Learning with Diversity: Self-Expanded Equalization for Better Generalized Deep Metric Learning

Jiexi Yan[1], Zhihui Yin[2], Erkun Yang[2], Yanhua Yang[1]*, and Heng Huang[3]
[1]School of Computer Science and Technology, Xidian University, China
[2]School of Electronic Engineering, Xidian University, China
[3]Department of Computer Science, University of Maryland College Park, USA
{jxyan1995,yzh.xdu,erkunyang}@gmail.com, yanhyang@xidian.edu.cn, henghuanghh@gmail.com

Abstract

Exploring good generalization ability is essential in deep metric learning (DML). Most existing DML methods focus on improving the model robustness against category shift to keep the performance on unseen categories. However, in addition to category shift, domain shift also widely exists in real-world scenarios. Therefore, learning better generalization ability for the DML model is still a challenging yet realistic problem. In this paper, we propose a new self-expanded equalization (SEE) method to effectively generalize the DML model to both unseen categories and domains. Specifically, we take a 'min-max' strategy combined with a proxy-based loss to adaptively augment diverse out-of-distribution samples that vastly expand the span of original training data. To take full advantage of the implicit cross-domain relations between source and augmented samples, we introduce a domain-aware equalization module to induce the domain-invariant distance metric by regularizing the feature distribution in the metric space. Extensive experiments on two benchmarks and a large-scale multi-domain dataset demonstrate the superiority of our SEE over the existing DML methods.

1. Introduction

Learning an effective metric to measure the visual similarities among examples is a fundamental problem in many computer vision tasks, such as image retrieval [15, 24], face recognition [10, 32] and person re-identification [54, 55, 22]. Metric learning aims to automatically learn a task-specific distance metric under which samples from the same class are encouraged to be closer than those from different classes. Taking advantage of deep learning technique [8, 14], deep metric learning (DML) methods employ deep neural networks (DNNs) [8, 52, 51] to extract more representative feature embeddings and demonstrate superior performance.

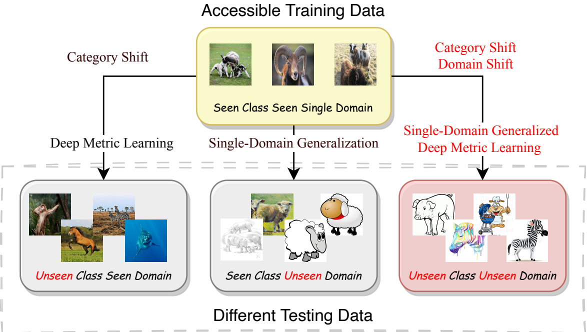

Figure 1. Comparisons of the single-domain generalized DML task with the conventional DML and single-domain generalization. The better generalization ability of the learned DML model is needed to adapt both unseen categories and domains.

In standard DML settings[23, 25, 33], we hope the trained metric model can better generalize to *unseen* classes in the testing phase, which aligns more with the scenarios in practical applications. To this end, many methods, such as XBM [41], DRML [59], and DCML [2], have been proposed to alleviate the impact of such *category shift* and learn discriminative metric that generalizes well to *unseen* classes. However, in many real-world applications, there exists not only *category shift* but also *domain shift* between the training data and testing data as shown in Figure 1. For example, the retrieval gallery in image retrieval includes images with different styles, but we can only use real images for training, as usual. Therefore, these DML methods suffer from poor generalization due to the impact of *domain shift*. Simultaneously ensuring better generalization ability to *unseen* categories and *unseen* domains for the metric model is a more challenging problem.

To further explore improving the generalization ability of DML, we propose a practical yet challenging task, namely

*Corresponding Author

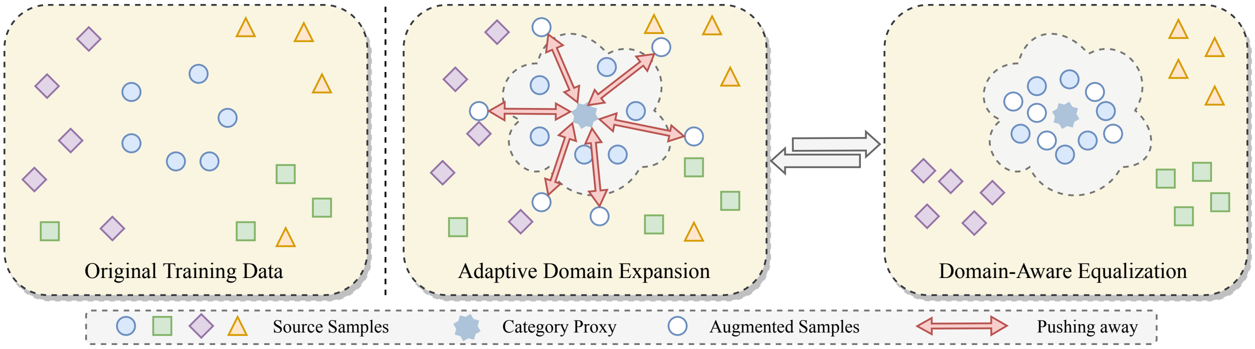

Figure 2. The simple illustration of our self-expanded equalization. We adopt the adaptive domain expansion module to augment out-of-distribution samples while utilizing the domain-aware equalization module to learn a more robust domain-invariant distance metric against domain shift.

single-domain generalized DML. As shown in Figure 1, giving a training dataset with seen categories and domains, we aim to train a metric model and generalize it to different unseen categories and domains. Different from domain generalization methods [45, 18, 53], we cannot obtain and use extra domain-related annotations, which are usually not accessible in practical applications. Therefore, domain generalization methods cannot be directly adopted in our task.

Learning adversarial augmentations is a general idea to improve the model's generalization ability, and some related methods [30, 42, 58] have been proposed for single-domain generalization. However, the generated samples should instead be considered adversarial perturbations that cannot effectively mimic the domain shift with diversity. Moreover, simply treating adversarial augmentations as the same training samples as the original training data will harm learning discriminative distance metric since the discarded domain-specific variations may be helpful to extract domain-invariant representations.

In this paper, we propose a self-expanded equalization (SEE) method for single-domain generalized DML, which consists of two main modules, *i.e.*, adaptive domain expansion (ADE) and domain-aware equalization (DAE). In the ADE module, we employ a 'min-max' strategy to modify the source data and adaptively conduct diverse domain expansion. We hope the generated augmentations keep consistency with the source data at pixel level while having a large margin with the proxy of this class learned by our proxy-based metric loss shown in Figure 2. To further improve the generalization ability, we utilize the DAE module to excavate the implicit semantic relations across different domain shifts and induce the domain-invariant distance metric that is more discriminative for unseen domains. Meanwhile, the generated augmentations will be regarded as hard samples to learn the discriminative metric for unseen classes. During training, the data distribution expansion and the model optimization are conducted alternatively in each iteration. The contributions are summarized as follows:

- We discuss and propose a more realistic yet challenging task, *i.e.*, single-domain generalized DML, which aims to generalize the metric model to unseen categories and domains.
- To handle this difficult task, we provide a new self-expanded equalization method to adaptively expand domain distribution with diversity and learn the consistent distance metric across different domain shifts.
- We perform experiments on multiple datasets and compare our method with state-of-the-art DML methods. Our SEE substantially outperforms all baselines across all benchmarks.

2. Related Work

Deep Metric Learning. In recent years, with the rapid development of deep learning, DML has been well-studied and widely used in many computer vision tasks [1, 32, 48, 46, 57, 50]. Most DML methods can be divided into two main categories: loss-based methods[6, 24, 40, 34, 11, 29, 33, 49] and hard mining methods [3, 4, 7, 44, 56, 47]. Loss-based methods impose a discriminative constraint on the image embeddings according to the label-driven similarity. For example, the contrastive loss [6] enforces the distance between positive pairs less than a threshold and between negative pairs greater than a threshold. The triplet loss [43, 32] uses a triplet including an anchor, a positive sample, and a negative sample, and enforces the anchor-negative distances to be larger than the anchor-positive distances by a fixed margin. Instead of directly measuring the similarity between sample pairs, proxy-based losses [11, 19, 29, 39, 60] consider the distance between samples and class-related proxies. Samples should be close to their ground-truth class proxies while far from the other class proxies. These methods are commonly introduced to address sampling complexity issues when sampling tuples.

Though these methods can perform well on unseen

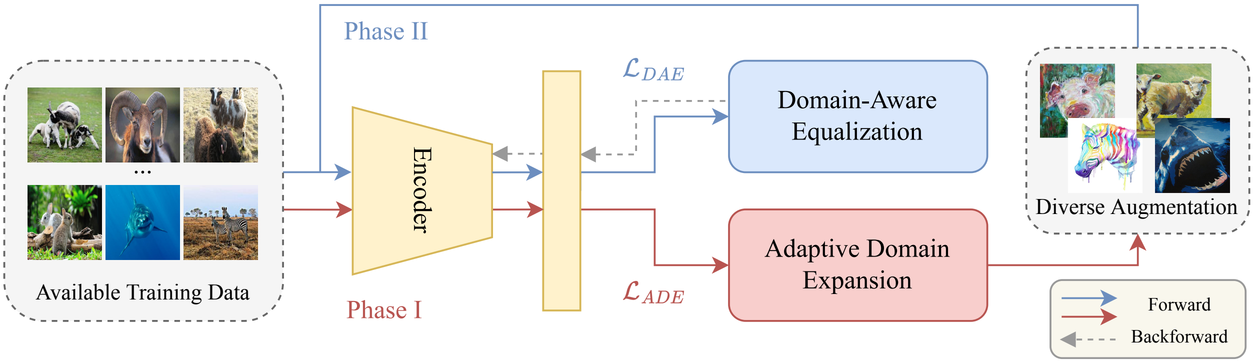

Figure 3. The overall framework architecture of the proposed self-expanded equalization method.

classes, performance degeneration exists when applied to unseen domains due to the domain shift. Therefore, exploring to learn better generalization ability for DML is still a challenging yet realistic problem.

Single-Domian Generalization. Domain discrepancy severely degrades the model performance when the existing domain shifts between training and testing data. To tackle this issue, many out-of-distribution generalization methods, including domain adaptation and domain generalization [5, 20, 53, 17] have been proposed.

Recently, a more challenging yet realistic task of single-domain generalization [37] is proposed, which aims to generalize a model trained on one source domain to many unseen target domains. To address this problem, a natural idea is to expand the source data and generate out-of-distribution samples. Based on this intuition, several methods, such as ADA [37], M-ADA [30], and ME-ADA [58], have been proposed. For example, ME-ADA [58] generates the adversarial samples by maximizing the entropy of the classifier.

Though these methods are effective for image classification, as the DML model utilizes similarity information among data rather than label information, these methods can not be directly applied to tackle the single-domain generalized DML problem.

3. Methodology

In this section, we formulate the proposed task of single-domain generalized DML and then introduce our self-expanded equalization framework in detail. As shown in Figure 3, two main phases are conducted alternatively. In phase I, the ADE module is provided to generate diverse augmentations and expand the source data distribution in a learnable manner. Simultaneously, we apply the DAE module to take full advantage of the augmentations with diversity to learn the domain-invariant metric that enhances the generalization ability of the metric model to unseen classes and domains.

3.1. Preliminary

Problem Formulation. We denote a training dataset with seen categories and domain by $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i$ and $y_i \in \{1, 2, \cdots, c\}$ are a sample and the corresponding label, respectively. In this paper, we hope to utilize the training dataset $\mathcal{D}$ to learn a DML model with strong generalization ability that can also perform well on testing dataset $\mathcal{T}$ with unseen classes and domains. To this end, DML will learn a discriminative feature embedding from the original space $\mathcal{X}$ to the d-dimensional metric space $\mathcal{Z}$ denoted as $f : \mathcal{X} \rightarrow \mathcal{Z} \subseteq \mathbb{R}^d$. In this way, we can calculate the similarity between samples in the metric space:

$$s(\mathbf{z}_i, \mathbf{z}_j) = \frac{< \mathbf{z}_i, \mathbf{z}_j >}{\|\mathbf{z}_i\|_2^2 \|\mathbf{z}_j\|_2^2}, \tag{1}$$

where $\mathbf{z}_i = f(\mathbf{x}_i)$, $\|\cdot\|_2$ denotes the ℓ_2 norm and $< \cdot, \cdot >$ denotes the inner product operation.

Proxy-based Loss. Considering the advantages of small sampling complexity and high convergence, we utilize a classical proxy-based loss [39] to optimize the DML model as follows:

$$\mathcal{L}_m = -\log \frac{e^{\alpha(s(\mathbf{z}_i, \mathbf{p}_{y_i}) - m)}}{e^{\alpha(s(\mathbf{z}_i, \mathbf{p}_{y_i}) - m)} + \sum_{j \neq y_i} e^{\alpha(s(\mathbf{z}_i, \mathbf{p}_j) - m)}}, \tag{2}$$

where $\mathbf{p}_j \in \{\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_c\}$ denotes the proxy of the j-th class, α is a scalar factor and m is a fixed margin.

3.2. Adaptive Domain Expansion

To effectively improve the generalization ability to unseen categories and domains, we hope the metric model can learn with diversity and broaden its horizons to wider data distribution. Inspired by this intuition, we propose an ADE module to dynamically augment the synthetic but reasonable out-of-distribution samples exploited to provide important semantic complements. The learned augmentations can effectively expand the source domain distribution and be regarded as hard samples enhancing the generalization ability to unseen data.

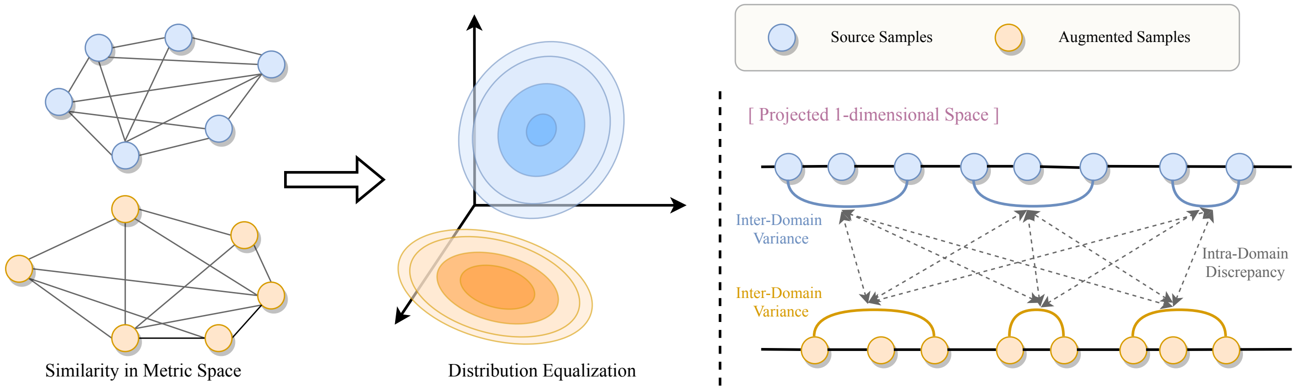

Figure 4. The simple illustration of the distribution equalization. (Left) The similarity graphs are employed to measure intra-domain data distributions for domain equalization. (Right) In a 1-dimensional sliced projection, we can model the inter-domain discrepancy according to the intra-domain variations.

In the ADE module, we take a 'min-max' strategy to adaptively update the source data samples to induce new augmented samples under different data distributions in a learnable manner. Specifically, we regard the pixels of the augmentation $\{\tilde{\mathbf{x}}_i\}_{i=1}^n$ as trainable parameters and utilize the corresponding source samples $\{\mathbf{x}_i\}_{i=1}^n$ to initialize them. And then, the augmentations can be adaptively optimized and updated during training.

High-level Similarity Minimum. On the one hand, we need to expand the source distribution and cover the distribution span of unseen data. Therefore, each augmented sample is enforced to keep far from its corresponding class proxy in the high-level semantic metric space:

$$\mathcal{L}_{high}^{(i)} = s(\tilde{\mathbf{z}}_i, \mathbf{p}_{y_i}), \tag{3}$$

where $\tilde{\mathbf{z}}_i = f(\tilde{\mathbf{x}}_i)$ is the embedded feature of the augmented sample $\tilde{\mathbf{x}}_i$ in the metric space, and $\mathbf{p}_{y_i}$ is the corresponding proxy vector learned by the proxy-based loss. In this way, the synthesized augmentations can capture more novel characteristics belonging to other unseen data distributions.

Low-level Similarity Maximum. On the other hand, we hope to keep the semantic consistency between the source and augmented data to preserve the semantic information. Accordingly, the mean squared error (MSE) loss in pixel level is adopted as follows:

$$\mathcal{L}_{low}^{(i)} = \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|_2^2. \tag{4}$$

To sum up, we can formulate the objective function for adaptive domain expansion as follows:

$$\mathcal{L}_{ADE} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{L}_{high}^{(i)} + \mathcal{L}_{low}^{(i)}\right). \tag{5}$$

3.3. Domain-Aware Equalization

After obtaining the augmented data $\mathcal{D}^+ = \{(\tilde{\mathbf{x}}_i, y_i)\}_i^n$, we can straightly employ the $\mathcal{A} = \mathcal{D} \cup \mathcal{D}^+$ as input to further fine-tune the metric model according to Eq. (2). However, treating all samples under different data distributions equally as input with corresponding labels discarded inter-domain variations that might contain helpful information to capture the domain-invariant characteristics and thus harm the generalization ability of the learned metric.

To address this issue, we apply the DAE module to use the diverse augmentations fully. Specifically, we analyze the distributions of the source and synthesized data according to the similarity among samples in the metric space and equalize the data distribution structure between source and augmented samples. Moreover, the learned augmentations can be regarded as hard samples to enhance the metric model's generalization ability.

Distribution-level Equalization. Data from different domains should have similar data distribution in the embedding space with the domain-invariant metric. Inspired by this intuition, we propose distribution equalization to penalize the inter-domain discrepancy and learn domain-invariant metric, which can further improve the generalization of the metric model. As illustrated in Figure 4, we first calculate the similarity graphs of source data and augmented data $\mathcal{G} = \{s(\mathbf{z}_i, \mathbf{z}_j)\}_{i,j=1}^n$ and $\mathcal{G}^+ = \{s(\tilde{\mathbf{z}}_k, \tilde{\mathbf{z}}_l)\}_{k,l=1}^n$, respectively. And then, we can indicate the inter-domain discrepancy between source and augmented data according to the discrepancy between intra-domain variations $\mathcal{G}$ and $\mathcal{G}^+$. In this procedure, we adopt the discrete optimal transport [28] to measure the inter-domain discrepancy since it can effectively induce the intrinsic geometrics of distributions. The corresponding Gromov Wasserstein distance between distributions $\mathcal{D}$ and $\mathcal{D}^+$ is formulated as follows:

$$\mathcal{W}(\mathcal{D}, \mathcal{D}^+) = \sum_{i,j,k,l} |s(\mathbf{z}_i, \mathbf{z}_j) - s(\tilde{\mathbf{z}}_k, \tilde{\mathbf{z}}_l)|^2 \Lambda_{ik}\Lambda_{jl}, \tag{6}$$

where $|\cdot|$ denotes the ℓ_1 norm, Λ_{ik} and Λ_{jl} are the corresponding items of coupling matrix $\Lambda \in \mathbb{R}^{n\times n}$ that is constrained to satisfy $\Lambda\mathbb{1}_n = \rho$ and $\Lambda^\top\mathbb{1}_n = \varrho$, where

$\mathbb{1}_n$ denotes a n-dimensional all-one vector and ρ, ϱ are weight vectors associated with $\mathbf{z}_i, \tilde{\mathbf{z}}_k$. In this paper, we set $\rho_i = 1/n, \varrho_k = 1/n, i, k \in [1, 2, c \ldots, n]$.

Considering that the solution of the distribution equalization described in Eq. (6) is a non-convex optimization problem, we adopt the sliced Gromov Wasserstein distance [36] to solve it. Specifically, we project the learned metric space into different 1-dimensional spaces with random directions. In this way, the sliced Gromov Wasserstein distance can be well approximated by capturing sample observations from the distributions shown in Figure 4. Formally, the sliced Gromov Wasserstein distance with T projection vectors $\{\pi_t\}_{t=1}^{T}$ is easy to calculated as follows:

$$\begin{aligned}\mathcal{L}_{de} = \frac{1}{T}\sum_{t=1}^{T}\sum_{i,j,k,l} &\left| s\left(\langle \mathbf{z}_i, \pi_t \rangle, \langle \mathbf{z}_j, \pi_t \rangle\right) \right. \\ &\left. - s\left(\langle \tilde{\mathbf{z}}_k, \pi_t \rangle, \langle \tilde{\mathbf{z}}_l, \pi_t \rangle\right) \right|^2 \mathit{\Lambda}_{ik}\mathit{\Lambda}_{jl}.\end{aligned} \tag{7}$$

Instance-level Equalization. In addition to inducing the domain-invariant metric in the distribution level, the learned augmentations can also be regarded as hard samples to achieve better generalization ability to unseen classes. The proxy-based loss enables a fast and safe convergence but does not consider the sample-to-sample relations, especially neglecting the rich semantic information of hard samples. By considering a large amount of hard negative pairs, the metric model can learn more discriminative feature embeddings. Specifically, given an augmented sample, it should be pushed closer to its corresponding source sample while pulling away from other augmented samples. The instance-level equalization loss is formulated as follows:

$$\mathcal{L}_{ie} = \frac{1}{n}\sum_{i=1}^{n} -\log \frac{e^{\alpha(s(\tilde{\mathbf{z}}_i, \mathbf{z}_i) - m)}}{\sum_{j \neq i} e^{\alpha(s(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j) - m)}}. \tag{8}$$

To sum up, we can formulate the objective function for domain-aware equalization as follows:

$$\mathcal{L}_{DAE} = \frac{1}{2n}\sum_{\mathcal{A}} \mathcal{L}_m + \tau_1 \mathcal{L}_{de} + \tau_2 \mathcal{L}_{ie}, \tag{9}$$

where τ_1, τ_2 are the hyperparameters.

3.4. Overall

To ensure the diversity of the domain expansion, we conduct the ADE module and DAE module alternatively. During training, these two modules are iteratively optimized according to Eq. (5) and Eq. (9), respectively. The augmentations will be updated every several epochs. In the beginning, we adopt the original training dataset $\mathcal{D}$ with the proxy-based loss described in Eq. (2) to warm up the metric model. The algorithm 1 details the approach of our SEE.

Algorithm 1 Training procedure of the proposed SEE.

```
Input: Training data $\mathcal{D}$, number of training epochs $E$, list
    of domain expansion epochs $\mathcal{N}$, and hyper-parameters
    $\alpha, \tau_1, \tau_2$.
Warm-up the metric model with parameter $\theta$.
for $e = 1, \ldots, E$ do
    ▷ ADE Step:
    if $e \in \mathcal{N}$ then
        $\mathcal{D}^+ = \emptyset$ ;
        $\tilde{\mathbf{x}}_i \leftarrow \mathbf{x}_i - \bigtriangledown \mathcal{L}_{ADE}$ according to Eq. (5);
        $\mathcal{D}^+ = \{\tilde{\mathbf{x}}_i\}_{i=1}^{n}$
    end if
    ▷ DAE Step:
    $\mathcal{A} = \mathcal{D} \cup \mathcal{D}^+$;
    $\theta \leftarrow \theta - \bigtriangledown \mathcal{L}_{DAE}$ according to Eq. (9);
end for
Output: A DML model with parameter $\theta$.
```

4. Experiments

In this section, we comprehensively evaluate the effectiveness of our SEE from different aspects. We first introduce the experimental settings and then evaluate the generalization ability of SEE compared with the state-of-the-art DML approaches. Finally, we present the ablation studies and qualitative results. More experimental results and analysis are provided in the *Supplementary Material*.

4.1. Experimental Settings

Datasets. To evaluate the performance of the proposed SEE, we conduct the experiments on two benchmarks, Cars196 [13] and CUB-200-2011 [38], and a large-scale real-world multi-domain dataset DomainNet [27]:

- **Cars196** [13] includes $16,185$ images from 196 car categories. The training set comprises $8,054$ car images from the first 98 categories, and the testing set comprises the remaining 8,131 images from the other 98 categories.
- **CUB-200-2011** [38] contains 11,788 images of 200 bird species. We use the first 100 species with 5,864 images for training and the remaining for testing.
- **DomainNet** [27] contains ~0.6 million natural images coming from six different data sources: Clipart (C), Infograph (I), Painting (P), Quickdraw (Q), Real (R), and Sketch (S) and 345 categories. Following the standard protocol [24], we utilize the first 173 categories for training and the rest for testing. Meanwhile, we train the model on real images and test it on data from the other domains.

To effectively evaluate the generalization ability, we use

Dataset	Training			Testing		
	# Samples	# Classes	Domain	# Samples	# Classes	Domain
CUB-200-2011	5,864	100	Real Image	5,924	100	Real Image
CUB-200-2011 Ext.	5,864	100	Real Image	5,924	100	Painting & Water-painting
Cars 196	8,054	98	Real Image	8,131	98	Real Image
Cars 196 Ext.	8,054	98	Real Image	8,131	98	Painting & Water-painting

Table 1. The statistics of two benchmark datasets in DML.

	Method	Real Image → Painting				Real Image → Water-painting			
		R@1	R@2	RP	MAP@R	R@1	R@2	RP	MAP@R
CUB-200-2011 Ext.	Contrastive loss [6]	31.93	42.94	15.11	7.35	23.63	32.61	11.39	4.74
	Triplet loss [9]	26.70	37.31	13.19	5.84	18.92	27.24	9.24	3.42
	Margin loss [44]	31.16	42.01	14.87	6.83	19.98	28.85	9.58	3.69
	Multi-similarity loss [40]	31.10	42.05	14.16	6.53	19.55	28.76	8.84	3.14
	Circle loss [34]	30.08	41.25	13.77	6.16	20.61	29.74	9.42	3.65
	ProxyNCA [19]	30.74	41.30	14.16	6.43	21.16	30.18	10.08	3.87
	CosFace [39]	31.55	41.78	14.78	7.34	22.75	32.74	11.58	5.03
	ProxyAnchor [11]	31.77	42.05	15.04	7.32	22.06	32.46	11.12	4.53
	Ours	**41.43**	**55.94**	**18.74**	**9.12**	**30.52**	**41.22**	**14.07**	**6.72**
Cars196 Ext.	Contrastive loss [6]	60.32	71.78	21.15	11.31	18.17	26.99	6.53	1.80
	Triplet loss [9]	33.35	45.18	15.24	6.13	11.29	17.19	3.85	0.76
	Margin loss [44]	41.84	54.10	15.82	7.00	16.03	24.34	5.95	1.56
	MS loss [40]	62.09	72.54	21.47	11.30	15.48	23.26	5.39	1.31
	Circle loss [34]	55.79	67.67	18.27	9.04	12.98	19.90	4.78	1.07
	ProxyNCA [19]	58.57	70.47	21.45	11.17	17.62	26.09	6.16	1.74
	CosFace [39]	58.66	70.36	21.76	11.87	15.63	22.69	5.32	1.42
	ProxyAnchor [11]	62.67	73.18	21.38	11.19	18.94	27.62	6.58	1.89
	Ours	**77.25**	**85.61**	**29.61**	**19.32**	**23.56**	**31.91**	**8.46**	**2.43**

Table 2. Experimental results (%) on CUB-200-2011 Ext. and Cars196 Ext. datasets.

generative models (Stylized Neural Painting [61] for painting and Paint Transformer [16] for water-painting) to extend the original benchmarks. The derived synthetic **Cars196 Ext.** and **CUB-200-2011 Ext.** have extra painting and water-painting images. We use the original real images for training, while the generated painting and water-painting images are adopted for testing. The statistics of these two benchmarks are summarized in Table 1. More details are given in the *Supplementary Material*.

Baselines. We compare our SEE with the state-of-the-art approaches including pair-based approaches contrastive loss [6], triplet loss [9], margin loss [44], multi-similarity loss (MS) [40], circle loss [34] and proxy-based approaches ProxyNCA (PA) [19], CosFace [39], and ProxyAnchor [11].

Evaluation Metrics. Following [21], to fairly and comprehensively evaluate different methods, we utilize the widely-used metric Recall@K (R@K), and two more stringent evaluation metrics, *i.e.*, R precision (RP), and MAP@R as evaluation metrics.

Implementation Details. In all experiments, the Pytorch [26] deep learning framework is used for implementation and the ADAM optimizer [12] is adopted to train the model. The ImageNet [31] pre-trained ResNet-50 [8] is adopted as the backbone model for fair comparison. In the image pre-processing procedure, we normalize all the

Method		*Contrastive [6]*	*Triplet [9]*	*Margin [44]*	*MS [40]*	*Circle [34]*	*CosFace [39]*	*ProxyNCA [19]*	*PA [11]*	*Ours*
R→S	Recall@1	51.30	48.47	48.52	53.10	51.61	49.59	54.14	54.88	**59.37**
	Recall@2	58.75	56.41	56.36	60.82	59.50	57.51	62.25	62.84	**67.68**
	R Precision	13.01	11.94	11.87	13.68	12.51	12.04	15.60	15.85	**19.43**
	MAP@R	6.48	5.73	5.66	6.83	6.13	5.88	8.19	8.53	**10.56**
R→I	Recall@1	33.46	33.05	33.06	33.99	33.61	33.51	34.86	35.46	**41.91**
	Recall@2	40.10	39.82	39.92	40.51	40.19	40.19	41.75	42.08	**51.47**
	R Precision	7.49	7.92	7.83	7.82	7.41	7.60	8.42	8.72	**10.32**
	MAP@R	3.38	3.56	3.48	3.69	3.42	3.51	3.86	4.20	**4.78**
R→P	Recall@1	57.79	55.86	55.83	58.49	58.17	57.42	58.32	58.07	**64.33**
	Recall@2	66.26	64.57	64.46	66.71	66.51	66.00	66.78	66.57	**71.82**
	R Precision	21.32	20.79	20.57	22.44	21.71	21.51	22.06	22.53	**25.19**
	MAP@R	13.13	12.54	12.34	14.22	13.65	14.18	13.51	14.13	**15.71**
R→Q	Recall@1	42.87	30.32	29.76	43.19	44.04	38.13	41.45	42.08	**50.19**
	Recall@2	55.24	40.95	39.92	54.84	56.05	49.75	52.90	53.64	**60.47**
	R Precision	14.90	9.49	9.20	15.16	15.23	12.68	14.56	14.80	**17.90**
	MAP@R	6.92	3.61	3.51	7.20	7.19	5.51	6.82	6.93	**9.27**
R→C	Recall@1	61.30	58.15	58.33	62.75	62.25	59.35	64.38	65.14	**69.94**
	Recall@2	69.12	66.29	66.61	70.90	70.21	67.32	72.27	72.93	**76.88**
	R Precision	18.68	17.62	17.49	20.09	18.74	17.21	21.73	22.82	**25.74**
	MAP@R	11.09	10.27	10.12	12.07	11.05	10.09	13.43	14.51	**17.28**
Average	Recall@1	49.34	45.17	45.10	50.31	49.94	47.60	50.62	51.13	**57.15**
	Recall@2	57.89	53.60	53.46	58.76	53.61	56.12	59.22	59.61	**65.66**
	R Precision	15.08	13.55	13.39	15.84	13.55	14.21	16.47	16.94	**19.72**
	MAP@R	8.20	7.14	7.02	8.80	8.29	7.65	9.17	9.66	**11.52**

Table 3. Experimental results (%) on DomainNet dataset.

images to 256×256 and then resize them to 224×224 by standard random cropping. For each iteration, we set the batch size to 120 on Cars196 Ext. and CUB-200-2011 Ext. datasets, and 512 on the DomainNet dataset. The embedding size is fixed to 512. During training, we update the augmentations every five epochs and iterate 50 times each augmentation generation. The learning rate is set to $1e-6$. We set $\tau_1 = 10, \tau_2 = 1$. The hyper-parameters in the proxy loss are set to the values searched by [21].

4.2. Experimental Results

Comparison on synthetic datasets. To verify the effectiveness of our proposed method, we make a comprehensive comparison of our SEE and the state-of-the-art methods on synthetic Cars196 Ext. [13] and CUB-200-2011 Ext. [38]. We use bold numbers to indicate the best results. The overall results are reported in Table 2, from which we can easily observe that the proposed SEE can outperform all baselines when testing on unseen categories and domains. In particular, our method outperforms the baseline proxy-based method [39] by $\sim$10%. This demonstrates that, through applying adaptive augmentation and equalization with diversity, our method can effectively improve the generalization ability of DML to unseen data distribution.

Comparison on the real-world dataset. To further demonstrate the generalization ability of our method un-

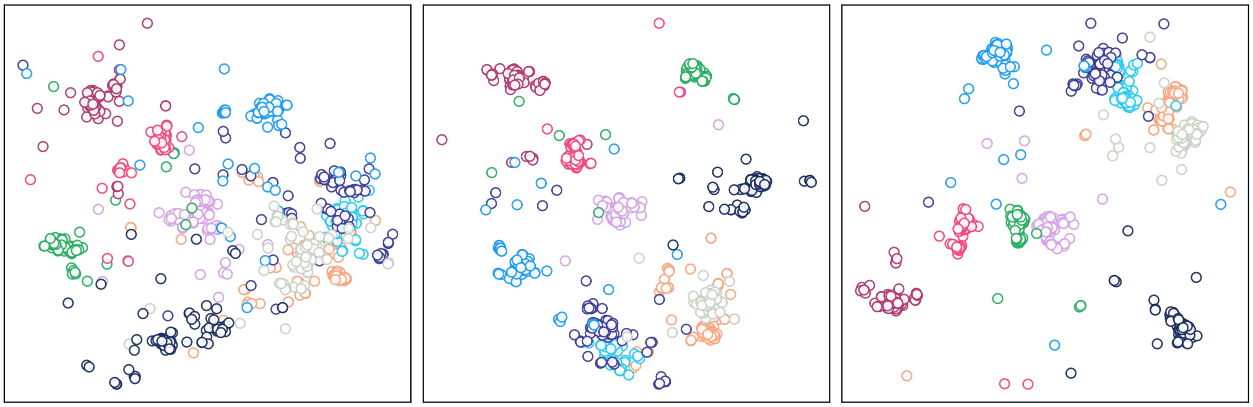
(a) Baseline (b) Baseline + ADE (c) Ours

Figure 5. The t-SNE [35] visualization results of the learned feature embeddings on the oil domain of the Cars196 Ext. dataset.

Method	Real Image → Painting				Real Image → Water-painting			
	R@1	R@2	RP	MAP@R	R@1	R@2	RP	MAP@R
Baseline [39]	31.55	41.78	14.78	7.34	22.75	32.74	11.58	5.03
Baseline with ADE module	36.14	46.29	16.47	8.26	26.51	38.19	12.21	5.54
Ours w/o $\mathcal{L}_{ie}$	39.82	52.78	17.41	8.84	29.20	40.11	13.23	6.38
Ours w/o $\mathcal{L}_{de}$	38.11	50.93	16.88	8.51	28.62	39.43	12.89	5.96
Ours	**41.43**	**55.94**	**18.74**	**9.12**	**30.52**	**41.22**	**14.07**	**6.72**

Table 4. Ablation analysis of our proposed self-expanded equalization on the CUB-200-2011 Ext. dataset.

der unknown data distribution, we also perform experiments on a real-world multi-domain dataset DomainNet [27]. We show the experimental results in Table 3. The bold numbers indicate the best results. Compared with the results when testing on the seen domain, the baseline DML methods suffer severe performance degradation. The reason may be that the discrepancy between the source and target domain is enormous, which makes the generalization more difficult. We can see that despite a large domain gap, our method outperforms these DML methods significantly. This generously supports our proposal that our method can effectively expand the training data and equalize inter-domain and intra-domain variance to improve the generalization of the DML model.

Visualization Results. Figure 5 shows the qualitative results of our method on the oil domain of the Cars196 Ext. dataset. We use the t-SNE [35] algorithm to visualize the learned metric space. We randomly chose 10 categories from the testing oil painting set of Cars196 Ext. dataset and visualize their feature embeddings learned by different methods. Compared with the baseline [39], our method learns the best metric space, demonstrating that our method can effectively improve the generalization ability of the DML model to unseen domains.

4.3. Ablation Study

Performance contributions of different components in the proposed method. We conduct an ablation study on different components of our method as shown in Table 4. Here, we adopt CosFase [39] as the baseline. We see that combining all elements achieves the best result, demonstrating each component's effectiveness.

Visualization Comparison. To better verify the effectiveness of each module in our method, we use the t-SNE [35] algorithm to visualize the metric space learned by different combinations of our methods as shown in Figure 5. It is easy to find that the proposed ADE module and DAE module both contribute to improving the model performance on unseen domains and categories.

5. Conclusion

In this paper, we focus on exploring the generalization ability of DML to unseen categories and domains. To further improve the generalization ability, we propose a more challenging yet realistic task, *i.e.*, single-domain generalized DML. To tackle this issue, we propose a self-expanded equalization method to expand the training data to various unseen distributions and effectively learn with diversity. Specifically, we take a 'min-max' strategy combined with our proxy-based metric loss to learn augmented samples in the ADE module adaptively. And then, we introduce the domain-aware equalization module to excavate the implicit inter-domain relations and make them into full use to learn the domain-invariant metric. Extensive experiments on two widely-used benchmark datasets in metric learning and a large-scale real-world multi-domain dataset demonstrate the effectiveness and superiority of SEE over the existing DML methods.

6. Acknowledgement

Our work was supported by Joint Fund of Ministry of Education of China (8091B022149), Key Research and Development Program of Shaanxi (2021ZDLGY01-03), National Natural Science Foundation of China (62132016, 62171343, and 62071361), and Fundamental Research Funds for the Central Universities (ZDRC2102). Erkun Yang was supported in part by the National Natural Science Foundation of China under Grant 62202365, Guangdong Basic and Applied Basic Research Foundation (2021A1515110026), and Natural Science Basic Research Program of Shaanxi (Program No.2022JQ-608).

References

[1] Shuo Chen, Gang Niu, Chen Gong, Jun Li, Jian Yang, and Masashi Sugiyama. Large-margin contrastive learning with distance polarization regularizer. In *ICML*, pages 1673–1683. PMLR, 2021.

[2] Xiang Deng and Zhongfei Zhang. Deep causal metric learning. In *ICML*, pages 4993–5006. PMLR, 2022.

[3] Yueqi Duan, Lei Chen, Jiwen Lu, and Jie Zhou. Deep embedding learning with discriminative sampling policy. In *CVPR*, pages 4964–4973, 2019.

[4] Yueqi Duan, Wenzhao Zheng, Xudong Lin, Jiwen Lu, and Jie Zhou. Deep adversarial metric learning. In *CVPR*, pages 2780–2789, 2018.

[5] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189. PMLR, 2015.

[6] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pages 1735–1742. IEEE, 2006.

[7] Ben Harwood, Vijay Kumar BG, Gustavo Carneiro, Ian Reid, and Tom Drummond. Smart mining for deep metric learning. In *ICCV*, pages 2821–2829, 2017.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[9] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[10] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *CVPR*, pages 1875–1882, 2014.

[11] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *CVPR*, pages 3238–3247, 2020.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[15] Zechao Li and Jinhui Tang. Weakly supervised deep metric learning for community-contributed image retrieval. *IEEE Transactions on Multimedia*, 17(11):1989–1999, 2015.

[16] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Ruifeng Deng, Xin Li, Errui Ding, and Hao Wang. Paint transformer: Feed forward neural painting with stroke prediction. In *ICCV*, pages 6598–6607, 2021.

[17] Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. In *NeurIPS*, 2022.

[18] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *CVPR*, pages 8046–8056, 2022.

[19] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *ICCV*, pages 360–368, 2017.

[20] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, pages 10–18. PMLR, 2013.

[21] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *ECCV*, pages 681–699. Springer, 2020.

[22] Bac Nguyen and Bernard De Baets. Kernel distance metric learning using pairwise constraints for person re-identification. *IEEE TIP*, 28(2):589–600, 2018.

[23] Hyun Oh Song, Stefanie Jegelka, Vivek Rathod, and Kevin Murphy. Deep metric learning via facility location. In *CVPR*, pages 5382–5390, 2017.

[24] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *ICCV*, pages 4004–4012, 2016.

[25] Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Deep metric learning with bier: Boosting independent embeddings robustly. *IEEE TPAMI*, 2018.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8026–8037, 2019.

[27] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019.

[28] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[29] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *CVPR*, pages 6450–6458, 2019.

[30] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *CVPR*, pages 12556–12565, 2020.

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

[32] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[33] Kihyuk Sohn. Improved deep metric learning with multiclass n-pair loss objective. In *NeurIPS*, pages 1857–1865, 2016.

[34] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6398–6407, 2020.

[35] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *JMLR*, 15(1):3221–3245, 2014.

[36] Titouan Vayer, Rémi Flamary, Romain Tavenard, Laetitia Chapel, and Nicolas Courty. Sliced gromov-wasserstein. In *NeurIPS*, volume 32, 2019.

[37] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *NeurIPS*, 31, 2018.

[38] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[39] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, pages 5265–5274, 2018.

[40] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*, pages 5022–5030, 2019.

[41] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *CVPR*, pages 6388–6397, 2020.

[42] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *ICCV*, pages 834–843, 2021.

[43] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *NeurIPS*, pages 1473–1480, 2006.

[44] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *CVPR*, pages 2840–2848, 2017.

[45] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, pages 14383–14392, 2021.

[46] Xinyi Xu, Zhengyang Wang, Cheng Deng, Hao Yuan, and Shuiwang Ji. Towards improved and interpretable deep metric learning via attentive grouping. *IEEE TPAMI*, 2022.

[47] Xinyi Xu, Yanhua Yang, Cheng Deng, and Feng Zheng. Deep asymmetric metric learning via rich relationship mining. In *CVPR*, pages 4076–4085, 2019.

[48] Jiexi Yan, Lei Luo, Cheng Deng, and Heng Huang. Unsupervised hyperbolic metric learning. In *CVPR*, pages 12465–12474, 2021.

[49] Jiexi Yan, Lei Luo, Cheng Deng, and Heng Huang. Adaptive hierarchical similarity metric learning with noisy labels. *IEEE TIP*, 32:1245–1256, 2023.

[50] Jiexi Yan, Erkun Yang, Cheng Deng, and Heng Huang. Metricformer: A unified perspective of correlation exploring in similarity learning. *NeurIPS*, 35:33414–33427, 2022.

[51] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *ECCV*, 2022.

[52] Xingyi Yang, Daquan Zhou, Songhua Liu, Jingwen Ye, and Xinchao Wang. Deep model reassembly. In *NeurIPS*, 2022.

[53] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *CVPR*, pages 7097–7107, 2022.

[54] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *ICPR*, pages 34–39. IEEE, 2014.

[55] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification. In *ECCV*, pages 188–204, 2018.

[56] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang. Hard-aware deeply cascaded embedding. In *ICCV*, pages 814–823, 2017.

[57] Borui Zhang, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Attributable visual similarity learning. *arXiv preprint arXiv:2203.14932*, 2022.

[58] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *NeurIPS*, 33:14435–14447, 2020.

[59] Wenzhao Zheng, Borui Zhang, Jiwen Lu, and Jie Zhou. Deep relational metric learning. In *CVPR*, pages 12065–12074, 2021.

[60] Yuehua Zhu, Muli Yang, Cheng Deng, and Wei Liu. Fewer is more: A deep graph metric learning perspective using fewer proxies. *arXiv preprint arXiv:2010.13636*, 2020.

[61] Zhengxia Zou, Tianyang Shi, Shuang Qiu, Yi Yuan, and Zhenwei Shi. Stylized neural painting. In *CVPR*, pages 15689–15698, 2021.