Confusion-Based Metric Learning for Regularizing Zero-Shot Image Retrieval and Clustering

Binghui Chen, Weihong Deng, *Member, IEEE*, Biao Wang, and Lei Zhang, *Fellow, IEEE*

***Abstract*— Deep metric learning turns to be attractive in zero-shot image retrieval and clustering (ZSRC) task in which a good embedding/metric is requested such that the unseen classes can be distinguished well. Most existing works deem this "good" embedding just to be the discriminative one and race to devise the powerful metric objectives or the hard-sample mining strategies for learning discriminative deep metrics. However, in this article, we first emphasize that the generalization ability is also a core ingredient of this "good" metric and it largely affects the metric performance in zero-shot settings as a matter of fact. Then, we propose the confusion-based metric learning (CML) framework to explicitly optimize a robust metric. It is mainly achieved by introducing two interesting regularization terms, i.e., the energy confusion (EC) and diversity confusion (DC) terms. These terms daringly break away from the traditional deep metric learning idea of designing discriminative objectives and instead seek to "confuse" the learned model. These two confusion terms focus on local and global feature distribution confusions, respectively. We train these confusion terms together with the conventional deep metric objective in an adversarial manner. Although it seems weird to "confuse" the model learning, we show that our CML indeed serves as an efficient regularization framework for deep metric learning and it is applicable to various conventional metric methods. This article empirically and experimentally demonstrates the importance of learning an embedding/metric with good generalization, achieving the state-of-the-art performances on the popular CUB, CARS, Stanford Online Products, and In-Shop datasets for ZSRC tasks.**

***Index Terms*— Confusion, generalization, image retrieval/clustering, regularization, zero-shot learning (ZSL).**

I. Introduction

SINCE the quest for algorithms that enable cognitive abilities is an important part of machine learning, zero-shot learning (ZSL) turns to be more attractive as it removes the limitation of category-consistency between training and testing sets. The model is required to learn concepts from seen classes and then to be capable of utilizing this learned knowledge to distinguish the unseen classes. ZSL has been widely explored in image classification [1]–[3] and retrieval tasks [4]–[6], etc. In this article, we focus on zero-shot image retrieval and clustering (ZSRC) tasks.

In order to accurately retrieve and cluster the unseen classes, most existing works employ deep metric learning to optimize a good embedding/metric, such as exploring tuple-based loss functions [6]–[13] and proposing efficient hard-sample mining strategies [9], [10], [14], etc. However, the above methods deem this "good" embedding/metric just to be the discriminative one and then concentrate on discriminative metric learning over the seen classes but neglect the importance of the generalization ability of the learned metric, which is more significant in ZSRC. As a result, without constraining the generalization/robustness these above methods are easily subject to the over-fitting problem on the seen classes, and some helpful or general knowledge for the unseen classes may have been left out with a high probability.

To be specific, in the ZSRC task, we emphasize that the generalization ability of the learned metric/embedding is seriously affected by the following two problems.

1) "The partial/biased learning behavior of deep models," specifically, as illustrated in Fig. 1(a),[1] for a functional learner parameterized by CNN, to correctly distinguish classes A and B, it will selectively learn the partial biased attribute concepts that are the easiest ones to reduce the current training loss over the seen classes (here head knowledge is enough to separate class A from B and thus is learned), instead of learning all-sided details and concepts. As a result, it yields over-fitting on the seen classes and generalizes worse to the unseen ones (i.e., classes C and D). In other words, deep networks easily learn to focus on surface statistical regularities rather than more general abstract concepts.
2) "Insufficient seen training data," as shown in Table I, zero-shot datasets are much smaller than the large-scale full-shot datasets (i.e., ImageNet), where there are only a few dozen or even several instances. When training deep models with a large number of parameters, this problem will further aggravate the over-fitting problem on the

Manuscript received 26 November 2020; revised 3 August 2021, 18 January 2022, and 2 April 2022; accepted 19 June 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62192784 and Grant 61871052. *(Corresponding author: Weihong Deng.)*

Binghui Chen and Biao Wang reside in Chaoyang District, Beijing 100022, China (e-mail: chenbinghui@bupt.cn; wangbiao225@foxmail.com).

Weihong Deng is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: whdeng@bupt.edu.cn).

Lei Zhang is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, SAR, China (e-mail: cslzhang@comp.polyu.edu.hk).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TNNLS.2022.3185668.

Digital Object Identifier 10.1109/TNNLS.2022.3185668

[1]In fact, the learned partial/biased knowledge is more complicated and cannot be easily illustrated in a figure; here for intuitive understanding, we show the toy example by some simple body-part knowledge.

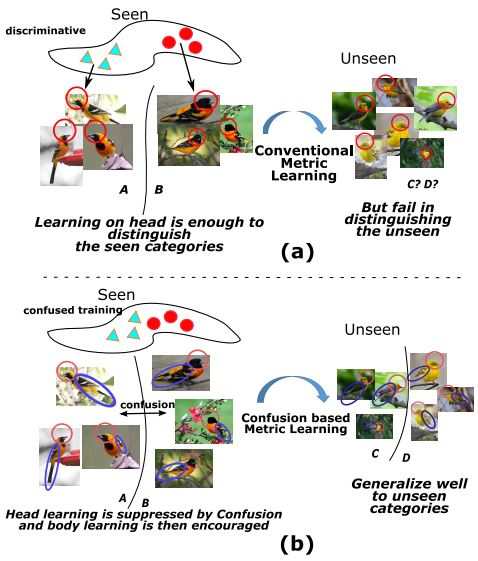

Fig. 1. Comparisons of conventional metric learning methods and our confusion-based metric learning (CML). In (a), deep model optimized by conventional metric learning will selectively learn head knowledge which is the easiest one to reduce the current training error and will omit some other helpful concepts. As a result, the testing classes cannot be distinguished well by the head. In (b), confusion terms are introduced to make the partial/biased head-based metric confused about itself, then as the training going, the confusion terms will regularize this metric to explore other complementary knowledge (even if this knowledge is not discriminative enough for the seen classes, it might be helpful for the unseen classes) and thus improve the generalization ability.

TABLE I

COMPARISON BETWEEN ZERO-SHOT DATASETS AND A LARGE-SCALE FULL-SHOT VISUAL CLASSIFICATION DATASET (IMAGENET). ZERO-SHOT DATASETS ARE MUCH SMALLER THAN IMAGENET

Dataset	ImageNet [19]	CUB [16]	CARS [15]	SOP [6]	In-Shop [17]
seen classes	1000	100	98	11318	3997
unseen classes	-	100	98	11316	3985
instances per class	1200	60	80	5	6
Zero-Shot	No	Yes	Yes	Yes	Yes

seen categories. Actually, the second issue can be mitigated by data augmentation, such as random mirroring and cropping, to some extent. In the following, we will dedicate to addressing the generalization problem caused by the first issue.

When optimizing the embedding/metric as in the aforementioned metric learning methods and without adopting the explicit robustness constraints, the poor generalization characteristics of the features are greatly exacerbated regardless of the efficacy of objective functions and hard-sample mining strategies. However, most ZSRC works ignore the importance of learning the robust descriptors. To this end, proposing an efficient regularization method for conventional metric learning to learn metrics with good generalization ability remains important, especially in ZSRC tasks.

In this article, we propose the confusion-based metric learning (CML) framework, an elegant regularization strategy, to alleviate the problem of over-fitting caused by the aforementioned first issue. It is mainly achieved by introducing two novel and simple confusion terms, i.e., the energy confusion (EC) term and the diversity confusion[2] (DC) term, which are "plug and play" and can be generally applied to many existing deep metric learning approaches. Both confusion terms play an adversarial role against the conventional metric learning objective. The EC term focuses primarily on the local feature confusion, while the DC term concentrates most on the global feature confusion; they are complementary to each other. Concretely, the EC term intends to minimize the expected value of the Euclidean distances between the paired images from two different categories that only present the local feature structure, while the DC term intends to minimize the diversity of the overall feature distribution (in other words, to make the overall feature distribution undiscriminating). As illustrated in Fig. 1(b), confusing the biased head-based metric will make the model less discriminative on the seen classes by reducing its dependence on head learning and thus give it chances of exploring other complementary and general knowledge. As a result, it can prevent over-fitting on the seen classes and improve the generalization ability of the embedding/metric in an adversarial manner. In other words, the confusion terms allow the SGD solver to escape from the "bad" local-minima region induced by the seen classes and to further explore more for the robust one. The main contributions of this work can be summarized as follows.

1) We emphasize that some crucial issue to ZSRC, i.e., the partial/biased learning behavior of deep model, is the key obstacle for improving the generalization ability of the learned embedding/metric.
2) We propose a (CML) framework to reinforce the robustness of the embedding/metric in an adversarial manner. The proposed EC term and the DC term focus on local and global feature confusion, respectively; they are "plug and play" and can work in conjunction with many existing metric methods.
3) Extensive experiments have been performed on several popular datasets for ZSRC, including CARS [15], CUB [16], Stanford Online Products [6], and In-shop [17].

This work is an extension to our earlier publication [18] which only concentrates on confusing the metric by the EC term from a local feature distribution confusion perspective. In this article, we further introduce a new DC term for confusing the embedding/metric from a global perspective which is complementary to the EC term. Employing either the EC term or the DC term can improve the performances of the unseen class retrieval/clustering, and when jointly optimizing both EC and DC terms, the performances can be further improved. The proposed CML framework is robust to the feature dimension and the adopted conventional metric objectives.

II. RELATED WORK

A. Zero-Shot Learning

ZSL removes the limitation of category consistency between training and testing sets and the model is required to learn

[2]Diversity is defined to describe the variance of the overall feature distribution.

concepts from seen classes and then to be capable of distinguishing the unseen classes. It turns to be attractive and has been widely explored in many computer vision tasks, such as classification [1], [2], [20] and retrieval [4], [5]. And recently, [21] focus on multi-label ZS classification. Ji *et al.* [22] aimed at cross-modal ZS hashing learning. Ji *et al.* [20] presented a triple discriminator GAN for ZS classification. Han *et al.* [23] proposed the hybrid feature generation model via instance-level and class-level supervision for classification task. Naeem *et al.* [24] proposed to use the graph structure of instances to enforce the relevant knowledge transfer from seen to unseen classes. Bo *et al.* [25] focused on the hardness sampling approach which is used to select and then train on samples from the unseen classes. A few works [26], [27] learned a more generalized visual-semantic mapping by iteratively training the model with a fixed number of pseudo labeled unseen-class samples, which is in the field of transductive ZSL. Liu *et al.* [28] proposed to learn features in hyperbolic space for capturing the hierarchical structure of semantic classes for zero-shot classification. Huynh and Elhamifar [29] proposed to use dense attribute-based attention mechanism to align each attribute-based feature with its attribute semantic vector. And [30] tries to learn redundancy-free features by projecting the original visual features into a new (redundancy-free) feature space and then restricting the statistical dependence between these two feature spaces. Most of these methods assume that they are capable of exploiting the extra auxiliary supervision information of the unseen classes (e.g., word representations of semantic class name and annotated attribute information); thus, they can explicitly align the extracted features of the unseen classes. However, in real-world applications, collecting and labeling these auxiliary information is time-consuming and impractical. Our CML concentrates on a more realistic scene where there are only seen class labels available. Therefore, improving the generalization of models remains important for ZSL.

B. Deep Metric Learning for ZSRC

Deep metric learning intends to optimize a feature representation of the input image that preserves the distance between similar data points small and dissimilar data points large in the feature space. The traditional contrastive [7] and triplet loss [10] have shown their powerful efficacy in broad applications. Additionally, there are some other deep metric learning works. Lifted [6] encourages each positive pair to compare the distances against all the negative pairs in one mini-batch, aiming to make full use of the mini-batch. Sampling-Matters [9] proposes a distance-weighted sampling strategy. Angular loss [11] optimizes a triangle-based angular function. Proxy-NCA [31] explains why popular classification loss works from a proxy-agent view, and its implementation is very similar to Softmax. ALMN [32] proposes to generate geometrical virtual negative point instead of employing hard-sample mining for learning discriminative embedding. DAML [33] utilize the generative adversary model to produce hard samples for the learning of discriminative embedding/metric. However, all the above methods are to cope with the metric by designing discriminative losses or exploring sample-mining strategies, thus suffer from the aforementioned issues easily. Additionally, HDC [8] employs the cascaded models and selects hard samples from different levels and models. BIER/ABIER losses [34], [35] adopts the online gradient boosting methods. ABE [36] and DeML [37] proposed to use attention-based ensemble metric. These methods try to improve the performances by resorting to the ensemble idea. However, different from all the above methods which aim at enhancing the discrimination of the metric, our CML has a clear object of improving the generalization ability of the learned metric via confusion.

C. Regularization Method

Regularization sometimes acts as an important role in deep learning since there are lots of parameters and the model is more likely to remember the training data instead of learning the more general abstract concepts. There are some interesting works aiming at improving the generalization, such as [38] introduce dropout to regularize the deep model training, [39], [40] add noise in the ReLU and Sigmoid activation functions, respectively, [41]–[43] add noise in weights and gradients, respectively. Moreover, some works intend to regularize deep models at the middle or the top layer. For example, [44]–[46] add the orthogonality regularization, [47] adds the uniform regularization, and [48], [49] add the L-2 regularization for the middle layers. For the top layer, [50] propose label-smoothing regularization technique for training deep models, [51] propose label-disturbing technique for improving the generalization ability of the deep models, [52] inject annealed noise into the softmax activations so as to boost the generalization ability by postponing the early Softmax saturation behavior, [53] propose to maximize the entropy over the class probabilities and later [54] provide a theoretical treatment of fine-grained classification based on maximum entropy. However, different from these methods which are performed on middle layers or mainly devised for the Softmax classifier layer, our CML mainly aims to promote the generalization ability of the deep metric learning by regularizing the output features.

III. Proposed Approach

In this section, we will first give insight about the conventional deep metric learning methods (see Section III-A), and then detail the proposed EC term (see Section III-B) and the DC term (see Section III-C), which concentrate on the local and global feature confusion, respectively. Finally we will provide the overall framework of our CML (see Section III-D). Notations are listed in Table XII in supplementary material.

A. Discriminative Deep Metric Learning

In general, deep metric learning follows the conventional metric learning idea to constrain the data similarity in the feature space, yet replaces the linear Mahalanobis matrix by a nonlinear function parameterized by a deep model. Now, denote the features encoded by the deep model by

$\{x_i\}_{i=1}^N \in \mathbb{R}^d$, d is the feature dimension, the corresponding label inputs are indicated by $\{y_i\}_{i=1}^N$, $y_i \in [1, \ldots, C]$, where C is the number of the seen classes. Then, the optimizing goal of deep metric learning is to make the distance measurement $D(x_i, x_j)$ in feature space as large as possible if $y_i \neq y_j$; otherwise, as small as possible, and then it can be formulated as

$$\theta_f = \arg\min_{\theta_f} L_m\big(\theta_f; T, D\big) \tag{1}$$

where L_m is some specific metric loss function, e.g., Hinge-like function and Softmax-like function, T indicates some instance-tuple, e.g., contrastive tuple $T(x_i, x_j)$ [7], triplet tuple $T(x_i, x_{i^+}, x_{i^-})$ [10], or N-pair tuple $T(x_i, x_{i^+}, x_{i_1^-}, \ldots, x_{i_{N-2}^-})$ [13], [32], D is the distance distribution measurement, e.g., Euclidean measurement [6], [8]–[10], [12] or inner-product measurement [13], [32], [34], and θ_f is the metric parameters to be learned.

As discussed in Section I, without taking the generalization ability into consideration explicitly, simply optimizing (1) with the well-designed objective function and instance-tuple as in most existing deep metric learning works would not lead a robust metric for ZSRC tasks, because: 1) the "partial/biased learning behavior of deep models" will mostly force the network to fit the surface statistical regularities rather than the more general abstract concepts, i.e., it will only highlight the concepts that are distinct for the seen classes instead of keeping all-sided information, resulting in over-fitting on the seen categories and limiting the generalization ability of the learned embedding; and 2) the "insufficient seen training data" will further deteriorate this over-fitting problem. Actually, the second issue can be mitigated by data augmentation, such as random mirroring and cropping. In the following, we will dedicate to addressing the generalization problem caused by the first issue.

Consider that the partial/biased learning behavior is actually induced by the nature of model training since in order to correctly distinguish the different seen classes, the deep metric has to be as confident about the feature distribution prediction over the current seen classes as possible (in other words, the output feature distribution should be discriminative enough where there is a large margin between classes, such that the training samples can be easily recognized), and as a result, only the partial biased knowledge that is discriminative for separating the seen categories, as shown in Fig. 1, are captured, while other potentially helpful knowledge are omitted.

To this end, a natural solution is to introduce an opposite optimizing objective, i.e., the feature distribution confusion term, into the conventional deep metric learning phase so as to "confuse" the network and reduce the discrimination of the output feature distribution over the seen classes. In other words, confusion could give model more chances of learning robust and general feature representations by continuously challenging model itself. As a result, model will not easily trust the surface statistical regularities and tend to focus more on the abstract general information. Motivated by this, we try to insert confusion learning into the current optimization objectives, i.e., (1). The confusion terms are optimized together with the conventional DML terms and the overall framework of our CML then can be formulated as

$$\min_{\theta_f} L = L_m\big(\theta_f; T, D\big) + \sum_i \lambda_i L_*\big(\theta_f\big) \tag{2}$$

where L_* indicates the confusion terms and λ_i means the hyper-parameter for balancing discrimination and generalization ability. The confusion terms should be good at perceiving feature manifold. We achieve this from both local and global feature confusion perspectives. Specifically, from local perspective, EC is proposed and it mainly try to confuse the model learning by making the feature representations from two different classes to be closer. Then from global perspective, DC is proposed and it mainly try to confuse the model learning by making the overall feature distribution with small diversity (less discriminative). These local and global confusions are mathematically complementary. Based on the definition of these two confusion terms, local confusion focuses more on the local feature space while the global confusion focuses more on the overall feature distribution. Therefore, they are complementary to each other though both target at making features less discriminative.

Below, we will describe these two confusion terms in detail.

B. Energy Confusion

For local feature distribution confusion, we would like to learn θ_f that make the output features from two different classes to be closer. It seems that the commonly adopted family of f-divergence for measuring the difference between two probability distributions might be a suitable choice, such as KL-divergence [55], Hellinger-distance [56], and total-variation-distance, however, we emphasize that they cannot be directly applied here since they mostly work with the probability measure (where $\sum_k x_{i,k} = 1$). Our confusion goal is based on the statistical distance between two random vectors following some probability distributions. To this end, we propose the EC term as follows:

$$\begin{aligned} L_{\text{ec}}\big(\theta_f; X_I, X_J\big) &= \mathbb{E}_{\widetilde{X_I}\widetilde{X_J}}\big(\|\widetilde{X_I} - \widetilde{X_J}\|_2^2\big) \\ &= \sum_{i,j} p_{i,j}\|x_i - x_j\|_2^2 \end{aligned} \tag{3}$$

where $\mathbb{E}$ indicates the expected value, X_I, X_J are two different class sets, $\widetilde{X_I}, \widetilde{X_J}$ are random feature vectors that obey some certain distribution, x_i, x_j are the corresponding feature observations, and $p_{i,j}$ is the joint probability. Since during training the samples are uniformly sampled and the classes are independent, we have $\widetilde{X_I} \sim \text{Uniform}(X_I)$, $\widetilde{X_J} \sim \text{Uniform}(X_J)$ and $p_{i,j} = p_i p_j = \frac{1}{N_I}\frac{1}{N_J}$ (N_* indicates the number of samples of the corresponding class).

From (3), one can observe that the EC term intends to minimize the expected distance between the randomly selected two different classes, so as to confuse the metric. This term serves as a local feature confusion term as it regularizes the feature distribution between the paired classes.

1) Discussion: Inferred from the above analysis, it seems that the commonly used statistical distance measurements general energy distance (GED) and maximum mean discrepancy (MMD) might be also useful here for confusing the network by

pushing different feature distributions closer. However, we will bridge our EC with these two methods, and illuminate the significance of our EC by theoretically accounting for why these two methods cannot be directly applied here.[3]

2) Relation to GED: Let $(\mathcal{Z}, \rho)$ be a semimetric space of negative type, and let $P, Q \in \mathcal{M}_+^1(\mathcal{Z}) \bigcap \mathcal{M}_\rho^1(\mathcal{Z})$, then the GED between P and Q, w.r.t. ρ is

$$D_{E,\rho}(P, Q) = 2\mathbb{E}_{\widetilde{P}\widetilde{Q}}\rho\big(\widetilde{P}, \widetilde{Q}\big) - \mathbb{E}_{\widetilde{P}\widetilde{P}'}\rho\big(\widetilde{P}, \widetilde{P}'\big) - \mathbb{E}_{\widetilde{Q}\widetilde{Q}'}\rho\big(\widetilde{Q}, \widetilde{Q}'\big) \tag{4}$$

where $\widetilde{P}, \widetilde{P}' \overset{\text{i.i.d.}}{\sim} P$ and $\widetilde{Q}, \widetilde{Q}' \overset{\text{i.i.d.}}{\sim} Q$. $D_{E,\rho}$ is a general extension of energy distance [57], [58] on metric space. Then we have:

Lemma 1: For two different class sets $X_I, X_J \in \mathcal{M}_+^1(\mathcal{Z}) \bigcap \mathcal{M}_\rho^1(\mathcal{Z})$, let ρ be squared Euclidean metric, i.e., $\|\cdot - \cdot\|_2^2$, then

$$L_{\text{ec}}\big(\theta_f; X_I, X_J\big) \geq \frac{1}{2} D_{E,\rho}(X_I, X_J).$$

Proof can be found in [18].

Remark: From Lemma 1, one can observe that our EC can be viewed as an upper bound of GED, minimizing this upper bound function is equivalent to optimizing GED to some extent. Moreover, it seems that directly optimizing GED with $\rho = \|\cdot - \cdot\|_2^2$ is reasonable as well, since GED itself is a statistical distance between two probability distributions. However, by comparing EC with GED, we emphasize that directly minimizing GED will additionally make $\mathbb{E}(\|\widetilde{X}_I - \widetilde{X}_I'\|_2^2) + \mathbb{E}(\|\widetilde{X}_J - \widetilde{X}_J'\|_2^2)$ large, i.e., making points in the same class be far away from each other which violates the basic discrimination criterion of metric learning and will degrade the model into a noisy counterpart, it is not what we desire. Therefore, GED cannot be directly applied here.

3) Relation to MMD: Let k be a kernel on $\mathcal{Z}$, and let $P, Q \in \mathcal{M}_+^1(\mathcal{Z}) \bigcap \mathcal{M}_k^{1/2}(\mathcal{Z})$. The MMD γ_k between P and Q is

$$\begin{aligned}\gamma_k^2(P, Q) &= \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}^2 = \big\|\mathbb{E}_{\widetilde{P}}k\big(\cdot, \widetilde{P}\big) - \mathbb{E}_{\widetilde{Q}}k\big(\cdot, \widetilde{Q}\big)\big\|_{\mathcal{H}_k}^2 \\ &= \mathbb{E}_{\widetilde{P}\widetilde{P}'}k\big(\widetilde{P}, \widetilde{P}'\big) + \mathbb{E}_{\widetilde{Q}\widetilde{Q}'}k\big(\widetilde{Q}, \widetilde{Q}'\big) - 2\mathbb{E}_{\widetilde{P}\widetilde{Q}}k\big(\widetilde{P}, \widetilde{Q}\big)\end{aligned} \tag{5}$$

where $\mu_k(*)$ is the kernel embedding, $\widetilde{P}, \widetilde{P}' \overset{\text{i.i.d.}}{\sim} P$ and $\widetilde{Q}, \widetilde{Q}' \overset{\text{i.i.d.}}{\sim} Q$. Then we have:

Lemma 2: For two different class sets $X_I, X_J \in \mathcal{M}_+^1(\mathcal{Z}) \bigcap \mathcal{M}_k^{1/2}(\mathcal{Z})$, let k be degree-1 homogeneous polynomial kernel, then

$$L_{\text{ec}}\big(\theta_f; X_I, X_J\big) \geq \gamma_k^2(X_I, X_J)$$

Proof can be found in [18].

Remark: From Lemma 2, one can observe that our EC can also be viewed as an upper bound of MMD. Moreover, it seems that directly optimizing MMD with degree-1 homogeneous polynomial kernel, i.e., $\gamma_k^2 = \|\mathbb{E}(\widetilde{X}_I) - \mathbb{E}(\widetilde{X}_J)\|_{\mathcal{H}_k}^2$, is reasonable as well, since many existing works employ this to pull two probability distributions closer, such as in transfer learning [59]–[61]. However, by expanding this γ_k^2, we have $\gamma_k^2 = \mathbb{E}(\widetilde{X}_I^T \widetilde{X}_I') + \mathbb{E}(\widetilde{X}_J^T \widetilde{X}_J') - 2\mathbb{E}(\widetilde{X}_I^T \widetilde{X}_J)$, and in this case, if we minimize γ_k^2 so as to pull different classes distributions closer and thus confuse the metric learning, we will additionally force $\mathbb{E}(\widetilde{X}_I^T \widetilde{X}_I') + \mathbb{E}(\widetilde{X}_J^T \widetilde{X}_J')$ to be small, which implicitly pushes the points within the same class further apart as their inner-products are getting small. These results also are not what we desire and will degrade the model into a noisy counterpart. Therefore, MMD cannot be directly applied here as well.

4) Remark Summary: We theoretically derive the relations between our EC and GED/MMD, and also reason about why GED/MMD cannot be directly applied here even if they have been widely adopted in many machine learning tasks for measuring probability distributions. Thus, we focus on locally "confusing" the metric learning via the EC term.

[3]Refer to Table XII in supplementary material or our original article [18] for some notations and preliminaries.

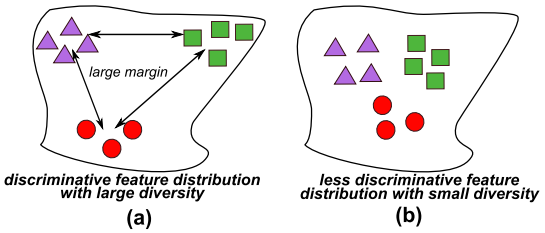

Fig. 2. Illustration of the diversity of feature distribution. (a) Since there is large margin between classes, the feature distribution is discriminative and has large diversity. (b) Margin between classes is small and all classes gather around a small region in the feature manifold, then the distribution is less discriminative and with small diversity.

C. Diversity Confusion

For global feature distribution confusion, we would like to learn θ_f that make the overall feature distribution with small diversity (less discriminative) as shown in Fig. 2(b). Specifically, to describe the diversity of the overall feature distribution, we have the DC term $\boldsymbol{\nu}$ as follows.

Definition 1: Given data distribution $X \sim p_x$, the diversity of feature distribution is

$$\boldsymbol{\nu}\big(\theta_f, p_x\big) \triangleq \sum_{k=1}^{d} \lambda_k \tag{6}$$

where

$$\{\lambda_k\}_{k=1}^d \text{ satisfy } \det\big(\Sigma^* - \lambda_k I_d\big) = 0$$

where Σ^* is the overall covariance matrix for the feature distribution and the eigenvalues $\{\lambda_1, \ldots, \lambda_d\}$ characterize the variance of the feature distribution across d dimensions. The definition of diversity in (6) is consistent with multivariate analysis, and is a common measure of the total variance of a data distribution [54], [62], [63].

From (6), one can observe that minimizing the diversity term intends to minimize the variance of the overall feature distribution so as to make the feature less discriminative as in Fig. 2(b) and thus confuse the metric. This term serves as a global feature confusion term as it regularizes the diversity of overall feature distribution.

However, directly minimizing (6) will significantly increase the computational cost, below, we will derive an equivalent term, an extremely simple one. Considering a commonly used hypothesis, we assume the deep feature x on the training seen classes follows a Gaussian mixture distribution with C Gaussian components where C is the number of classes, then it can be expressed as:

$$x \sim \sum_{c=1}^{C} \gamma_c \mathcal{N}(\mu_c, \Sigma_c), \quad \text{where } \forall c, \gamma_c \geq 0, \sum_{c=1}^{C} \gamma_c = 1$$

for each class c, Σ_c is the d-dimensional covariance matrices and μ_c is the corresponding mean vector. Without loss of generality, the above Gaussian mixture distribution has been re-centered to be of zero-mean, i.e., $\mathbb{E}_{X\sim p_X}[x] = 0$ and $\mu^* = \sum_{c=1}^{C} \gamma_c \mu_c = 0$. For this distribution, the equivalent covariance matrix can be given by

$$\Sigma^* \triangleq \sum_{c=1}^{C} \gamma_c \big(\mu_c \mu_c^T + \Sigma_c\big). \tag{7}$$

Proof: Using the law of conditional expectation, we have

$$\begin{aligned}\Sigma^* &= \mathrm{Cov}(x) = \mathbb{E}\big(xx^T\big) - \mathbb{E}(x)\mathbb{E}\big(x^T\big) \\ &= \mathbb{E}_c\mathbb{E}\big(xx^T|c\big) - 0 = \sum_{c}^{C} \gamma_c \mathbb{E}\big(xx^T|c\big) \\ &= \sum_{c}^{C} \gamma_c \big(\mu_c \mu_c^T + \Sigma_c\big)\end{aligned}$$

■

Moreover, we have the following Lemmas.

Lemma 3: For a d-dimensional multivariate normal distribution $z \sim \mathcal{N}(\mu, \Sigma)$, we have

$$\mathbb{E}\big(\|z\|_2^2\big) = \mathrm{tr}(\Sigma) + \|\mu\|_2^2 \tag{8}$$

Proof:

$$\begin{aligned}\mathbb{E}\big(\|z\|_2^2\big) &= \mathbb{E}\left(\sum_{i=1}^{d} z_i^2\right) = \sum_{i=1}^{d} \mathbb{E}\big(z_i^2\big) = \sum_{i=1}^{d} \big(Var(z_i) + \big(\mathbb{E}(z_i)^2\big)\big) \\ &= \mathrm{tr}(\Sigma) + \sum_{i=1}^{d} (\mathbb{E}(z_i))^2 = \mathrm{tr}(\Sigma) + \|\mu\|_2^2\end{aligned}$$

■

Lemma 4: For a random vector z that follows d-dimensional Gaussian mixture distribution with C Gaussian components, i.e., $z \sim \sum_{c=1}^{C} \gamma_c \mathcal{N}(\mu_c, \Sigma_c)$ where $\forall c\ \gamma_c \geq 0$ and $\sum_{c=1}^{C} \gamma_c = 1$, we have:

$$\mathbb{E}\big(\|z\|_2^2\big) = \sum_{c=1}^{C} \gamma_c \big(\mathrm{tr}(\Sigma_c) + \|\mu_c\|_2^2\big). \tag{9}$$

Proof: Using the law of conditional expectation, we have

$$\mathbb{E}\big(\|z\|_2^2\big) = \mathbb{E}_c\mathbb{E}\big(\|z\|_2^2|c\big) = \sum_{c=1}^{C} \gamma_c \mathbb{E}\big(\|z\|_2^2|c\big).$$

As the conditional distribution is a d-dimensional Gaussian distribution, i.e., $\mathcal{N}(\mu_c, \Sigma_c)$, then from Lemma 3 we can obtain the following:

$$= \sum_{c=1}^{C} \gamma_c \big(\mathrm{tr}(\Sigma_c) + \|\mu_c\|_2^2\big).$$

■

Then, based on (7) and Lemma 4, we can derive Definition 1 to the following equivalent counterpart.

Definition 2: Given data distribution $X \sim p_X$, the diversity of the d-dimensional Gaussian mixture feature distribution is

$$\begin{aligned}\boldsymbol{\nu}\big(\theta_f, p_X\big) &\triangleq \sum_{k=1}^{d} \lambda_k = \mathrm{tr}\big(\Sigma^*\big) = \mathrm{tr}\left(\sum_{c=1}^{C} \gamma_c \big(\mu_c \mu_c^T + \Sigma_c\big)\right) \\ &= \sum_{c=1}^{C} \gamma_c \big(\mathrm{tr}(\Sigma_c) + \|\mu_c\|_2^2\big) = \mathbb{E}\big(\|x\|_2^2\big).\end{aligned} \tag{10}$$

As shown in Definition 2, the computation of diversity can be simplified, and we will employ (10) as the final DC term throughout our experiments. Moreover, the deep model is actually optimized by mini-batch SGD, in other words, during each iteration, we instead compute the empirical average $\mathbb{E}_{X\sim p_B}(\|x\|_2^2)$ where B indicate the dataset in one mini-batch. And the estimate error can be bounded as follows:

Lemma 5: For mini-batch B of size b and actually $\|x\|_2^2 \in [0, \zeta]$ (ζ is the largest value over the training set), with probability at least $1 - \delta/2$, we have

$$\left|\mathbb{E}_{X\sim p_B}\left(\|x\|_2^2\right) - \mathbb{E}_{X\sim p_X}\left(\|x\|_2^2\right)\right| \leq \zeta \sqrt{\frac{1}{2b} \log \frac{4}{\delta}}$$

Proof:

$$\mathbb{E}_{X\sim p_X}\left(\|x\|_2^2\right) = \int_{X\sim p_X} \|x\|_2^2 p_X dx$$

$$\mathbb{E}_{X\sim p_B}\left(\|x\|_2^2\right) = \frac{1}{b} \sum_{i=1}^{b} \|x_i\|_2^2.$$

As B has i.i.d. samples of X, we have

$$\mathbb{E}_{X\sim p_X}\left(\mathbb{E}_{X\sim p_B}\left(\|x\|_2^2\right)\right) = \frac{1}{b} \sum_{i=1}^{b} \mathbb{E}_{X\sim p_X}\left(\|x\|_2^2\right) = \mathbb{E}_{X\sim p_X}\left(\|x\|_2^2\right).$$

Then, from Hoeffding Inequality, we have

$$P\Big\{\left|\mathbb{E}_{X\sim p_B}\left(\|x\|_2^2\right) - \mathbb{E}_{X\sim p_X}\left(\mathbb{E}_{X\sim p_B}\left(\|x\|_2^2\right)\right)\right| \geq t\Big\} \leq 2 \exp \frac{-2b^2t^2}{b\zeta^2}.$$

Setting RHS as $\delta/2$, thus we have with probability at least $1 - \delta/2$

$$\left|\mathbb{E}_{X\sim p_B}\left(\|x\|_2^2\right) - \mathbb{E}_{X\sim p_X}\left(\|x\|_2^2\right)\right| \leq \zeta \sqrt{\frac{1}{2b} \log \frac{4}{\delta}}.$$

■

Remark: Although we employ $\mathbb{E}_{X\sim p_B}(\|x\|_2^2)$ to minimize the diversity of the overall feature distribution, the estimation error can be bounded by Lemma 5. Thus, we can globally "confuse" the deep metric learning via this DC term.

D. Confusion-Based Metric Learning

As discussed above, the learned metric represents the learned concepts to some extent, and the more discriminative the feature distribution is on the seen classes, the greater the risk of concept over-fitting. The proposed two confusion terms are complementary to each other and serve as the regularization techniques that insist to prevent the feature being over-discriminative about the seen classes and to mitigate the biased learning issue by avoiding the learner being stuck in the training-data-specific concepts. In other words, the deep metric learning is regularized by explicitly reducing model's dependence on the partial/biased knowledge, and this is mainly achieved by the idea of feature distribution confusion. Additionally, "confusing" also gives SGD solver chances of escaping from the "partial" and "bad" local-minima that induced by the seen classes, and then exploring other solution regions for the more "general" ones.

Then, employing (3) and (10) for locally and globally confusing the deep metric learning, respectively, the framework of CML can be generally applied to various deep metric learning objectives, where we simultaneously train our confusion terms and the distance metric term as follows:

$$\min_{\theta_f} L = L_m\big(\theta_f; T, D\big) + \lambda_1 L_{\text{ec}}\big(\theta_f; X_I, X_J\big) + \lambda_2 \boldsymbol{\nu}\big(\theta_f, p_x\big) \tag{11}$$

where λ_1, λ_2 are the tradeoff hyper-parameters, the class sets X_I, X_J are randomly chosen from the current mini-batch. In order to demonstrate the effectiveness of the proposed CML framework, we develop various SOTA metric learning objective functions here, i.e., $L_m(\theta_f; T, D)$:

1) CML (Tri): For triplet-tuple T and Euclidean measurement D, we employ [10], [64]

$$L_m\big(\theta_f; T, D\big) = \sum_i^N \big[\|x_i - x_{i^+}\|_2^2 - \|x_i - x_{i^-}\|_2^2 + m\big]_+ \tag{12}$$

where the objective limits the distances of negative pairs larger than that of the positive pairs by margin m and features x_i is first L2 normalized, we experimentally find $m = 0.1$ performs best.

2) CML (N-Pair): For N-tuple T and inner-product measurement D, we employ [13]

$$L_m\big(\theta_f; T, D\big) = \sum_{i=1}^N \log\left(1 + \sum_{j=1, y_j \neq y_i}^N \exp\big(x_i^T x_j - x_i^T x_{i^+}\big)\right) \tag{13}$$

where the objective limits the inner product of each negative pair $x_i^T x_j$ smaller than that of the positive pair $x_i^T x_{i^+}$.

3) CML (Binomial): For contrastive-tuple T and cosine measurement D, we employ [34], [65]

$$L_m\big(\theta_f; T, D\big) = \sum_{i,j} \log\Big(1 + e^{-(2s_{ij}-1)\alpha(D_{ij}-\beta)\eta_{ij}}\Big) \tag{14}$$

where $s_{ij} = 1$ if x_i, x_j are from the same class; otherwise $s_{ij} = 0$, $\alpha = 2, \beta = 0.5$ are the scaling and translation parameters, respectively, η_{ij} is the penalty coefficient and is set to 1 if $s_{ij} = 1$; otherwise $\eta_{ij} = 25$, $D_{ij} = (x_i^T x_j / \|x_i\| \|x_j\|)$.

Remark: As in (11), our CML is**** achieved by jointly training the conventional deep metric objective and the proposed confusion terms. They form an adversarial learning scheme by optimizing the opposite objective functions. Specifically, L_m acts as a "defender" and $\{L_{\text{ec}}(\theta_f; X_I, X_J), \boldsymbol{\nu}(\theta_f, p_x)\}$ act as the "attacker"; the attacker intends to make the feature distribution less discriminative by minimizing the local inter-class distances and the global feature diversity, so as to confuse the metric and make it confound with the training data, while in order to correctly distinguish the training data, the defender has to learn more "general" and complementary concepts. As the defending-attacking going, the learned metric will be less likely to the prejudiced concepts, thus successfully prevent the "partial/biased learning behavior" and improve the generalization ability.

IV. Experiments and Results

A. Implementation Details

Following many other works, e.g., [6], [13], [18], we choose the pre-trained GooglenetV1 [66] as our backbone CNN, add an new fully-connected layer after the Global-average-pooling layer and randomly initialize it. If not specified, we set the feature dimension to be 512 throughout our experiments. We also adopt exactly the same data preprocessing method [6] so as to make fair comparisons with other works. For training, the optimizer is Adam [67] with learning rate $1e-5$ and weight decay $2e-4$. The training iterations are $5k$ (CUB), $10k$ (CARS), $20k$ (Stanford Online Products and In-Shop), respectively. The new fc-layer is optimized with ten times learning rate for fast convergence. Moreover, for fair comparison, we use minibatch of size 128 throughout our experiments, which is composed of 64 random selected classes with two instances each class. Our work is implemented by Caffe [68] framework.

For CARS, CUB, Stanford Online Products, and In-Shop datasets, the used values of λ_1/λ_2 are shown in Tables II–V.

B. Evaluation and Datasets

The same as many other works, the retrieval performance is evaluated by Recall@K metric. And following [6], we evaluate the clustering performances via normalized mutual information (NMI) and F_1 metrics. The input of NMI is a set of clusters $\Omega = \{\omega_1, \ldots, \omega_K\}$ and the ground truth classes $\mathbb{C} = \{c_1, \ldots, c_K\}$, where ω_i represents the samples that belong to the ith cluster, and c_j is the set of samples with label j. NMI is defined as the ratio of mutual information and the mean entropy of clusters and the ground truth, $\text{NMI}(\Omega, \mathbb{C}) = (2I(\Omega, \mathbb{C})/H(\Omega) + H(\mathbb{C}))$, and F_1 metric is the harmonic mean of precision and recall as follows $F_1 = (2PR/P + R)$. Then our CML is evaluated over the widely used benchmarks with the standard zero-shot evaluation protocol [6]. 1)

1) CARS [15] contains 16 185 car images from 196 classes. We split the first 98 classes for training (8054 images) and the rest 98 classes for testing (8131 images).

TABLE II

COMPARISONS(%) WITH STATE-OF-THE-ART METHODS ON CARS [15] DATASET. THE VALUE OF λ_1 USED OVER (TRI, N-PAIR,BINOMIAL) ARE {0.02, 0.3, 0.13}, RESP, AND VALUE OF λ_2 USED OVER (TRI,N-PAIR,BINOMIAL) ARE {0.04, 0.02, 0.03}, RESPECTIVELY

CARS					
Method	R@1	R@2	R@4	NMI	F1
Lifted [6]	49.0	60.3	72.1	55.1	21.5
Clustering [69]	58.1	70.6	80.3	59.0	-
Angular [11]	71.3	80.7	87.0	62.4	31.8
ALMN [32]	71.6	81.3	88.2	62.0	29.4
DAML [33]	75.1	83.8	89.7	66.0	36.4
HORDE* [70]	80.7	88.1	92.6	67.9	37.2
Triplet	68.4±0.15	78.3±0.07	86.2±0.01	58.8±0.25	26.7±0.18
CML-e (Tri)	*81.2±0.17*	*88.0±0.06*	*92.8±0.00*	*66.0±0.19*	*32.8±0.23*
CML-d (Tri)	*82.4±0.14*	*89.5±0.07*	*93.5±0.00*	*67.8±0.25*	*36.9±0.16*
CML (Tri)	***83.0±0.16***	***89.7±0.03***	***93.8±0.00***	***67.7±0.12***	***38.3±0.07***
N-Pair	74.5±0.26	83.7±0.11	90.2±0.08	60.7±0.26	29.7±0.22
CML-e (N-Pair)	*80.3±0.18*	*88.2±0.06*	*92.4±0.00*	*64.7±0.15*	*32.2±0.08*
CML-d (N-Pair)	*78.9±0.14*	*86.9±0.07*	*91.9±0.00*	*66.3±0.12*	*36.1±0.10*
CML (N-Pair)	***81.2±0.15***	***88.6±0.05***	***92.7±0.00***	***66.7±0.17***	***36.6±0.13***
Binomial	74.2±0.12	83.1±0.04	86.7±0.00	61.7±0.09	28.8±0.04
CML-e (Binomial)	*84.4±0.12*	*90.4±0.05*	*93.8±0.00*	*68.4±0.11*	*38.2±0.05*
CML-d (Binomial)	*83.4±0.14*	*89.7±0.02*	*93.6±0.00*	*68.4±0.09*	*37.2±0.14*
CML (Binomial)	***85.1±0.14***	***90.8±0.04***	***94.0±0.00***	***69.3±0.14***	***39.1±0.08***

TABLE III

COMPARISONS(%) WITH STATE-OF-THE-ART METHODS ON STANFORD ONLINE PRODUCTS [6]. THE VALUE OF λ_1 USED OVER (TRI, N-PAIR, BINOMIAL) ARE {0.002, 0.03, 0.013} AND VALUE OF λ_2 USED OVER (TRI, N-PAIR, BINOMIAL) ARE {0.01, 0.02, 0.01}, RESPECTIVELY

Stanford Online Products					
Method	R@1	R@10	R@100	NMI	F1
Lifted [6]	62.1	79.8	91.3	87.4	24.7
Clustering [69]	67.0	83.7	93.2	89.5	-
Angular [11]	70.9	85.0	93.5	87.8	26.5
ALMN [32]	69.9	84.8	92.8	-	-
DAML [33]	68.4	83.5	92.3	89.4	32.4
HORDE* [70]	71.4	85.6	93.6	89.7	32.5
Triplet	57.8±0.03	75.6±0.01	88.5±0.00	86.8±0.05	20.8±0.04
CML-e (Tri)	*64.9±0.02*	*80.0±0.01*	*90.5±0.00*	*86.8±0.06*	*23.4±0.07*
CML-d (Tri)	*65.1±0.01*	*80.2±0.01*	*90.7±0.00*	*87.2±0.12*	*23.0±0.05*
CML (Tri)	***66.3±0.01***	***80.8±0.00***	***91.0±0.00***	***87.3±0.05***	***24.3±0.07***
N-Pair	68.0±0.01	84.0±0.00	93.1±0.00	87.5±0.08	25.8±0.02
CML-e (N-Pair)	*69.8±0.01*	*84.7±0.00*	*93.2±0.00*	*88.1±0.03*	*27.5±0.06*
CML-d (N-Pair)	*69.8±0.00*	*85.1±0.00*	*93.4±0.00*	*88.1±0.01*	*27.8±0.00*
CML (N-Pair)	***70.5±0.01***	***85.2±0.00***	***93.4±0.00***	***88.1±0.02***	***28.0±0.03***
Binomial	68.5±0.02	84.0±0.00	93.1±0.00	88.5±0.08	29.9±0.05
CML-e (Binomial)	*71.3±0.03*	*85.6±0.01*	*93.6±0.01*	*89.9±0.16*	***32.7±0.14***
CML-d (Binomial)	*71.2±0.02*	*85.4±0.01*	*93.5±0.00*	*89.7±0.15*	*32.6±0.09*
CML (Binomial)	***72.2±0.05***	***85.9±0.02***	***93.7±0.01***	***89.9±0.11***	*32.6±0.05*

2) CUB [16] includes 11 788 bird images from 200 classes. We use the first 100 classes for training (5864 images) and the rest 100 classes for testing (5924 images).
3) Stanford Online Products [6] has 11 318 classes for training (59 551 images) and the other 11 316 classes for testing (60 502 images).
4) In-Shop [17] contains 3997 classes for training (25 882 images) and the resting 3985 classes for testing (28 760 images). The test set is partitioned into the query set of 3985 classes (14 218 images) and the retrieval database set of 3985 classes (12 612 images).

For ZSRC task, there is no intersection of classes between training and testing sets.

Notation: We use "CML" to indicate the proposed method. Moreover, to demonstrate the effectiveness of each confusion term, we use "CML-e/CML-d" to indicate using the EC/DC term alone.[4]

C. Comparison With State-of-the-Art

To highlight the significance of the proposed CML framework, we compare it with the aforementioned corresponding baseline methods, i.e., the wildly used triplet loss [10], N-Pair loss [13], and binomial loss [65]. Moreover, we also compare CML with other state-of-the-art deep metric methods, including Lifted loss [6], Clustering [69], Angular [11], ALMN [32], DAML [33], and HORDE [70]. The experimental results over CARS [15], CUB [16], Stanford Online Products [6] and In-shop Clothes [17] are in Tables II–V respectively, bold number indicates the improvement over the baseline method. For our implemented results, we re-run each model five times and report the mean value and standard deviation in tables.

From these tables, one can first observe that both the proposed CML-e/CML-d methods can consistently improve the performances of the conventional deep metric learning methods (i.e., triplet loss, N-pair loss, and binomial loss) on all the benchmark datasets by a large margin, demonstrating the necessity of explicitly enhancing the generalization ability of the learned deep metric and validating the universality and effectiveness of the proposed local/global confusion terms. Furthermore, by jointly employing the local and global confusion terms, i.e., the EC and DC terms which are complementary to each other, the overall CML framework can further improve the final performances on these datasets. Meanwhile, the proposed CML (binomial) also surpasses all the listed state-of-the-art deep metric learning approaches on these four benchmarks, demonstrating the superiority of our method and the importance of learning robust feature descriptors. In summary, learning "general" concepts by avoiding the biased learning behavior is more important in ZSRC tasks, and the generalization ability of the optimized metric heavily affects

Notation: We use 'CML' to indicate the proposed method. Moreover, to demonstrate the effectiveness of each confusion term, we use 'CML-e/CML-d' to indicate using the EC/DC term alone [4].

TABLE IV

COMPARISONS(%) WITH STATE-OF-THE-ART METHODS ON CUB [16] DATASET. THE VALUE OF λ_1 USED OVER (TRI,N-PAIR,BINOMIAL) ARE {0.02, 0.3, 0.13}, RESPECTIVELY, AND VALUE OF λ_2 USED OVER (TRI,N-PAIR,BINOMIAL) ARE {0.01, 0.02, 0.03}, RESPECTIVELY

CUB					
Method	R@1	R@2	R@4	NMI	F1
Lifted [6]	47.2	58.9	70.2	56.2	22.7
Clustering [69]	48.2	61.4	71.8	59.2	-
Angular [11]	53.6	65.0	75.3	61.0	30.2
ALMN [32]	52.4	64.8	75.4	60.7	28.5
DAML [33]	52.7	65.4	75.5	61.3	29.5
HORDE* [70]	55.7	66.9	77.0	61.8	31.1
Triplet	48.9±0.24	62.1±0.18	73.0±0.07	56.9±0.33	24.7±0.25
CML-e (Tri)	*53.8±0.22*	*64.6±0.13*	*75.2±0.08*	*59.4±0.14*	*27.0±0.21*
CML-d (Tri)	*55.2±0.19*	*66.8±0.09*	*76.3±0.04*	*60.8±0.22*	*29.0±0.27*
CML (Tri)	***55.3±0.13***	***67.1±0.04***	***76.5±0.01***	***61.5±0.25***	***29.2±0.18***
N-Pair	51.2±0.27	63.1±0.15	73.8±0.03	58.4±0.34	26.8±0.19
CML-e (N-Pair)	*53.4±0.22*	*65.0±0.10*	*75.5±0.05*	*60.1±0.21*	*28.6±0.18*
CML-d (N-Pair)	*53.3±0.24*	*65.8±0.14*	*76.4±0.05*	*60.3±0.22*	*28.6±0.16*
CML (N-Pair)	***54.6±0.17***	***66.1±0.10***	***76.8±0.03***	***60.9±0.23***	***29.3±0.21***
Binomial	53.1±0.13	65.1±0.09	75.3±0.02	59.4±0.28	26.5±0.24
CML-e (Binomial)	*55.7±0.17*	*66.5±0.08*	*76.7±0.04*	*61.8±0.25*	*30.5±0.16*
CML-d (Binomial)	*56.2±0.16*	*66.9±0.04*	*76.9±0.01*	*61.3±0.18*	*31.1±0.12*
CML (Binomial)	***56.7±0.13***	***67.3±0.02***	***77.1±0.00***	***62.8±0.12***	***31.7±0.09***

[4]In our earlier publication [18], "CML-e" is called "ECAML."

TABLE V

COMPARISONS(%) WITH STATE-OF-THE-ART METHODS ON IN-SHOP CLOTHES [17] DATASETS. THE VALUE OF λ_1 USED OVER (TRI,N-PAIR,BINOMIAL) ARE {0.002, 0.03, 0.013}, RESPECTIVELY, AND VALUE OF λ_2 USED OVER (TRI,N-PAIR,BINOMIAL) ARE {0.01, 0.02, 0.01}, RESPECTIVELY

	In-Shop				
Method	R@1	R@10	R@20	R@30	R@40
Joints [17]	41	64	68	71	73
Poselets [17]	42	65	70	72	74
FashionNet [17]	53	73	76	77	79
HDC [8]	62.1	84.9	89.0	91.2	92.3
BIER [34]	76.9	92.8	95.2	96.2	96.7
HORDE* [70]	83.9	94.8	96.5	97.5	97.7
Triplet	64.6±0.17	87.3±0.10	90.9±0.05	92.7±0.04	93.9±0.01
CML-e (Tri)	*68.3±0.22*	*90.0±0.14*	*93.3±0.04*	*94.8±0.01*	*95.7±0.00*
CML-d (Tri)	*69.6±0.17*	*90.2±0.09*	*94.1±0.04*	*95.4±0.01*	*95.8±0.01*
CML (Tri)	***70.4±0.15***	***90.6±0.04***	***94.2±0.03***	***95.4±0.01***	***95.8±0.00***
N-Pair	77.9±0.18	94.0±0.09	96.1±0.06	96.9±0.02	97.4±0.00
CML-e (N-Pair)	*79.5±0.21*	*94.6±0.16*	*96.1±0.11*	*97.0±0.01*	*97.4±0.00*
CML-d (N-Pair)	*78.7±0.13*	*94.0±0.08*	*95.9±0.03*	*96.9±0.03*	*97.3±0.00*
CML (N-Pair)	***79.6±0.17***	***94.5±0.13***	***96.2±0.08***	***97.1±0.04***	***97.5±0.01***
Binomial	82.0±0.23	94.6±0.14	96.2±0.08	97.2±0.02	97.6±0.01
CML-e (Binomial)	*83.9±0.17*	*95.2±0.11*	*96.5±0.06*	*97.3±0.05*	*97.7±0.02*
CML-d (Binomial)	*83.2±0.22*	*95.0±0.15*	*96.4±0.11*	*97.3±0.06*	*97.7±0.02*
CML (Binomial)	***84.4±0.18***	***95.3±0.14***	***96.7±0.06***	***97.5±0.04***	***97.7±0.00***

the performance of the conventional deep metric learning methods.

One might observe that the effect of the proposed method varies between different datasets. It is because the effect differences over different datasets are influenced by the dataset itself, e.g., dataset scale, data structure, object pose. For example, CARS dataset is a relatively easier and small dataset, there are 98 classes with 80 instances per class for training, it is easy to perform local and global confusion. However, In-Shop Clothes dataset is a large scale and harder dataset, where there are more classes (3997) but with few instances (6) per class for training, which will introduce difficulties in either local or global confusion learning. Therefore, it is normal that the effect of the proposed method varies between different datasets.

Moreover, both EC and DC terms seem to both make the features less discriminative. Thus, the partial/selective learning behavior of deep model can be mitigated, i.e., suppressing the over-fitting on the training data. In summary, both of these two terms intend to achieve the similar goal, i.e., to make the feature less discriminative. While, how to make the feature less discriminative is what we need to consider. In our previous publication [18], we propose EC term, which just focuses the local feature confusion, and now in this extension article, we further propose DC which intends to further confuse the feature distribution from a global perspective. Since these two terms have the same target, i.e., making the feature less discriminative, when performing DC over the EC, the performance improvements are only a few if compared with the improvements over baseline methods. This phenomenon is reasonable because their targets are similar. But comparing the results of CML with that of CML-e (our previous work), the performances are indeed improved and significant, e.g., on CARS dataset, we have CML(tri) [R@1 = 83.0%] vs CML-e(tri)[R@1 = 81.2%], on CUB dataset, we have CML(tri) [R@1 = 55.3%] versus CML-e(tri)[R@1 = 53.8%], on In-Shop dataset, we have CML(tri) [R@1 = 70.4%] versus

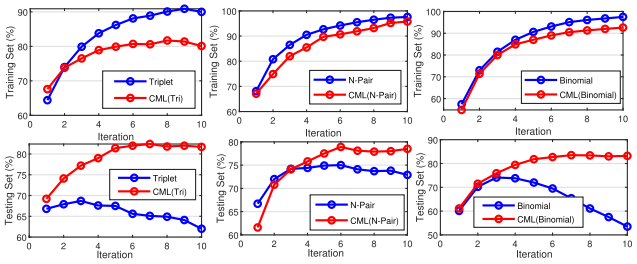

Fig. 3. Recall@1 curves on training (for the seen classes, top fig) and testing (for the unseen classes, bottom fig) sets over CARS dataset.

CML-e(tri)[R@1 = 68.3%] and so on. These demonstrate the effectiveness of our global confusion term, i.e., DC.

DAML [33] also employs the adversarial training scheme, but our CML can be trained in a direct manner unlike DAML which first needs to train a good sample-generator and then optimizes the deep embedding with the produced samples by this generator (this two-stage training procedure is more performance-sensitive and training-complex, thus performs worse than our CML as shown in Tables II–IV). DAML starts from the idea of improving the discrimination of the deep metric, not trying to prevent the "partial/biased learning behavior of deep model" and to improve the generalization ability of the deep metric.

D. Discussion

1) Regularization Ability: To demonstrate the regularization ability of our CML, we plot the R@1 retrieval result curves on training (seen) and testing (unseen) sets, respectively, as shown in Fig. 3. Specifically, for example, from the two figures in the left column, one can observe that the training curve (blue) of the conventional triplet method rises quickly to a relatively high level, but its testing curve (blue) only rises a little at first and then starts dropping to quite a low level, showing that in zero-shot settings, the deep metric learned by the conventional triplet loss is more likely to over-fit the seen classes and generalizes worse to the unseen classes. Conversely, after employing our local/global confusion terms, the training curve (red) of CML (tri) rises much slower than the original triplet and stops rising at a relatively lower level (80% versus 90%); however, the testing curve of our CML (tri) rises fast to quite a high level, more than 80% which is much better the original triplet loss (only nearly 70%), implying that our CML (tri) indeed serves as a regularization technique and improves the generalization ability of the learned deep metric by suppressing the learning of partial/biased metric over the seen classes. Moreover, a similar phenomenon can be observed from the comparisons between CML (N-pair)/CML (binomial) and N-pair/binomial.

In summary, for the conventional deep metric objective, the large gaps between training and testing show the deficiency of discriminative training in ZSRC tasks. As the model is more likely to focus on the surface statistical regularities rather than the more general abstract concepts during the training, extra regularization constraints should be imposed explicitly for the learning of robust metric. Therefore, we propose to

TABLE VI

COMPARISON WITH DROPOUT [38] ON CARS DATASETS. WE HAVE EXPERIMENTED DROPOUT WITH $\{0.1, 0.25, 0.4\}$ RATIO. THE UNDERLINED NUMBERS REPRESENT THE IMPROVEMENTS BY DROPOUT, THE SLANT NUMBERS INDICATE THE IMPROVEMENTS BY OUR LOCAL/GLOBAL CONFUSION TERMS AND THE BOLD NUMBERS INDICATE THE FINAL IMPROVEMENTS BY CML

	CARS			
	R@1	R@2	R@4	R@8
Binomial	74.2±0.12	83.1±0.04	86.7±0.00	92.9±0.00
Dropout(Binomial, 0.1)	73.3±0.26	82.2±0.18	88.7±0.07	92.5±0.04
Dropout(Binomial, 0.25)	74.8±0.31	83.6±0.24	85.9±0.15	92.6±0.10
Dropout(Binomial, 0.4)	72.3±0.44	81.0±0.27	87.3±0.22	92.4±0.16
CML-e (Binomial)	*84.4±0.12*	*90.4±0.05*	*93.8±0.00*	*96.6±0.00*
CML-d (Binomial)	*83.4±0.14*	*89.7±0.02*	*93.6±0.00*	*96.2±0.00*
CML (Binomial)	**85.1±0.14**	**90.8±0.04**	**94.0±0.00**	**96.7±0.00**

TABLE VII

ABLATION STUDY ON PARAMETER λ_1 WHEN SETTING $\lambda_2 = 0$

	CARS R@1					
λ_1	0 (Triplet)	0.001	0.01	0.02	0.1	1
CML-e (Tri)	68.3	74.6	80.1	**81.2**	72.3	59.3
λ_1	0 (N-Pair)	0.1	0.2	0.3	0.4	0.5
CML-e (N-Pair)	74.3	77.4	79.6	**80.3**	78.6	73.7
λ_1	0 (Binomial)	0.01	0.1	0.13	0.15	0.5
CML-e (Binomial)	74.2	76.3	83.1	**84.4**	84.3	69.7

TABLE VIII

ABLATION STUDY ON PARAMETER λ_2 WHEN SETTING $\lambda_1 = 0$

	CARS R@1					
λ_2	0 (Triplet)	0.01	0.02	0.03	0.04	0.05
CML-d (Tri)	68.3	80.5	81.3	82.0	**82.4**	81.4
λ_2	0 (N-Pair)	0.01	0.02	0.03	0.04	0.05
CML-d (N-Pair)	74.3	77.0	**78.9**	78.1	76.7	73.4
λ_2	0 (Binomial)	0.01	0.02	0.03	0.04	0.05
CML-d (Binomial)	74.2	78.4	82.2	**83.4**	80.1	79.5

use local/global feature distribution confusion terms to prevent the partial/biased learning problem and enhance the richness of the learned knowledge in an adversary manner. And to some extent, the comparison results in Fig. 3 validate the effect and importance of our confusion method.

2) Ablation Study on Regularization Method: As aforementioned, our work is to introduce regularization constraint for the learning of robust deep metric which could be benefit to the ZSRC tasks. In fact, there are some other works aiming at imposing regularization constraints at the top layer of the whole network, such as label-smoothing [50], label-disturbing [51], Noisy-Softmax [52], and maximum entropy [54]. However, these methods are all designed for the Softmax classifier layer and cannot be applied in the deep metric learning approaches.

To this end, in order to show the effectiveness of our CML in the deep metric learning community, we compare CML with the commonly used "Dropout" method [38]. The dropout layer is placed before the feature layer, and we test it with three different dropout ratios, including {0.1, 0.25, 0.4}, on CARS benchmark. From Table VI, one can observe that although using dropout with ratio 0.25 improves the results of R@1 and R@2, and using dropout with ratio 0.1 improves the results of R@4, the relative improvements are limited and not worthy of attention, and other results might be decreased (e.g., R@8). However, in contrast to Dropout, our CML-e and CML-d can both significantly surpass the baseline model by a large margin, e.g., CML-e (R@1 = 84.4) versus binomial (R@1 = 74.2), and CML-d (R@1 = 83.4) versus binomial (R@1 = 74.2), validating the effectiveness of each of our confusion terms. And when jointly regularizing the deep metric learning by the local and global feature distribution confusion terms, the performances of ZSRC can be further improved, achieving the best results shown by numbers in red.

We conjecture that is because dropout actually is not specially designed for the metric learning and the tested datasets are all fine-grained datasets,[5] in which case simply depressing the neurons of the deep network to zero will heavily affect the output distributions of these fine-grained classes regardless of the used value of dropout ratio as a result of small inter-class variations (for example, training the deep metric with a relatively smaller ratio, i.e., 0.1, the performance still cannot be improved.). In other words, dropout approach does not take into account the "semantic" feature relations between samples in the feature space and just coarsely depresses the output of neurons so as to regularize the deep metric learning. This is independent with respect to the optimization goal of metric learning, where relations between samples are semantically focused on; thus, it is reasonable that dropout does not perform the best. However, our regularization techniques, i.e., EC and DC terms, derive from the local and global feature distribution confusions, respectively, which consider the semantic relations between samples and are directly against the optimization goal of metric learning; as a result, they can successfully regularize the metric learning.

In summary, our CML framework is specially designed for the deep metric learning and indeed performs well.

3) Ablation Study on λ_1 & λ_2: To show the effects of parameters λ_1 and λ_2, we provide the results of CML-e (tri, N-pair, binomial) and CML-d (tri, N-pair, binomial) when using different λ_1 and λ_2 on CARS benchmark. Results are in Tables VII and VIII. It can be observed that when $\lambda_1 = \lambda_2 = 0$ our CML degenerates into the corresponding conventional metric learning method and the performance is unsatisfactory, while as λ_1 or λ_2 increases, the performances of CML-e (tri, N-pair, binomial) or CML-d (tri, N-pair, binomial) will peak around $\{0.02, 0.3, 0.13\}$ or $\{0.04, 0.02, 0.03\}$ respectively, and outperform the baselines (triplet, N-pair, binomial) by a large margin, validating the effectiveness and importance of our CML-e/CML-d.

Moreover, the best value combination of λ_1 and λ_2 will be slightly different from the values learned separately, but the difference is not large because the EC and DC terms are complementary to each other. The final results of our CML reported later are achieved by simply using the separately learned λ_1 and λ_2, and the performance improvements can be observed, demonstrating the effectiveness of our idea.

4) Ablation Study on Embedding Size: We also conduct quantitative experiments on embedding size with CML (binomial). From Table IX, it can be observed that for the conventional binomial metric learning method, most of the

[5]In the used four datasets, different classes have subtle differences, thus it is more challenging to recognize them.

TABLE IX

ABLATION EXPERIMENTAL RESULTS ON EMBEDDING SIZE

CARS					
Methods	R@1	R@2	R@4	NMI	F1
Binomial-128	70.4±0.23	80.6±0.10	87.2±0.06	60.8±0.21	29.1±0.18
CML (Binomial)-128	**79.7±0.26**	**87.7±0.11**	**91.9±0.05**	**64.5±0.17**	**32.8±0.21**
Binomial-256	73.3±0.19	82.4±0.04	88.5±0.01	61.5±0.15	30.0±0.19
CML (Binomial)-256	**82.8±0.23**	**88.9±0.02**	**92.7±0.01**	**66.9±0.14**	**35.7±0.12**
Binomial-384	74.0±0.15	82.5±0.03	88.8±0.01	61.9±0.18	30.6±0.21
CML (Binomial)-384	**84.0±0.12**	**90.1±0.06**	**93.9±0.02**	**67.7±0.15**	**36.5±0.09**
Binomial-512	74.2±0.12	83.1±0.04	86.7±0.00	61.7±0.09	28.8±0.04
CML (Binomial)-512	**85.1±0.14**	**90.8±0.04**	**94.0±0.00**	**69.3±0.14**	**39.1±0.08**

evaluation indexes' results (e.g., R@4, R@8, NMI, and F_1) actually do not increase with the embedding size (from 128 to 512-dim) and even have a decreasing trend, showing that the risk of over-fitting increases with feature size and without robustness learning the performances of the learned metric cannot be guaranteed even if its theoretical representation ability increases with the feature size. By employing our CML, the performances can be consistently improved and indeed increase with embedding size, showing that the proposed CML indeed regularizes the deep metric learning.

5) Comparison Between DC and Weight Decay: From Definition 2, one can observe that our final DC term is formally similar to the commonly used parameter regularization method weight decay (i.e., $\|W\|_2^2$). However, for Weight Decay, it only regularizes the parameters in each layer, but the final output feature distribution has not been explicitly regularized. Due to the powerful nonlinear-mapping ability of the cascaded multineuron-layers, the output feature distribution can be arbitrarily changed even if the parameter has been regularized. Therefore, when training with the conventional deep metric objectives, the output feature still has a high probability of being over-discriminative (with large diversity) for the seen classes and the biased learning behavior of deep model exists as well. This can be validated by our baseline experiments, for example, all of our baseline methods (triplet, N-pair, binomial) are actually trained with weight decay ($2e-4$), but from Fig. 3, it can be observed that these baseline methods are more likely to be over-fitting on the seen classes (with a large gap between training and testing results) and the performances on unseen classes are disappointing as well, implying that the weight decay is not well designed for feature distribution regularization and has little effect on improving the generalization of the learned metric. In contrast, when explicitly regularizing the feature distribution with the proposed DC term, CML-d (tri, N-pair, binomial)[6] can learn more robust feature descriptors and surpass the baseline methods (regularized by weight decay) by a large margin.

Moreover, it seems that [48], [49] also use L-2 regularization. However, different from [48] which simply tries L-2 regularization and [49] which derives L-2 regularization from a simplified analytical form of classification loss, our DC term comes from the rigorous definition of data diversity under the Gaussian mixture distribution. Our DC is proposed independently since we originally intend to regularize the diversity of data distribution, but just obtain the similar target.

[6]For fair comparison with baseline methods, we also use weight decay in CML-d and CML.

TABLE X

DIVERSITY COMPARISON OF FEATURE DISTRIBUTION

Diversity on CARS			
Method	Seen Classes	Unseen Classes	R@1(%)
Binomial	1.67	1.24	74.2
CML-d (Binomial)	1.54 ↓	1.49 ↑	83.5 ↑

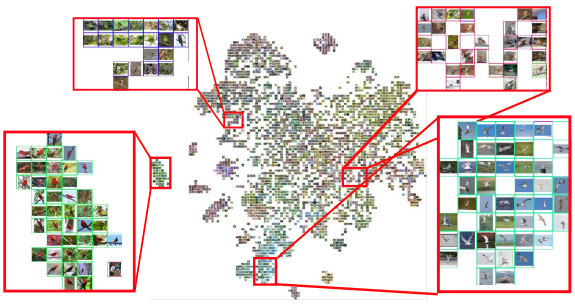

Fig. 4. Visualization of the proposed CML (Binomial) with t-SNE [71] on the testing set of CUB dataset, where the color of the border for each image indicates the image label, best viewed by zooming in.

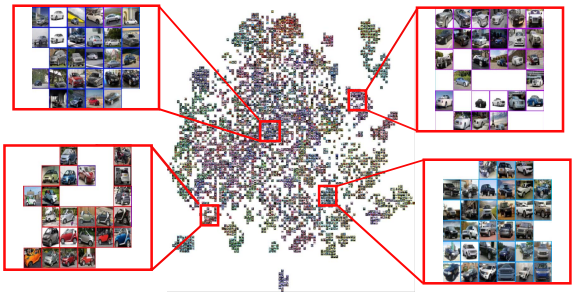

Fig. 5. Visualization of the proposed CML (binomial) with t-SNE [71] on the testing set of CARS dataset, where the color of the border for each image indicates the image label, best viewed by zooming in.

And from this perspective, our work can be regarded as another explanation for the success of [48] and [49], which is also meaningful.

6) Diversity Comparison: Different from the EC term, which is easier and more intuitive to understand, the DC term derives from a global and abstract perspective. To this end, we conduct quantitative experiments to provide intuitive understanding. From Table X, one can observe that the diversity of feature distribution is large on the seen classes but small on the unseen classes when using the conventional metric approaches (i.e., binomial), however, by depressing the diversity of feature distribution over the seen classes via our DC term, the diversity of feature distribution over the seen classes is reduced while the diversity over the unseen classes is increased and the corresponding R@1 retrieval result is improved as well, demonstrating that our DC term indeed works and regularizes the deep metric learning.

7) Embedding Visualizations: Moreover, we also provide visualizations of the learned embedding in Figs. 4 and 5, which show the visualization results of our CML (binomial) using t-SNE [71] on the testing sets of CUB and CARS

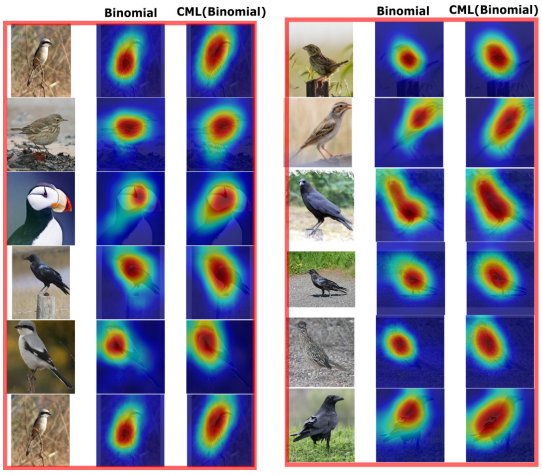

Fig. 6. Informative region attention comparisons between the conventional Binomial loss and our CML (Binomial). It can be observed that our CML (Binomial) all can focus on more informative object-parts.

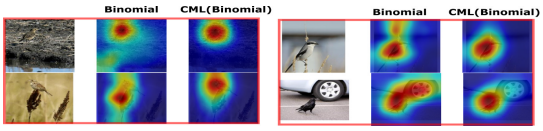

Fig. 7. Noisy background suppressing comparisons between the conventional binomial loss and our CML (binomial). It can be observed that our CML (binomial) can also suppress the noisy background learning.

datasets, respectively. We zoomed-in view the specific regions to highlight the representative categories in each figure. From these figures, it can be observed that although these classes are unseen in training phase, the proposed CML (binomial) is still able to group the similar instances. These visualization results also intuitively demonstrate the effectiveness of our CML.

E. Case Study

To give an intuitive understanding of the proposed overall CML framework, we provide some case study (attention visualization) on the testing images as in Figs. 6 and 7. The attention map is produced by averaging the convolutional output maps at the last convolutional layer. From Fig. 6, one can observe that the attention region induced by our CML (binomial) is larger than that induced by binomial, implying that only using the conventional binomial loss the attention region will focus most on the partial/biased object parts while by suppressing this partial/biased learning behavior with our confusion terms, more informative parts can be focused on by CML (binomial).

Moreover, from Fig. 7, one can observe that by using our confusion terms, the noisy background information can also be suppressed, showing the good generalization and discrimination of our CML framework.

F. Compatibility With Random-Erasing (RE)

As our work mainly focuses on the level of feature representation, we further test the compatibility of our method with the image-level regularization method random-erasing (RE) [72]. The results of CML(binomial) on CARS dataset are in Table XI, one can observe that using both CML and RE can further improve the performances over CML, demonstrating the compatibility of our method with RE.

TABLE XI
COMPATIBILITY WITH RE

	CARS					
Methods	R@1	R@2	R@4	R@8	NMI	F1
RE [72]	82.1±0.29	89.3±0.15	93.5±0.02	96.0±0.01	68.4±0.23	37.9±0.17
CML	85.1±0.14	90.8±0.04	94.0±0.00	96.7±0.00	69.3±0.14	39.1±0.08
CML+RE	**86.8±0.20**	**91.0±0.13**	**94.0±0.01**	**96.7±0.00**	**69.7±0.21**	**39.5±0.17**

V. CONCLUSION

In this article, we propose the CML framework, a generally applicable method to various conventional deep metric learning approaches, for ZSRC tasks by explicitly intensifying the generalization ability within the learned embedding with the help of our EC and DC terms. The CML breaks away from the traditional metric learning idea of only devising discriminative objective functions since this convention idea is not capable of addressing the biased learning behavior of deep model which will produce the prejudiced (biased) metric on seen classes and is the key obstacle of performance improvement in zero-shot settings. Extensive experiments on the popular ZSRC benchmarks (CARS, CUB, Stanford Online Products, and In-Shop) demonstrate the significance and necessity of our idea of learning metric with good generalization by confusion.

CML has the ability to improve the model generalization in zero-shot settings. It does not need the extra annotations during training or testing (many existing zero-shot works need these information to train their models). CML contains two complementary terms, i.e., the local term EC and the global term DC, which can consistently improve the performances. CML also can be directly applied to many existing metric losses. However, there is also one weakness: the value of hyper-parameter λ_1 and λ_2 seems to be a little performance-sensitive and a little hard to adjust. Therefore, we would like to use AutoML method to automatically learn or adjust these hyper-parameters, or explore other more powerful confusion terms that can simultaneously handle both local and global structures in future work. Moreover, forcing model to gain more rich and robust knowledge is important to improve generalization. Thus, devising methods that can intrinsically perceive images and patterns instead of biased remembering surface statistic information is another possible and valuable trend in future works.

ACKNOWLEDGMENT

The authors hereby give specific thanks to Alibaba Group for their contribution to this article.

REFERENCES

[1] Y. Fu, T. M. Hospedales, T, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2332–2345, Nov. 2015.

[2] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4166–4174.

[3] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2021–2030.

[4] J. Dalton, J. Allan, and P. Mirajkar, "Zero-shot video retrieval using content and concepts," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 1857–1860.

[5] Y. Shen, L. Liu, F. Shen, and L. Shao, "Zero-shot sketch-image hashing," 2018, *arXiv:1803.02284*.

[6] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4004–4012.

[7] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 1988–1996.

[8] Y. Yuan, K. Yang, and C. Zhang, "Hard-aware deeply cascaded embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 814–823.

[9] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2840–2848.

[10] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[11] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," 2017, *arXiv:1708.01682*.

[12] C. Huang, C. C. Loy, and X. Tang, "Local similarity-aware deep feature embedding," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 1262–1270.

[13] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 1857–1865.

[14] V. B. Kumar, B. Harwood, G. Carneiro, I. Reid, and T. Drummond, "Smart mining for deep metric learning," 2017, *arXiv:1704.01285*.

[15] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Dec. 2013, pp. 554–561.

[16] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds200–2011 dataset," California Inst. Technol. Pasadena, CA, USA, Tech. Rep. 2010-001, 2011.

[17] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.

[18] B. Chen and W. Deng, "Energy confused adversarial metric learning for zero-shot image retrieval and clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8134–8141.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[20] Z. Ji, J. Yan, Q. Wang, Y. Pang, and X. Li, "Triple discriminator generative adversarial network for zero-shot image classification," *Sci. China Inf. Sci.*, vol. 64, no. 2, pp. 1–14, Feb. 2021.

[21] Z. Ji *et al.*, "Deep ranking for image zero-shot multi-label classification," *IEEE Trans. Image Process.*, vol. 29, pp. 6549–6560, 2020.

[22] Z. Ji, Y. Sun, Y. Yu, Y. Pang, and J. Han, "Attribute-guided network for cross-modal zero-shot hashing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 321–330, Jan. 2020.

[23] Z. Han, Z. Fu, S. Chen, and J. Yang, "Contrastive embedding for generalized zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 2371–2381.

[24] M. F. Naeem, Y. Xian, F. Tombari, and Z. Akata, "Learning graph embeddings for compositional zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 953–962.

[25] L. Bo, Q. Dong, and Z. Hu, "Hardness sampling for self-training based transductive zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16499–16508.

[26] M. Ye and Y. Guo, "Progressive ensemble networks for zero-shot recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11728–11736.

[27] K. Li, M. R. Min, and Y. Fu, "Rethinking zero-shot learning: A conditional visual classification perspective," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3583–3592.

[28] S. Liu, J. Chen, L. Pan, C.-W. Ngo, T.-S. Chua, and Y.-G. Jiang, "Hyperbolic visual embedding learning for zero-shot recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9273–9281.

[29] D. Huynh and E. Elhamifar, "Fine-grained generalized zero-shot learning via dense attribute-based attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4483–4493.

[30] Z. Han, Z. Fu, and J. Yang, "Learning the redundancy-free features for generalized zero-shot object recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12865–12874.

[31] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 360–368.

[32] B. Chen and W. Deng, "ALMN: Deep embedding learning with geometrical virtual point generating," 2018, *arXiv:1806.00974*.

[33] Y. Duan, W. Zheng, X. Lin, J. Lu, and J. Zhou, "Deep adversarial metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2780–2789.

[34] M. Opitz, G. Waltner, H. Possegger, and H. Bischof, "BIER—Boosting independent embeddings robustly," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5189–5198.

[35] M. Opitz, G. Waltner, H. Possegger, and H. Bischof, "Deep metric learning with bier: Boosting independent embeddings robustly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 276–290, Feb. 2018.

[36] W. Kim, B. Goyal, K. Chawla, J. Lee, and K. Kwon, "Attention-based ensemble for deep metric learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 736–751.

[37] B. Chen, W. Deng, J. Hu, and H. Shen, "Hybrid-attention based decoupled metric learning for zero-shot image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2750–2759.

[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[39] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," 2013, *arXiv:1308.3432*.

[40] C. Gulcehre, M. Moczulski, M. Denil, and Y. Bengio, "Noisy activation functions," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 3059–3068.

[41] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," 2015, *arXiv:1505.05424*.

[42] A. Graves, "Practical variational inference for neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2348–2356.

[43] A. Neelakantan *et al.*, "Adding gradient noise improves learning for very deep networks," 2015, *arXiv:1511.06807*.

[44] T. Chen *et al.*, "ABD-Net: Attentive but diverse person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8351–8361.

[45] N. Bansal, X. Chen, and Z. Wang, "Can we gain more from orthogonality regularizations in training deep networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4261–4271.

[46] M. Lezcano-Casado and D. Martínez-Rubio, "Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group," 2019, *arXiv:1901.08428*.

[47] W. Liu *et al.*, "Learning towards minimum hyperspherical energy," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6222–6233.

[48] H. Peng, L. Mou, G. Li, Y. Chen, Y. Lu, and Z. Jin, "A comparative study on regularization strategies for embedding-based neural networks," 2015, *arXiv:1508.03721*.

[49] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3018–3027.

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[51] L. Xie, J. Wang, Z. Wei, M. Wang, and Q. Tian, "DisturbLabel: Regularizing CNN on the loss layer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4753–4762.

[52] B. Chen, W. Deng, and J. Du, "Noisy Softmax: Improving the generalization ability of DCNN via postponing the early softmax saturation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5372–5381.

[53] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," 2017, *arXiv:1701.06548*.

[54] A. Dubey, O. Gupta, R. Raskar, and N. Naik, "Maximum entropy fine-grained classification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2018, pp. 635–645.
[55] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
[56] E. Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen Veränderlichen," *J. für die reine und Angew. Math.*, vol. 1909, no. 136, pp. 210–271, Jul. 1909.
[57] G. J. Székely and M. L. Rizzo, "Testing for equal distributions in high dimension," *InterStat*, vol. 5, no. 16, pp. 1249–1272, 2004.
[58] G. J. Székely and M. L. Rizzo, "A new test for multivariate normality," *J. Multivariate Anal.*, vol. 93, no. 1, pp. 58–80, 2005.
[59] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," 2015, *arXiv:1502.02791*.
[60] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," 2016, *arXiv:1605.06636*.
[61] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.
[62] G. A. F. Seber, *Multivariate Observations* (Wiley Series in Probability and Statistics), Apr. 1984.
[63] D. Jonsson, "Some limit theorems for the eigenvalues of a sample covariance matrix," *J. Multivariate Anal.*, vol. 12, no. 1, pp. 1–38, Mar. 1982.
[64] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2794–2802.
[65] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 34–39.
[66] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[68] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[69] H. O. Song, S. Jegelka, V. Rathod, and K. Murphy, "Deep metric learning via facility location," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5382–5390.
[70] P. Jacob, D. Picard, A. Histace, and E. Klein, "Metric learning with HORDE: High-order regularizer for deep embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6539–6548.
[71] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[72] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13001–13008.

Binghui Chen received the B.E. and Ph.D. degrees in telecommunication engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2015 and 2020, respectively.

His research interests include computer vision, deep learning, face recognition, deep embedding learning, and machine learning.

Weihong Deng (Member, IEEE) received the B.E. degree in information engineering and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2004 and 2009, respectively.

From October 2007 to December 2008, he was a Postgraduate Exchange Student with the School of Information Technologies, The University of Sydney, Sydney, NSW, Australia, under the support of the China Scholarship Council. He is currently a Professor with the School of Information and Telecommunications Engineering, BUPT. His research interests include statistical pattern recognition and computer vision, with a particular emphasis in face recognition.

Biao Wang received the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2013.

He currently works with the DAMO Academy, Alibaba Group, Beijing, as a Senior Researcher, leading the Industry Visual Intelligence team. He has published over 20 papers in top tier academic conferences and journals. His current research interest include image classification, object detection, action recognition, and unsupervised learning.

Lei Zhang (Fellow, IEEE) received the B.Sc. degree from the Shenyang Institute of Aeronautical Engineering, Shenyang, China, in 1995, and the M.Sc. and Ph.D. degrees in control theory and engineering from Northwestern Polytechnical University, Xi'an, China, in 1998 and 2001, respectively.

From 2001 to 2002, he was a Research Associate with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. From January 2003 to January 2006, he worked as a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada. In 2006, he joined the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, as an Assistant Professor. Since July 2017, he has been a Chair Professor with the Department of Computing. His research interests include computer vision, image and video analysis, pattern recognition, and biometrics.

Prof. Zhang is also a Senior Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING. He is/was an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *SIAM Journal of Imaging Sciences*, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (CSVT), and *Image and Vision Computing*. He is a "Clarivate Analytics Highly Cited Researcher" from 2015 to 2021.