Jiwen Lu, Junlin Hu, and Jie Zhou

# Deep Metric Learning for Visual Understanding

*An overview of recent advances*

©ISTOCKPHOTO.COM/ZAPP2PHOTO

Metric learning aims to learn a distance function to measure the similarity of samples, which plays an important role in many visual understanding applications. Generally, the optimal similarity functions for different visual understanding tasks are task specific because the distributions for data used in different tasks are usually different. It is generally believed that learning a metric from training data can obtain more encouraging performances than handcrafted metrics [1]–[3], e.g., the Euclidean and cosine distances. A variety of metric learning methods have been proposed in the literature [2]–[5], and many of them have been successfully employed in visual understanding tasks such as face recognition [6], [7], image classification [2], [3], visual search [8], [9], visual tracking [10], [11], person reidentification [12], cross-modal matching [13], image set classification [14], and image-based geolocalization [15]–[17].

Metric learning techniques are usually classified into two categories: unsupervised [4] and supervised [4]. Unsupervised metric learning attempts to learn a low-dimensional subspace to preserve the useful geometrical information of the samples. Supervised metric learning, which is the mainstream metric learning technique and the focus in this article, seeks an appropriate metric by formulating an optimization objective function to exploit supervised information of the training samples, where the objective functions are designed for different specific tasks. However, most conventional metric learning methods usually learn a linear mapping to project samples into a new feature space, which suffer from the nonlinear relationship of data points in metric learning. While the kernel trick can be adopted to address this nonlinearity problem, this type of method suffers from the scalability problem because the kernel trick has two major issues: 1) choosing a kernel is typically difficult and quite empirical and 2) the expression power of kernel functions is often not flexible enough to capture the nonlinearity in the data. Motivated by the fact that deep learning is an effective solution to model the nonlinearity of samples, several deep metric learning (DML) methods [6]–[10], [12], [14], [18]–[34] have been proposed in recent years. The key idea

*Digital Object Identifier 10.1109/MSP.2017.2732900*
*Date of publication: 13 November 2017*

of DML is to explicitly learn a set of hierarchical nonlinear transformations to map data points into other feature space for comparing or matching by exploiting the architecture of neural networks in deep learning, which unifies feature learning and metric learning into a joint learning framework. The goal of this article is to provide an overview of recent advances in DML techniques and their various applications in different visual understanding tasks.

### Mathematical background

To have a deep understanding of the concept of metric learning, we briefly introduce some necessary mathematical background. This section simply introduces the basic definitions of a metric space and how to find a well-defined metric (or pseudo-metric) over the original inputs by finding a mapping into a Euclidean space.

#### Definition 1

A metric over a set $\mathcal{X}$ is a mapping $d: \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ and this mapping $d$ satisfies the following properties (axioms) for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$:

1) $d(\mathbf{x}, \mathbf{y}) \geq 0$
2) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
3) $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$
4) $d(\mathbf{x}, \mathbf{x}) = 0$
5) $d(\mathbf{x}, \mathbf{y}) = 0 \Longleftrightarrow \mathbf{x} = \mathbf{y}$.

In Definition 1, axiom 1) is called the *nonnegativity axiom*, axiom 2) is known as the *symmetry axiom*, axiom 3) is called the *triangle inequality axiom*, axiom 4) is referred to as the *identity axiom*, and axiom 5) is known as the *identity of indiscernibles axiom*. A pair $(\mathcal{X}, d)$, in which $\mathcal{X}$ is a set and $d$ is a metric, is called a *metric space*.

#### Definition 2

A pseudo-metric over a set $\mathcal{X}$ is a mapping $d: \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ satisfying the following properties (axioms) for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$:

1) $d(\mathbf{x}, \mathbf{y}) \geq 0$
2) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
3) $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$
4) $d(\mathbf{x}, \mathbf{x}) = 0$.

A pair $(\mathcal{X}, d)$, in which $\mathcal{X}$ is a set and $d$ is a pseudo-metric, is called a *pseudo-metric space*. We find that the pseudo-metric doesn't need to satisfy the identity of indiscernibles axiom of the metric. In metric learning, we may consider the pseudo-metrics sometimes instead of metrics and refer to them as *metrics*.

The Euclidean distance is a widely used metric, which is usually adopted to measure the dissimilarity of data points. Give two data points $\mathbf{x}$ and $\mathbf{y}$, the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$d(\mathbf{x},\mathbf{y}) = \| \mathbf{x} - \mathbf{y} \|_2 = \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})}, \tag{1}$$

in which a large distance means the dissimilarity of $\mathbf{x}$ and $\mathbf{y}$, and a small distance denotes the similarity of $\mathbf{x}$ and $\mathbf{y}$.

The main objective of metric learning is to learn a metric over the input data points. One widely used method to learn a metric is to first map the input data points of the original space into a Euclidean metric space and then compute the Euclidean distance after the mapping. The following lemma declares this method.

#### Lemma 1

Let $\mathcal{X} = \{\mathbf{x}, \mathbf{y}, \mathbf{z}, \cdots\}$ be a set, $f: \mathcal{X} \to \mathbb{R}^n$ be any well-defined mapping, and $d: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ be the Euclidean metric over $\mathbb{R}^n$, then $d_f: \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ defined by $d_f(\mathbf{x}, \mathbf{y}) = d(f(\mathbf{x}), f(\mathbf{y})) = \| f(\mathbf{x}) - f(\mathbf{y}) \|_2$ is a well-defined pseudo-metric over $\mathcal{X}$.

As Definition 2 (pseudo-metric) keeps for all data points $f(\mathbf{x})$, $f(\mathbf{y})$, $f(\mathbf{z})$ and it is independent of the selection of mapping $f$, Lemma 1 is verified.

With Lemma 1, metric learning is the procedure of learning the mapping function $f$. In addition, from the perspective of feature representation, the goal of metric learning can be obviously interpreted as finding a new feature representation $\mathbf{h} = f(\mathbf{x})$ of the data point $\mathbf{x}$ to better suit the Euclidean space. Thus, the objective of metric learning is to find mapping $f$ under various loss functions and constraints.

### An illustration

To simply illustrate how metric learning works, we conducted an experiment on the MNIST data set [36]. We sampled 150 samples from three classes of handwritten digits: four, seven, and nine, where each class contains 50 samples. Each digit sample is a 28 × 28 grayscale image, and we lexicographically converted it into a 784-dimensional feature vector. We employed the linear discriminant analysis (LDA) as a metric learning method to project data points from the original space to the transformed space. Figure 1 shows an example of how metric learning works on this real-world data set. As seen, samples from different classes are mixed in the original space, and they are well separated in the transformed space.

In this article, we focus on DML, which explicitly learns a nonlinear mapping $f$ to map data points into a new feature space by exploiting the architecture of deep neural networks, in which the nonlinear mapping $f$ is parameterized by the weights and biases of deep neural network.

## DML

In this section, we introduce the basic concepts of DML, and discuss the similarities and differences among the existing DML methods.

### Basic concepts

From Lemma 1, DML is to explicitly learn a nonlinear mapping $f$ to map data points into a new feature space by exploiting the architecture of deep neural networks, in which the nonlinear mapping $f$ is parameterized by the weights and biases of deep neural network.

Given a simple neural network architecture as shown in Figure 2, for an input $\mathbf{x} \in \mathbb{R}^{r^{(0)}}$, its output of the first layer is $\mathbf{h}^{(1)} = \varphi(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \in \mathbb{R}^{r^{(1)}}$, and its output of the $m$th layer is $\mathbf{h}^{(m)} = \varphi(\mathbf{W}^{(m)}\mathbf{h}^{(m-1)} + \mathbf{b}^{(m)}) \in \mathbb{R}^{r^{(m)}}$, $1 \leq m \leq M$, $\mathbf{h}^{(0)} = \mathbf{x}$, where matrix $\mathbf{W}^{(m)} \in \mathbb{R}^{r^{(m)} \times r^{(m-1)}}$ and vector $\mathbf{b}^{(m)} \in \mathbb{R}^{r^{(m)}}$ are weights and biases of this neural network, $M$ is the total number

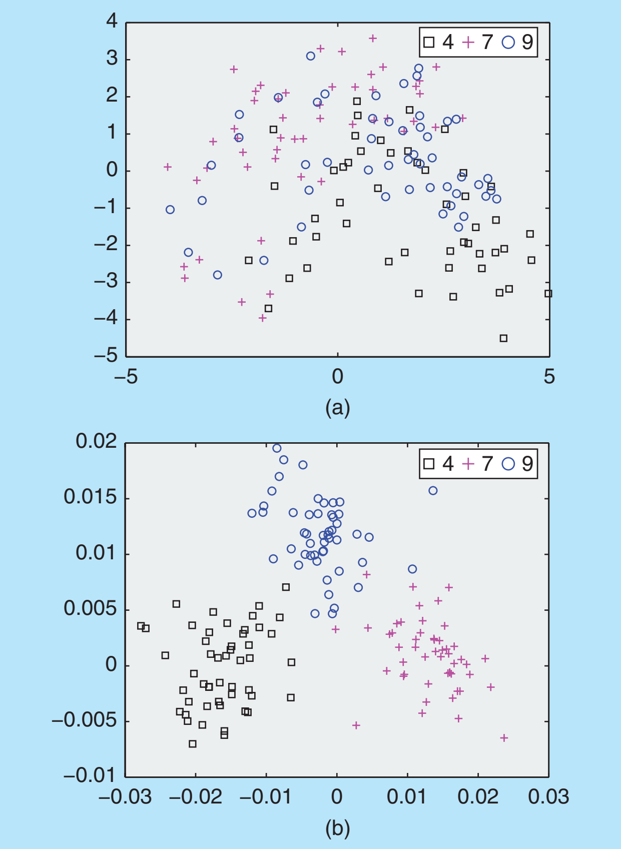

**FIGURE 1.** An example on the MNIST data set to illustrate how metric learning works. For ease of visualization, these samples are embedded into the two-dimensional feature spaces (a) and (b) by principal component analysis and LDA, respectively.

of layers, $r^{(m)}$ is the number of neural units in the $m$th layer, $\varphi:\mathbb{R}\mapsto\mathbb{R}$ is a nonlinear activation function (e.g., sigmoid and tanh). In this way, the output of this neural network at the most top layer can be represented as:

$$f(\mathbf{x}) = \mathbf{h}^{(M)} = \varphi(\mathbf{W}^{(M)}\mathbf{h}^{(M-1)} + \mathbf{b}^{(M)}) \in \mathbb{R}^{r^{(M)}}$$
$$\mathbf{h}^{(1)} = \varphi(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}), \tag{2}$$

where the mapping $f:\mathbb{R}^{r^{(0)}}\to\mathbb{R}^{r^{(M)}}$ is a parametric nonlinear function which is determined by a set of parameters $\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{m=1}^{M}$.

Let $f$ be the mapping function of a neural network. For an input $\mathbf{x}$, $f(\mathbf{x})$ is its output through this neural network. According to Lemma 1, the distance of data points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the deep metric space is to calculate the Euclidean distance between $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ as:

$$d_f(\mathbf{x}_i, \mathbf{x}_j) = d(f(\mathbf{x}_i), f(\mathbf{x}_j)) = \| f(\mathbf{x}_i) - f(\mathbf{x}_i) \|_2. \tag{3}$$

The goal of DML is to learn the mapping $f$ under certain constraints, where $f$ is parameterized by the weights and biases of the neural network.

Figure 3 shows another widely used architecture of neural network, called a *convolutional neural network* (*CNN*), which has been employed by many DML algorithms recently. Generally, CNNs comprise several convolutional layers, subsampling layers, and fully connected layers. Specifically, the feed-forward network in Figure 2 is the fully connected part of CNN architecture in Figure 3.

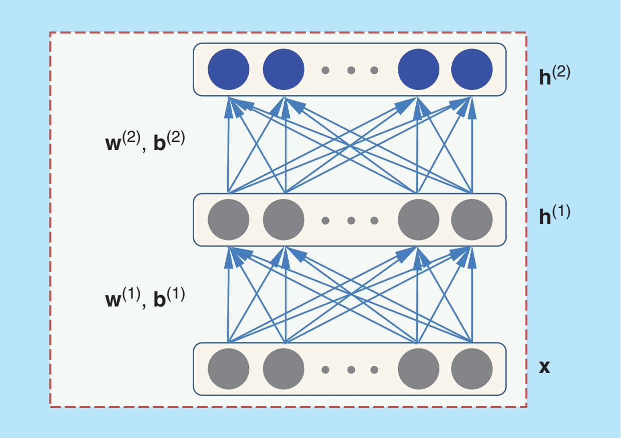
$\mathbf{h}^{(2)}$
$\mathbf{w}^{(2)}, \mathbf{b}^{(2)}$
$\mathbf{h}^{(1)}$
$\mathbf{w}^{(1)}, \mathbf{b}^{(1)}$
$\mathbf{x}$

**FIGURE 2.** A simple illustration of a feed-forward neural network architecture used in many DML methods [23]. The input to the network is $\mathbf{x}$, and the output of the hidden layer and the top layer is $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$, respectively, in which $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$ are weights and biases of this neural network, $1 \le m \le 2$.

### DML via Siamese networks

Typically, there are two main types of neural networks used in DML methods: Siamese networks and triplet networks. Figure 4 shows the diagrams of Siamese networks and triplet networks for DML. For a pair of data points $(\mathbf{x}_i, \mathbf{x}_j)$, we say they are a *similar pair* (or *positive pair*) if $\mathbf{x}_i$ and $\mathbf{x}_j$ are semantically similar, and they are called a *dissimilar pair* (or *negative pair*) if they are semantically dissimilar. Let $\mathcal{S} = \{(i, j)\}$ be an index set consisting of similar pairs, and $\mathcal{D} = \{(i, j)\}$ be an index set consisting of dissimilar pairs, respectively. The Siamese networks-based DML framework is trained by minimizing a contrastive loss function:

$$L(\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{m=1}^{M}) = \sum_{(i,j)\in\mathcal{S}} h(d_f(\mathbf{x}_i, \mathbf{x}_j) - \tau_1) + \sum_{(i,j)\in\mathcal{D}} h(\tau_2 - d_f(\mathbf{x}_i, \mathbf{x}_j)), \tag{4}$$

where $h(x) = \max(0, x)$ is the hinge loss function, and $\tau_1$ and $\tau_2$ are two positive thresholds, $\tau_1 < \tau_2$. By minimizing this contrastive loss function, we expect the distance $d_f(\mathbf{x}_i, \mathbf{x}_j)$ for a positive pair to be less than a smaller parameter $\tau_1$ and that of a negative pair to be larger than a larger parameter $\tau_2$. Figure 5 shows the key idea of such DML methods.

Dimensionality reduction by learning an invariant mapping (DrLIM) [18], [19] is an important work on DML via Siamese networks for face verification. DrLIM exploited discriminative information from neighborhood relationships of samples to learn the mapping function. There are four characteristics in their method: 1) it only needs neighborhood relationships between training samples; 2) it learns distance functions that are robust to nonlinear transformations of the input signals; 3) the learned function can handle the unseen classes problem so that the new coming testing samples can also be used with the learned metric; and 4) the mappings generated by the function is smooth and coherent in the output space.

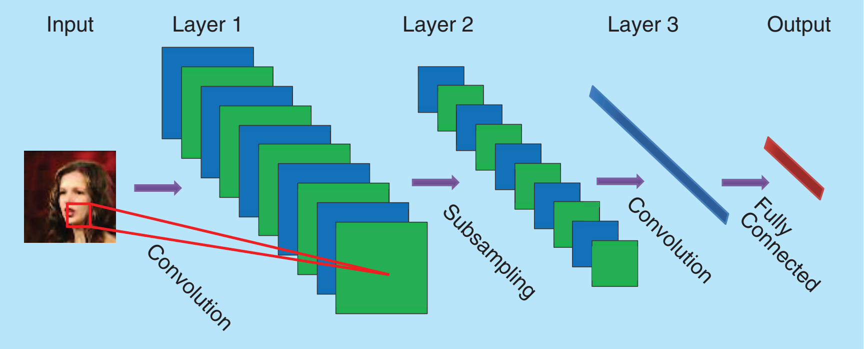

**FIGURE 3.** An illustration of a CNN architecture. This CNN comprises two convolutional layers $C_1$ and $C_3$, a subsampling layer $S_2$, and a fully connected layer $F_3$.

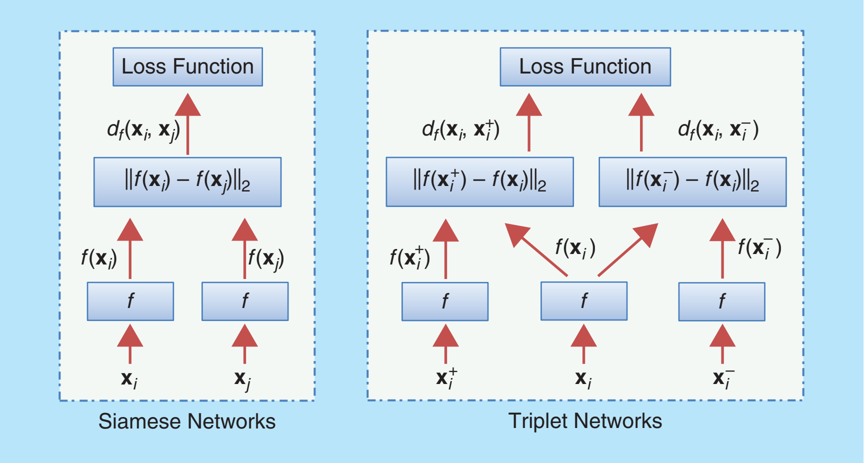

**FIGURE 4.** Diagrams of Siamese networks and triplet networks for DML. Siamese networks are composed of two same neural networks $f$ with shared parameters, where $(\mathbf{x}_i, \mathbf{x}_j)$ is a similar/dissimilar pair. Triplet networks consist of three same neural networks $f$ with shared parameters, where $(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)$ is a triplet, $\mathbf{x}_i$ is a reference, $\mathbf{x}_i^+$ and $\mathbf{x}_i^-$ are similar and dissimilar examples to $\mathbf{x}_i$.

Cai et al. [20] introduced a deep nonlinear metric learning (DNLML) method by using a deep independent subspace analysis (ISA) network, called DNLML-ISA for face verification. ISA is an unsupervised learning algorithm and a two-layer neural network, where different active functions in the first and second layers were used, respectively. Specifically, DNLML-ISA employed the ISA network to transform features from the original space to another feature subspace. To identify discriminative features, DNLML-ISA combined the side information constraints for metric learning with ISA, and stacked the ISA networks into a deep architecture. Since DNLML-ISA is trained layer by layer, it cannot use the backpropagation algorithm to update the model and also cannot fully exploit the discriminative information.

Hu et al. [6] introduced a discriminative DML (DDML) method for face verification. Unlike the stacked model used in DNLML-ISA, DDML employed a fully connected deep neural network to learn multiple nonlinear transformations to map face samples into a discriminative distance space, where the similarity of each positive pair is enlarged and that of each negative pair is reduced, respectively. The denoising autoencoder was used as the initialization of the parameters of each layer and then the backpropagation was used to update the model. The key advantage of DDML is that it can be trained on a small size of training data set and without using the extensive outside labeled data.

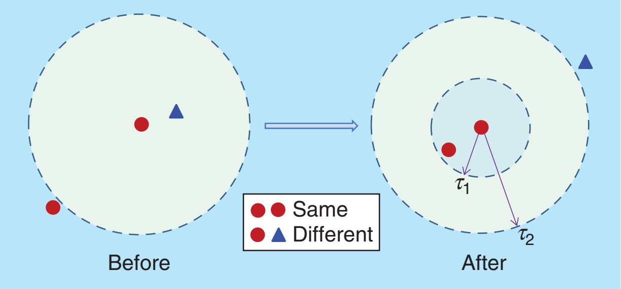

**FIGURE 5.** The basic idea of DML methods via Siamese network using (4) [6]. At the top layer of the network, the distance $d_f(\mathbf{x}_i, \mathbf{x}_j)$ for a positive pair is less than a smaller parameter $\tau_1$ and that of a negative pair is larger than a larger parameter $\tau_2$, respectively.

Taigman et al. [21] introduced a DeepFace method by employing an end-to-end metric learning method with the Siamese network for face recognition. Unlike DDML, where only the metrics were learned at the fully connected layers, DeepFace performed discriminative learning with the convolutional, pooling, and fully connected layers so that more labeled training samples were used to train the model. Finally, the parameters of the Siamese network were trained by the standard cross-entropy loss and backpropagation method.

Sun et al. [7] used carefully designed deep convolutional networks (deep ConvNets) by making use of both the verification and identification information to learn the deep identification-verification features (DeepID2) [7] for face verification. Specifically, their method extracted deep features with two signals: the first is the identification signal, which was achieved by following the DeepID2 layer with an *n*-way softmax layer. The network was trained by minimizing the cross-entropy identification loss. The other one is the verification signal, which enforced that DeepID2 features extracted from the same class are as similar as possible. Their method showed that both the identification and verification signal contributed to the final discriminative feature learning.

Yi et al. [12] proposed a DML method with a Siamese deep neural network to learn a similarity metric from image pixels directly for person reidentification. Their method jointly learned discriminative features and similarity measures under a unified deep framework. The network has a symmetrical structure, where two subnetworks were connected by a cosine similarity layer. There are two convolutional layers and a full connected layer for each subnetwork. Their method has two key advantages: 1) it can learn a similarity metric from image pixels directly; 2) it can learn multichannel filters to capture both the color and texture information from body images simultaneously.

Most DML methods assume that the training and testing samples are collected in similar scenarios and the same distribution assumption is usually made. This assumption does not hold in many real world applications, especially when samples are captured across different data sets. To address this, Hu et al. [23] proposed a deep transfer metric learning (DTML) method to learn hierarchical nonlinear transformations for cross-domain visual recognition, which learned transferrable discriminative knowledge from the labeled source domain to the unlabeled target domain. Specifically, DTML learned a deep metric network by maximizing the interclass variations and minimizing the intraclass variations, and minimizing the distribution divergence between the source domain and the target domain at the top layer of the network. To better exploit the discriminative information from the source domain, they also considered exploiting discriminative information from the middle layers of the deep network so that more discriminative information can be exploited.

Recently, Lu et al. [14] introduced a multimanifold DML (MMDML) method to recognize objects form different viewpoints or under different illuminations. Specifically, MMDML jointly learns multiple nonlinear feed-forward neural networks, one for each object class, to explicitly project the instances from each image set into a common feature space at the top layer of all networks, where the maximal manifold margin constraint is enforced. In this way, class-specific discriminative information can be effectively exploited for classification. The authors' method achieved competitive performance on five widely used image set data sets.

Table 1 shows basic characteristics of several Siamese networks-based DML methods. In this table, the strongly supervised setting means that the class labels of training data are used to train neural networks, and the weakly supervised setting denotes that only the pairwise labels of similar pairs and dissimilar pairs are used to train neural networks.

**Table 1. Characteristics of several DML methods using Siamese networks.**

| Method | Setting | End to End? | Convolutional Architecture? |
|---|---|---|---|
| DrLIM [19] | Strongly supervised | Yes | Yes |
| DNLML-ISA [20] | Weakly supervised | No | No |
| DDML [6] | Weakly supervised | No | No |
| DeepFace [21] | Strongly supervised | Yes | Yes |
| DeepID2 [7] | Strongly supervised | Yes | Yes |
| DTML [23] | Strongly supervised | No | No |
| MMDML [14] | Strongly supervised | No | No |

### *DML via triplet networks*

DML using triplet networks was trained by minimizing a triplet loss function, which exploits labels of training data to generate triplets. Given a triplet $(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)$, $\mathbf{x}_i^+$ is a similar example to the reference $\mathbf{x}_i$, and $\mathbf{x}_i^-$ is a dissimilar example to the $\mathbf{x}_i$. A triplet $(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)$ means that $\mathbf{x}_i$ is more similar to $\mathbf{x}_i^+$ in contrast to $\mathbf{x}_i^-$, i.e., $d_f(\mathbf{x}_i, \mathbf{x}_i^+) < d_f(\mathbf{x}_i, \mathbf{x}_i^-)$. DML via triplet networks aims to minimize the following loss function for triplets:

$$L(\{\mathbf{W}^{(m)}, \mathbf{b}^{(m)}\}_{m=1}^{M}) = \sum_i h(\tau + d_f(\mathbf{x}_i, \mathbf{x}_i^+) - d_f(\mathbf{x}_i, \mathbf{x}_i^-)), \quad (5)$$

where $h(x) = \max(0, x)$ is the hinge loss function, and $\tau > 0$ is a margin between $d_f(\mathbf{x}_i, \mathbf{x}_i^+)$ and $d_f(\mathbf{x}_i, \mathbf{x}_i^-)$. The triplet network pulls the similar example close to reference and pushes dissimilar example further away.

Wang et al. [9] proposed a deep ranking model with the triplet-based hinge loss functions to learn similarity metric from raw images. Specifically, they employed a multiscale neural network architecture to capture both the global visual properties and the image semantics. An efficient online triplet sampling method was presented to generate a large amount of training data to learn the parameters of the network.

Hoffer et al. [26] employed a triplet network architecture for DML, which aims to learn useful representations by distance comparisons. Their method is similar to the approach in [9] that learned a deep ranking similarity function for image retrieval. Their method made a comprehensive study of the triplet architecture, and demonstrated that the triplet approach is a strong competitor to the Siamese approach.

Schroff et al. [24] introduced a FaceNet deep model that directly learns a mapping from the original sample space to a compact Euclidean space. Once this space is produced, face recognition and clustering can be easily implemented under the network. Specifically, FaceNet used a deep convolutional network to directly optimize the embedding itself rather than using an intermediate bottleneck layer. Triplets of roughly aligned matching/nonmatching face patches were generated for training with an online triplet mining method.

Bell and Bala [27] proposed learning visual similarity for product design with the CNNs, which exploit communities of users to help each other answering questions about products in images. Their method contains two different domains of product images: products cropped from internet scenes, and products in their iconic form. With the help of a multidomain deep embedding, it can deal with several applications of visual search including identifying products in scenes and finding stylistically similar products.

Song et al. [28] introduced a DML method via lifted structured feature embedding (LiftedStruct) to learn semantic feature embeddings where similar examples are mapped close to each other and dissimilar examples are mapped farther apart. Their method took full advantage of the training batches in the network training stage by lifting the vector of pairwise distances within the batch to the matrix of pairwise distances. This step enabled the method to learn the state of the art feature embedding by optimizing a new structured prediction objective on the lifted problem. Experiments on three large-scale data sets demonstrated significant improvements over existing deep feature embedding methods.

Cui et al. [29] presented an iterative framework for fine-grained visual categorization with humans in the loop information. Their method can handle three challenges in existing fine-grained visual categorization methods: lacking of training data, large number of fine-grained categories, and high intraclass versus low interclass variance. Using DML with humans in the loop, a low-dimensional feature embedding with anchor points on manifolds was learned for each category, where these anchor points captured intraclass variances and remained discriminative among different classes. In each round, images with high confidence scores from our model were sent to humans for labeling. By comparing these images with exemplar images, labelers marked each candidate image as either a true positive or a false positive. True positives were added into the current data set and false positives were considered as hard negatives for the DML model. Then the model was retrained with an expanded data set and hard negatives for the next round iteration. The proposed DML method was evaluated on two fine-grained data sets. Experimental evaluations showed that their method achieved significant performance gain over state-of-the-art methods.

Shi et al. [31] proposed a deep metric embedding method with triplet loss for person reidentification. Their method introduced a positive sample mining method to train robust CNN for person reidentification. In addition, a metric weight constraint was used to improve the learning, so that the learned metric has a better generalization ability. They empirically found that both of these tricks improve the reidentification performance.

Lim et al. [33] proposed a competitive approach for style similarity learning of three-dimensional (3-D) shapes using DML, which made use of recent advances in triplet based metric learning with neural networks. The key advantages of their method are four aspects:

- it explored DML techniques for perceived style similarities of 3-D shapes
- it showed that rendered images of 3-D geometry from multiple viewpoints were an appropriate representation and how salient views can be selected
- it used a triplet sampling method that does not rely on style class labels and allows for an efficient learning procedure
- it showed how heterogeneous data sources in the form of 3-D geometry and annotated photographs found online can be integrated into the DML method.

### DML via other networks

There are also some DML methods via other networks. For example, Batchelor and Green [22] proposed using DML on CNNs to learn features with good locality for object recognition. In particular, they considered two metric learning methods: neighborhood components analysis and mean square error's gradient minimization (MEGM). They utilized a nonlinear form of MEGM as an alternative to neighborhood components analysis and proposed some stochastic sampling methods to apply them to larger data sets with a minibatch stochastic gradient descent algorithm.

Sohn [32] proposed a DML method using multiclass $N$-pair loss [32]. Their method first generated triplet loss by allowing joint comparison among more than one negative example. Then, $N-1$ negative examples were considered to reduce the computational burden of evaluating deep embedding vectors. They demonstrated the superiority of their method over other competing loss functions for a variety of tasks such as fine-grained object recognition and verification, image clustering and retrieval, and face verification and identification.

## Visual understanding applications

In this section, we show various visual understanding applications via DML, including face recognition, image classification, visual search, person reidentification, visual tracking, cross-modal matching, and image set classification.

### Face recognition

Chopra et al. [18] learned a similarity metric for face verification. Their approach learned a CNN-based mapping from the input space to the target space, where the $L_1$ norm can directly approximate the semantic distance. Cai et al. [20] learned a nonlinear metric using the deep ISA network. Compared with kernel-based methods, deep models present strong discriminative power and better exploit the nature of the data set. Sun et al. [7] proposed a DeepID2 method to increase the interpersonal variations with the identification signals, and reduce the intrapersonal distances with the verification signals. Taigman et al. [21] presented the DeepFace network by exploiting a 3-D face model and training a nine-layer CNN network. Hu et al. [6] presented a DDML method by learning a set of hierarchical nonlinear transformations, where the distance between positive pairs is smaller than negative pairs by a threshold. They also proposed a DTML [23] for cross-data set face recognition. DTML transferred the information from the labeled source domain to the unlabeled target domain, and minimized their distribution divergence. Schroff et al. [24] proposed a FaceNet method by learning a projection to map facial images to a compact Euclidean space. With the learned embedding, feature vectors can be directly used to measure the similarity of faces. Most these DML methods achieved the state-of-the-art performance on the widely used LFW and YouTube Face data sets.

### Image classification

Batchelor and Green [22] utilized CNN architecture to learn a deep nonlinear metric, where the learned features with good locality show good performance and generalization for image classification. Hoffer and Ailon [26] utilized a triplet-based network to learn deep metrics by distance comparisons. The triplet network contains three instances of networks with shared parameters, where three samples with a positive pair and a negative pair can be simultaneously fed into the network. Cui et al. [29] learned a deep metric for fine-grained categorization. Human helps to label high confidence images in each loop to expand data sets and hard negatives, where the network was further retrained in the next loop.

### Visual search

Wu et al. [8] proposed an online multimodal deep similarity learning for visual search. They applied deep-learning techniques to

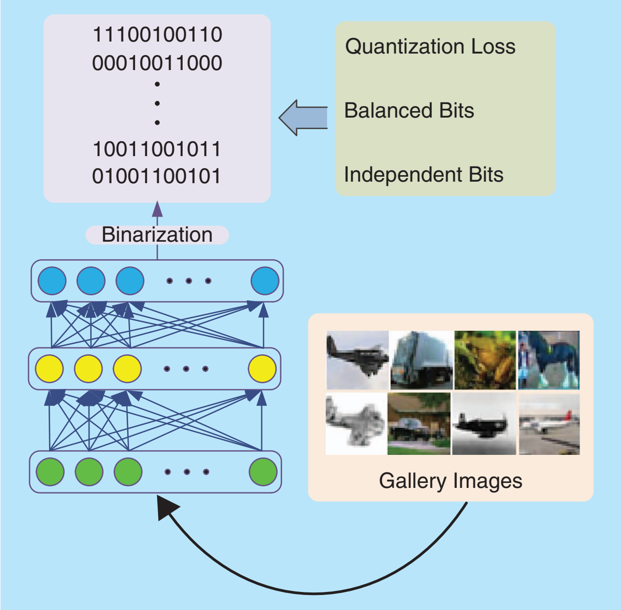

**FIGURE 6.** The basic idea of deep hashing for large scale visual search [35], which employed a feed-forward neural network to map each gallery image into a compact Hamming feature space with three criteria.

obtain a flexible nonlinear similarity metric from images, which have multimodal feature representations via an efficient and scalable online learning method. Their proposed technique achieved encouraging results on several multimodal images retrieval tasks. Wang et al. [9] proposed a ranking-based deep-learning method for fine-grained image search, which employed a triplet-based hinge loss ranking function and a multiscale neural network. Their method outperformed existing hand-crafted features and deep models in numerous experiments. Liong et al. [35] proposed a deep hashing approach for large scale visual search. Their method learned compact binary codes to exploit nonlinear relationship of samples, and Figure 6 shows the key idea of their method.

### Person reidentification

Yi et al. [12] proposed a DML method for person reidentification, which learned a similarity metric from image pixels directly with a Siamese networks. Hu et al. [23] proposed a DTML method for cross data set person reidentification, where the discriminative information exploited from the source domain was transferred to the target domain with limited labeled samples. Shi et al. [31] proposed a deep embedding metric method for person reidentification, which used a moderate positive sample mining method for robust CNN training, and improved the leaning procedure with a metric weight constraint.

### Visual tracking

Hu et al. [10] employed DML for visual object tracking. Their DML tracker adopts the marginal fisher analysis criterion to characterize the separability of the positive samples and negative samples by maximizing the variance of interclass negative sample pairs. They first learned a multilayer nonlinear feed-forward neural network to map both the sampled templates and particles into a discriminative feature space to minimize the intraclass variations of positive sample pairs and maximize the interclass variations of negative sample pairs at the top layer of the network. Then, they selected the candidate which is most similar to the template in the learned deep network and considered it as the target in the current predicted frame. Experimental results demonstrated that their DML tracker achieved very competitive performance on a challenging benchmark data set. Figure 7 shows the main procedure of their proposed DML tracker.

### Cross-modal matching

Liong et al. [13] proposed a deep coupled metric learning (DCML) method for cross-modal matching. Unlike existing cross-modal

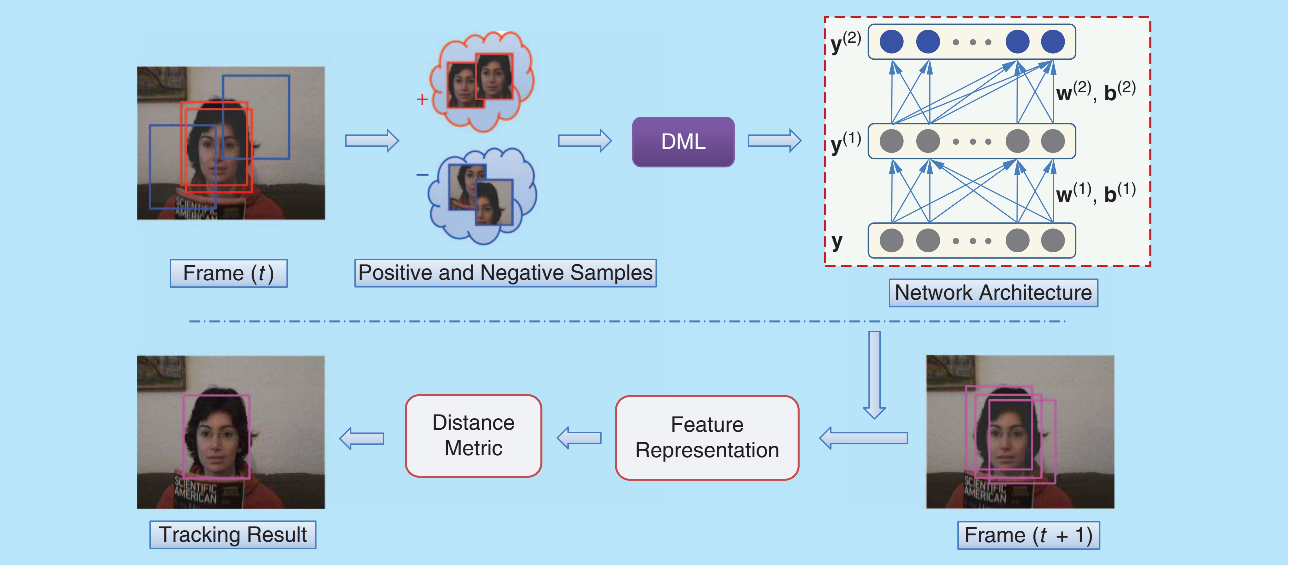

**FIGURE 7.** The key procedure of the DML tracker [10]. This tracker sampled some positive and negative samples to construct a training set and learn a deep metric with these samples at the *t*th frame, For the coming (*t* + 1)th frame, their tracker computed the similarity between each candidate and the template and find the candidate which has the maximal similarity score as the foreground the (*t* + 1)th frame.

learning methods such as canonical correlation analysis and partial least squares, which learn a single linear latent space to reduce the modality gap, their DCML designs two neural networks to learn two sets of hierarchical nonlinear transformations (one set for each modality) to nonlinearly map data samples into a shared feature subspace, under which the intraclass variation is minimized and the interclass variation is maximized, and the difference of each sample pair captured from two modalities of the same class is minimized, respectively. Experimental results on three different cross-modal matching applications including text-image matching, tag-image retrieval, and heterogeneous face recognition demonstrated the effectiveness of the proposed method. Lin et al. [15], Workman et al. [16], and Vo and Hays [17] employed DML techniques to address the cross-view matching problem for image-based geolocalization, in which these methods were used to localize a ground-level query image by matching to a reference database of aerial/overhead images.

### *Image set classification*

Lu et al. [14] presented an MMDML method to recognize objects form different viewpoints or under different illuminations. Specifically, MMDML jointly learns multiple nonlinear feedforward neural networks, one for each object class, to explicitly project the instances from each image set into a common feature space at the top layer of all networks, where the maximal manifold margin constraint is enforced. In this way, class-specific discriminative information can be effectively exploited for classification. The authors' method achieved competitive performance on five widely used image set data sets. Figure 8 shows the key idea of their MMDML method.

## Summary and future research directions

In this article, we have summarized the recent trends of DML and shown their wide applications of various visual understanding tasks including face recognition, image classification, visual search, person reidentification, visual tracking, cross-modal matching, and image set classification. Empirical results have clearly demonstrated that DML can significantly improve the state of the art in these visual understanding tasks.

There are five interesting directions of DML for future research:

1) Most existing DML methods learn one neural network from a single feature representation and cannot deal with multiple feature representations directly. In many visual understanding applications, it is easy to extract multiple features for each sample for multiple feature fusion. However, these features extracted from the same sample are usually highly correlated to each other even if they could characterize samples from different aspects. For multiple feature fusion, this highly correlated information should be preserved because it usually reflects the intrinsic information of samples. How to perform DML with multiview feature representation to preserve the correlation of different features and further improve the performance is a desirable future work.
2) Most existing DML methods assume that high-quality and clean samples are usually obtained so that the learned metrics are employed for visual understanding. In many real-world applications, visual data are usually captured in wild conditions so that many noisy and low-quality samples are usually collected, so that it is desirable to develop robust DML methods that can well measure the similarity of these noisy and low-quality samples. Hence, how to develop robust DML methods is another interesting future direction for research.
3) Most existing DML methods are developed for a single specific task, which means that a large amount of labeled data for this task are usually required to exploit the supervision information. In some real applications, it is difficult to collect extensive labeled data for a specific task. Therefore, it is desirable to conduct multitask DML which can leverage labeled samples from multiple different yet related tasks so that it is much easier to obtain more labeled samples for DML, which is also an interesting future research direction.
4) Most existing DML methods are supervised. In many real applications, it is easier to collect an extensive unlabeled data rather than labeled data for practical applications. Hence, how to develop more effective unsupervised or semisupervised DML is an important future direction.
5) Most existing DML methods utilize the contrastive and triplet loss functions to train deep models. To complete the family of DML, employing other loss functions (e.g., quadruplet loss [37]) is also a promising path to the development of DML.

## Acknowledgments

We would like to thank the anonymous reviewers for their insightful suggestions for improving the article. This work was supported in part by the National Key Research and Development

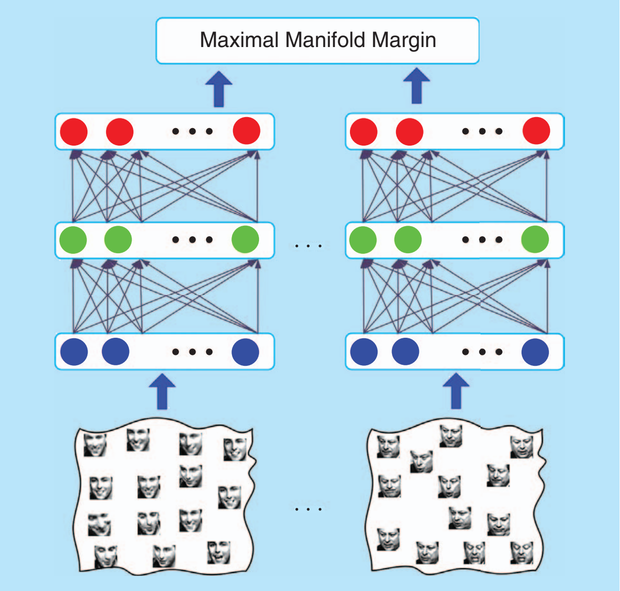

**FIGURE 8.** The basic idea of MMDML for image set classification [14]. MMDML models each image set as a nonlinear manifold and employs a feed-forward neural network to nonlinearly map it into a feature space. Assume there are *C* classes, MMDML designs *C* feed-forward neural networks (one for each manifold). At the top layer of the network, the manifold margin is maximized so that the parameters of these manifolds can be updated with backpropagation. Finally, the testing image set is fed to each network and the smallest distance between it and the training class is used for classification.

Program of China under grant 2016YFB1001001 and the National Natural Science Foundation of China under grant 61672306.

## Authors

***Jiwen Lu*** (lujiwen@tsinghua.edu.cn) received his B.Eng. degree in mechanical engineering and his M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively. He received his Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an associate professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and machine learning. He is an associate editor of four international journals including *Pattern Recognition*, and an elected member of the IEEE Technical Committees of the IEEE Circuits and Systems Society and the IEEE Signal Processing Society. He is a Senior Member of the IEEE.

***Junlin Hu*** (jhu007@e.ntu.edu.sg) received his B.Eng. degree from the Xi'an University of Technology, Xi'an, China, in 2008, and the M.Eng. degree from Beijing Normal University, China, in 2012. He is currently pursuing his Ph.D. degree with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include computer vision, pattern recognition, and biometrics.

***Jie Zhou*** (jzhou@tsinghua.edu.cn) received his B.S. and M.S. degrees from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and his Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995. From 1995 to 1997, he served as a postdoctoral fellow with the Department of Automation, Tsinghua University, Beijing, China, where he has been a full professor, since 2003. His current research interests include computer vision, pattern recognition, and image processing. He received the National Outstanding Youth Foundation of China Award. He is a Senior Member of the IEEE.

## References

[1] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell, "Distance metric learning with application to clustering with side-information," in *Proc. Neural Information Processing Systems*, 2002, pp. 505–512.

[2] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. Int. Conf. Machine Learning*, 2007, pp. 209–216.

[3] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learning Res.*, vol. 10, pp. 207–244, 2009.

[4] B. Kulis, "Metric learning: A survey," *Found. Trends Mach. Learning*, vol. 5, no. 4, pp. 287–364, 2013.

[5] M. Harandi, M. Salzmann, and R. Hartley, "Joint dimensionality reduction and metric learning: A geometric take," in *Proc. Int. Conf. Mach. Learning*, 2017, pp. 1404–1413.

[6] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2014, pp. 1875–1882.

[7] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Neural Information Processing Systems*, 2014, pp. 1988–1996.

[8] P. Wu, S. C. H. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao, "Online multimodal deep similarity learning with application to image retrieval," in *Proc. ACM Conf. Multimedia*, 2013, pp. 153–162.

[9] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2014, pp. 1386–1393.

[10] J. Hu, J. Lu, and Y.-P. Tan, "Deep metric learning for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 11, pp. 2056–2068, 2016.

[11] H. Li, Y. Li, and F. Porikli, "Deeptrack: Learning discriminative feature representations online for robust visual tracking," *IEEE Trans. Image Proc.*, vol. 25, no. 4, pp. 1834–1848, 2016.

[12] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. Int. Conf. Pattern Recognition*, 2014, pp. 34–39.

[13] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1234–1244, 2017.

[14] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou, "Multi-manifold deep metric learning for image set classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1137–1145.

[15] T.-Y. Lin, Y. Cui, S. J. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 5007–5015.

[16] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 3961–3969.

[17] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *Proc. European Conf. Computer Vision*, 2016, pp. 494–509.

[18] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, pp. 539–546.

[19] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006, pp. 1735–1742.

[20] X. Cai, C. Wang, B. Xiao, X. Chen, and J. Zhou, "Deep nonlinear metric learning with independent subspace analysis for face verification," in *Proc. ACM Conf. Multimedia*, 2012, pp. 749–752.

[21] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[22] O. Batchelor and R. D. Green, "Object recognition by stochastic metric learning," in *Proc. Int. Conf. Simulated Evolution and Learning*, 2014, pp. 798–809.

[23] J. Hu, J. Lu, and Y.-P. Tan, "Deep transfer metric learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 325–333.

[24] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[25] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3279–3286.

[26] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Similarity-Based Pattern Recognition, Third Int. Workshop*, 2015, pp. 84–92.

[27] S. Bell and K. Bala, "Learning visual similarity for product design with convolutional neural networks," *ACM Trans. Graphics*, vol. 34, no. 4, pp. 98, 2015.

[28] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4004–4012.

[29] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained categorization and data set bootstrapping using deep metric learning with humans in the loop," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 1153–1162.

[30] J. Hu, J. Lu, Y.-P. Tan, and J. Zhou, "Deep transfer metric learning," *IEEE Trans. Image Proc.*, vol. 25, no. 12, pp. 5576–5588, 2016.

[31] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W.-S. Zheng, and S. Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. European Conf. Computer Vision*, 2016, pp. 732–748.

[32] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proc. Neural Information Processing Systems*, 2016, pp. 1849–1857.

[33] I. Lim, A. Gehre, and L. Kobbelt, "Identifying style of 3-D shapes using deep metric learning," *Comput. Graphics Forum*, vol. 35, no. 5, pp. 207–215, 2016.

[34] J. Lu, J. Hu, and Y.-P. Tan, "Discriminative deep metric learning for face and kinship verification," *IEEE Trans. Image Processing*, vol. 26, no. 9, pp. 4269–4282, 2017.

[35] V. E. Liong, J. Lu, G. Wang, P. Moulin, and J. Zhou, "Deep hashing for compact binary codes learning," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2015, pp. 2475–2483.

[36] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[37] M. T. Law, N. Thome, and M. Cord, "Learning a distance metric from relative comparisons between quadruplets of images," *Int. J. Comput. Vision*, vol. 121, no. 1, pp. 65–94, 2017.