Towards Improved Proxy-Based Deep Metric Learning via Data-Augmented Domain Adaptation

Li Ren, Chen Chen, Liqiang Wang, Kien Hua

University of Central Florida
{Li.Ren, Chen.Chen, Liqiang.Wang, Kien.Hua}@ucf.edu

Abstract

Deep Metric Learning (DML) plays an important role in modern computer vision research, where we learn a distance metric for a set of image representations. Recent DML techniques utilize the *proxy* to interact with the corresponding image samples in the embedding space. However, existing proxy-based DML methods focus on learning individual proxy-to-sample distance, while the overall distribution of samples and proxies lacks attention. In this paper, we present a novel proxy-based DML framework that focuses on aligning the sample and proxy distributions to improve the efficiency of proxy-based DML losses. Specifically, we propose the **D**ata-**A**ugmented **D**omain **A**daptation (**DADA**) method to adapt the domain gap between the group of samples and proxies. To the best of our knowledge, we are the first to leverage domain adaptation to boost the performance of proxy-based DML. We show that our method can be easily plugged into existing proxy-based DML losses. Our experiments on benchmarks, including the popular CUB-200-2011, CARS196, Stanford Online Products, and In-Shop Clothes Retrieval, show that our learning algorithm significantly improves the existing proxy losses and achieves superior results compared to the existing methods. The code and Appendix are available at: https://github.com/Noahsark/DADA

Introduction

The fundamental task of Deep Metric Learning (DML) focuses on learning deep representation with a known similarity metric. DML is a crucial topic in computer vision since it has a wide range of applications, including image retrieval (Lee, Jin, and Jain 2008; Yang et al. 2018; Ren et al. 2021), person re-identification (Yi et al. 2014; Wojke and Bewley 2018; Dai et al. 2019), and image localization (Lu et al. 2015; Ge et al. 2020). Modern DML techniques utilize deep neural networks (DNN) to project image samples into a hidden space where similar data points are grouped within short distances while the dissimilar points are separated. The majority of DML approaches focus on optimizing the similarities between pairwise samples with various loss functions, ranging from contrastive losses (Hadsell, Chopra, and LeCun 2006), triplet losses (Schroff, Kalenichenko, and Philbin 2015) to cross-entropy losses (Boudiaf et al. 2020).

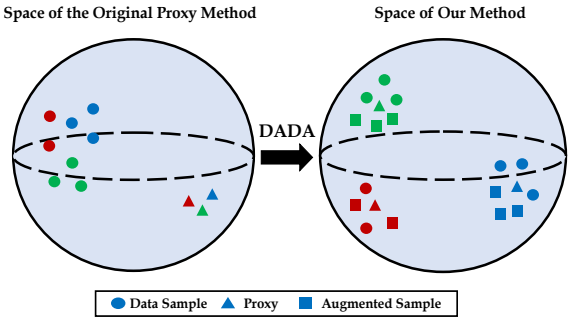

Figure 1: The intuition of our Data-Augmented Domain Adaptation (DADA). The classes are labeled with unique colors. The initial distribution gap between the data samples and corresponding proxies causes ambiguity for proxy-based deep metric learning. Our proposed method solves this problem by aligning the data samples and proxies, assuming they are from different data domains. We further augment the data to a dense manifold with mixed features to support this alignment.

With the increasing number of samples in the deep learning tasks, the basic pair or triplet losses face the difficulty of high computational complexity. Some approaches select informative samples by mining the hard or semi-hard samples (Wu et al. 2017; Katharopoulos and Fleuret 2018) while another group is devoted to comparing the sample clusters (Oh Song et al. 2017) or the statistics of the samples (Rippel et al. 2016).

Unlike the pair-based DML methods, the proxy-based approaches try to learn a group of trainable vectors, named *proxy*, instead of sweeping all sample pairs within the mini-batch or cluster (Movshovitz-Attias et al. 2017; Kim et al. 2020). Thus, the proxies capture the semantic information about the classes and optimize the uninformative sample-sample comparison with the proxy-sample relations. Based on the efficient proxy-sample distance metrics, later works further select the most informative proxy (Zhu et al. 2020) or assign each class with multiple proxies (Qian et al. 2019) to capture the intra-class structures. However, those existing proxy-based approaches simply guide the proxies by measuring their similarity with data samples where the learning process still faces a fundamental problem: *the colossal*

distribution gap between the proxies and the data samples, since the proxies are initially sampled from a normal distribution that does not contain any semantic information. The distribution gap would slow the convergence speed and cause ambiguity and bias in the learning process. Initializing the proxy with representations of the data sample is one straightforward solution to this problem. However, the distribution of proxies still differs dramatically between the early and late training stages due to the poor quality of sample representations at the early stage. Additionally, it takes a significant amount of extra time and space to calculate the representations for every class in each iteration.

In this paper, we introduce a novel framework to solve these problems by aligning the distributions of the proxies and the data samples (as illustrated in Figure 1). Specifically, we utilize Adversarial Domain Adaptation (Wang and Deng 2018) techniques to minimize their distribution gap. To align those distributions, we propose a *domain-level discriminator*, which is a classifier to separate their domain properties. Note that the single domain discriminator would cause *mode collapse* (Goodfellow et al. 2020; Che et al. 2017) where the majority of data points are constrained to a local area so that their discriminative information is lost. To endorse their discriminative information, we leverage one additional *category-level discriminator* to evaluate the consistency of their class properties. We show that with these discriminators, the adversarial training signal can efficiently align the distributions of the data samples and the proxies.

However, there are still two difficulties in learning the distribution of the proxy space: (1) the limited number and diversity of the proxies and (2) the large initial gap between the proxies and the data samples. The limited number of proxies causes difficulties for discriminators in capturing the inter-class manifold structure, and the large domain gap further hinders their learning efficiency. To overcome these challenges, we propose a novel data-augmented domain as a bridge where the data samples and the proxies are evenly mixed to conduct an intermediate domain. This domain contains rich mixing samples holding information and statistics from both sides. We also propose to create mixture samples within the same categories to increase the density of the manifold. We demonstrate the mechanisms of our method in Figure 2. Our experiments show that the proposed method can easily plug into existing proxy-based losses to boost their performance dramatically. Our main contributions are three-fold:

- We propose a novel adversarial learning framework to optimize the existing proxy-based DML by aligning the overall distributions of the data samples and the proxies at both domain and category levels.
- We propose an additional data-augmented domain that contains mixup representations from both sides to further bridge the distribution gap. We show that our combined discriminators efficiently guide the proxies and the data samples to a hidden space under the same distribution, in which the proxy-based loss significantly increases its learning efficiency.
- Our experiments demonstrate the effectiveness of our adversarial adaptation method on the image data samples and the proxies. We show that our approach increases the performance of existing proxy-based DML loss by a large margin, and our best result outperforms the state-of-the-art methods on four popular benchmarks.

Related Work

Pair-based DML. Metric Learning in the computer vision area aims to learn a metric that measures the distance between a pair of image samples. Initially, the image samples inside a class and out of a class are regarded as *positive* and *negative* samples; and they are learned and projected to a low dimensional space (Hadsell, Chopra, and LeCun 2006; Oh Song et al. 2016). The samples in different classes are paired and measured with the *contrastive loss* (Chopra, Hadsell, and LeCun 2005; Hadsell, Chopra, and LeCun 2006). To further compare the ranking relation between pairs of samples, an additional sample is selected as an anchor to compare with both positive and negative samples with the *triplet loss* (Weinberger and Saul 2009; Wang et al. 2014; Cheng et al. 2016; Hermans, Beyer, and Leibe 2017) where the positive sample is ensured to be closer than the negative samples. Based on the triple loss, Sohn et al.(2016) propose SoftMax cross-entropy to compare the group of pairs to improve pair sampling.

The computational cost of these pair-based works is always high due to the workload of comparing each sample with all other samples within a given batch. Additionally, these methods reveal sensitivity to the size of the batch, where their performance may significantly drop if the size is too small.

Proxy-based DML. To further accelerate the sampling and clustering process, Movshovitz et al. (2017) leverage the *proxy*, a group of learnable representations, to compare data samples via the Neighbourhood component analysis (NCA) loss (Roweis, Hinton, and Salakhutdinov 2004). The motivation is to set image samples as anchors to compare with proxies of different classes instead of corresponding samples to reduce sampling times. Teh et al. (2020) further improve the ProxyNCA by scaling the gradient of proxies. Zhu et al. (2020) propose to sample the most informative negative proxies to improve the performance, while Kim et al. (2020) set the proxies as anchors instead of the samples to learn the inter-class structure. Yang et al. (2022) develop hierarchical-based proxy loss to boost learning efficiency. Roth et al. (2022) regular the distribution of samples around the proxies following a non-isotropic distribution. In contrast to these methods that compare the single sample-proxy pair, our method further refines the manifold structure by aligning the whole distributions between proxies and image samples via a novel adversarial domain adaptation framework.

Domain Adaptation and Adversarial Learning. Domain Adaptation initially aims to solve the lack of labeled data where the learned feature is domain-invariant so that classifiers can be easily shifted to the new data distribution. The basic idea is to match the feature distributions to decrease their *domain shift* between the source and target datasets (Quiñonero-Candela et al. 2008; Torralba and Efros 2011). One important branch of domain adaptation

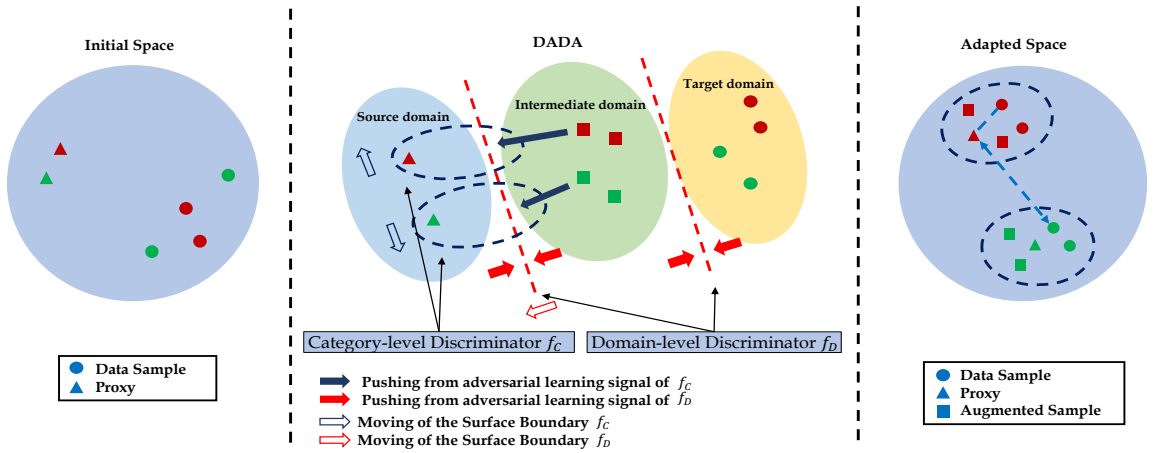

Figure 2: Demonstrate the mechanisms of our adversarial learning. Each class is labeled with a unique color. Left: Illustrate the Initial Space. Mid: Illustrate the training mechanisms and progress of our proposed method. Right: Illustrate the Adapted Space after training. The surface boundaries of the classifiers are trained to discriminate the domains with Domain-level Discriminators, and sample classes with Category-level Discriminators in the discriminator training phase. In the generator training phase, the samples and proxies are pushed to fool the Domain-level Discriminators from the adversarial learning signals while the class predictions from Category-level Discriminators are maintained.

is Adversarial Learning (Goodfellow et al. 2020; Hassan-Pour Zonoozi and Seydi 2022), where two or more models take part in the min-max game to generate domain-invariant features.

Ganin et al. (2015) first generate domain invariant features with adversarial training on neural networks. Tzeng et al. (2017) improve the discriminator that does not share the weight with the feature generator. Pei et al. (2018) utilize multiple discriminators assigned for each class to improve performance. Saito et al. (2018) minimize the prediction discrepancy of two discriminators on the target domain, while Lee et al. (2019) improve the method to compare their sliced Wasserstein distance instead. The primary application of adversarial learning is to produce synthetic textual or image data (Kingma and Welling 2013; Radford, Metz, and Chintala 2015; Isola et al. 2017). Ren et al. (2018; 2019) also applied this technique to enhance the quality of image captioning.

Recent studies have investigated the application of domain adaptation in image or textual retrieval tasks (Laradji and Babanezhad 2020; Pinheiro 2018). Wang et al. (2017) employ domain adaptation to align image and textural data using a single discriminator, whereas Ren et al. (2021) utilize multiple discriminators to get improved performance. In contrast to previous efforts, we propose aligning the distributions of data representations and proxies within the same image modality.

Proposed Method

We propose a new framework to close the gap between the distributions of the data samples and the proxies for proxy-based DML losses that are already in place. We utilize the adversarial domain adaptation technique to transfer data samples and proxies to domain invariant feature space. To overcome the limitation of the number of proxies, we also conduct a novel strategy to augment data as a bridge between the samples and proxies, which demonstrates a smooth learning process.

Preliminary

Deep Metric Learning (DML) Consider a set of data samples $\mathcal{S} = \{I_i, y_i\}_{i=1}^N$ with raw images I_i and its corresponding class label $y_i \in \{1, \dots, C\}$; we learn a projection function $f_G : \mathcal{S} \xrightarrow{f} \mathcal{X}$, which project the input data samples to a hidden embedding space (or metric space) $\mathcal{X}$. We define the projected features set as $X = \{x_i \in \mathbb{R}^d\}_{i=1}^N$. The primal goal of Deep Metric Learning (DML) is to refine the projector function $f_G(\cdot)$, which is usually constructed with *convolutional deep neural networks* (CNN) as the backbone, to generate the projected features that can be easily measured with defined distance metric $d(x_i, x_j)$ based on the semantic similarity between sample I_i and I_j. Here we adopt the distance metric $d(\cdot)$ as the *cosine similarity*. Before delivering features to any loss, we use L2 normalization to eliminate the effect of differing magnitudes.

Proxy-based DML To boost the learning efficiency, a group of DML methods pre-define a set of learnable representations $P = \{p_i \in \mathbb{R}^d\}_{i=1}^C$, named *proxy*, to represent subsets or categories of data samples. Typically there is one proxy for each class so that the number of proxies is the same as the number of classes C. The proxies are also optimized with other network parameters. The first proxy-based method, Proxy-NCA (Movshovitz-Attias et al. 2017), or its improved version Proxy-NCA++ (Teh, DeVries, and Taylor 2020), utilizes the Neighborhood Component Analysis (NCA) (Goldberger et al. 2004) loss to conduct this optimization. The later loss Proxy-Anchor (PA) (Kim et al. 2020) inversely sets the proxy as the anchor and measures

all proxies for each minibatch of samples. The PA loss $\mathcal{L}_{proxy}(X, P)$ can be presented as

$$\mathcal{L}_{proxy}(X,P) = \frac{1}{|P^+|}\sum_{p\in P^+}\log\left(1+\sum_{x\in X_p^+} e^{-\tau d(x,p)+\delta}\right) + \frac{1}{|P|}\sum_{p\in P}\log\left(1+\sum_{x\in X_p^-} e^{\tau d(x,p)+\delta}\right) \tag{1}$$

where X_p^+ denotes the set of positive samples for a proxy p; X_p^- is its complement set; τ is the scale factor; and δ is the margin. Since the PA updates all proxies for each mini-batch, the model has higher learning efficiency in capturing the structure of samples beyond the mini-batches. We propose these two fundamental proxy-based losses (PNCA++ and PA) that achieve competitive results as our baselines.

Domain Data Augmentation

To reduce the distribution gap between the data samples and the proxies, we transform the proxy-based DML into a domain adaptation problem. We regard the data samples as data points in the source domain, while the initialed proxies are data points in the target domain. We noticed that the number of proxies in the target domain is especially limited compared to the data samples because the basic proxy method only assigns a single proxy for each class. The unbalanced samples and proxies would cause learning biases in modeling the distributions. Also, the proxies are initialized from a normal distribution and do not contain any related semantic information, which also causes difficulty in aligning their distribution to the data sample domain.

To overcome these difficulties, we propose a novel data augmentation strategy to create an intermediate domain to balance the amount of data points for domain adaptation. Specifically, we interpolate the space with mixed features from data set X and proxy set P. For each data sample x_i and its corresponding proxy p_i, we create a data feature $\hat{d}$:

$$\hat{d}_i = \{\lambda x_i + (1-\lambda)p_i\}, \tag{2}$$

where $\lambda \sim Beta(\alpha, \beta)$ is the linear interpolation coefficient that sampled from *beta* distribution with $\alpha > 0$ and $\beta > 0$ that decide its probability density function. The new data contains semantic information between the data sample and proxies and shares their distribution statistics. Therefore pushing $\hat{d}$ is equal to pushing both samples x and proxies p, and their distribution is closer to the data sample domain than the original proxies.

In addition, we further propose extending the number of training instances by augmenting the data-proxy pairs within the same class. For each pair of samples (x_i, x_j) and their corresponding augmented data $(\hat{d}_i, \hat{d}_j)$ inside the mini-batch, we propose the following mixing:

$$\begin{aligned}\tilde{x}_i &= \{\mu_1 x_i + (1-\mu_1)x_j\}\\ \tilde{d}_i &= \{\mu_2 \hat{d}_i + (1-\mu_2)\hat{d}_j\},\end{aligned} \tag{3}$$

where μ_1, μ_2 are also sampled from $Beta$ distribution. Then we mix the new samples $\tilde{x}_i$ and $\tilde{d}_i$ inside the mini-batch to ensure the number of data samples with the same label $n \geq 2$. Combined with the original mini-batch, the augmented data sample set and the augmented proxy set are noted as $\tilde{X} = X \cup \{\tilde{x}_1, \tilde{x}_2 \cdots\}$ and $\tilde{D} = \{\hat{d}_1, \hat{d}_2 \cdots\} \cup \{\tilde{d}_1, \tilde{d}_2 \cdots\}$, and the size of mini-batch is also extended accordingly. We then normalize the composed features in $\tilde{X}$ and $\tilde{D}$ with L2 normalization to constrain them on a unit hypersphere embedding space where the magnitude is fixed to 1.

Domain-level Discriminator

Based on the augmented data, our goal is to refine the set $\tilde{X}$, $\tilde{D}$ and the original proxy set P to *domain invariant* representations that share the same distribution to help the proxy-based losses. We follow the principle idea of adversarial domain adaptation (Ganin et al. 2016) to estimate the domain divergence by learning a domain-level discriminator. Specifically, we learn a classifier $f_D(\cdot)$ that minimizes the risk of *domain prediction* (to predict if the data comes from a unique domain) between the set $\tilde{X}$, $\tilde{D}$ and P.

Generally, we would label the data from a specific domain with the one-hot label as the prediction target. Since we have three different domains including the augmented data domain and our labeling space is symmetric, we would simply assume the features $\tilde{x} \in \tilde{X}$ are labeled as $y_0 = \overline{001}$ while $\tilde{d} \in \tilde{D}$ are labeled as $y_1 = \overline{010}$, and the initial proxies P are labeled as $y_2 = \overline{100}$ for convenience. Specifically, we estimate the domain classifier $f_D(\cdot)$ as an MLP with a single hidden layer and a ReLU function. The hidden layer is then projected to a 3-dimensional head as the logits prediction of the domains. To optimize the $f_D(\cdot)$ with a low prediction risk on the labeling space, we conduct the cross-entropy objective $\mathcal{L}_{adv}$ as follows

$$\begin{aligned}\mathcal{L}_{adv}(\tilde{X}, \tilde{D}, P) &= \sum_i^{\tilde{N}} \mathcal{L}_{ce}(f_D(\tilde{x}_i), y_0)\\ &+ \sum_i^{\tilde{N}} \mathcal{L}_{ce}(f_D(\tilde{d}_i), y_1) + \sum_i^{C} \mathcal{L}_{ce}(f_D(p_i), y_2)\end{aligned} \tag{4}$$

where $\mathcal{L}_{ce}$ is the cross entropy loss and $\tilde{N}$ is the total number of samples after the data augmentation. The parameters of the classifier $f_D(\cdot)$ are optimized to minimize the adversarial loss $\mathcal{L}_{adv}$ in training. Recall that the feature x is generated from the projection function $f_G(\cdot)$. Thus, the parameters of generator $f_G(\cdot)$ are optimized to fool the discriminator $f_D(\cdot)$ in the opposite direction. Since $\tilde{D}$ in the target domain contains features that mixed from X and proxies P, optimizing the $\tilde{D}$ equals optimizing the generator $f_G(\cdot)$ in the source domain while updating the original proxies P. Thus, the adversarial learning signal of $\mathcal{L}_{adv}$ would help both generator $f_G(\cdot)$ and the original proxies P to maintain the domain invariant representations to fool the classifier.

Category-level Discriminator

One drawback of the domain-level discriminator described above is that the discriminate information, especially the inter-class correlation, is ignored in the optimization process. Losing the discriminative information will cause all

data points to be concentrated on a local area or a surface, which would cause inter-class ambiguity and confuse the metric learning losses. To solve this problem, we further propose a category-level discriminator that learns to predict the class of data samples and compare the discrepancy of predictions between the data samples and mixture proxies.

Specifically, we optimize a classifier $f_C(\cdot)$ with the feature generator $f_G(\cdot)$ to predict the category label $Y = \{y_0, y_1, \dots\}$ from mixture data samples $\tilde{X}$ with the classification loss $\mathcal{L}_{cls}(\tilde{X}, Y)$ as

$$\mathcal{L}_{cls}(\tilde{X}, Y) = \frac{1}{\tilde{N}} \sum_{i}^{\tilde{N}} \mathcal{L}_{ce}(f_C(\tilde{x_i}), y_i). \tag{5}$$

The cross-entropy loss $\mathcal{L}_{ce}$ would provide a supervised learning signal to $f_G(\cdot)$ to maintain the discriminative information during the DML training process.

We note that the data samples that share the distributions would also share the labeling space with the target proxy domain. To further align the distributions, we propose to constrain the samples from the source domain and augmented data from the target domain to have a low discrepancy of predictions from our category classifier $f_C(\cdot)$. Thus, one additional goal of $f_C(\cdot)$ is to learn the maximized discrepancy of the category prediction between the data samples $\tilde{X}$ and mixture proxies $\tilde{D}$ while the $\tilde{D}$ are later optimized to minimize this discrepancy.

To measure the discrepancy of the category probabilities, we empirically adopt the discrepancy introduced in (Chen et al. 2022) that utilizes the *Nuclear-norm Wasserstein Distance (NWD)*. The NWD is demonstrated to be the upper bound of the *Frobenius-norm*, which estimates the correlations of the predictions (Cui et al. 2020). Thus, we compare the NWD between the logistic predictions of $f_C(\cdot)$ from the augmented samples $\tilde{X}$ and data $\tilde{D}$. The loss $\mathcal{L}_d(\tilde{X}, \tilde{D})$, which measures the NWD can be described as,

$$\mathcal{L}_d(\tilde{X}, \tilde{D}) = \frac{1}{\tilde{N}} (\sum_{i}^{\tilde{N}} ||f_C(\tilde{X})||_\star - \sum_{i}^{N} ||f_C(\tilde{D})||_\star), \tag{6}$$

where $||x||_\star = \sum \sigma(x)$ denotes the *nuclear-norm* of x, which is defined as the sum of its singular values.

The Combined Loss and Training Progress

We adopt the paradigm of adversarial learning to alternatively update the gradient of our feature generator $f_G(\cdot)$ and the discriminators $f_D(\cdot)$ and $f_C(\cdot)$ discussed above. To this end, we train our combined loss by playing the *min-max game* as follows,

$$\min_{f_G, f_C} \{\eta(\mathcal{L}_{cls} + \max_{f_C} \mathcal{L}_d)\} + (1-\eta) \min_{f_D} \max_{f_G} \mathcal{L}_{adv}, \tag{7}$$

where η is the pre-defined hyperparameter that balances the contribution between the domain-level and category-level discriminators. Empirically, we do not set another weight between classification loss $\mathcal{L}_{cls}$ and discrepancy loss $\mathcal{L}_d$. We also need the original proxy-based loss $\mathcal{L}_{proxy}$ in Eq. 1 to do the basic DML of the sample-proxy pair in training. Note that the augmented data set $\tilde{D}$ is only for domain adaptation progress; the original $\mathcal{L}_{proxy}$ only operates $\tilde{X}$ and original proxies P. Thus, our combined training progress can be described as the following two sub-processes:

$$(\theta_{f_D}, \theta_{f_C}) = \arg \min_{f_D, f_C} \{\eta(\mathcal{L}_{cls} - \mathcal{L}_d) + (1-\eta)\mathcal{L}_{adv}\}, \tag{8}$$

$$(\theta_{f_G}, P) = \arg \min_{f_G, P, \tilde{D}} \{\eta(\mathcal{L}_{cls} + \mathcal{L}_d) - (1-\eta)\mathcal{L}_{adv} + \gamma \mathcal{L}_{proxy}\}, \tag{9}$$

where parameters θ_{f_D} and θ_{f_C} are updated in first phase and θ_{f_G} and the proxies P are updated with $\tilde{D}$ in the second phase. Even if gradient reversal layers are accepted for achieving adversarial training in earlier domain adaptation works, we empirically conclude that a separate training phase would be more feasible for us in our search for stable training parameters. The full training progress can be referred to in Algorithm 1.

Algorithm 1: Data-Augmented Domain Adaptation (DADA) for Proxy-based Deep Metric Learning

Input: Training Set $\mathcal{S} = \{I_i, y_i\}_{i=1}^N$
Initialization: θ_{f_G}, θ_{f_C}, θ_{f_D}, and proxies P
while *stop criteria is not satisfied* **do**
  Obtain a batch $\{I_i, y_i\}_{i=1}^n$ from S
  Select proxies $P = \{p_i\}_{i=1}^n$ according the labels Y
  Embedding features $X = \{x_i\}_{i=1}^n \leftarrow f_G(I)$
  /* *Prepare the mixture data domain* */
  Sample $\lambda, \mu_1, \mu_2 \sim Beta$ distribution
  Compose $\hat{P} \leftarrow \{\lambda X + (1-\lambda)P\}$
  Compose $\tilde{X} \leftarrow X \cup \{\mu_1 x_i + (1-\mu_1)x_j\}$
  Compose $\tilde{P} \leftarrow \hat{P} \cup \{\mu_2 \hat{p}_i + (1-\mu_2)\hat{p}_j\}$
  /* *Discriminator Training Phase* begin */
  Cal $\Delta\theta_{f_D}, \Delta\theta_{f_C} \leftarrow \eta \frac{\partial(\mathcal{L}_{cls}(\tilde{X},Y) - \mathcal{L}_d(\tilde{X},\tilde{P}))}{\Delta\theta_{f_D}, \Delta\theta_{f_C}}$
  Cal $\Delta\theta_{f_D}, \Delta\theta_{f_C} \leftarrow (1-\eta) \frac{\partial \mathcal{L}_{adv}(\tilde{X},\tilde{P},P)}{\Delta\theta_{f_D}, \Delta\theta_{f_C}}$
  Update $\theta_{f_D}, \theta_{f_C} \leftarrow Adam\{\Delta\theta_{f_D}, \Delta\theta_{f_C}\}$
  /* *Generator Training Phase* begin */
  Cal $\Delta\theta_{f_G}, \Delta P \leftarrow \eta \frac{\partial(\mathcal{L}_{cls}(\tilde{X},Y) + \mathcal{L}_d(\tilde{X},\tilde{P}))}{\Delta\theta_{f_G}, \Delta P}$
  Cal $\Delta\theta_{f_G}, \Delta P \leftarrow -(1-\eta) \frac{\partial \mathcal{L}_{adv}(\tilde{X},\tilde{P},P)}{\Delta\theta_{f_G}, \Delta P}$
  Cal $\Delta\theta_{f_G}, \Delta P \leftarrow \gamma \frac{\partial \mathcal{L}_{proxy}(\tilde{X},P)}{\Delta\theta_{f_G}, \Delta P}$
  Update $\theta_{f_G}, P \leftarrow Adam\{\Delta\theta_{f_G}, \Delta P\}$

Experiments

We present our performance study and discuss the experimental results in this section.

Datasets and Metrics

We use the standard benchmarks CUB-200-2011 (**CUB200**) (Wah et al. 2011) with 11,788 bird images and 200 classes, and **CARS196** (Krause et al. 2013) that contains 16,185 car images and 196 classes. We also evaluate our method on larger Stanford Online Products (**SOP**) (Oh Song et al. 2016) benchmark that includes 120,053 images with 22,634 product classes, and In-shop Clothes Retrieval (**In-Shop**) (Liu

Method	Reference	Settings	CUB-200			CARS-196			SOP		
		Arch	R@1	R@2	R@4	R@1	R@2	R@4	R@1	R@10	R@100
PNCA (2017)	CVPR17'	BN	49.2	61.9	67.9	73.2	82.4	86.4	73.7	–	–
DiVA(2020)	ECCV20'	R50	69.2	79.3	–	87.6	92.9	–	79.6	91.2	–
S2SD(2021)	ICML21'	R50	70.1	79.7	71.6	89.5	93.9	72.9	80.0	91.4	–
DCML-Proxy†(2021a)	CVPR21'	R50	65.2	76.4	84.8	81.2	89.8	94.6	–	–	–
DRML(2021b)	ICCV21'	BN	68.7	78.6	86.3	86.9	92.1	95.2	71.5	85.2	93.0
PA+AVSL†(2022)	CVPR22'	R50	71.9	81.7	88.1	91.5	95.0	97.0	79.6	91.4	96.4
PA+NIR†(2022)	CVPR22'	R50	69.1	79.6	–	87.7	92.5	–	80.7	91.5	–
HIST(2022)	CVPR22'	R50	71.4	81.1	88.1	89.6	93.9	96.4	**81.4**	92.0	**96.7**
DAS(2022)	ECCV22'	R50	69.2	79.3	87.0	87.8	93.2	96.0	80.6	91.8	**96.7**
MS+CRT(2022)	NeurIPS22'	R50	64.2	75.5	84.1	83.3	89.8	93.9	79.0	91.1	96.5
▷PNCA++(2020)	ECCV20'	R50	69.0	79.8	87.3	86.5	92.5	95.7	80.7	92.0	**96.7**
PNCA+**DADA**(R50)	Ours	R50	71.4	81.1	87.6	90.5	93.4	96.8	81.2	91.8	96.5
▷PA(2020)	CVPR20'	BN	68.4	79.2	86.8	86.1	91.7	95.0	79.1	90.8	96.2
PA+**DADA**(BN)	Ours	BN	69.8	80.4	87.1	89.4	92.1	96.2	79.6	91.0	96.3
▷PA (R50) ∗ (2020)	CVPR20'	R50	69.7	80.0	87.0	87.7	92.9	95.8	80.0	91.7	96.6
PA+**DADA**(R50)	Ours	R50	**72.9**	**81.9**	**88.3**	**92.1**	**95.2**	**97.1**	81.0	**92.1**	96.2

Table 1: Comparison with the state-of-the-art litterateurs on CUB200-2011 (2011), CARS196 (2013), Stanford Online Products (SOP) (2016). The works are sorted by their published date. The second column shows the same architecture of the backbone we selected to compare with our proposed method. R50 represents the ResNet50 and BN for InceptionBN and GN for GoogleNet backbones. † denotes the methods based on proxy-based DML, and ▷ labels the works on which our method is based. We adopt the experimental results of PA(R50) from the third papers (2022). The Bold represents the best score.

et al. 2016) dataset with 25,882 images and 7982 classes. We follow the data split that is consistent with the standard settings of existing DML works (Teh, DeVries, and Taylor 2020; Kim et al. 2020; Venkataramanan et al. 2022; Zheng et al. 2021b; Roth, Vinyals, and Akata 2022; Lim et al. 2022; Zhang et al. 2022). We adopt the **Recall@K** (K=1,2,4 in CUB200 and CARS196, K=1,10,100 in SOP, and K=1,10,20,30 in In-Shop) proposed in existing works to evaluate the accuracy of ranking. We also evaluate it with Mean Average Precision at R (**MAP@R**) that based on the ideas of MAP and R-precision, which is a more informative DML metric (Musgrave, Belongie, and Lim 2020).

Implementation Details

We train our model in a machine that contains a single RTX3090 GPU with 24GB memory. The Implementation is based on the existing RDML (Roth et al. 2020)

Backbones and Preprocessing. In this paper, we propose two basic backbones to evaluate our learning algorithm: the ResNet50(He et al. 2016) and the InceptionBN (Ioffe and Szegedy 2015). They are pre-trained on ImageNet1K(Deng et al. 2009) and are widely used in DML works for performance evaluation, where we resize the image to 224×224, do random resized cropping, and random horizontal flipping. In the test phase, the images are first resized to 256×256, then cropped back to 224×224. A linear head embeds the feature from the second last layer of the backbones to a 512-dimension hidden space. We follow the standard preprocessing introduced in other deep metric learning works (Venkataramanan et al. 2022; Zheng et al. 2021b; Roth, Vinyals, and Akata 2022; Lim et al. 2022; Zhang et al. 2022). We also adopt global max and average pooling with layer normalization on CNN backbones suggested by Teh et al. (Teh, DeVries, and Taylor 2020) to further improve the generalization of features.

In-Shop Clothes Retrieval (In-Shop)				
Methods	Arch	R@1	R@10	R@20
MS (2019)	BN	89.7	97.9	98.5
SHM (2019)	BN	90.7	97.8	98.5
SCT (2020)	R50	90.0	97.5	98.1
XBM (2020)	BN	89.9	97.6	98.4
IBC (2021)	R50	92.8	**98.5**	**99.1**
PA† (2020)	BN	90.4	98.1	98.8
PA+Mix† (2022)	R50	91.9	98.2	98.8
PNCA++†(2020)	R50	90.4	98.1	98.8
PNCA + **DADA** (ours)	R50	91.7	98.2	98.6
PA + **DADA** (ours)	R50	**93.0**	**98.5**	98.9

Table 2: Compare with the existing state-of-the-art DML works on the In-Shop (2016) dataset. The Bold represents the best score.

Training Details. Our optimization is done using Adam ($\beta_1 = 0.5, \beta_2 = 0.999$) (Kingma and Ba 2015) with a decay of $1 \cdot 10^{-3}$. We set the learning rate at $1.2 \cdot 10^{-4}$ for the feature generator $f_G(\cdot)$ and $5 \cdot 10^{-4}$ for our discriminators. We adopt the learning rate $4 \cdot 10^{-2}$ for the proxies as suggested in (Roth, Vinyals, and Akata 2022). For most of the experiments, we fixed the batch size to 90 as a default setting, which is consistent with (Kim et al. 2020). Empirically we apply *batch normalization* on the domain-level discriminator to reduce its correlation variance within the batch. For all experiments, the first layer of $f_C(\cdot)$ is set to 512. For the second layer, we assigned 128 dimensions to the CUB200 and CARS196 datasets, 8192 dimensions to the SOP datasets, and 4096 dimensions to the In-Shop

	ProxyAnchor		ProxyNCA++	
Settings	R@1	MAP@R	R@1	MAP@R
Baseline	69.1	26.5	68.4	25.8
+Aug	69.3	26.5	68.5	25.9
$+\mathcal{L}_{adv}$	70.2	27.3	69.2	26.4
$+\mathcal{L}_{adv}$+Aug	70.9	27.8	69.8	26.8
$+\mathcal{L}_{cls}$	69.3	27.0	68.9	26.2
$+\mathcal{L}_{cls} + \mathcal{L}_{d}$	69.9	27.4	69.5	26.6
$+\mathcal{L}_{cls} + \mathcal{L}_{d}$+Aug	70.4	27.8	69.4	26.7
$+\mathcal{L}_{adv} + \mathcal{L}_{cls}$	71.4	28.2	69.3	27.1
$+\mathcal{L}_{adv} + \mathcal{L}_{cls} + \mathcal{L}_{1}$	71.6	28.2	69.4	27.0
$+\mathcal{L}_{adv} + \mathcal{L}_{cls} + \mathcal{L}_{d}$	72.0	28.9	69.9	27.7
$+\mathcal{L}_{adv} + \mathcal{L}_{cls} + \mathcal{L}_{d}$ + Aug (**ours**)	**72.9**	**29.9**	**70.2**	**28.0**

Table 3: Study the contribution of each component of our method and loss function on CUB200. We reproduce the result of ProxyAnchor, which has a batch size of 90, and ProxyNCA++, which has a batch size of 32, as the baseline of our method. Aug represents the alignment of the augmented data and samples we introduced in Sec . We denote the difference in percentage point (pp) compared with our baseline in the bracket.

datasets. We set $\{\eta = 0.005, \gamma = 0.0075\}$ for CUB200, and $\{\eta = 0.01, \gamma = 0.0075\}$ for CARS196. We select $\{\eta = 0.01, \gamma = 0.005\}$ for both SOP and In-Shop datasets.

Qualitative Results

Comparing with Proxy Baselines. We compare the performance of our approach with the existing proxy-based metric learning methods and the recent *state-of-the-art* metric learning methods on the popular benchmarks introduced above (refer to Table 1). We observe that our DADA frameworks can significantly improve the performance of the original proxy-based DML methods (marked with $\triangleright$) by a large margin. Specifically, comparing with the original PA method on ResNet50, our proposed PA+DADA outperforms $3.2pp$ (4.6%) on the recall@1 of CUB200 and $4.4pp$ (5.0%) on the recall@1 of CARS196. On the larger datasets (SOP and In-Shop), our method is also better than the original PA and PNCA++.

Comparing with state-of-the-art. We further compare the performance of our method with the state-of-the-art methods based on the CNN backbones as listed in Table 1 and 2. For the CARS196 dataset, our method reaches 92.1 on Recall@1, which has a $0.9pp$ improvement over the previous state-of-the-art AVSL (Zhang et al. 2022) on the ResNet50 backbone. For CUB200, our method outperforms the previous state-of-the-art AVSL $0.6pp$ on Recall@1, $0.2pp$ on Recall@2, and $0.1pp$ on Recall@4. We observe that our performance on SOP and In-Shop is limited but very close to the previous state-of-the-art IBC (Seidenschwarz, Elezi, and Leal-Taixé 2021), CRT (Kan et al. 2022), and HIST (Lim et al. 2022) on a few metrics. The lesser improvement in the high-value recall of these two datasets is mainly due to the large number of classes (11318 and 3997) and the limited number of samples in each class (less than 10). This causes some difficulty for our category-level discriminator to learn the discriminative information. Nevertheless, our method still achieves good performance comparable to those of the state-of-the-art methods in all metrics and outperforms other proxy-related methods on these two datasets. We will investigate techniques to overcome this limitation in our future works.

Ablation Study

Contributions of the Objective Components. We analyze the ablation study to evaluate the contribution of each objective component of our proposed framework based on both ProxyAnchor (Kim et al. 2020) and ProxyNCA++ (Teh, DeVries, and Taylor 2020) on the CUB200 as listed in Table 3. We first notice that the data augmentation strategy (Aug) does not improve our baseline significantly in the absence of $\mathcal{L}_{adv}$ and $\mathcal{L}_{cls}$. This is because, without those regularization losses, Aug simply boosts some redundant positive samples and the mixed features do not take part in training. We conclude that the domain-level discriminator with $\mathcal{L}_{adv}$ has higher efficiency when the category-level discriminator with $\mathcal{L}_{cls}$ helps regularize the space and avoid the inter-class ambiguity. It increases the improvement to $+2.3pp$ on R@1 and $+1.7pp$ on MAP@R from $+1.1pp$ on R@1 and $+0.8pp$ on MAP@R in comparison with the single $\mathcal{L}_{adv}$ setting. We also demonstrate that the efficiency of the category-level classifier ($+\mathcal{L}_{cls}$) can be further improved by comparing the discrepancy of class prediction ($+\mathcal{L}_{d}$) between the source data and target proxies in adversarial learning. Comparing the general discrepancy L1 distance ($+\mathcal{L}_{1}$), the proposed NWD also show increasing performances on both R@1 and MAP@R. A similar conclusion can also be driven by the results based on ProxyNCA. Therefore, we conclude that the combination of the domain and the category-level discriminator is more suitable for proxy-based DML than the settings with any single discriminator. We also study the impact of our hyperparameters and the combination of data groups that apply domain adaptation in the Appendix.

Conclusion

In this paper, we present an adversarial domain adaptation method with data augmentation to optimize the hidden space of the data and the proxies. We overcome the initial distribution gap between them to boost the learning efficiency of deep metric learning. We propose to align the domains of the data and the initial proxies by optimizing two classifiers at different levels, and training the embedding function and the proxies against them. To enhance the density of the manifold, we propose a strategy to conduct a mixture space by mixing the features from both domains. Our experimental results based on four popular deep metric learning benchmarks demonstrate that our learning method and mixed space efficiently boost the learning efficiency of existing proxy-based methods. While our framework focuses on solving the challenge of proxy-based DML methods, we believe it can be easily extended to other related metric learning methods, and it can also benefit zero-shot and self-supervised learning works. These are interesting and challenging works for future study.

References

Boudiaf, M.; Rony, J.; Ziko, I. M.; Granger, E.; Pedersoli, M.; Piantanida, P.; and Ayed, I. B. 2020. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In *ECCV*, 548–564. Springer.

Che, T.; Li, Y.; Jacob, A. P.; Bengio, Y.; and Li, W. 2017. Mode regularized generative adversarial networks. *ICLR*.

Chen, L.; Chen, H.; Wei, Z.; Jin, X.; Tan, X.; Jin, Y.; and Chen, E. 2022. Reusing the Task-specific Classifier as a Discriminator: Discriminator-free Adversarial Domain Adaptation. In *CVPR*, 7181–7190.

Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; and Zheng, N. 2016. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 1335–1344.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, volume 1, 539–546. IEEE.

Cui, S.; Wang, S.; Zhuo, J.; Li, L.; Huang, Q.; and Tian, Q. 2020. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *CVPR*, 3941–3950.

Dai, Z.; Chen, M.; Gu, X.; Zhu, S.; and Tan, P. 2019. Batch dropblock network for person re-identification and beyond. In *ICCV*, 3691–3701.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. Ieee.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*, 1180–1189. PMLR.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

Ge, Y.; Wang, H.; Zhu, F.; Zhao, R.; and Li, H. 2020. Self-supervising fine-grained region similarities for large-scale image localization. In *ECCV*, 369–386. Springer.

Goldberger, J.; Hinton, G. E.; Roweis, S.; and Salakhutdinov, R. R. 2004. Neighbourhood components analysis. *NIPS*, 17.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, 1735–1742. IEEE.

HassanPour Zonoozi, M.; and Seydi, V. 2022. A Survey on Adversarial Domain Adaptation. *Neural Processing Letters*, 1–41.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 448–456. PMLR.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*, 1125–1134.

Kan, S.; Liang, Y.; Li, M.; Cen, Y.; Wang, J.; and He, Z. 2022. Coded Residual Transform for Generalizable Deep Metric Learning. *NeurIPS*.

Katharopoulos, A.; and Fleuret, F. 2018. Not all samples are created equal: Deep learning with importance sampling. In *ICML*, 2525–2534. PMLR.

Kim, S.; Kim, D.; Cho, M.; and Kwak, S. 2020. Proxy anchor loss for deep metric learning. In *CVPR*, 3238–3247.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. *ICLR*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *ICCV workshop*, 554–561.

Laradji, I. H.; and Babanezhad, R. 2020. M-adda: Unsupervised domain adaptation with deep metric learning. *Domain adaptation for visual understanding*, 17–31.

Lee, C.-Y.; Batra, T.; Baig, M. H.; and Ulbricht, D. 2019. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *CVPR*, 10285–10295.

Lee, J.-E.; Jin, R.; and Jain, A. K. 2008. Rank-based distance metric learning: An application to image retrieval. In *CVPR*, 1–8. IEEE.

Lim, J.; Yun, S.; Park, S.; and Choi, J. Y. 2022. Hypergraph-Induced Semantic Tuplet Loss for Deep Metric Learning. In *CVPR*, 212–222.

Liu, L.; Huang, S.; Zhuang, Z.; Yang, R.; Tan, M.; and Wang, Y. 2022. DAS: Densely-Anchored Sampling for Deep Metric Learning. *ECCV*.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 1096–1104.

Lu, G.; Yan, Y.; Ren, L.; Song, J.; Sebe, N.; and Kambhamettu, C. 2015. Localize me anywhere, anytime: a multi-task point-retrieval approach. In *ICCV*, 2434–2442.

Milbich, T.; Roth, K.; Bharadhwaj, H.; Sinha, S.; Bengio, Y.; Ommer, B.; and Cohen, J. P. 2020. Diva: Diverse visual feature aggregation for deep metric learning. In *ECCV*, 590–607. Springer.

Movshovitz-Attias, Y.; Toshev, A.; Leung, T. K.; Ioffe, S.; and Singh, S. 2017. No fuss distance metric learning using proxies. In *CVPR*, 360–368.

Musgrave, K.; Belongie, S.; and Lim, S.-N. 2020. A metric learning reality check. In *ECCV*, 681–699. Springer.

Oh Song, H.; Jegelka, S.; Rathod, V.; and Murphy, K. 2017. Deep metric learning via facility location. In *CVPR*, 5382–5390.

Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *CVPR*, 4004–4012.

Pei, Z.; Cao, Z.; Long, M.; and Wang, J. 2018. Multi-adversarial domain adaptation. In *AAAI*.

Pinheiro, P. O. 2018. Unsupervised domain adaptation with similarity learning. In *CVPR*, 8004–8013.

Qian, Q.; Shang, L.; Sun, B.; Hu, J.; Li, H.; and Jin, R. 2019. Softtriple loss: Deep metric learning without triplet sampling. In *ICCV*, 6450–6458.

Quiñonero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N. 2008. Covariate shift and local learning by distribution matching.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Ren, L.; and Hua, K. 2018. Improved image description via embedded object structure graph and semantic feature matching. In *ISM*, 73–80. IEEE.

Ren, L.; Li, K.; Wang, L.; and Hua, K. 2021. Beyond the deep metric learning: enhance the cross-modal matching with adversarial discriminative domain regularization. In *ICPR*, 10165–10172. IEEE.

Ren, L.; Qi, G.-J.; and Hua, K. 2019. Improving diversity of image captioning through variational autoencoders and adversarial learning. In *WACV*, 263–272. IEEE.

Rippel, O.; Paluri, M.; Dollar, P.; and Bourdev, L. 2016. Metric learning with adaptive density discrimination. *ICLR*.

Roth, K.; Milbich, T.; Ommer, B.; Cohen, J. P.; and Ghassemi, M. 2021. Simultaneous similarity-based self-distillation for deep metric learning. In *ICML*, 9095–9106. PMLR.

Roth, K.; Milbich, T.; Sinha, S.; Gupta, P.; Ommer, B.; and Cohen, J. P. 2020. Revisiting training strategies and generalization performance in deep metric learning. In *ICML*, 8242–8252. PMLR.

Roth, K.; Vinyals, O.; and Akata, Z. 2022. Non-isotropy Regularization for Proxy-based Deep Metric Learning. In *CVPR*, 7420–7430.

Roweis, S.; Hinton, G.; and Salakhutdinov, R. 2004. Neighbourhood component analysis. *NIPS*, 17(513-520): 4.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 3723–3732.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.

Seidenschwarz, J. D.; Elezi, I.; and Leal-Taixé, L. 2021. Learning intra-batch connections for deep metric learning. In *ICML*, 9410–9421. PMLR.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. *NIPS*, 29.

Suh, Y.; Han, B.; Kim, W.; and Lee, K. M. 2019. Stochastic class-based hard example mining for deep metric learning. In *CVPR*, 7251–7259.

Teh, E. W.; DeVries, T.; and Taylor, G. W. 2020. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In *ECCV*, 448–464. Springer.

Torralba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR*, 1521–1528. IEEE.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *CVPR*, 7167–7176.

Venkataramanan, S.; Psomas, B.; Avrithis, Y.; Kijak, E.; Amsaleg, L.; and Karantzalos, K. 2022. It takes two to tango: Mixup for deep metric learning. *ICLR*.

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

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial cross-modal retrieval. In *Multimedia*, 154–162.

Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; and Wu, Y. 2014. Learning fine-grained image similarity with deep ranking. In *CVPR*, 1386–1393.

Wang, M.; and Deng, W. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153.

Wang, X.; Han, X.; Huang, W.; Dong, D.; and Scott, M. R. 2019. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*, 5022–5030.

Wang, X.; Zhang, H.; Huang, W.; and Scott, M. R. 2020. Cross-batch memory for embedding learning. In *CVPR*, 6388–6397.

Weinberger, K. Q.; and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2).

Wojke, N.; and Bewley, A. 2018. Deep cosine metric learning for person re-identification. In *WACV*, 748–756. IEEE.

Wu, C.-Y.; Manmatha, R.; Smola, A. J.; and Krahenbuhl, P. 2017. Sampling matters in deep embedding learning. In *CVPR*, 2840–2848.

Xuan, H.; Stylianou, A.; Liu, X.; and Pless, R. 2020. Hard negative examples are hard, but useful. In *ECCV*, 126–142. Springer.

Yang, J.; She, D.; Lai, Y.-K.; and Yang, M.-H. 2018. Retrieving and classifying affective images via deep metric learning. In *AAAI*, volume 32.

Yang, Z.; Bastan, M.; Zhu, X.; Gray, D.; and Samaras, D. 2022. Hierarchical proxy-based loss for deep metric learning. In *WACV*, 1859–1868.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Deep metric learning for person re-identification. In *ICPR*, 34–39. IEEE.

Zhang, B.; Zheng, W.; Zhou, J.; and Lu, J. 2022. Attributable Visual Similarity Learning. In *CVPR*, 7532–7541.

Zheng, W.; Wang, C.; Lu, J.; and Zhou, J. 2021a. Deep compositional metric learning. In *CVPR*, 9320–9329.

Zheng, W.; Zhang, B.; Lu, J.; and Zhou, J. 2021b. Deep relational metric learning. In *ICCV*, 12065–12074.

Zhu, Y.; Yang, M.; Deng, C.; and Liu, W. 2020. Fewer is more: A deep graph metric learning perspective using fewer proxies. *NeurIPS*, 33: 17792–17803.