class separability of the DML representation by capturing different abstract levels of invariant features.

In this paper, we propose a Distance Metric Learning with Joint Representation Diversification (JRD). In particular, we propose an easy-to-compute Joint Representation Similarity (JRS) regularizer, which penalizes inter-class distributional similarities of hidden activations in multiple layers, capturing different abstract levels of invariant features, and therefore promotes diversification of DML representations. Specifically, the JRS regularizer computes inner products between covariance-operator-based joint distribution embeddings (Baker, 1973; Fukumizu et al., 2004) in a tensor product space induced by kernel mean embeddings (Smola et al., 2007) in reproducing kernel Hilbert spaces (RKHSs). We give an interpretation of JRS with translation invariant kernels in terms of inner products of characteristic functions in $L^2$ spaces.The JRS regularizer promotes the AMSoftmax loss by learning representations at a better balancing point of intra-class compactness and inter-class separability, allowing more flexible information retention. Experiments on three challenging fine-grained deep DML benchmark datasets (Cub-200-2011 (Wah et al., 2011), Cars-196 (Krause et al., 2013), and Stanford Online Products (Oh Song et al., 2016)) show that the proposed JRD model is competitive to the state-of-the-art methods.

## 2. Motivation

In DML, a mapping from the original feature space to a DML representation space is trained by instances of seen

classes. The learned mapping can generate DML representations for seen class instances, denoted by $\mathcal{D}_s$, as well as representations for unseen class instances, denoted by $\mathcal{D}_u$. The generalization error bound on $\mathcal{D}_u$ in a hypothesis space $\mathcal{H}$ is a natural metric to quantify the generalizability of the DML representation. The following result (Ben-David et al., 2010) provides an estimate for the error bound.

**Theorem 1** (Ben-David et al. (2010))**.** *Let $\mathcal{H}$ be a hypothesis space. Denote by $\epsilon_s$ and $\epsilon_u$ the generalization errors on $\mathcal{D}_s$ and $\mathcal{D}_u$, then for every $h \in \mathcal{H}$:*

$$\epsilon_u(h) \leq \epsilon_s(h) + \hat{d}_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_u) + \lambda. \tag{1}$$

*The $\hat{d}_{\mathcal{H}}$ in (1) is the $\mathcal{H}$-divergence measuring a sense of distance between $\mathcal{D}_s$ and $\mathcal{D}_u$ in $\mathcal{H}$, specifically,*

$$\hat{d}_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_u) = 1 - \min_{h \in \mathcal{H}} \epsilon^*(h), \tag{2}$$

*where $\epsilon^*(h)$ is the error of a binary classifier $h \in \mathcal{H}$ that discriminates between instances from $\mathcal{D}_s$ and $\mathcal{D}_u$.*

*The $\lambda$ in (1) is the error of an ideal joint hypothesis $h^* = \arg\min_{h \in \mathcal{H}} \epsilon_s(h) + \epsilon_u(h)$, that is,*

$$\lambda = \epsilon_s\left(h^*\right) + \epsilon_u\left(h^*\right). \tag{3}$$

We empirically investigate the influence of additional explicit penalization on intra-class distances of representations for AMSotmax (classification) loss (Wang et al., 2018), by examining the $\mathcal{H}$-divergence and $\lambda$ w.r.t. different magnitude of additional penalizations with the CUB-200-2011 dataset (Wah et al., 2011). We first train the DML mapping with a DNN (detailed setting in 5.1) by the loss below,

$$\mathcal{L}_{AMS} - \alpha \sum_I \frac{1}{N^I_{pairs}} \sum_{\mathbf{x}^I_i, \mathbf{x}^I_j \in \mathcal{T}_I} e^{-\frac{1}{2}(\mathbf{x}^I_i - \mathbf{x}^I_j)^2}, \tag{4}$$

where $\mathcal{L}_{AMS}$ is the AMSoftmax loss (12), $\mathcal{T}_I$ is the set of examples from class $I$, $N^I_{pairs}$ is the number of different sample pairs in $\mathcal{T}_I$, and $\alpha > 0$ controls the magnitude of penalization on intra-class distances. Then we generate DML representations $\{\mathcal{D}^\alpha_s, \mathcal{D}^\alpha_u\}$ with the trained DML mapping. With a prespecified labeled and unlabeled splits of $\mathcal{D}^\alpha_s$ and $\mathcal{D}^\alpha_u$ respectively, the $\mathcal{H}$-divergence is computed by a binary classifier discriminating instances from $\mathcal{D}^\alpha_s$ and $\mathcal{D}^\alpha_u$. The ideal joint hypothesis $h^*$ is found by a multi-class classifier on both labeled data of $\mathcal{D}^\alpha_s$ and $\mathcal{D}^\alpha_u$ to compute $\lambda$. Both classifiers are fitted on the labeled splits, following practice Liu et al. (2019). The above process is replicated several times with different values of $\alpha$. For DML, we consider nearest neighbor classifiers as the hypothesis space. The results are shown in Figure 2. As the magnitude of penalization on intra-class distances of representations $\alpha$ grows, the $\lambda$ gets inflated more acutely than the declining tendency of $\mathcal{H}$-divergence, suggesting a larger error bound.

For open-set classification tasks, the information used to discriminate seen classes may not totally coincide with the information used to discriminate unseen classes. Explicitly penalizing intra-class distances of representations violates the Maximum entropy principle (MEP) (Jaynes, 1957). Additional explicit penalization on intra-class distances might cause the DNNs to filter out necessary information for discriminating unseen classes. As a result, the DML representation $\mathcal{D}_s$ and $\mathcal{D}_u$ may tend to contain information only about seen classes. Though $\mathcal{H}$-divergence might decline, the error of an ideal joint hypothesis $\lambda$ is likely to get inflated acutely, leading to a worse generalization error bound. This observation motivates a better balancing point between intra-class compactness and inter-class separability by putting more emphasis on inter-class separability with a *a diversified DML representation*, which allows more flexible information retaining. To promote a diversified DML representation, a natural choice is encouraging the inter-class distributions to be dissimilar. In particular, we take advantage of the hierarchical representations of DNNs and propose the joint representation similarity (JRS) regularizer.

## 3. Preliminary

To measure the similarity between distributions, we seek help from their representers in the reproducing kernel spaces (RKHSs). The kernel embeddings of distributions is an effective and powerful tool to study marginal distributions and joint distributions in RKHSs for many applications such as independent component analysis (Bach & Jordan, 2002), two-sample test (Gretton et al., 2012), domain adaption (Long et al., 2015; 2017) and generative modeling (Li et al., 2015; 2017). For better clarification of the proposed approach, we review the necessary preliminaries.

**Definition 1** (reproducing kernel Hilbert space)**.** *A Hilbert space $\mathcal{RKHS}$ of real-valued functions on non-empty $\mathcal{X}$ is a reproducing kernel Hilbert space if there is a symmetric and positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that*

- $\forall \boldsymbol{x} \in \mathcal{X}$, $k(\cdot, \boldsymbol{x}) \in \mathcal{RKHS}$,
- *(The reproducing property)* $\forall \boldsymbol{x} \in \mathcal{X}$, $\forall f \in \mathcal{RKHS}$, $\langle f(\cdot), k(\cdot, \boldsymbol{x}) \rangle_{\mathcal{RKHS}} = f(\boldsymbol{x})$.

*In particular, $k$ is the reproducing kernel of $\mathcal{RKHS}$, and for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$, $k(\boldsymbol{x}, \boldsymbol{y}) = \langle k(\cdot, \boldsymbol{x}), k(\cdot, \boldsymbol{y}) \rangle_{\mathcal{RKHS}}$.*

We can define a representer $\mu_{\mathbb{P}}$ in $\mathcal{RKHS}$ of a probability measure $\mathbb{P}$. The existence of $\mu_{\mathbb{P}}$ is assured if $\mathbb{E}_{\mathbf{X} \sim \mathbb{P}}[\sqrt{k(\mathbf{X}, \mathbf{X})}] < \infty$ (Smola et al., 2007). Formally,

**Definition 2** (kernel mean embedding, (Smola et al., 2007; Berlinet & Thomas-Agnan, 2011))**.** *Let $M^1_+(\mathcal{X})$ be the space of all probability measures $\mathbb{P}$ on a measurable space $(\mathcal{X}, \Sigma)$. The kernel mean embedding of probability measures into $\mathcal{RKHS}$ is defined by the mapping* $\mu : M^1_+(\mathcal{X}) \longrightarrow \mathcal{RKHS}, \quad \mathbb{P} \longmapsto \int k(\cdot, \boldsymbol{x}) \mathrm{d}\mathbb{P}(\boldsymbol{x}) \triangleq \mu_{\mathbb{P}}$.

When there are multiple measurable spaces $\{\mathcal{X}^1, \ldots, \mathcal{X}^L\}$, the kernel embedding in Definition 2 defined on one measurable space can be generalized to joint distributions by defining cross-covariance operators on RKHSs (Baker, 1973). Let $\mathbf{X}^{1:L} = (\mathbf{X}^1, \ldots, \mathbf{X}^L)$ be a random variable taking values on Cartesian product $\times_{l=1}^L \mathcal{X}^l = \mathcal{X}^1 \times \cdots \times \mathcal{X}^L$. For $l \in \{1, \cdots, L\}$, let $(\mathcal{RKHS}^l, k^l)$ be the RKHS with a measurable kernel on $\mathcal{X}^l$ respectively. Throughout, we assume the integrability $\mathbb{E}_{\mathbf{X}^l}[k^l(\mathbf{X}, \mathbf{X})] < \infty$ which ensures that $\mathcal{RKHS}^l \subset L^2(\mathbb{P}_{\mathbf{X}^l})$.

**Definition 3** (cross-covariance operator, (Baker, 1973; Fukumizu et al., 2004)). *Suppose that $M_+^1(\times_{l=1}^L \mathcal{X}^l)$ is the space of all probability measures $\mathbb{P}$ on $\times_{l=1}^L \mathcal{X}^l$. The (uncentered) covariance operator of measures in $M_+^1(\times_{l=1}^L \mathcal{X}^l)$ into a tensor product space $\otimes_{l=1}^L \mathcal{RKHS}^l = \mathcal{RKHS}^1 \otimes \cdots \otimes \mathcal{RKHS}^L$ is defined by the mapping, $\mathcal{C}_{\boldsymbol{X}^{1:L}} : M_+^1(\times_{l=1}^L \mathcal{X}^l) \longrightarrow \otimes_{l=1}^L \mathcal{RKHS}^l$, $\mathbb{P} \mapsto \int_{\times_{l=1}^L \boldsymbol{X}^l} (\otimes_{l=1}^L k^l(\cdot, \boldsymbol{x}^l)) \mathrm{d}\mathbb{P}(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^L) \triangleq \mathcal{C}_{\boldsymbol{X}^{1:L}}(\mathbb{P})$.*

## 4. Proposed Approach

**Problem Formulation**: Assume that the training set of size $N$ is represented by $\{\mathbf{z}_i, y_i\}_{i=1}^N$, in which $\mathbf{z}_i$ and $y_i \in \{1, \ldots, C\}$ denote the training instance and its class label from population $(\mathbf{Z}_{train}, \mathbf{Y}_{train})$. Let $f_\theta$ be the neural network that accounts for the DML mapping. Denote by $\mathbf{x}_i = f_\theta(\mathbf{z}_i)$ the DML representation of some instance $\mathbf{z}_i$. The goal of DML is learning a $\hat{\theta}$ by $\{\mathbf{z}_i, y_i\}_{i=1}^N$ such that for any testing examples $(\mathbf{z}_s, y_0)$,$(\mathbf{z}'_s, y_0)$ and $(\mathbf{z}_t, y_1)$, possibly $\{y_0, y_1\} \cap \{1, \ldots, C\} = \emptyset$, if $y_0 \neq y_1$, then

$$\begin{aligned} \mathbf{d}(\mathbf{x}_s, \mathbf{x}'_s) &= \mathbf{d}(f_{\hat{\theta}}(\mathbf{z}_s), f_{\hat{\theta}}(\mathbf{z}'_s)) \\ &< \mathbf{d}(f_{\hat{\theta}}(\mathbf{z}_s), f_{\hat{\theta}}(\mathbf{z}_t)) = \mathbf{d}(\mathbf{x}_s, \mathbf{x}_t). \end{aligned} \tag{5}$$

The prespecified metric $\mathbf{d}(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+ \cup \{0\}$ is usually taken as the euclidean distance or the chord distance induced by cosine similarity: $\mathbf{d}(\mathbf{s}, \mathbf{t}) = 2(1 - \frac{\mathbf{s} \cdot \mathbf{t}}{\|\mathbf{s}\|_{\ell^2} \|\mathbf{t}\|_{\ell^2}})$.[1]

**Framework Overview**: We first give an overview of the proposed approach in Figure 3. Denote by $\mathbf{X}^I$ the *random variable* of the class $I$ DML representation $\mathbf{X}|\mathbf{Y} = I$. Following standard deep DML practice Schroff et al. (2015), we start by extracting a pooling representation $\mathbf{X}^-$ using a CNN. The DML representation $\mathbf{X}$ is then obtained by an affine transform of the pooling representation $\mathbf{X}^-$. Thus $\mathbf{X}$ can be viewed as a representation capturing more abstract concepts as a linear combination of lower-level concepts. In addition, a classification loss computes a class level representation vector $\mathbf{X}^+ \in \mathbb{R}^C$ with each entry denoting the tendency of an input instance falling into a particular class. This class level representation is capturing even more abstract concepts. For two different class $I$ and $J$, the proposed joint representation similarity (JRS) regularizer promotes the inter-class separability between $\mathbf{X}^I$ and $\mathbf{X}^J$ by penalizing the similarity between probability measures $\mathbb{P}(\mathbf{X}^{I-}, \mathbf{X}^I, \mathbf{X}^{I+})$ and $\mathbb{Q}(\mathbf{X}^{J-}, \mathbf{X}^J, \mathbf{X}^{J+})$.

[1] If $\|\mathbf{s}\|_{\ell^2} = \|\mathbf{t}\|_{\ell^2} = 1$, the Euclidean distance and the chord distance are equivalent.

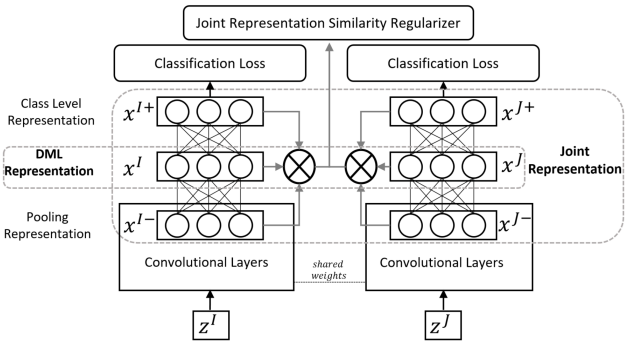



*Figure 3.* The architecture of distance metric learning with joint representation diversification (JRD). Denote by $\mathbf{x}$ and $\mathbf{z}$ the instantiations of random variable $\mathbf{X}$ and $\mathbf{Z}$. For class $I$ and $J$, a joint representation similarity regularizer penalizes the similarity between probability measures $\mathbb{P}(\mathbf{X}^{I-}, \mathbf{X}^I, \mathbf{X}^{I+})$ and $\mathbb{Q}(\mathbf{X}^{J-}, \mathbf{X}^J, \mathbf{X}^{J+})$.

### 4.1. Joint Representation Similarity Regularizer

A natural choice to measure the similarity between two joint representations $\mathbb{P}$ and $\mathbb{Q}$ is by measuring the similarity of their kernel embeddings defined in Definition 3. Formally,

**Definition 4** (Joint Representation Similarity). *Suppose that $\mathbb{P}(\boldsymbol{X}^1, \ldots, \boldsymbol{X}^L)$ and $\mathbb{Q}(\boldsymbol{X}'^1, \ldots, \boldsymbol{X}'^L)$ are probability measures on $\times_{l=1}^L \mathcal{X}^l$. Given $L$ reproducing kernels $\{k^l\}_{l=1}^L$, the joint representation similarity between $\mathbb{P}$ and $\mathbb{Q}$ is defined as the inner product of $\mathcal{C}_{\boldsymbol{X}^{1:L}}(\mathbb{P})$ and $\mathcal{C}_{\boldsymbol{X}'^{1:L}}(\mathbb{Q})$ in $\otimes_{l=1}^L \mathcal{RKHS}^l$, i.e.,*

$$\mathcal{S}_{JRS}(\mathbb{P}, \mathbb{Q}) \triangleq \langle \mathcal{C}_{\boldsymbol{X}^{1:L}}(\mathbb{P}), \mathcal{C}_{\boldsymbol{X}'^{1:L}}(\mathbb{Q}) \rangle_{\otimes_{l=1}^L \mathcal{RKHS}^l} \tag{6}$$

Usually the true distribution $\mathbb{P}(\mathbf{X}^1, \ldots, \mathbf{X}^L)$ is unknown, in practice we can estimate $\mathcal{C}_{\mathbf{X}^{1:L}}(\mathbb{P}) = \int_{\times_{l=1}^L \mathbf{X}^l} (\otimes_{l=1}^L k^l(\cdot, \mathbf{x}^l)) \mathrm{d}\mathbb{P}(\mathbf{x}^1, \ldots, \mathbf{x}^L)$ using a finite sample $\{\mathbf{x}_i^{1:L}\}_{i=1}^n \sim \mathbb{P}$ (Song & Dai, 2013),

$$\widehat{\mathcal{C}_{\mathbf{X}^{1:L}}}(\mathbb{P}) = \frac{1}{n} \sum_{i=1}^n \otimes_{l=1}^L k^l(\cdot, \mathbf{x}_i^l). \tag{7}$$

This empirical estimate converges to its population counterpart with a rate of $O(n^{-\frac{1}{2}})$ (Berlinet & Thomas-Agnan, 2011; Muandet et al., 2017), as a consequence of the $\sqrt{n}$-consistency of the empirical kernel mean embedding (Lopez-Paz et al., 2015).

The reproducing property gratifies a rather easy-to-compute empirical estimate of $\mathcal{S}_{JRS}(\mathbb{P}, \mathbb{Q})$ using two sets of samples

$$\{\mathbf{x}_i^{1:L}\}_{i=1}^n \sim \mathbb{P} \text{ and } \{\mathbf{x}_j'^{1:L}\}_{j=1}^{n'} \sim \mathbb{Q},$$

$$\begin{aligned}\widehat{\mathcal{S}_{JRS}}(\mathbb{P},\mathbb{Q}) &\triangleq \langle \widehat{\mathcal{C}_{\mathbf{X}^{1:L}}}(\mathbb{P}), \widehat{\mathcal{C}_{\mathbf{X}'^{1:L}}}(\mathbb{Q})\rangle_{\otimes_{l=1}^L \mathcal{RKHS}^l} \\ &= \frac{1}{nn'}\sum_{i=1}^n\sum_{j=1}^{n'}\langle \otimes_{l=1}^L k^l(\cdot,\mathbf{x}_i^l), \otimes_{l=1}^L k^l(\cdot,\mathbf{x}_j'^l)\rangle_{\otimes_{l=1}^L \mathcal{RKHS}^l} \\ &= \frac{1}{nn'}\sum_{i=1}^n\sum_{j=1}^{n'}\prod_{l=1}^L k^l(\mathbf{x}_i^l,\mathbf{x}_j'^l).\end{aligned} \tag{8}$$

With the definition of joint representation similarity and its empirical estimate, we next introduce the construction of the JRS regularizer in terms of pooling representation $\mathbf{X}^-$, DML representation $\mathbf{X}$ and class level representation $\mathbf{X}^+$, in which case $L=3$ and $\mathbf{X}^{1:3}=(\mathbf{X}^-,\mathbf{X},\mathbf{X}^+)$.

Provided a training mini-batch $\{\mathbf{z}_i,y_i\}_{i=1}^M$, denote by $\{\mathbf{x}_i^-\}_{i=1}^M$, $\{\mathbf{x}_i\}_{i=1}^M$ and $\{\mathbf{x}_i^+\}_{i=1}^M$ the sets of instances of corresponding representations respectively. Specifically, in the proposed DML approach, we let

$$\mathbf{x}_i = \frac{\mathbf{W}^T\mathbf{x}_i^- + \mathbf{b}}{\|\mathbf{W}^\top\mathbf{x}_i^- + \mathbf{b}\|_{\ell^2}}, \mathbf{x}_i^+ = \left(\frac{\mathbf{w}_1}{\|\mathbf{w}_1\|_{\ell^2}}\cdots\frac{\mathbf{w}_C}{\|\mathbf{w}_C\|_{\ell^2}}\right)^T\mathbf{x}_i, \tag{9}$$

where $\mathbf{W}\in\mathbb{R}^{D\times d}$, $\mathbf{b},\mathbf{w}_1,\ldots,\mathbf{w}_C\in\mathbb{R}^d$ are parameters to be optimized. We perform $\ell^2$ normalization for $\{\mathbf{x}_i\}_{i=1}^M$ and $\{\mathbf{w}_j\}_{j=1}^C$ for better robustness against noise such as illumination, rotation, and scaling (Wang et al., 2018).

Suppose that $\bigcup_{I=1}^C\{\mathbf{x}_i^I\}_{i=1}^{n_I}$ is the partition of set $\{\mathbf{x}_i\}_{i=1}^M$ concerning the same-class equivalence relation, where $n^I$ is the number of class $I$ examples in the training set. Similarly we have partitions $\bigcup_{I=1}^C\{\mathbf{x}_i^{I-}\}_{i=1}^{n_I}$ of set $\{\mathbf{x}_i^-\}_{i=1}^M$ and $\bigcup_{I=1}^C\{\mathbf{x}_i^{I+}\}_{i=1}^{n_I}$ of set $\{\mathbf{x}_i^+\}_{i=1}^M$ .

The JRS regularizer is defined as follows.

**Definition 5** (joint representation similarity (JRS) regularizer)**.** *The joint representation similarity regularizer $\mathcal{L}_{JRS}$ penalizes the empirical joint representation similarities for all class pairs $\{I,J\}\subseteq\{1,\ldots,C\}$. Specifically,*

$$\begin{aligned}\mathcal{L}_{JRS} &\triangleq \sum_{I\neq J} n^I n^J \widehat{\mathcal{S}_{JRS}}(\mathbb{P}^I,\mathbb{P}^J) \\ &= \sum_{I\neq J}\sum_{i=1}^{n^I}\sum_{j=1}^{n^J} k^-(\boldsymbol{x}_i^{I-},\boldsymbol{x}_j^{J-})k(\boldsymbol{x}_i^I,\boldsymbol{x}_j^J)k^+(\boldsymbol{x}_i^{I+},\boldsymbol{x}_j^{J+}),\end{aligned} \tag{10}$$

*where $k^-$, $k$ and $k^+$ are reproducing kernels, $n^I n^J$ reweights class pair $(I,J)$ according to its credibility.*

### 4.2. Deep DML with Joint Representation Diversification

Since a deep DML model is often over-parameterized, only a JRS regularizer might not suffice to train a deep DML model. However, with the JRS regularizer, we may find a better balancing point of intra-class compactness and inter-class separability when combined with classification losses such as AMSoftmax loss (Wang et al., 2018) that does not penalize $\mathbf{d}(\mathbf{x}_i^I,\mathbf{x}_j^I)$ explicitly. To this end, we propose deep distance metric learning with joint representation diversification (JRD), which allows more flexible information retaining than classification losses. The optimization objective of our proposed JRD takes the form

$$\mathcal{L}_{JRD} = \mathcal{L}_{AMS} + \alpha\frac{1}{N_{pairs}}\mathcal{L}_{JRS}, \tag{11}$$

where $N_{pairs}$ denotes the number of pairs of instances from different classes in a mini-batch, and the hyperparameter $\alpha>0$ controls the trade-off between the diversification of DML representation and the intra-class compactness. The AMSoftmax loss $\mathcal{L}_{AMS}$ is written as

$$\mathcal{L}_{AMS} = \frac{1}{M}\sum_i^M -\log\frac{e^{s(\mathbf{w}_{y_i}'^\top\mathbf{x}_i-m)}}{e^{s(\mathbf{w}_{y_i}'^\top\mathbf{x}_i-m)}+\sum_{j\neq y_i}^C e^{s\mathbf{w}_j'^\top\mathbf{x}_i}}. \tag{12}$$

The $\{\mathbf{w}_j'\}_{j=1}^C$ are $\ell^2$-normalizations of $\{\mathbf{w}_j\}_{j=1}^C$, and can be viewed as the proxies (Movshovitz-Attias et al., 2017) or prototypes (Snell et al., 2017) of C classes in the training set. In (9), $\mathbf{x}_i$s are $\ell^2$-normalized. Thus $\mathbf{w}_{y_i}'^\top\mathbf{x}_i$ is the cosine similarity between $\mathbf{x}_i$ and the correct proxy corresponding to the $y_i$ class, and $\mathbf{w}_j'^\top\mathbf{x}_i$ are cosine similarities between $\mathbf{x}_i$ and proxies of other classes. The margin hyperparameter $m\in(0,1)$ encourages $\mathbf{w}_{y_i}'^\top\mathbf{x}_i$ being larger than all other cosine similarities $\mathbf{w}_j'^\top\mathbf{x}_i$ by a margin $m$. Hyperparameter $s$ is controlling the scale of cosine similarities.

### 4.3. Theoretical Analysis

#### 4.3.1. Interpretation of JRS

The nature and implications of the proposed JRS regularizer are not apparent. However, with the famous Bochner's theorem (Bochner, 1959), we may give some explanations for $\mathcal{S}_{JRS}(\mathbb{P},\mathbb{Q}) = \langle\mathcal{C}_{\mathbf{X}^{1:L}}(\mathbb{P}),\mathcal{C}_{\mathbf{X}'^{1:L}}(\mathbb{Q})\rangle_{\otimes_{l=1}^L\mathcal{RKHS}^l}$ when the kernels are translation invariant kernels $k^l(\mathbf{x},\mathbf{x}') = \psi^l(\mathbf{x}-\mathbf{x}')$ on $\mathbb{R}^d$. Note that translation invariant kernels cover many useful classes of characteristic kernels such as Gaussian and Laplace kernels.

**Proposition 1.** *Suppose that $\{k^l(\boldsymbol{x},\boldsymbol{x}')=\psi^l(\boldsymbol{x}-\boldsymbol{x}')\}_{l=1}^L$ on $\mathbb{R}^d$ are bounded, continuous reproducing kernels. Let $P^l\triangleq\mathbb{P}(\boldsymbol{X}^l|\boldsymbol{X}^{1:l-1})$ for $l=1,\ldots,L$ with $P^1=\mathbb{P}(\boldsymbol{X}^1)$. Then for any $\mathbb{P}(\boldsymbol{X}^1,\ldots,\boldsymbol{X}^L),\mathbb{Q}(\boldsymbol{X}'^1,\ldots,\boldsymbol{X}'^L)\in M_+^1(\times_{l=1}^L\mathcal{X}^l)$,*

$$\mathcal{S}_{JRS}(\mathbb{P},\mathbb{Q}) = \prod_{l=1}^L\langle\phi_{P^l}(\omega),\phi_{Q^l}(\omega)\rangle_{L^2(\mathbb{R}^d,\Lambda^l)}, \tag{13}$$

*where $\phi_{P^l}(\omega)$ and $\phi_{Q^l}(\omega)$ are the characteristic functions*

*of the distributions $P^l$ and $Q^l$, and $\Lambda^l$ is a (normalized) non-negative Borel measure characterized by $\psi^l(\boldsymbol{x}-\boldsymbol{x}')$.*

(Proof in Appendix A.) When the kernels are translation invariant, the joint representation similarity is equivalent to a product of inner products of characteristic functions (of conditional distributions) in $L^2$ space. The inner products are associated with a set of measures $\{\Lambda^l\}_{i=1}^L$ that re-weight the Lebesgue measure in $\mathbb{R}^d$. Moreover, the set of measures $\{\Lambda^l\}_{i=1}^L$ relies on the choice of translation invariant kernels.

#### 4.3.2. Inner Product v.s. Norm Regularizer

To penalize the similarities between kernel embeddings in the RKHSs, another possible solution is encouraging the norm distance, maximum mean discrepancy (MMD) (Gretton et al., 2007; 2012), between kernel embeddings of distributions from different classes. For ease of illustration, we only consider the case of a marginal distribution from one measurable space $\mathcal{X}$. Suppose that the conditions for the existence of kernel embeddings are satisfied, the MMD distance between probability measures $\mathbb{P}$ and $\mathbb{Q}$ is written as (Gretton et al., 2007),

$$\begin{aligned}&\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}}-\mu_{\mathbb{Q}}\|^2_{\mathcal{RKHS}}\\&= \|\mu_{\mathbb{P}}\|^2_{\mathcal{RKHS}} + \|\mu_{\mathbb{Q}}\|^2_{\mathcal{RKHS}} - 2\langle\mu_{\mathbb{P}},\mu_{\mathbb{Q}}\rangle_{\mathcal{RKHS}}\end{aligned} \tag{14}$$

Thus maximizing the MMD distance, $\mathrm{MMD}^2(\mathbb{P},\mathbb{Q})$, is equivalent to minimizing the inner product similarity $\langle\mu_{\mathbb{P}},\mu_{\mathbb{Q}}\rangle_{\mathcal{RKHS}}$ plus maximizing the norm of kernel embeddings $\|\mu_{\mathbb{P}}\|^2_{\mathcal{RKHS}}$ and $\|\mu_{\mathbb{Q}}\|^2_{\mathcal{RKHS}}$.

In the context of DML, maximizing norm of kernel embeddings might lead to less generalizable DML representation when the kernels are translation invariant. This can be seen from the empirical estimate by a sample $\{\mathbf{x}_i\}_{i=1}^m \sim \mathbb{P}$, that is, $\frac{1}{m^2}\sum_{i,j=1}^m k(\mathbf{x}_i,\mathbf{x}_j)$. When $k$ is a translation invariant kernel, say a Gaussian kernel $k(\mathbf{x},\mathbf{y}) = e^{-\frac{1}{2}(\mathbf{x}-\mathbf{y})^2}$, maxmimizing the norm estimate $\frac{1}{m^2}\sum_{i,j=1}^m k(\mathbf{x}_i,\mathbf{x}_j)$ impose explicit penalization on intra-class distances of DML representations and might lead to a less generalizable representation from the discussion in Section 2.

## 5. Experiments

### 5.1. Experimental Settings

**Datasets**: We conduct our experiments[2] on three benchmark datasets: CUB-200-2011 (CUB) (Wah et al., 2011), Cars-196 (CARS) (Krause et al., 2013), Stanford Online Products (SOP) (Oh Song et al., 2016). We adopt the standard data split protocol. For the CUB dataset that consists of 200 classes with 11,788 images, we use the first 100 classes with 5,864 images for training and the remaining 100 classes with 5,924 images for testing. CARS dataset is composed of 16,185 car images belonging to 196 classes. The first 98 classes are used for training, and the rest classes are used for testing. SOP dataset contains 120,053 product images from 22,634 classes, as the first 11,318 classes are used for training and the rest 11,316 classes for testing.

**Kernel design**: We consider a mixture of $K$ Gaussian kernels $k(\mathbf{x},\mathbf{x}') = \frac{1}{K}\sum_{k=1}^K e^{\frac{-(\mathbf{x}-\mathbf{x}')^2}{\sigma_k^2}}$, with varying bandwidth $\sigma_k^2$s for better capturing short-range and long-range dependencies (Wenliang et al., 2019). In this paper, we fix $K=3$ for pooling representation and DML representation, with $\{\sigma_1^2,\sigma_2^2,\sigma_3^2\} = \{0.5\tau,\tau,2\tau\}$, and $K'=1$ for class level representation with $\sigma^2=\tau$. $\tau$ is a self-adaptive parameter computed by averages of $\ell^2$-distances in the exponent part.

**Implementation details**: Our method is implemented by Pytorch on four Nvidia RTX8000s. We use the Inception Network (Szegedy et al., 2015) with BN (Ioffe & Szegedy, 2015) as the backbone, which is pre-trained on the ImageNet ILSVRC 2012 dataset (Russakovsky et al., 2015). Following practice Qian et al. (2019), BN layers of the backbone network is frozen during training. An affine transform is added on top of the global pooling layer to extract the DML representation, where the embedding size is fixed as 512 throughout the experiments. All images are cropped to 224×224 before feeding into the network. Random crop and random horizontal mirroring are used for data augmentation during training and single-center crop for testing. The $s$ in cosine softmax is fixed as 20. Adam optimizer (Kingma & Ba, 2014) is used for optimization. The number of epochs is set to be 50 (80 for SOP). The initial rates for parameters in the model and the softmax loss are 1e-4 and 1e-2 (1e-1 for SOP), respectively, and are divided by 10 every 20 (40 for SOP) epochs. The hyperparameters $\alpha, m$ and batch size $M$ are selected by 10-fold cross-validation, which is $\alpha=1$ for CUB and CARS, $\alpha=0.4$ for SOP; $m=0.1$ for CUB and SOP, $m=0.05$ for CARS; and batch size (100,50,120) for (CUB,CARS,SOP), respectively.

### 5.2. Comparison to the State-of-the-art

To emphasize the meliority of our DML framework JRD, we compare it with some state-of-the-art methods. For a fair comparison, we only consider methods that use the same backbone network and embedding size as ours. The methods with different backbones such as Wang et al. (2019b); Sanakoyeu et al. (2019), of different embedding sizes such as Roth et al. (2019); Lu et al. (2019), or adding additional convolutional layers such as Jacob et al. (2019) are not considered for comparison. Performances of all methods are evaluated using Recall@K metric on image retrieval task. Recall@K indicates the proportion of test images (queries) whose K nearest neighbors retrieved from the test set include samples from the same class.

[2]Codes are avalable at github.com/YangLin122/JRD

*Table 1.* Retrieval results Recall@K (%) of JRD and the state-of-the-art methods with the same backbone network and embedding size of 512 on the standard test split of three datasets.

| | CUB | | | | CARS | | | | SOP | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Recall@K(%) | 1 | 2 | 4 | 8 | 1 | 2 | 4 | 8 | 1 | 10 | 100 |
| DE_DSP(Duan et al., 2019) | 53.6 | 65.5 | 76.9 | - | 72.9 | 81.6 | 88.8 | - | 68.9 | 84.0 | 92.6 |
| HDML(Zheng et al., 2019) | 53.7 | 65.7 | 76.7 | 85.7 | 79.1 | 87.1 | 92.1 | 95.5 | 68.7 | 83.2 | 92.4 |
| DAMLRRM(Xu et al., 2019) | 55.1 | 66.5 | 76.8 | 85.3 | 73.5 | 82.6 | 89.1 | 93.5 | 69.7 | 85.2 | 93.2 |
| ECAML(Chen & Deng, 2019a) | 55.7 | 66.5 | 76.7 | 85.1 | 84.5 | 90.4 | 93.8 | 96.6 | 71.3 | 85.6 | 93.6 |
| DeML (Chen & Deng, 2019b) | 65.4 | 75.3 | 83.7 | 89.5 | **86.3** | **91.2** | 94.3 | 97.0 | 76.1 | 88.4 | 94.9 |
| SoftTriple Loss(Qian et al., 2019) | 65.4 | 76.4 | 84.5 | 90.4 | 84.5 | 90.7 | **94.5** | 96.9 | 78.3 | 90.3 | 95.9 |
| MS(Wang et al., 2019a) | 65.7 | 77.0 | **86.3** | 91.2 | 84.1 | 90.4 | 94.0 | 96.5 | 78.2 | **90.5** | **96.0** |
| JRD | **67.9** | **78.7** | 86.2 | **91.3** | 84.7 | 90.7 | 94.4 | **97.2** | **79.2** | **90.5** | **96.0** |

*Table 2.* Retrieval results Recall@K (%) with 95% confidence intervals of JRD and the variants of JRD in terms of different constituents of JRS regularizer on three datasets. The results on SOP is evaluated on the standard test split.

| | CUB | | | | CARS | | | | SOP | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Recall@K(%) | 1 | 2 | 4 | 8 | 1 | 2 | 4 | 8 | 1 | 10 | 100 |
| JRD | 50.7(1.1) | 63.7(1.1) | 74.8(1.2) | 84.1(1.2) | 61.2(1.3) | 72.6(0.9) | 82.2(0.6) | 89.2(0.7) | 79.2 | 90.5 | 96.0 |
| MRD | 49.4(1.1) | 62.3(1.1) | 74.5(1.2) | 83.6(1.2) | 59.8(1.3) | 71.5(1.2) | 80.6(0.9) | 88.0(0.9) | 78.8 | 90.4 | 95.9 |
| JRD-C | 48.6(1.5) | 61.4(1.4) | 73.4(1.5) | 83.0(1.4) | 58.5(1.5) | 69.6(1.3) | 79.1(0.7) | 86.6(0.9) | 77.7 | 89.8 | 95.6 |
| JRD-Pooling | 49.4(1.2) | 62.2(1.0) | 74.1(1.2) | 83.3(1.0) | 59.1(1.5) | 70.7(1.2) | 80.3(0.5) | 87.7(0.6) | 79.0 | 90.4 | 95.9 |

The Table 1 reports the results on three benchmark datasets. On the more challenging CUB and SOP datasets, our JRD model achieves state-of-the-art performances on all Recall@Ks. Particularly, JRD increases Recall@1 by 2.2% over the best baseline on the CUB dataset. As for the CARS dataset, JRD obtains best or second-best in terms of all metrics. In a word, our proposed JRD model achieves the best performance or performance on par with the state-of-the-art on all three different benchmark datasets, denoting the effectiveness and generalization ability of our model. Furthermore, since all the compared baselines are based on DNNs, the proposed modularized easy-to-compute JRS regularizer is compatible with all those methods. It is possible to achieve further improvements when combined with other insights, which we leave as a direction of future works.

### 5.3. Analysis of the JRS Regularizer

**Effects of joint representations**: We analyze the effects of modeling joint representations by comparing the JRD model with the following variants: a) MRD (M for Marginal): The model using only DML representation in the regularizer. b) JRD-C: The model using pooling representation and DML representation in the regularizer. c) JRD-pooling: The model using DML representation and class level representation in the regularizer.

We divide the test sets in CUB and CARS into ten disjoint smaller sets, with each smaller set contains examples from all the classes. This partition is not applicable to the SOP dataset since the average number of examples in each class is less than ten, and there are some classes containing only two examples. We report the mean retrieval results with 95% confidence intervals in Table 2. The results demonstrate that JRD outperforms all its variants, indicating the effectiveness of modeling joint representation. In particular, the JRD model outperforms the MRD model, denoting the superiority of joint representation over marginal representation. Besides, JRD-C and JRD-pooling also perform worse than the JRD model, which demonstrates that using pooling representation and class level representation can both increase the effectiveness of JRS regularizer.

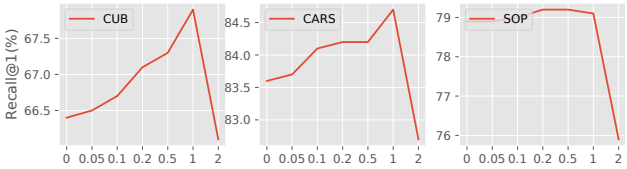



*Figure 4.* Recall@1 w.r.t. $\alpha$ on three datasets.

**Sensitivity of $\alpha$**: We study the effects of the magnitudes of JRS regularizer by varying $\alpha$ from 0 to 2. The Recall@1 on three datasets are shown in Figure 4. The results of different datasets are similar. We can observe bell-shaped curves that confirm the motivation of regularization, controlling trade-offs between intra-class compactness and inter-class separability. The regularization effects on SOP are not as significant as on the other two, possibly because there are many classes (11,318) for training, and the average number of examples in each class is less than ten. Thus the proba-

bility of sampling more than two examples from the same class in a mini-batch is small and resulting in a relatively small magnitude of effective gradients of the regularizer.

**Effect of explicit penalization on intra-class distances of representations**: To investigate the effect of additional explicit penalization on intra-class distances of representations and to substantiate claims in Section 4.3.2, we compare the proposed inner-product based JRS regularizer with a norm-based Joint Maximum Mean Discrepancy (JMMD) regularizer (Long et al., 2017). We also compare JRD and JMMD with a model regularized by a Joint Intra-class (JIntra) regularizer. The JMMD regularizer and JIntra regularizer takes the form as follows.

$$\widehat{\mathcal{D}_{JMMD}}(\mathbb{P},\mathbb{Q}) \triangleq -\|\widehat{\mathcal{C}_{\mathbf{X}^{1:L}}}(\mathbb{P}),\widehat{\mathcal{C}_{\mathbf{X}'^{1:L}}}(\mathbb{Q})\|_{\otimes_{l=1}^{L}\mathcal{RKHS}^l}, \tag{15}$$

$$\widehat{\mathcal{D}_{JIntra}}(\mathbb{P},\mathbb{Q}) \triangleq -\|\widehat{\mathcal{C}_{\mathbf{X}^{1:L}}}(\mathbb{P})\|_{\otimes_{l=1}^{L}\mathcal{RKHS}^l} - \|\widehat{\mathcal{C}_{\mathbf{X}'^{1:L}}}(\mathbb{Q})\|_{\otimes_{l=1}^{L}\mathcal{RKHS}^l}. \tag{16}$$

Using the same kernel design as in Section 5.1, JRD considers inner product similarity and promote presentation diversification. JIntra is the regularizer considering only the norm of kernel embeddings and promoting intra-class compactness by imposing explicit penalization on intra-class distances of representations. JMMD is considering both inner product similarity and the norm of kernel embeddings. We train models with JMMD regularizer, models with JIntra regularizer, and models with JRS regularizer on the CUB dataset with different values of $\alpha$. The Recall@1 with 95% confidence bands w.r.t. $\alpha$ is shown in Figure 5. Only JRD has a positive impact on performance. For trained models JRD with $\alpha = 1$ and JMMD with $\alpha = 0.1$, we extract DML representations of seen classes and unseen classes, respectively. Similar to Section 2, we compute the error of an ideal joint hypothesis $\lambda^{NN}$ and the $\mathcal{H}$-divergence $\hat{d}_{\mathcal{H}^{NN}}$) of two groups of representations for nearest neighbor classifiers. The results with 95% confidence intervals are shown in Table 3. The representations produced by the JRS regularizer has smaller $(\lambda^{NN} + \hat{d}_{\mathcal{H}^{NN}})$, which coincides with the result of Recall@1 and indicates a lower generalization error bound. Optimizing with a JMMD regularizer imposes explicit penalization on intra-class distances of DML representations, which potentially filtered out discriminating information for unseen classes and resulting in an inflated error of an ideal joint hypothesis.

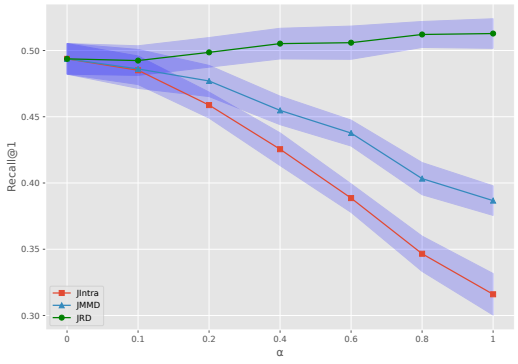


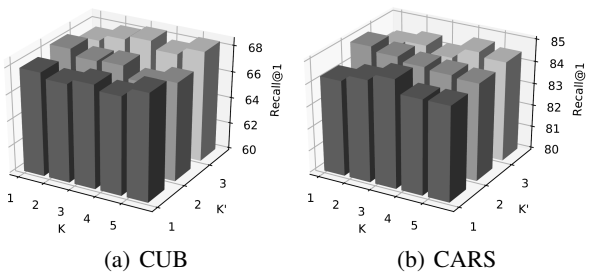

*Figure 5.* Recall@1 with 95% confidence bands w.r.t. $\alpha$ of JRD, JMMD, and JIntra on the CUB dataset. *Best viewed in color.*

*Table 3.* Recall@1, $\lambda$ and $\mathcal{H}$-divergence with 95% confidence intervals of JMMD, and JRM on the CUB dataset.

| Regularizers | Recall@1 | $\lambda^{NN}$ | $\hat{d}_{\mathcal{H}^{NN}}$ |
|---|---|---|---|
| JMMD($\alpha$@0.1) | 0.486(0.015) | 0.321(0.006) | 0.9275(0.003) |
| JRD($\alpha$@1) | 0.506(0.013) | 0.310(0.006) | 0.934(0.004) |

**On kernel design**: We use a mixture of $K$ Gaussian kernels as the reproducing kernel, with $K = 3$ for pooling representation and DML representation and $K' = 1$ for class level representation. We study the influence of $(K, K')$ on Recall@1 for the CUB and CARS datasets. The results are shown in Figure 6. In both datasets, the mixture of Gaussian kernels outperforms the single Gaussian kernel with $(K, K') = (1, 1)$. On the other hand, the optimal values of $(K, K')$ are inconsistent for two datasets.

We study the influence of kernel functions $k(\mathbf{x}, \mathbf{x}')$. We compare several common reproducing kernels, including the single Gaussian kernel, the single Laplace Kernel, the degree-$p$ inhomogeneous polynomial kernel for $p = 2$ and $p = 5$, and the kernel inducing moment generating function. The retrieval results on the CUB dataset are shown in Table 4. Among these kernels, Gaussian kernels and Laplace kernels are better ones. We guess this is because the kernel embedding of Gaussian kernels and Laplace kernels are injective, implying no information loss when mapping the distribution into the Hilbert space. However, which kernel is optimal for distance metric learning remains an open question, which we leave as a future research direction.




(a) CUB (b) CARS

*Figure 6.* Recall@1 for different values of $(K, K')$ in the mixture of Gaussian kernels on the CUB and CARS datasets.

*Table 4.* Recall@K (%) of JRD for some common reproducing kernels on the CUB dataset. The magnitude parameter of the JRS regularizer is searched and set to $x$, denoted as $\alpha@x$ for different reproducing kernels.

| kernel | $k(\mathbf{x}, \mathbf{x}')$ | R@1(%) | R@2(%) | R@4(%) | R@8(%) |
|---|---|---|---|---|---|
| Gaussian | $exp(-\frac{(\mathbf{x}-\mathbf{x}')^2}{\sigma^2})$ ($\alpha$@1) | 67.9 | 78.5 | 86.1 | 91.2 |
| Laplace | $exp(-\frac{\|\mathbf{x}-\mathbf{x}'\|_1}{\sigma})$ ($\alpha$@1) | 68.1 | 78.2 | 86.4 | 91.8 |
| degree-p Inhomogeneous polynomial kernel | $(\mathbf{x} \cdot \mathbf{x}' + 1)^2$ ($\alpha$@1e-3) | 66.1 | 77.0 | 85.3 | 90.9 |
| (p=2,p=5) | $(\mathbf{x} \cdot \mathbf{x}' + 1)^5$ ($\alpha$@1e-3) | 65.2 | 76.2 | 86.4 | 90.7 |
| Kernel inducing moment generating function | $exp(\mathbf{x} \cdot \mathbf{x}')$ ($\alpha$@1e-3) | 66.1 | 76.7 | 85.4 | 91.1 |

## 6. Related Work

**Distance metric learning**: With given input feature vectors, conventional distance metric learning methods (Xing et al., 2003; Weinberger et al., 2006; Davis et al., 2007) learn a mapping matrix from the input space to a representation space. To avoid overfitting, regularizing techniques such as penalizing $\ell^1$-norm (Niu et al., 2014), penalizing trace norm (Liu et al., 2015), and recently, orthogonality-promoting regularizations (Wang et al., 2012; Xie, 2015; Carreira-Perpinán & Raziperchikolaei, 2016; Chen et al., 2017; Xie et al., 2018) are proposed. The orthogonality-promoting regularizations encourage the projection vectors to be orthogonal, thus can be viewed as a class of techniques promoting central moment diversification.

**Deep DML**: For complex tasks with high-dimensional original input, more data-driven deep DML methods are developed. The triplet loss (Schroff et al., 2015) improves the pioneering contrastive loss (Chopra et al., 2005) by penalizing the relative intra-class distances compared with inter-class distances in the sampled triplets. However, during the optimization of triplet loss, the vast majority of sampled triplets produce gradients with near-zero magnitude and depress the performance. To generate useful gradients for training, several lines of efforts are a) hard negative mining methods (Harwood et al., 2017; Wang et al., 2019a; Duan et al., 2019) that sample or weighting informative examples, b) losses (Ustinova & Lempitsky, 2016; Huang et al., 2016; Sohn, 2016; Oh Song et al., 2016) that constrain on more examples than triplets, c) classification losses (Movshovitz-Attias et al., 2017; Wang et al., 2018; Qian et al., 2019) methods that get rid of sampling by learning a proxy for each class and constraining distances between examples and proxies. There are also regularization methods proposed to control overfitting (Chen & Deng, 2019a; Jacob et al., 2019). An energy confusion regularization term is proposed in Chen & Deng (2019a) within an adversarial framework, which can be viewed as a particular case of JRS regularizer when only considering the DML representation with a degree-1 homogeneous polynomial kernel. A high-order moment regularizer is proposed in (Jacob et al., 2019). The JRS regularizer can generalize the high-order moment regularizer by considering infinite-order moment with a kernel $k(\mathbf{x}, \mathbf{x}') = e^{(\mathbf{x} \cdot \mathbf{x}')}$, which lead to the kernel mean embedding being the moment-generating function.

**Diversity-Promoting regularization**: There are many recent literature on diversity-promoting regularization from the larger picture of latent space models (LSMs). Examples are a regularizer encouraging pairwise dissimilarity of latent components (Xie et al., 2015), a regularizer encouraging larger volume of the parallelepiped formed by latent components (Kwok & Adams, 2012), a regularizer penalizing the covariances of hidden activations in the DNNs (Cogswell et al., 2016), and a regularizer promoting uncorrelation as well as evenness of latent components (Xie et al., 2017). In contrast, we propose a JRS regularizer that takes advantage of deep architectures of DNNs and promotes dissimilar joint representations across multiple layers.

## 7. Conclusion

In this paper, we have attempted to address two issues of existing Deep DML methods, including how to leverage the hierarchical representations of DNNs to improve the DML representation, and how to better balance between the intra-class compactness and the inter-class separability. We propose an easy-to-compute JRS regularizer for deep DML that captures different abstract levels of invariant features and promotes inter-class separability, by diversifying the joint representation across three layers of hidden activations. Combining the JRS regularizer with the classification loss (AMSoftmax) allows a better balancing point between intra-class compactness and inter-class separability, and thus more flexible information retention of the original inputs.

## 8. Acknowledgements

The work was supported by the National Key Research and Development Program of China (2016YFB1001200), the Project 2019BD005 supported by PKU-Baidu fund, and the National Natural Science Foundation of China (61772045).

## References


Bach, F. R. and Jordan, M. I. Kernel independent component analysis. *Journal of Machine Learning Research*, 3(Jul): 1–48, 2002.

Baker, C. R. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186: 273–289, 1973.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8):1798–1828, 2013.

Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Bochner, S. *Lectures on Fourier integrals*, volume 42. Princeton University Press, 1959.

Carreira-Perpinán, M. A. and Raziperchikolaei, R. An ensemble diversity approach to supervised binary hashing. In *Advances in Neural Information Processing Systems*, pp. 757–765, 2016.

Chen, B. and Deng, W. Energy confused adversarial metric learning for zero-shot image retrieval and clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8134–8141, 2019a.

Chen, B. and Deng, W. Hybrid-attention based decoupled metric learning for zero-shot image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2750–2759, 2019b.

Chen, Y., Zhang, H., Tong, Y., and Lu, M. Diversity regularized latent semantic match for hashing. *Neurocomputing*, 230:77–87, 2017.

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

Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., and Batra, D. Reducing overfitting in deep networks by decorrelating representations. In *Proceedings of International Conference on Learning Representations*, 2016.

Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. Information-theoretic metric learning. In *International Conference on Machine Learning*, pp. 209–216, 2007.

Duan, Y., Chen, L., Lu, J., and Zhou, J. Deep embedding learning with discriminative sampling policy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4964–4973, 2019.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pp. 2121–2129, 2013.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5 (Jan):73–99, 2004.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pp. 513–520, 2007.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Harwood, B., Kumar, B., Carneiro, G., Reid, I., Drummond, T., et al. Smart mining for deep metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2821–2829, 2017.

Huang, C., Loy, C. C., and Tang, X. Local similarity-aware deep feature embedding. In *Advances in Neural Information Processing Systems*, pp. 1262–1270, 2016.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Jacob, P., Picard, D., Histace, A., and Klein, E. Metric learning with horde: High-order regularizer for deep embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6539–6548, 2019.

Jaynes, E. T. Information theory and statistical mechanics. *Physical Review*, 106(4):620, 1957.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Kwok, J. T. and Adams, R. P. Priors for diversity in generative latent variable models. In *Advances in Neural Information Processing Systems*, pp. 2996–3004, 2012.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pp. 2203–2213, 2017.

Li, Y., Swersky, K., and Zemel, R. Generative moment matching networks. In *International Conference on Machine Learning*, pp. 1718–1727, 2015.

Liu, H., Long, M., Wang, J., and Jordan, M. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, pp. 4013–4022, 2019.

Liu, W., Mu, C., Ji, R., Ma, S., Smith, J. R., and Chang, S.-F. Low-rank similarity metric learning in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Liu, W., Wen, Y., Yu, Z., and Yang, M. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Learning*, volume 2, 2016.

Long, M., Cao, Y., Wang, J., and Jordan, M. I. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, volume 37, pp. 97–105, 2015.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning*, volume 70, pp. 2208–2217, 2017.

Lopez-Paz, D., Muandet, K., Schölkopf, B., and Tolstikhin, I. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pp. 1452–1461, 2015.

Lu, J., Xu, C., Zhang, W., Duan, L.-Y., and Mei, T. Sampling wisely: Deep image embedding by top-k precision optimization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7961–7970, 2019.

Movshovitz-Attias, Y., Toshev, A., Leung, T. K., Ioffe, S., and Singh, S. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 360–368, 2017.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Niu, G., Dai, B., Yamada, M., and Sugiyama, M. Information-theoretic semi-supervised metric learning via entropy regularization. *Neural Computation*, 26(8): 1717–1762, 2014.

Oh Song, H., Xiang, Y., Jegelka, S., and Savarese, S. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4004–4012, 2016.

Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., and Jin, R. Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6450–6458, 2019.

Roth, K., Brattoli, B., and Ommer, B. Mic: Mining interclass characteristics for improved metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8000–8009, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015.

Sanakoyeu, A., Tschernezki, V., Buchler, U., and Ommer, B. Divide and conquer the embedding space for metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 471–480, 2019.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pp. 1857–1865, 2016.

Song, L. and Dai, B. Robust low rank kernel embeddings of multivariate distributions. In *Advances in Neural Information Processing Systems*, pp. 3228–3236, 2013.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

Ustinova, E. and Lempitsky, V. Learning deep embeddings with histogram loss. In *Advances in Neural Information Processing Systems*, pp. 4170–4178, 2016.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

Wang, J., Kumar, S., and Chang, S.-F. Semi-supervised hashing for large-scale search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2393–2406, 2012.

Wang, X., Han, X., Huang, W., Dong, D., and Scott, M. R. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5022–5030, 2019a.

Wang, X., Hua, Y., Kodirov, E., Hu, G., Garnier, R., and Robertson, N. M. Ranked list loss for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5207–5216, 2019b.

Weinberger, K. Q., Blitzer, J., and Saul, L. K. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pp. 1473–1480, 2006.

Wendland, H. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

Wenliang, L., Sutherland, D., Strathmann, H., and Gretton, A. Learning deep kernels for exponential family densities. In *International Conference on Machine Learning*, pp. 6737–6746, 2019.

Weston, J., Bengio, S., and Usunier, N. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

Xie, P. Learning compact and effective distance metrics with diversity regularization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 610–624. Springer, 2015.

Xie, P., Deng, Y., and Xing, E. Diversifying restricted boltzmann machine for document modeling. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1315–1324, 2015.

Xie, P., Singh, A., and Xing, E. P. Uncorrelation and evenness: a new diversity-promoting regularizer. In *International Conference on Machine Learning*, pp. 3811–3820, 2017.

Xie, P., Wu, W., Zhu, Y., and Xing, E. Orthogonality-promoting distance metric learning: Convex relaxation and theoretical analysis. In *International Conference on Machine Learning*, pp. 5403–5412, 2018.

Xing, E. P., Jordan, M. I., Russell, S. J., and Ng, A. Y. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, pp. 521–528, 2003.

Xu, X., Yang, Y., Deng, C., and Zheng, F. Deep asymmetric metric learning via rich relationship mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4076–4085, 2019.

Yang, E., Deng, C., Li, C., Liu, W., Li, J., and Tao, D. Shared predictive cross-modal deep quantization. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):5292–5303, 2018.

Zheng, W., Chen, Z., Lu, J., and Zhou, J. Hardness-aware deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 72–81, 2019.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, pp. 321–328, 2004.