Energy Confused Adversarial Metric Learning for Zero-Shot Image Retrieval and Clustering

Binghui Chen, Weihong Deng

Beijing University of Posts and Telecommunications

chenbinghui@bupt.edu.cn, whdeng@bupt.edu.cn

Abstract

Deep metric learning has been widely applied in many computer vision tasks, and recently, it is more attractive in *zero-shot image retrieval and clustering* (ZSRC) where a good embedding 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 thus race to devise powerful metric objectives or hard-sample mining strategies for leaning discriminative embedding. However, in this paper, we first emphasize that the generalization ability is a core ingredient of this 'good' embedding as well and largely affects the metric performance in zero-shot settings as a matter of fact. Then, we propose the Energy Confused Adversarial Metric Learning (ECAML) framework to explicitly optimize a robust metric. It is mainly achieved by introducing an interesting Energy Confusion regularization term, which daringly breaks away from the traditional metric learning idea of discriminative objective devising, and seeks to 'confuse' the learned model so as to encourage its generalization ability by reducing overfitting on the seen classes. We train this confusion term together with the conventional metric objective in an adversarial manner. Although it seems weird to 'confuse' the network, we show that our ECAML indeed serves as an efficient regularization technique for metric learning and is applicable to various conventional metric methods. This paper empirically and experimentally demonstrates the importance of learning embedding with good generalization, achieving state-of-the-art performances on the popular CUB, CARS, Stanford Online Products and In-Shop datasets for ZSRC tasks. Code available at http://www.bhchen.cn/.

1. Introduction

Since *zero-shot learning* (ZSL) removes the limitation of category-consistency between training and testing sets, it turns to be more attractive where the model is required to learn concepts from *seen* classes and then enables to distinguish the *unseen* classes. ZSL has been widely explored in image classification (Changpinyo et al. 2016; Fu et al. 2015) and retrieval tasks (Dalton, Allan, and Mirajkar 2013; Shen et al. 2018; Oh Song et al. 2016), *etc*. In this paper, we focus on *zero-shot image retrieval and clustering* tasks (ZSRC). In order to accurately retrieve
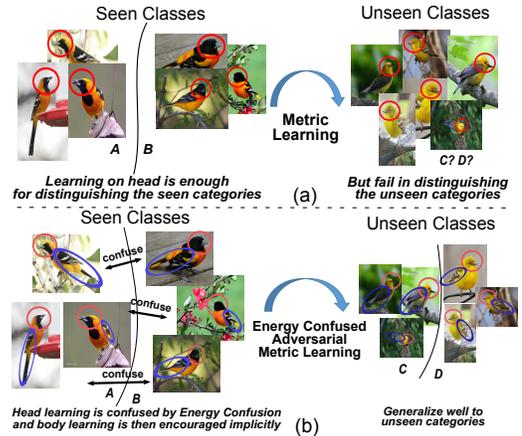
Figure 1: Comparisons of conventional metric learning methods and our Energy Confused Adversarial Metric Learning (ECAML). In (a), the deep model optimized by conventional metric learning will selectively learn head knowledge which is the easiest one to reduce the current training error and omit other helpful concepts, but the testing instances cannot be distinguished well by the head. In (b), the Energy Confusion(EC) term among different classes is introduced so as to make the biased head-based metric confused about itself, then as the training going, EC 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.

and cluster the *unseen* classes, most existing works employ *Deep Metric Learning* to optimize a good embedding, such as exploring tuple-based loss functions (Sun et al. 2014; Yuan, Yang, and Zhang 2017; Wu et al. 2017; Schroff 2015; Oh Song et al. 2016; Wang et al. 2017; Huang, Loy, and Tang 2016; Sohn 2016) and proposing efficient hard-sample mining strategies (Kumar et al. 2017; Wu et al. 2017; Schroff 2015), *etc*. However, the above methods deem this 'good' embedding just to be the discriminative one and then concentrate on the discriminative learning over the *seen* classes, but neglect the importance of the generalization ability of the learned metric which is significant in ZSRC as well, as a result, without robustness constraining they are

easily subject to concepts overfitting problem on the *seen* classes and some helpful or general knowledge for *unseen* classes may have been left out with a high probability.

To be specific, in ZSRC, we emphasize that the generalization ability of the learned embedding is seriously affected by the following problem: "the biased learning behavior of deep models", concretely, 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 attributes concepts that are the easiest ones to reduce the current training loss over the *seen* classes (here head knowledge is enough to sperate class A from B and thus is learned), instead of learning all-sided details and concepts, thus yielding over-fitting on *seen* classes and generalizing worse to *unseen* ones (classes C and D). In another word, in order to correctly recognize classes, deep networks easily learn to focus on surface statistical regularities rather than more general abstract concepts.

Therefore, when learning the embedding as in the aforementioned conventional metric learning methods, this issue objectively exists and impedes the learning of the desired good embedding. And without explicit and benign robustness constraint, the learned embedding is unable to generalize well to the *unseen* classes. Most ZSRC works ignore the importance to learn robust descriptors. To this end, proposing efficient regularization method for conventional metric learning to learn metrics with good generalization is important, especially in ZSRC tasks.

In this paper, we propose the ***Energy Confused Adversarial Metric Learning*** (ECAML) framework, an elegant regularization strategy, to alleviate the problem of generalization in ZSRC tasks by randomly confusing the learned metric during each iteration. It is mainly achieved by a novel and simple *Energy Confusion* (EC) term which is 'plug and play' and can be generally applied to many existing deep metric learning approaches. Concretely, this confusion term plays an adversary role against the conventional metric learning objective, which intends to minimize the expected value of the Euclidean distances between the paired images from two different categories. 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, preventing overfitting on the *seen* classes and improving the generalization ability of the embedding in an adversarial manner. In another word, the EC term allows the SGD solver to escape from the 'bad' local-minima region induced by the *seen* classes and to explore more for the robust one. The main contributions of this work can be summarized as follows:

- We emphasize that the crucial issue to ZSRC, i.e. *the biased learning behavior of deep model*, is the key stumbling block of improving the generalization ability of the learned embedding.

- We propose ***Energy Confused Adversarial Metric Learning***(ECAML) framework to reinforce the robustness of embedding in an adversarial manner. The Energy Confusion(EC) term

---

[1]In fact, the learned partial biased knowledge is more complicated and cannot be easily illustrated in figure, here for intuitive understanding, we translate it into some single body-part knowledge.

is 'plug and play' and can work in conjunction with many existing metric methods. To our knowledge, it is the first work to introduce confusion for deep metric learning.

- Extensive experiments have been performed on several popular datasets for ZSRC, including CARS (Krause et al. 2013), CUB (Wah et al. 2011), Stanford Online Products (Oh Song et al. 2016) and In-shop(Liu et al. 2016), achieving state-of-the-art performances.

2. Related Work

Zero-shot: ZSL has been widely explored in many computer vision tasks, such as image classification (Changpinyo et al. 2016; Fu et al. 2015) and image retrieval (Dalton, Allan, and Mirajkar 2013; Shen et al. 2018). Most of these ZSL methods are capable of exploiting the extra auxiliary supervision information of the *unseen* classes (e.g. word representations of semantic name), thus aligning the learned features in an explicit manner. However in real applications, collecting and labelling these auxiliary information is time-consuming and impractical. Our ECAML concentrates on a more actual scene where there are only *seen* class labels available.

Deep metric learning for ZSRC: The commonly used contrastive (Sun et al. 2014) and triplet loss (Schroff 2015) have been broadly studied. Additionally, there are some other deep metric learning works: Smart-mining (Kumar et al. 2017) combines local triplet loss and global loss to optimize the deep metric with hard-samples mining. Sampling-Matters (Wu et al. 2017) proposes distance weighted sampling strategy. Angular loss (Wang et al. 2017) optimizes a triangle-based angular function. Proxy-NCA (Movshovitz-Attias et al. 2017) explains why popular classification loss works from a proxy-agent view, and its implementation is very similar to Softmax. ALMN (Chen and Deng 2018) proposes to generate geometrical virtual negative point instead of employing hard-sample mining for learning discriminative embedding. 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 issue easily. Additionally, HDC (Yuan, Yang, and Zhang 2017) employs the cascaded models and selects hard-samples from different levels and models. BIER loss (Opitz et al. 2017; 2018) adopts the online gradients boosting methods. These methods try to improve the performances by resorting to the ensemble idea. Different from all these methods, our ECAML has a clear object of improving the generalization ability of the learned metric by introducing the Energy Confusion regularization term.

Regularization technique: Regularization methods sometimes are important for deep models as the deep models are more likely to be data-driven. There are some works injecting random noise into deep nets so as to ensure the robust training, such as Bengio, Léonard, and Courville (2013) and Gulcehre et al. (2016) add noise in the ReLU and Sigmoid activation functions respectively, Blundell et al. (2015), Graves (2011) and Neelakantan et al. (2015) add noise in weights and gradients respectively. Moreover, some research works intend to regularize the deep models at the top layer, i.e. Softmax classifier layer, for example, Szegedy et al. (2016) propose

label-smoothing regularization technique for training deep models, Xie et al. (2016) propose label-disturbing technique for improving the generalization ability of the deep models and, Chen, Deng, and Du (2017) inject annealed noise into the softmax activations so as to boost the generalization ability by postponing the early Softmax saturation behavior. However, different from these above methods which are mainly devised for classification tasks and applicable to the Softmax classifier layer, our ECAML aims to promote the generalization ability of the metric learning in ZSRC tasks, and it is achieved by training the EC term in an adversarial manner.

3 Notations and Preliminaries

In this section, we review some notations and the necessary preliminaries on the relation between semimetric and RKHS kernels for later convenience, which will be used to interpret the differences between our EC (Sec.4.1) and some other existing methods, i.e. *general energy distance* and *maximum mean discrepancy*.

If not specified, we will assume that $\mathcal{Z}$ is any topological space where the Borel measures can be defined. Denote by $\mathcal{M}(\mathcal{Z})$ the set of all finite signed Borel measures on $\mathcal{Z}$, and by $\mathcal{M}^1_+(\mathcal{Z})$ the set of all Borel probability measures on $\mathcal{Z}$.

Definition 1. *(RKHS) Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{Z}$. A function $k : \mathcal{Z} \times \mathcal{Z} \to R$ is called a reproducing kernel of $\mathcal{H}$, if (i) $\forall z \in \mathcal{Z}$, $k(\cdot, z) \in \mathcal{H}$, and (ii) $\forall z \in \mathcal{Z}$, $\forall f \in \mathcal{H}$, $< f, k(\cdot, z) >_{\mathcal{H}} = f(z)$. If $\mathcal{H}$ has a reproducing kernel, it is called a* reproducing kernel Hilbert space *(RKHS).*

Definition 2. *(Semimetric) Let $\mathcal{Z}$ be a nonempty set and let $\rho : \mathcal{Z} \times \mathcal{Z} \to [0, +\infty)$ be a function such that $\forall z, z^{'} \in \mathcal{Z}$, (i) $\rho(z, z^{'}) = 0$ iff $z = z^{'}$ and (ii) $\rho(z, z^{'}) = \rho(z^{'}, z)$. $(\mathcal{Z}, \rho)$ is called a semimetric space and ρ is a semimetric.*

Definition 3. *(Negative type) Semimetric space $(\mathcal{Z}, \rho)$ is said to have negative type if $\forall n \geq 2, z_1, \ldots, z_n \in \mathcal{Z}$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^{n} \alpha_i = 0$, $\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho(z_i, z_j) \leq 0$.*

Then we have the following propositions, which are derived from (Van Den Berg, Christensen, and Ressel 2012).

Proposition 1. *If ρ satisfies Def.3, then so does ρ^q, where $0 < q < 1$.*

Proposition 2. *ρ is a semimetric of negative type iff there exists a $\mathcal{H}$ and an injective map $\varphi : \mathcal{Z} \to \mathcal{H}$, such that*

$$\rho(z, z^{'}) = \|\varphi(z) - \varphi(z^{'})\|^2_{\mathcal{H}} \tag{1}$$

This shows that $(\mathbb{R}^d, \| \cdot - \cdot \|^2)$ is of negative type, and by taking $q = 1/2$, we conclude that all Euclidean spaces are of negative type (Sejdinovic et al. 2012; 2013) , which will be used to reason our Energy Confusion term. Then we also show that the semimetrics of negative type and symmetric positive definite kernels are in fact closely related by the following Lemma (for more details please refer to Van Den Berg, Christensen, and Ressel 2012).

Lemma 1. *For a nonempty $\mathcal{Z}$, let ρ be a semimetric on $\mathcal{Z}$. Let $z_0 \in \mathcal{Z}$, and denote $k(z, z^{'}) = \frac{1}{2}(\rho(z, z_0) + \rho(z^{'}, z_0) - \rho(z, z^{'}))$. Then k is positive definite iff ρ is of negative type.*

We call the kernel defined above *distance-induced kernel* and, it is induced by the semimetric ρ and centered at z_0. By varying the point at the center z_0, we obtain a kernel family $\mathcal{K}_\rho = \frac{1}{2}[\rho(z, z_0) + \rho(z^{'}, z_0) - \rho(z, z^{'})]_{z_0 \in \mathcal{Z}}$, induced by ρ. Then we can always express Eq.1 in terms of the canonical feature map for RKHS $\mathcal{H}_k$ as the following proposition.

Proposition 3. *Let $(\mathcal{Z}, \rho)$ be a semimetric space of negative type, and $k \in \mathcal{K}_\rho$. Then:*

1. *k is nondegenerate, i.e. the Aronszajn map $z \to k(\cdot, z)$ is injective.*

2. *$\rho(z, z^{'}) = k(z, z) + k(z^{'}, z^{'}) - 2k(z, z^{'}) = \|k(\cdot, z) - k(\cdot, z^{'})\|^2_{\mathcal{H}_k}$*

For the above valid ρ, we say that k generates ρ. And the above proposition implies that the Aronszajn map $z \to k(\cdot, z)$ is an isometric embedding of a metric space $(\mathcal{Z}, \rho^{1/2})$ into $\mathcal{H}_k$, for each $k \in \mathcal{K}_\rho$. Lem.1 and Prop.3 reveal the general link between semimetrics of negative type and RKHS kernels in different views. By taking some special cases of ρ and k, we are able to elucidate our EC in the following sections.

4. Proposed Approach

4.1 Energy Confusion

As discussed in (Sec.1), without taking the generalization ability into consideration explicitly, simply optimizing a discriminative objective metric functions or applying hard-sample mining strategies like in most existing metric learning works wouldn't lead a robust metric for ZSRC tasks, since the '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 discriminative for the *seen* classes instead of keeping all-sided information, resulting in overfitting on the *seen* categories and limiting the generalization ability of the learned embedding.

Consider that the biased learning behavior is actually induced by the nature of model training since in order to correctly distinguish different *seen* classes, the deep metric has to be confident about the feature distribution prediction over the current *seen* classes as far as possible(e.g. features of different classes should be far away from each other) and as a result, only the partial biased knowledge that are discriminative to separate *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. a feature distribution confusion term, into the conventional metric learning phase so as to 'confuse' the network and reduce the over-confident predictions of distances between feature distributions on the *seen* classes**. Specifically, denote the input features by $\{x_i\}_{i=1}^N$, the corresponding label inputs by $\{y_i\}_{i=1}^N, y_i \in [1 \ldots C]$, where C is the number of *seen* classes. The conventional metric optimizing goal is to make

the distance measurement $D(x_i, x_j)$ as large as possible if $y_i \neq y_j$, otherwise as small as possible, and it can be formulated as:

$$\theta_f = \arg\min_{\theta_f} L_m(\theta_f; T, D) \qquad (2)$$

where L_m is some specific metric loss function, T indicates some instance-tuple, e.g. contrastive tuple $T(x_i, x_j)$ (Sun et al. 2014), triplet tuple $T(x_i, x_{i+}, x_{i-})$ (Schroff 2015) or N-Pair tuple $T(x_i, x_{i+}, x_{i_1^-}, \ldots, x_{i_{N-2}^-})$ (Sohn 2016), D is the distance distribution measurement, e.g. Euclidean measurement (Oh Song et al. 2016; Yuan, Yang, and Zhang 2017; Huang, Loy, and Tang 2016; Schroff 2015; Wu et al. 2017) or inner-product measurement (Opitz et al. 2017; Sohn 2016), and θ_f is the metric parameters to be learned.

Therefore, in order to prevent the biased learning behavior by confusing the feature distribution learning, ***we would like to learn θ_f that make the feature distributions from different classes closer*** when under some specific $\{L, T, D\}$. It seems that the commonly adopted family of $f\text{-}divergence$ for measuring the difference between two probability distributions might be a suitable choice, such as KL-divergence (Kullback and Leibler 1951), Hellinger-distance (Hellinger 1909) and Total-variation-distance, however, we emphasize that they cannot be directly applied here since they mostly work with the probability measure (where $\sum_k x_{i,k} = 1$) but our confusion goal is based on the statistical distance between two random vectors following some probability distributions. To this end, we propose the ***Energy Confusion*** term as follows:

$$L_{ec}(\theta_f; X_I, X_J) = \mathbb{E}_{\widetilde{X_I}\widetilde{X_J}}(\|\widetilde{X_I} - \widetilde{X_J}\|_2^2)$$
$$= \sum_{i,j} p_{i,j} \|x_i - x_j\|_2^2 \qquad (3)$$

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 which 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 $\widetilde{X_I} \sim Uniform(X_I), \widetilde{X_J} \sim Uniform(X_J)$ and $p_{i,j} = p_i p_j = \frac{1}{N_I}\frac{1}{N_J}$. In this case, $\{L, T, D\}$ are expected value function, contrastive tuple and Euclidean measurement respectively.

From Eq.3, one can observe that the EC term intends to minimize the distance expected value between different classes so as to confuse the metric. As discussed above, the learned embedding represents the learned concepts to some extend, and the more accurate the prediction of distance on the *seen* classes, the greater the risk of concepts overfitting. EC serves as a regularization term that would like to prevent the model being over-confident about the *seen* classes and mitigate the biased learning issue by avoiding the learner being stuck in the training-data-specific concepts. In another word, the 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. Moreover, 'confusing' also gives SGD solver chances of escaping from the 'partial' and 'bad' local-minima induced

by the *seen* instances, and then exploring other solution regions for the more 'general' ones.

Discussion: Inferred from the above analysis, it seems that the commonly used *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.

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 *general energy distance*(GED) between P and Q, w.r.t ρ is:

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

where $\widetilde{P}, \widetilde{P}' \overset{i.i.d}{\sim} P$ and $\widetilde{Q}, \widetilde{Q}' \overset{i.i.d}{\sim} Q$. $D_{E,\rho}$ is a general extension of *energy distance*(Székely and Rizzo 2004; 2005) on metric space. Then we have:

Lemma 2. *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:*

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

Proof. from Prop.2, if ρ is the squared Euclidean metric, we have $(\mathcal{Z}, \rho)$ is of negative type, thus from Eq.4

$$\frac{1}{2}D_{E,\rho}(X_I, X_J) = \mathbb{E}(\|\widetilde{X}_I - \widetilde{X}_J\|_2^2) - \frac{1}{2}\{\mathbb{E}(\|\widetilde{X}_I - \widetilde{X}_I'\|_2^2)$$
$$+ \mathbb{E}(\|\widetilde{X}_J - \widetilde{X}_J'\|_2^2)\}$$

since $\mathbb{E}(\|\widetilde{X}_* - \widetilde{X}_*'\|_2^2) \geq 0$ always holds, we have

$$\frac{1}{2}D_{E,\rho}(X_I, X_J) \leq \mathbb{E}(\|\widetilde{X}_I - \widetilde{X}_J\|_2^2)$$

by substituting Eq.3 here, the proof is completed. $\qquad \square$

Remark: From Lem.2, 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 extend. 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 isn't what we desire. Therefore, GED cannot be directly applied here.

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 *maximum mean discrepancy*(MMD) γ_k between P and Q is:

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

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

Lemma 3. *For two different class sets* $X_I, X_J \in \mathcal{M}^1_+(\mathcal{Z}) \bigcap \mathcal{M}^{1/2}_k(\mathcal{Z})$, *let* k *be degree-1 homogeneous polynomial kernel, then:*

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

Proof. Insert the *distance-induced kernel* k by corresponding ρ from Lem.1 into Eq.5, and cancel out the terms dependant on a single random variable, we have:

$$\gamma_k^2(X_I, X_J) = \frac{1}{2}\mathbb{E}_{\widetilde{X}_I\widetilde{X}'_I}[\rho(\widetilde{X}_I, z_0) + \rho(\widetilde{X}'_I, z_0) - \rho(\widetilde{X}_I, \widetilde{X}'_I)]$$
$$+ \frac{1}{2}\mathbb{E}_{\widetilde{X}_J\widetilde{X}'_J}[\rho(\widetilde{X}_J, z_0) + \rho(\widetilde{X}'_J, z_0) - \rho(\widetilde{X}_J, \widetilde{X}'_J)]$$
$$- \mathbb{E}_{\widetilde{X}_I\widetilde{X}_J}[\rho(\widetilde{X}_I, z_0) + \rho(\widetilde{X}_J, z_0) - \rho(\widetilde{X}_I, \widetilde{X}_J)]$$
$$= \mathbb{E}_{X_IX_J}\rho(X_I, X_J) - \frac{1}{2}\mathbb{E}_{X_IX'_I}\rho(X_I, X'_I) - \frac{1}{2}\mathbb{E}_{X_JX'_J}\rho(X_J, X'_J)$$

(6)

i.e. $\gamma_k^2(X_I, X_J) = \frac{1}{2}D_{E,\rho}(X_I, X_J)$, since k is *degree*-1 homogeneous polynomial kernel, from Prop.3 we have the corresponding generated $\rho = \|\cdot - \cdot\|_2^2$, then by using Lem.2, we have $L_{ec}(\theta_f; X_I, X_J) \geq \gamma_k^2(X_I, X_J)$. □

Remark: From Lem.3, 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 (Long et al. 2015; 2016; Tzeng et al. 2014). 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. This results also aren't what we desire and will degrade the model into a noisy counterpart. Therefore, MMD cannot be directly applied here as well.

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

4.2 Energy Confused Adversarial Metric Learning

The framework of ECAML can be generally applied to various metric learning objective functions, where we simultaneously train our Energy Confusion term and the distance metric term as follows:

$$\min_{\theta_f} L = L_m(\theta_f; T, D) + \lambda \sum_{I,J,I\neq J} L_{ec}(\theta_f; X_I, X_J) \quad (7)$$

where λ is the trade-off hyper-parameter and class sets X_I, X_J are randomly chosen in the current minibatch. In order to demonstrate the effectiveness of the proposed ECAML framework, we develop various SOTA metric learning objective functions here, i.e. $L_m(\theta_f; T, D)$:

ECAML(Tri): For triplet-tuple T and Euclidean measurement D, we employ (Schroff 2015; Wang and Gupta 2015):

$$L_m(\theta_f; T, D) = \sum_i^N [\|x_i - x_{i+}\|_2^2 - \|x_i - x_{i-}\|_2^2 + m]_+ \quad (8)$$

where the objective limits the distances of negative pairs larger than that of the positive pairs by margin m and features x_i is assumed to be on unit sphere, we experimentally find $m = 0.1$ performs best.

ECAML(N-Pair): For N-tuple T and inner-product measurement D, we employ (Sohn 2016):

$$L_m(\theta_f; T, D) = \sum_{i=1}^N \log(1 + \sum_{j=1, y_j \neq y_i}^N exp(x_i^T x_j - x_i^T x_{i+})) \quad (9)$$

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

ECAML(Binomial): For contrastive-tuple T and cosine measurement D, we employ (Yi et al. 2014; Opitz et al. 2017):

$$L_m(\theta_f; T, D) = \sum_{i,j} \log(1 + e^{-(2s_{ij}-1)\alpha(D_{ij}-\beta)\eta_{ij}}) \quad (10)$$

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 *resp*, η_{ij} is the penalty coefficient and is set to 1 if $s_{ij} = 1$, otherwise $\eta_{ij} = 25$, $D_{ij} = \frac{x_i^T x_j}{\|x_i\|\|x_j\|}$.

Moreover, for numerical stability, we extend our EC to a logarithmic counterpart and thus Eq.7 becomes:

$$\min_{\theta_f} L = L_m(\theta_f; T, D) + \lambda \sum_{I,J,I\neq J} \log(1 + L_{ec}(\theta_f; X_I, X_J))$$

(11)

Discussion: From Eq.11, our ECAML is achieved by jointly training the conventional metric objective and the proposed Energy Confusion goal. These two terms form an adversarial learning scheme by optimizing the opposite objective functions. Specifically, L_m acts as a 'defender' and L_{ec} acts as an 'attacker', the attacker intends to confuse the metric so as to 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 embedding will be less likely to the prejudiced concepts and, thus successfully prevent the biased learning behavior and improve the generalization ability. Moreover, we experimentally find that the overfitting mainly appears at the fc layer, thus our EC term is only used to constrain the learning of fc layer.

5. Experiments and Results

Implementation details: Following many other works, e.g. (Oh Song et al. 2016; Sohn 2016), we choose the pretrained *GooglenetV1* (Szegedy et al. 2014) as our bedrock CNN and randomly initialized an added fully connected layer. If not specified, we set the embedding size as 512 throughout our experiments. We also adopt exactly the same data preprocessing method (Oh Song et al. 2016) so as to make fair comparisons with other works[2]. For training, the optimizer is

---

[2] Only the images in CARS dataset are preprocessed differently, see the detail underneath Tab.3
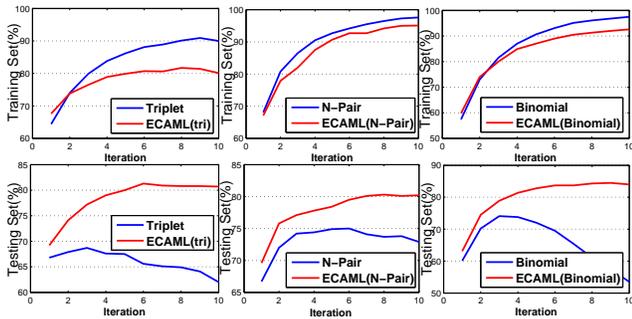
Figure 2: Recall@1 curves on training(seen classes, top fig) and testing(unseen classes, bottom fig) sets over CARS dataset.

CARS R@1						
λ	0 (Triplet)	0.001	0.01	0.02	0.1	1
ECAML(tri)	68.3	74.6	80.1	**81.0**	72.3	59.3
λ	0 (N-Pair)	0.1	0.2	0.3	0.4	0.5
ECAML(N-Pair)	74.3	77.4	79.6	**80.4**	78.6	73.7
λ	0 (Binomial)	0.01	0.1	0.13	0.15	0.5
ECAML(Binomial)	74.2	76.3	83.1	**84.5**	84.3	69.7

Table 1: Ablation experimental results on parameter λ.

	CARS			
	R@1	R@2	R@3	R@4
Binomial	74.2	83.1	86.7	92.9
Dropout(Binomial,0.1)	73.1	82.1	88.6	92.6
Dropout(Binomial,0.25)	74.5	83.3	85.9	92.6
Dropout(Binomial,0.4)	72.4	81.4	87.5	92.5
ECAML(Binomial)	84.5	90.4	93.8	96.6

Table 2: Comparison with Dropout on CARS datasets. We have experimented Dropout with {0.1,0.25,0.4} ratio.

Adam (Kingma and Ba 2014) 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), *resp*. The new fc-layer is optimized with 10 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 (Jia et al. 2014).

Evaluation and datasets: The same as many other works, the retrieval performance is evaluated by Recall@K metric. And following (Oh Song et al. 2016), 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, NMI$(\Omega, \mathbb{C}) = \frac{2I(\Omega, \mathbb{C})}{H(\Omega)+H(\mathbb{C})}$, and F_1 metric is the harmonic mean of precision and recall as follows $F_1 = \frac{2PR}{P+R}$. Then our ECAML is evaluated over the widely used benchmarks with the standard *zero-shot* evaluation protocol (Oh Song et al. 2016) on **CARS** (Krause et al. 2013), **CUB** (Wah et al. 2011), **Stanford Online Products** (Oh Song et al. 2016) and **In-Shop** (Liu et al. 2016).

5.1 Ablation Experiments

Regularization ability: To demonstrate the regularization ability of our ECAML, we plot the R@1 retrieval result curves on training(*seen*) and testing(*unseen*) sets *resp*, as in Fig.2. Specifically, for example, from the figures in left column, one can observe that the training curve of the conventional Triplet method rises quickly to a relatively high level but its testing curve only rises a little at first and then starts dropping to quite a low level, showing that the metric learned by conventional Triplet are more likely to over-fit the *seen* classes and generalize worse to the *unseen* classes in *zero-shot* settings. Conversely, by employing our ECAML(Tri), the training result curve rises much slower than the original Triplet and stops rising at a relatively lower level (80% *vs.* 90%), however, the testing cure of our ECAML(Tri) rises

fast to quite a high level, more than 80%, implying that our ECAML(Tri) indeed serves as a regularization method and improves the generalization ability of the learned metric by suppressing the learning of biased metric over *seen* classes caused by 'biased learning behavior'. Moreover, the similar phenomenon can be observed by ECAML(N-Pair,Binomial). **Ablation on λ**: From Tab.1, it can be observed that when $\lambda = 0$ our ECAML degenerates into the corresponding conventional metric learning method and the performance is unsatisfactory, and as λ increasing, the performances of ECAML(tri,N-Pair,Binomial) peak around {0.02, 0.3, 0.13} *resp* and outperform the baselines (Triplet, N-Pair, Binomial) by a large margin, validating the effectiveness and importance of our ECAML.

Ablation on regularization method There are some other research works aiming at imposing regularization in the top layer of the whole network, such as label-smothing(Szegedy et al. 2016), label-disturbing(Xie et al. 2016) and Noisy-Softmax(Chen, Deng, and Du 2017). However these methods are all designed for Softmax classifier layer and cannot be applied in the metric learning methods. Then, in order to show the effectiveness of our ECAML in the metric learning framework, we compare it with the commonly used 'Dropout' method. The dropout layer is placed after the CNN model. From Tab.2, one can observe that although the dropout with ratio 0.25 improves most of the performances over the baseline, the improvements are limited and not worthy of attention. However, in contrast to Dropout, our ECAML significantly surpasses the baseline model by a large margin. We conjecture that is because Dropout is not specially designed for the metric learning and the tested datasets are all fine-grained datasets in which simply depressing the neurons to be zero will largely affects the estimated distributions of these fine-grained classes regardless of the ratio value due to the small inter-class variations (for example, by using a smaller ratio (e.g. 0.1) the performance will still be reduced). In summary, our ECAML regularization method is specially designed for the deep metric learning and indeed performs well.

CARS						
Method	R@1	R@2	R@4	R@8	NMI	F1
Lifted(Oh Song et al. 2016)	49.0	60.3	72.1	81.5	55.1	21.5
Clustering(Song et al. 2017)	58.1	70.6	80.3	87.8	59.0	-
Angular(Wang et al. 2017)	71.3	80.7	87.0	91.8	62.4	31.8
ALMN(Chen and Deng 2018)	71.6	81.3	88.2	93.4	62.0	29.4
DAML(Duan et al. 2018)	75.1	83.8	89.7	93.5	66.0	36.4
Triplet	68.3	78.3	86.2	91.7	59.2	26.2
ECAML(Tri)	*81.0*	*88.2*	*92.8*	*96.0*	*65.7*	*33.0*
N-Pair	74.3	83.6	90.2	93.1	61.8	29.9
ECAML(N-Pair)	*80.4*	*88.2*	*92.4*	*95.8*	*64.6*	*32.7*
Binomial	74.2	83.1	86.7	92.9	61.5	28.8
ECAML(Binomial)	*84.5*	*90.4*	*93.8*	*96.6*	*68.4*	*38.4*

Table 3: Comparisons(%) with state-of-the-arts on CARS(Krause et al. 2013). λ for ECAML(tri, N-Pair, Binomial) are $\{0.02, 0.3, 0.13\}$. Here, the images are directly resized to 256x256, which are different from(Oh Song et al. 2016), then a 227x227 region is cropped.

CUB						
Method	R@1	R@2	R@4	R@8	NMI	F1
Lifted(Oh Song et al. 2016)	47.2	58.9	70.2	80.2	56.2	22.7
Clustering(Song et al. 2017)	48.2	61.4	71.8	81.9	59.2	-
Angular(Wang et al. 2017)	53.6	65.0	75.3	83.7	61.0	30.2
ALMN(Chen and Deng 2018)	52.4	64.8	75.4	84.3	60.7	28.5
DAML(Duan et al. 2018)	52.7	65.4	75.5	84.3	61.3	29.5
Triplet	49.5	61.7	73.2	82.5	57.2	24.1
ECAML(Tri)	*53.4*	*64.7*	*75.1*	*84.7*	*60.1*	*26.9*
N-Pair	51.9	63.3	73.9	83.0	59.7	26.5
ECAML(N-Pair)	*53.2*	*65.1*	*75.9*	*84.9*	*60.4*	*28.5*
Binomial	52.9	65.0	75.4	83.6	59.0	26.5
ECAML(Binomial)	*55.7*	*66.5*	*76.7*	*85.1*	*61.8*	*30.5*

Table 4: Comparisons(%) with state-of-the-arts on CUB(Wah et al. 2011). λ for ECAML(tri, N-Pair, Binomial) are $\{0.02, 0.3, 0.13\}$.

Stanford Online Products					
Method	R@1	R@10	R@100	R@1000	NMI	F1
Lifted(Oh Song et al. 2016)	62.1	79.8	91.3	97.4	87.4	24.7
Clustering(Song et al. 2017)	67.0	83.7	93.2	-	89.5	-
Angular(Wang et al. 2017)	70.9	85.0	93.5	98.0	87.8	26.5
ALMN(Chen and Deng 2018)	69.9	84.8	92.8	-	-	-
DAML(Duan et al. 2018)	68.4	83.5	92.3	-	89.4	32.4
Triplet	57.9	75.6	88.5	96.3	86.4	20.8
ECAML(Tri)	*64.9*	*80.0*	*90.5*	*96.9*	*87.0*	*23.3*
N-Pair	68.0	84.0	93.1	97.8	86.9	25.8
ECAML(N-Pair)	*69.8*	*84.7*	*93.2*	*97.8*	*88.0*	*27.2*
Binomial	68.5	84.0	93.1	97.7	88.5	29.9
ECAML(Binomial)	*71.3*	*85.6*	*93.6*	*98.0*	*89.9*	*32.8*

Table 5: Comparisons(%) with state-of-the-arts on Stanford Online Products(Oh Song et al. 2016). λ for ECAML(tri, N-Pair, Binomial) are $\{0.002, 0.03, 0.013\}$ *resp*.

In-Shop					
Method	R@1	R@10	R@20	R@30	R@40	R@50
FashionNet(Liu et al. 2016)	53	73	76	77	79	80
HDC(Yuan, Yang, and Zhang 2017)	62.1	84.9	89.0	91.2	92.3	93.1
BIER(Opitz et al. 2017)	76.9	92.8	95.2	96.2	96.7	97.1
Triplet	64.4	87.1	91.0	92.7	93.9	94.8
ECAML(Tri)	*68.0*	*89.9*	*93.3*	*94.8*	*95.7*	*96.3*
N-Pair	78.2	94.3	96.0	96.9	97.4	97.7
ECAML(N-Pair)	*79.8*	*94.6*	*96.1*	*97.0*	*97.4*	*97.7*
Binomial	81.7	94.5	96.2	97.2	97.6	97.9
ECAML(Binomial)	*83.8*	*95.1*	*96.6*	*97.3*	*97.7*	*98.0*

Table 6: Comparisons(%) with state-of-the-arts on In-shop (Liu et al. 2016). λ for ECAML(tri, N-Pair, Binomial) are $\{0.002, 0.03, 0.013\}$ *resp*.

5.2 Comparison with State-of-the-art

To highlight the significance of our ECAML framework, we compare with the aforementioned corresponding baseline methods, i.e. the wildly used Triplet (Schroff 2015), N-Pair (Sohn 2016) and Binomial (Yi et al. 2014), moreover, we also compare ECAML with other SOTA methods. The experimental results over CUB, CARS, Stanford Online Products and In-shop are in Tab.3-Tab.6 *resp*, bold number indicates improvement over baseline method. From these tables, one can observe that our ECAML consistently improves the performances of original metric learning methods (i.e. Triplet, N-Pair and Binomial) on all the benchmark datasets by a large margin, demonstrating the necessity of explicitly enhancing the generalization ability of the learned metric and validating the universality and effectiveness of our ECAML. Furthermore, our ECAML(Binomial) also surpasses all the listed state-of-the-art approches. 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 the performance of conventional metric learning methods.

6. Conclusion

In this paper, we propose the *Energy Confused Adversarial Metric Learning* (ECAML) framework, a generally applicable methods to various conventional metric learning approaches, for ZSRC tasks by explicitly intensifying the generalization ability within the learned embedding with the help of our Energy Confusion term. Extensive experiments on the popular ZSRC benchmarks (CUB, CARS, Stanford Online Products and In-Shop) demonstrate the significance and necessity of our idea of learning metric with good generalization by energy confusion.

References

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

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*.

Changpinyo, S.; Chao, W.-L.; Gong, B.; and Sha, F. 2016. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5327–5336.

Chen, B., and Deng, W. 2018. Almn: Deep embedding learning with geometrical virtual point generating. *arXiv preprint arXiv:1806.00974*.

Chen, B.; Deng, W.; and Du, J. 2017. Noisy softmax: Improving the generalization ability of dcnn via postponing the early softmax saturation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dalton, J.; Allan, J.; and Mirajkar, P. 2013. Zero-shot video retrieval using content and concepts. In *Proceedings of the 22nd ACM*

international conference on Information & Knowledge Management, 1857–1860. ACM.

Duan, Y.; Zheng, W.; Lin, X.; Lu, J.; and Zhou, J. 2018. Deep adversarial metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2780–2789.

Fu, Y.; Hospedales, T. M.; Xiang, T.; and Gong, S. 2015. Transductive multi-view zero-shot learning. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)* 37(11):2332–2345.

Graves, A. 2011. Practical variational inference for neural networks. In *Advances in neural information processing systems*, 2348–2356.

Gulcehre, C.; Moczulski, M.; Denil, M.; and Bengio, Y. 2016. Noisy activation functions. In *International Conference on Machine Learning (ICML)*, 3059–3068.

Hellinger, E. 1909. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik* 136:210–271.

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

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, 675–678. ACM.

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

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

Kullback, S., and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics* 22(1):79–86.

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

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1096–1104.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Deep transfer learning with joint adaptation networks. *arXiv preprint arXiv:1605.06636*.

Movshovitz-Attias, Y.; Toshev, A.; Leung, T. K.; Ioffe, S.; and Singh, S. 2017. No fuss distance metric learning using proxies. In *The IEEE International Conference on Computer Vision (ICCV)*.

Neelakantan, A.; Vilnis, L.; Le, Q. V.; Sutskever, I.; Kaiser, L.; Kurach, K.; and Martens, J. 2015. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*.

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

Opitz, M.; Waltner, G.; Possegger, H.; and Bischof, H. 2017. Bier - boosting independent embeddings robustly. In *The IEEE International Conference on Computer Vision (ICCV)*.

Opitz, M.; Waltner, G.; Possegger, H.; and Bischof, H. 2018. Deep metric learning with bier: Boosting independent embeddings robustly. *arXiv preprint arXiv:1801.04815*.

Schroff, F. e. a. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823.

Sejdinovic, D.; Gretton, A.; Sriperumbudur, B.; and Fukumizu, K. 2012. Hypothesis testing using pairwise distances and associated kernels (with appendix). *arXiv preprint arXiv:1205.0411*.

Sejdinovic, D.; Sriperumbudur, B.; Gretton, A.; and Fukumizu, K. 2013. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics* 2263–2291.

Shen, Y.; Liu, L.; Shen, F.; and Shao, L. 2018. Zero-shot sketch-image hashing. *arXiv preprint arXiv:1803.02284*.

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

Song, H. O.; Jegelka, S.; Rathod, V.; and Murphy, K. 2017. Deep metric learning via facility location. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun, Y.; Chen, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems (NIPS)*, 1988–1996.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2014. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2818–2826.

Székely, G. J., and Rizzo, M. L. 2004. Testing for equal distributions in high dimension. *InterStat* 5(16.10):1249–1272.

Székely, G. J., and Rizzo, M. L. 2005. A new test for multivariate normality. *Journal of Multivariate Analysis* 93(1):58–80.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Van Den Berg, C.; Christensen, J.; and Ressel, P. 2012. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*, volume 100. Springer Science & Business Media.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds200-2011 dataset. *California Institute of Technology*.

Wang, X., and Gupta, A. 2015. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802.

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

Wu, C.-Y.; Manmatha, R.; Smola, A. J.; and Krahenbuhl, P. 2017. Sampling matters in deep embedding learning. In *The IEEE International Conference on Computer Vision (ICCV)*.

Xie, L.; Wang, J.; Wei, Z.; Wang, M.; and Tian, Q. 2016. Disturblabel: Regularizing cnn on the loss layer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4753–4762.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Deep metric learning for person re-identification. In *International Conference on Pattern Recognition (ICPR)*, 34–39. IEEE.

Yuan, Y.; Yang, K.; and Zhang, C. 2017. Hard-aware deeply cascaded embedding. In *The IEEE International Conference on Computer Vision (ICCV)*.