symmetry

MDPI

Review

Deep Metric Learning: A Survey

Mahmut KAYA [1,*] and Hasan Şakir BİLGE [2]

[1] Department of Computer Engineering, Engineering Faculty, Siirt University, Siirt 56100, Turkey
[2] Department of Electrical - Electronics Engineering, Engineering Faculty, Gazi University, Ankara 06570, Turkey
* Correspondence: mahmutkaya@siirt.edu.tr

Received: 23 July 2019; Accepted: 20 August 2019; Published: 21 August 2019

Abstract: Metric learning aims to measure the similarity among samples while using an optimal distance metric for learning tasks. Metric learning methods, which generally use a linear projection, are limited in solving real-world problems demonstrating non-linear characteristics. Kernel approaches are utilized in metric learning to address this problem. In recent years, deep metric learning, which provides a better solution for nonlinear data through activation functions, has attracted researchers' attention in many different areas. This article aims to reveal the importance of deep metric learning and the problems dealt with in this field in the light of recent studies. As far as the research conducted in this field are concerned, most existing studies that are inspired by Siamese and Triplet networks are commonly used to correlate among samples while using shared weights in deep metric learning. The success of these networks is based on their capacity to understand the similarity relationship among samples. Moreover, sampling strategy, appropriate distance metric, and the structure of the network are the challenging factors for researchers to improve the performance of the network model. This article is considered to be important, as it is the first comprehensive study in which these factors are systematically analyzed and evaluated as a whole and supported by comparing the quantitative results of the methods.

Keywords: metric learning; deep metric learning; similarity; siamese network; triplet network

1. Introduction

The concept of machine learning, which allows for computers to learn clearly without being directly programmed, emerged after computers gained the ability to recognize objects [1]. Today, we can benefit from machine learning in many areas to make our life much easier. The fields where it is used include places such as face recognition, medical diagnosis, intrusion detection systems, speech recognition, voice recognition, text mining, object recognition, and so on. Thus, machine learning can be claimed to offer successful solutions for complex problems and large amounts of data [2,3].

Machine learning algorithms may produce a very successful classification model that is based on the available data. The proposed model is expected to yield successful results in both test and training data. However, it may not always be possible to expect the proposed model to produce the desired result in any problem, because each data has its own basic problems that need to be solved. For example, when you deal with a face recognition problem, you may deal with such factors as pose variations, illumination differences, scaling, background, occlusion, and expression, each of which causes various problems in the classification of the data. Therefore, to address these factors, the distinguishing characteristics of the data must be well defined in order to correctly classify the data.

k-nearest neighbor, support vector machines, and Naïve Bayes classifiers can be used in machine learning applications. Although these algorithms have a certain classification performance, it is possible to represent the data in a better way. These algorithms do not transform an original dataset

to a new space. The effect of each feature is not equal in terms of classification. Therefore, feature weighting might be used before classification. The dataset can also be transformed from original space to a new space. To realize this, data transformation algorithms, like Principal Component Analysis and Linear Discriminant Analysis, have been benefited. The dataset can be classified after the data transformation using these algorithms.

Metric learning is an approach based directly on a distance metric that aims to establish similarity or dissimilarity between objects. While metric learning aims to reduce the distance between similar objects, it also aims to increase the distance between dissimilar objects. For this reason, there are approaches, such as k-nearest neighbors, which calculate distance information, and approaches where the data is transformed into a new representation. While the metric learning approaches are moved to the transformation space with distance information, the method is basically based on a W projection matrix. Current studies are directly related to Mahalanobis distance in general [4–6]. When Mahalanobis distance is transformed into the Euclidean distance, the metric learning approach is presented based on the decomposition of the covariance matrix and the use of symmetric positive definite matrices while performing these operations.

Data volume has been increasing day by day, which provides significant advantages for more accurate classification. However, this also brings a large number of calculations with it. It is essential to carry out operations in pieces and together due to many computing requirements. Therefore, it is possible to come up with fast and successful solutions in machine learning thanks to the power of parallel computing. Deep learning with multi-layered structure has become one of the most popular topics of our time in computer science with the development of GPU technology in recent years [7]. Deep learning, which provides a new representation of the data over raw data, obtains the automatic extraction of features where the goal is to achieve higher abstraction levels when transforming data [8,9]. Deep learning offers a compact structure on its own and it includes the classification in the architecture. It has also a nonlinear structure that can provide us more realistic approaches to detect real-world problems with the power of activation functions.

In the last few years, deep learning and metric learning have been brought together to introduce the concept of deep metric learning [10]. Deep metric learning is based on the principle of similarity between samples. In 2017, Lu et al. [10] summarized the concept of deep metric learning for visual understanding tasks. The concept of deep metric learning has been shown in Figure 1. In this study, more recent approaches for Image, Video, Speech, and Text tasks were dealt with. The network structure, loss function, and sample selection are important factors for the success of the network in deep metric learning. All of the details of these main factors were mentioned in light of recent studies. Besides, a general framework was presented with quantitative experiments in comparing the methods.

In this study, firstly metric learning was focused on. After giving some details about the background of metric learning in Section 2, recent improvements in deep learning were discussed. Subsequently, the relationship between deep learning and metric learning was addressed. After that, deep metric learning problems, sample selection, and metric loss functions were explained in detail in Section 3. Finally, some conclusions about the current situation and the future of deep metric learning were presented.

This paper attempts to put forward the significance of deep metric learning and the issues handled in this field. This research is considered to be important, as it is the first comprehensive study where these factors are systematically examined and evaluated as a whole while comparing the quantitative results of the methods.

2. Metric Learning

Each dataset has specific problems in terms of classification and clustering. Distance metrics that do not have a good learning ability independent of the problem can be claimed not to yield successful results in the classification of data. Therefore, a good distance metric is required to achieve successful results on the input data [11,12]. To address this problem, several works have been conducted

while using metric learning approaches [6,11–14]. Metric learning provides a new distance metric by analyzing data. A metric learning approach that performs the learning process on the data will have a higher ability to distinguish the sample data. The main purpose of metric learning is to aim to learn a new metric to reduce the distances between samples of the same class and increase the distances between the samples of different class [15]. As can be seen in Figure 1c, while metric learning aims to bring similar objects closer, it increases the distance between dissimilar objects.

Concerning the studies of metric learning in the literature, it can be seen that the studies are directly related to Mahalanobis distance metric. Let $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{d \times N}$ be the training samples, where $x_i \in \mathbb{R}^d$ is ith training example and N is the total number of training samples. The distance between x_i and x_j is calculated as:

$$d_M(\mathrm{x_i}, \mathrm{x_j}) = \sqrt{(\mathrm{x_i} - \mathrm{x_j})^{\mathrm{T}} \mathrm{M} (\mathrm{x_i} - \mathrm{x_j})} \tag{1}$$

$d_M(\mathrm{x_i}, \mathrm{x_j})$ is a distance metric, it must have the properties of nonnegativity, the identity of indiscernibles, symmetry, and the triangle inequality. M needs to be symmetric and positive semidefinite. All of the eigenvalues or determinants of M must be positive or zero to be positive semidefinite. When we decompose M, as follows:

$$M = W^T W \tag{2}$$

$$\begin{aligned} d_M(\mathrm{x_i}, \mathrm{x_j}) &= \sqrt{(\mathrm{x_i} - \mathrm{x_j})^{\mathrm{T}} \mathrm{M} (\mathrm{x_i} - \mathrm{x_j})} \\ &= \sqrt{(\mathrm{x_i} - \mathrm{x_j})^{\mathrm{T}} \mathrm{W}^T \mathrm{W} (\mathrm{x_i} - \mathrm{x_j})} \\ &= \|Wx_i - Wx_j\|_2 \end{aligned} \tag{3}$$

As can be seen from Equation (3), W has a linear transformation property. Thanks to this property, Euclidean distance in the transformed space is equal to Mahalanobis distance in original space for two samples. This linear transformation shows us the reality in the infrastructure of metric learning.

Obtaining a better representation capability for data will certainly enable us to make more accurate predictions possible in classification or clustering problems [11,12]. Metric learning aims to learn a good distance metric from data. The distance metric provides a new data representation that has more meaningful and powerful discrimination using similarity relationship between samples. When we discuss metric learning, it is useful to mention a linear transformation at first. Linear metric learning approaches provide more flexible constraints in the transformed data space and improves learning performance. These approaches have some advantages, such as convex formulations and robustness to overfitting [16]. Although linear approaches help us to learn a good metric, it is possible to gain better representation capabilities over the data. To better interpret the data, it is necessary to act in accordance with its nature. The linear transformation has a limited ability to achieve optimum performance over the new representation of data, because they have poor performance to capture nonlinear feature structure. Higher performance is aimed to be achieved by carrying the problem to a non-linear space through kernel methods in metric learning in order to overcome this problem [6]. Although these nonlinear approaches are practical to solve non-linear problems, they may have a negative effect against overfitting. In recent years, with the interest in deep metric learning, it is possible to propose a more compact solution to overcome such problems that exist in both approaches.

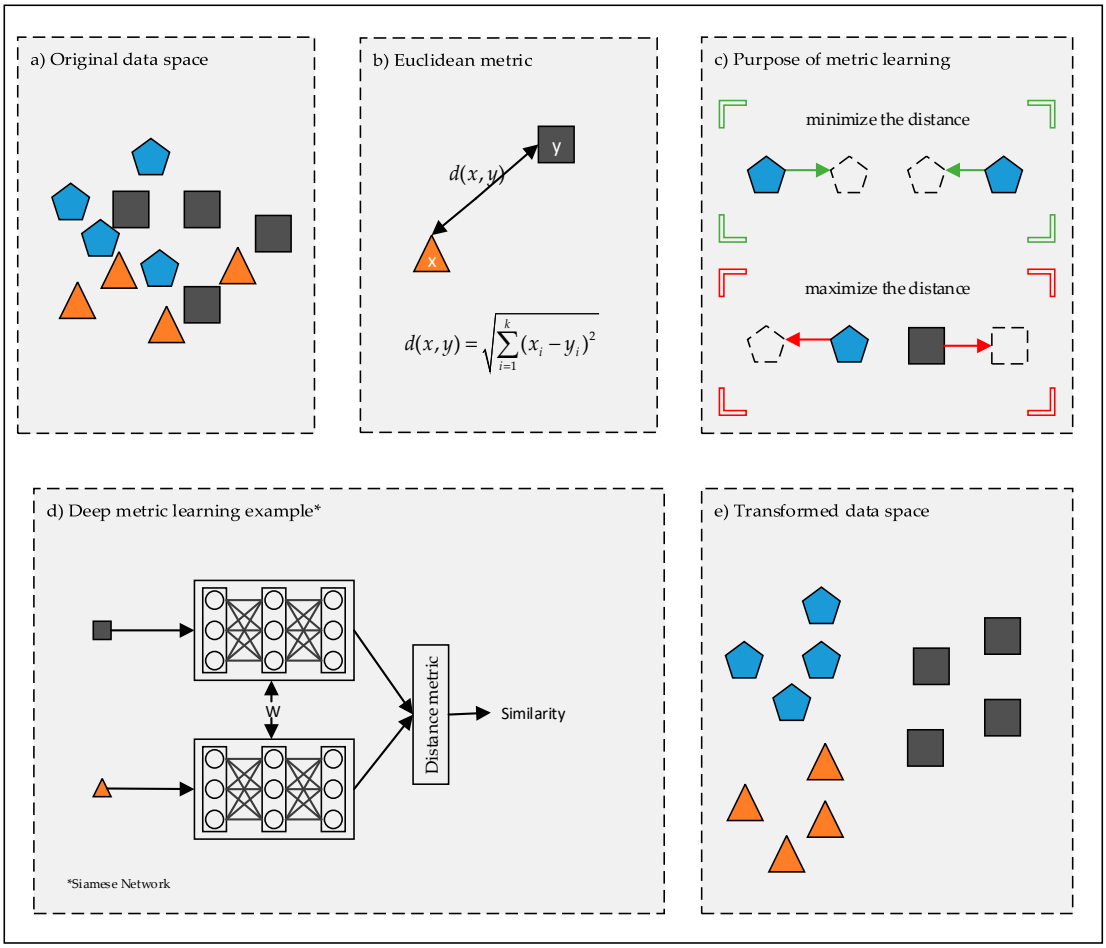


Figure 1. Deep Metric Learning.

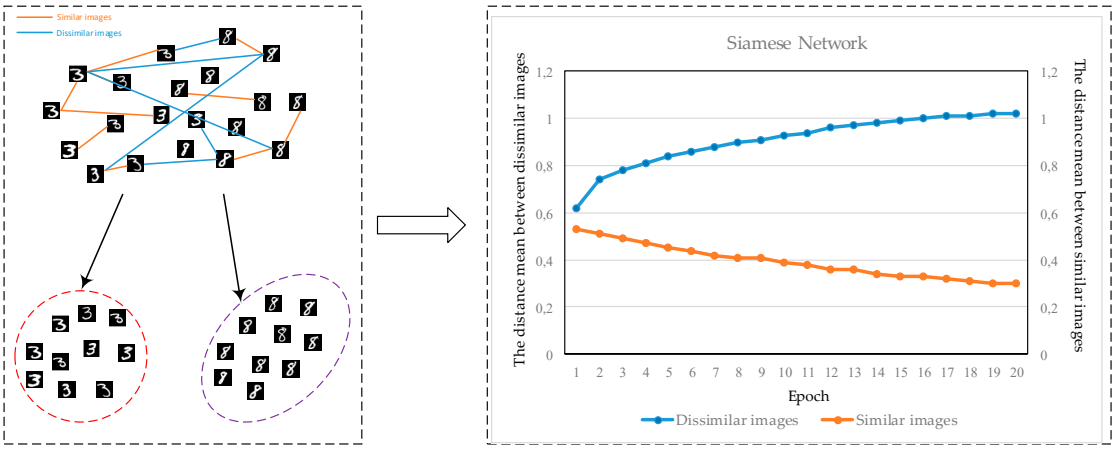
3. Deep Metric Learning

Traditional machine learning techniques are limited by their ability to process data on raw data. Therefore, they need feature engineering, such as preprocessing and feature extraction steps before classification or clustering tasks. All of these steps require expertise and they are not directly within the classification structure. However, deep learning learns the higher level of data directly in the classification structure. This perspective shows the fundamental difference between traditional machine learning methods and deep learning. Unlike traditional machine learning methods, deep learning needs a high size of data to achieve successful results, since it is not successful enough in low data size. Besides, deep learning algorithms require a lot of time to train data because of the high data size and a large number of parameters of the algorithm. Therefore, NVIDIA introduced a cuDNN GPU-accelerated library for deep neural networks to perform these high-performance calculations. Thanks to this library, a lot of deep learning frameworks, such as Caffe, Caffe2, Chainer, Microsoft CNTK, Matlab, Mxnet, PaddlePaddle, PyTorch, TensorFlow, and Theano [17], have been developed while using the power of GPU.

Basic similarity metrics that are used for data classification are the distances of Euclidean, Mahalanobis, Matusita [18], Bhattacharyya [19], and Kullback-Leibler [20]. However, these pre-defined metrics have limited capabilities in data classification. Hence, an approach based on the Mahalanobis metric was proposed to classify the data into traditional metric learning to address this problem. In this approach, the data is transformed into a new feature space with higher discrimination power. Usually, metric learning approaches are related to the linear transformation of the data without any kernel function. However, these approaches are not successful enough to reveal nonlinear knowledge of the

data [21]. For this reason, the expected outcomes could not be obtained while using metric learning. Although a solution with a kernel-based approach was provided to overcome this problem, there is no obvious success due to some issues such as scaling [22]. Unlike traditional metric learning methods, deep learning solves this problem using activation functions that have nonlinear structure.

Most of the existing deep learning approaches are based on the deep architectural background rather than the distance metric in a new representation space of the data. However, distance-based approaches have recently become one of the most interesting topics in deep learning [15,23–26]. While decreasing the distance between dissimilar samples [27,28], deep metric learning, which aims to increase the distance between similar samples, is directly related to the distance between samples. To execute this process, the metric loss function has been benefited in deep learning. While aiming to bring the samples from the same classes closer to each other, we pushed the samples from different classes apart from each other (Figure 2a). To illustrate this process with a figure, some experiments on the MNIST image dataset were conducted while using contrastive loss [29]. Distance values represent the mean of distances among similar or dissimilar images in Figure 2b. As can be seen in the Figure, the distance value for similar images decreased step by step after each epoch. On the other hand, the distance value for dissimilar images also increased at the same. The distance relationship for Siamese network has been successfully applied in each epoch for similar or dissimilar images (Figure 2b). This experiment proves to us that the purpose of the approach can be successfully implemented.



(**a**) (**b**)

Figure 2. Distance Relationship for a Siamese Network (**a**) Desired handwritten data discrimination for three and eight digits (**b**) after Siamese network applied to MNIST data for three and eight digits.

3.1. Deep Metric Learning Problems

Deep metric learning, which utilizes deep architectures by obtaining embedded feature similarity through nonlinear subspace learning, develops problem-based solutions that are caused by learning from raw data. When the scope of deep metric learning is considered, it has a wide range from video understanding to others including person re-identification, medical problems, three-dimensional (3D) modelling [23,30], face verification and recognition [27,31,32], and signature verification [33].

There are many kinds of problems in understanding videos, including video annotation, recommendation, and search. When encountered with such problems, it is possible to make use of a metric space to come up with solutions. To illustrate, Lee et al. extracted audio and visual features from videos to benefit from these useful contents at first [34]. After feature extraction, they presented a deep neural network embedding model that is based on triplet learning, which is also a source of inspiration for similar studies. The purpose of [35] is to learn a metric that is based on deep metric learning for moving human localization in video surveillance where the authors conducted a deep multi-channel

residual networks-based metric learning for this task. When the method was compared with popular deep metric learning methods, it was found to outperform the others. Predefined distance metrics may be insufficient for visual tasks because of significant variations on visual objects. To address this problem, Hu et al. [36] also used deep metric learning based on a distance metric instead of employing a predefined similarity metric to maximize the distance between the positive samples and minimize the distance between negative samples for visual tracking. The success of these studies indicates the advantages of working in metric space.

Another important problem in machine learning is person re-identification. Convolutional neural network-based approaches have come into question with the success of deep learning methods in recent years [37]. Person re-identification tasks aim to identify different images of the same person taken in different situations. Thus, it is possible to learn a suitable distance metric to solve these issues [38,39]. There are some kinds of benchmark person re-identification datasets, such as CUHK01, CUHK03, Market-1501, MARS, and VIPER, in the literature. Deep metric learning for person re-identification provides us to use end-to-end learning between the input image and the transformed feature space [40]. In [41], the authors obtained a score value for body similarity of the two images based on metric similarity. The model that is based on this approach begins with two layers of tied convolution and maximum pooling. Subsequently, cross-input neighborhood differences are computed. In the last step, patch summation attributes, cross-patch attributes, and softmax function are utilized to identify a person as the same or different. In another study, Ding et al. [42] aimed to maximize the relative distance between two dissimilar images for triplet loss. However, one image could be included in several triplet units, which can increase the number of triplet units. For this reason, they improved the effective triplet generation scheme and optimized gradient descent algorithm, depending on the number of original images instead of the number of triplets.

Disease patterns should be similar in medical images to make a diagnosis of a patient. Recently, deep learning has rapidly become a trending topic to solve the problems of medical images. There are some medical image problems, such as classification, detection, segmentation, and registration [43]. Deep metric learning algorithms that are based on the similarity approach can help to solve these problems. Using a deep metric, a higher representation level of data could be provided for the analysis of medical images. Unlike Triplet Network, the authors in [44] take the global structure of the embedding space and overlapping labels into account while using ML2 loss. The proposed approach allows us to distinguish normal radiological images from abnormal ones through the metric loss function.

Currently, studies have tended towards deep metric learning, which has an efficient discrimination power for 3D shape retrieval [23,30,45–48]. Both sketch and 3D shape of images benefited for 3D shape retrieval while using shared weights and a metric loss function in [23,45,47]. A network model that was based on CNN+Siamese network on three large datasets aimed to achieve efficient 3D image retrieval. To carry out 3D image retrieval, a metric loss that combines correlation and discrimination loss was used in [23,46]. Different from [23], the metric loss was also used in the hidden layer during training. The authors in [48] attempted to find a more robust and discriminative feature that was embedded with a novel loss function that combined triplet loss and center loss for 3D image retrieval task. In another 3D image retrieval task, Lim et al. [30] used the triplet network model to detect the styles of 3D images. They compared the triplet loss value with the distances of similar and dissimilar images.

Deep metric learning introduces the-state-of-the-art methods that have very discriminative information for face recognition and verification in recent years [27,31,32,49,50]. Hu et al. [31] proposed a novel deep discriminative metric learning model that had a hierarchical nonlinear transformation with face pairs taking advantage of neural network for face verification. In addition, the same authors aimed to reveal the kinship relation between people by examining the image of two different faces while using model-based discriminative deep metric learning method [49]. The authors in [32] suggested a system, called FaceNet, using a novel online triplet learning model by focusing on face similarity under the cover of Euclidean space. The system was designed to deal with face problems, such as

verification, face recognition, and face clustering. There are also studies regarding face tasks, such as facial expression recognition [51], and facial age estimation [52], with deep metric learning.

Though studies on deep learning are generally conducted in the field of computer vision, there are considerable studies on text understanding and information retrieval in the literature. To give an example, Mueller and Thyagarajan used the Siamese network with LSTM architecture to identify the semantic similarities between the sentences [53]. Although Benajiba et al. [54] employed a similar network model to learn the similarity of semantic patterns, they utilized a regression function that is the mean squared error from the SQL structure distance to train the network model unlike [53]. The authors in [55] proposed a dependency-based Siamese LSTM network model, where they used the main and supporting components in sentence representation to create a difference for learning sentence representation. The authors in [56] aimed to learn thematic similarity between the sentences. First of all, they generated weakly-supervised triplet sentences from Wikipedia articles. Afterwards, they used the Triplet network to cluster the Wikipedia sentences with high-quality sentence embeddings.

Another field where deep metric learning has achieved successful results is the processing of audio signals [50]. The authors in [57] exploited Triplet and Quadruple networks for speaker diarization. They utilized different sampling strategies and margin parameter selection to observe their effect on diarization performance. According to the authors, although semi-hard negative mining usually obtains successful results on computer vision applications, this strategy only has successful results with fixed parameters and triplet loss for speaker diarization. Wang et al. [58] carried out some experiments while using prototypical network loss [59] and triplet loss for speaker verification and speaker identification tasks. The authors obtained very successful results for speaker recognition in two different data sets. According to the authors, the results of the prototypical network loss were better than those of the triplet loss and they had a faster training time.

Although there are studies categorized above in different disciplines on deep metric learning, it is possible to find more studies carried out by researchers in other disciplines where some issues regarding music similarity [60], crowdedness regression [61], similar region search [62], volumetric image recognition [63], instance segmentation [64], edge detection [65], pan-sharpening [66], and so on were examined. Therefore, deep metric learning can be claimed to offer invaluable contributions to the literature, due to its high performance in different fields. Looking at the number of academic publications on "deep metric learning", as presented in Figure 3, it is understood that the topic has received growing attention.

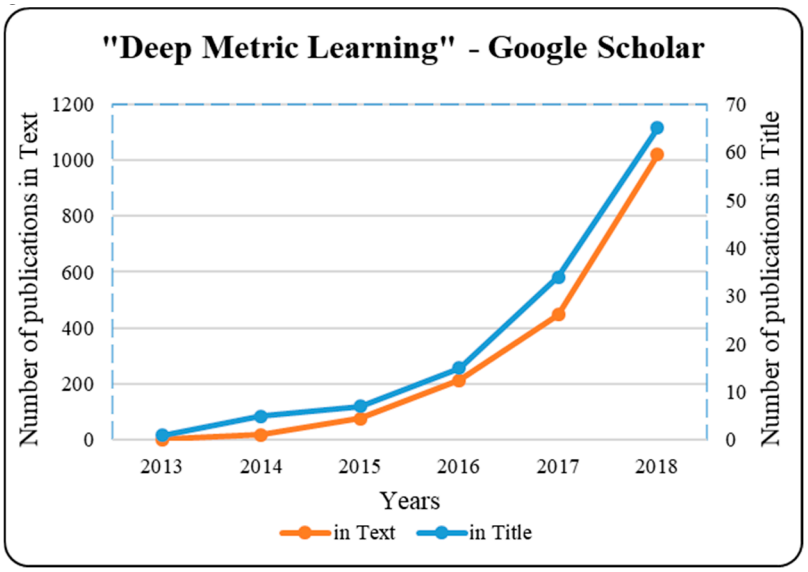


Figure 3. The number of academic publications on deep metric learning.

Deep metric learning provides us with very effective results on a wide variety of topics. Table 1 was created to demonstrate the studies published in the top journals/conferences in recent years and a similar evaluation protocol for the benchmark datasets were used for different topics. The comparative results presented in Table 1 indicate that deep metric learning has attained very successful results in many different disciplines. As can be seen in Table 1, there are specific evaluation metrics for each problem in the literature. For example, F1, which takes both false positive and false negative into account, and Normalized Mutual Information (NMI), which determines the quality of clustering using the entropy of class labels and the entropy of cluster labels, have been used for image clustering tasks; on the other hand, recall @R (rank accuracy) performance metric has been used for image retrieval tasks [67]. While the rank values are evaluated to measure the performance of person re-Identification tasks, accuracy is utilized to measure face verification. To measure 3D shape retrieval performances, First Tier (FT), Nearest Neighbor (NN), which is the percentage of the first-closest matches belonging to the query class, Emeasure (E), Second Tier (ST), Discounted Cumulated Gain (DCG), and mean Average Precision (mAP), which calculate the mean of average precision for each query, are preferred [48]. FT is the ratio of the relevant matches to the size of the query class C when the number of retrieved models is |C| for top K matches. If K=2|C|, ST is obtained. E-measure is equal to the 1-F1 measure. Semantic textual similarity task uses Pearson correlation (r), which is a measure of the linear correlation, Spearman's (ρ), which assesses monotonic relationships of two variables, and mean squared error (MSE) to evaluate the performance of the task [53]. Equal Error Rate (EER), which shows that the rate of false acceptances is equal to the rate of false rejections, and Minimum Decision Cost Function (MDCF), which is a simultaneous measure of discrimination and calibration are used to measure the performance of the speaker verification task [50].

Table 1. Comparison of benchmark datasets for deep metric learning problems.

Dataset	Reference	Task and Comparative Results						Evaluation Protocol	CNN LSTM	Year
		Image Clustering (%)		*Image Retrieval*						
				Recall@R (%)						
		NMI	F1	R=1	R=2	R=4	R=8			
CUB-200-2011 [68]	Song et al. [69]	56.2	22.7	46.5	58.1	69.8	80.2	200 classes with 11788 images	CNN	2016
	Sohn et al. [70]	60.3	27.2	50.9	63.3	74.2	83.2	the first 100 classes for training	CNN	2016
	Wang et al. [67]	**61.1**	**29.4**	54.7	66.3	76.0	83.9	(5864 images)	CNN	2017
	Song et al. [71]	59.2	-	48.1	61.4	71.8	81.9	the rest of the classes for testing	CNN	2017
	Ge et al. [72]	-	-	**57.1**	**68.8**	**78.7**	**86.5**	(5,924 images)	CNN	2018
		NMI	F1	R=1	R=2	R=4	R=8			
CAR-196 [73]	Song et al. [69]	55.1	21.5	48.3	61.1	71.8	81.1	198 classes with 16,185 images	CNN	2016
	Sohn et al. [70]	**63.9**	**33.5**	71.1	79.7	86.4	91.6	the first 98 classes for training	CNN	2016
	Wang et al. [67]	63.2	32.2	71.4	81.4	87.5	92.1	(8,054 images)	CNN	2017
	Song et al. [71]	59.0	-	58.1	70.6	80.2	87.8	the other 98 classes for testing	CNN	2017
	Ge et al. [72]	-	-	**81.4**	**88.0**	**92.7**	**95.7**	(8,131 images)	CNN	2018
		NMI	F1	R=1	R=10	R=100	R=1000			
Online Products [69]	Song et al. [69]	87.4	24.7	63.0	80.5	91.7	97.5	22634 products with 120053 images. the first	CNN	2016
	Sohn et al. [70]	88.1	28.1	67.7	83.7	92.9	97.8	11318 product categories for training	CNN	2016
	Wang et al. [67]	88.6	**29.9**	70.9	85.0	93.5	98.0	(59,551 images)	CNN	2017
	Song et al. [71]	**89.4**	-	67.0	83.6	93.2	-	the other 11316 product categories for testing	CNN	2017
	Ge et al. [72]	-	-	**74.8**	**88.3**	**94.8**	**98.4**	(60,502 images)	CNN	2018
		Person re-Identification								
		R=1				R=5				
Market-1501 [74]	Ustinova et al. [75]	59.47				80.73		1501 identities in total 750 identities for training and 751 identities for test	CNN	2016
	Chen et al. [38]	83.55				92.37		-	CNN	2018
	Yang et al. [39]	84.26				**93.59**		1501 identities in total 750 identities for training and 751 identities for test	CNN	2019
	Yao et al. [76]	**88.20**				-		1501 identities in total 750 identities for training and 751 identities for test	CNN	2019

Table 1. *Cont.*

Dataset	Reference	Task and Comparative Results						Evaluation Protocol	CNN LSTM	Year
		R=1				R=5				
CUHK03 [77]	Ustinova et al. [75]	65.77				92.85		1360 identities in total 1160 identities for training and 100 for test	CNN	2016
	Chen et al. [38]	68.63				92.28		-	CNN	2018
	Yang et al. [39]	39.64				-		1367 identities in total 767 identities for training and 700 for test	CNN	2019
	Yao et al. [76]	**82.75**				**96.59**		1360 identities in total 1160 identities for training and 100 for test	CNN	2019
		3D Shape Retrieval								
		NN	FT	ST	E	DCG	Map			
SHREC'13 [78]	Dai et al. [23]	65.0	63.4	71.9	34.8	76.6	67.4	1258 shapes and 7200 sketches, grouped into 90 classes.	CNN	2017
	Dai et al. [46]	73.0	71.5	77.3	36.8	81.6	74.4	the number of sketches for each class is equal to 80.	CNN	2018
	He et al. [48]	**76.3**	**78.7**	**84.9**	**39.2**	**85.4**	**80.7**	50 sketches for training and 30 for testing for each group.	CNN	2018
		NN	FT	ST	E	DCG	Map			
SHREC'14 [79]	Dai et al. [23]	27.2	27.5	34.5	17.1	49.8	28.6	13680 sketches and 8987 3D models, grouped into 171 classes.	CNN	2017
	Dai et al. [46]	40.3	32.9	39.4	20.1	54.4	33.6	the number of sketches for each class is equal to 80.	CNN	2018
	He et al. [48]	**58.5**	**45.5**	**53.9**	**27.5**	**66.6**	**47.7**	50 sketches for training and 30 for testing for each group.	CNN	2018
		Face verification								
		Accuracy								
LFW [80]	Hue et al. [31]	90.68						***10 folds:*** each fold has 300 matched pairs and 300 mismatched pairs	-	2014
	Lu et al. [49]	**94.50**							-	2017
	Hue at al. [81]	93.27						***Image restricted***	-	2018

Table 1. *Cont.*

Dataset	Reference	Task and Comparative Results			Evaluation Protocol	CNN LSTM	Year
		Accuracy					
YTF [82]	Hue et al. [31]	82.34			***10 folds:*** each fold has 250 intra-personal pairs and 250 inter-personal pairs	-	2014
	Lu et al. [49]	**82.50**			***Image restricted***	-	2017
		Semantic Textual Similarity					
		r	ρ	MSE			
SICK [83]	Mueller et al. [53]	**0.88**	**0.83**	**0.22**	9927 sentence pairs 5000 for training and 4927 for testing	LSTM	2016
	Zhu et al. [55]	0.83	0.77	0.34		LSTM	2018
	Ein-Dor et al. [56]	0.81	0.72	0.33		LSTM	2018
		Speaker Verification					
		EER (%)		MDCF			
NIST i-vector [84]	Triplet Network [85]	2.85		0.30	1306 speakers recorded with 5 i-vectors each. Total 9634 test i-vector and 12582004 trial. randomly divided train subset and test subset. All i-vectors have 600 dimensions	-	2015
	Chen et al. [50]	**2.69**		**0.27**		-	2019
		EER (%)		MDCF			
VCTK [86]	Triplet Network [85]	12.26		-	the first 90 speakers were divided into training, validation and test sets. 18 speakers were used as an "unseen" set	LSTM	2015
	Wang et al. [58]	**10.77**		-		LSTM	2019
		EER (%)		MDCF			
VoxCeleb2 [87]	Triplet Network [85]	15.92		-	selected a subset containing 101 speakers. 71 speakers for training and validation other 30 speakers are used as the "unseen" set	LSTM	2015
	Wang et al. [58]	**13.68**		-		LSTM	2019

3.2. Sample Selection

Deep metric learning consists of three main parts, which are informative input samples, the structure of the network model, and a metric loss function. Although deep metric learning especially deals with metric loss function, informative sample selection also plays a very important role in classification or clustering. Informative samples are one of the most substantial elements that increase the success of deep metric learning. The sampling strategy is capable of increasing both the success of the network and the training speed of the network. The easiest way to determine train samples in contrastive loss is by means of randomly chosen positive or negative pairs of objects. In the beginning, some papers tend to use easy sample pairs for the Siamese network in embedding learning [29,88]. However, the authors in [89] emphasized that the learning process could be slowed down and negatively affected after the network reached an acceptable performance level. To address this problem, more discriminative models were obtained while using hard negative mining [89,90]. Triplet network uses an anchor, a positive, and a negative sample to train a network for classification. In [91], it was observed that some easy triplets had no effect in updating a model due to their poor discriminative power. These triplets cause a waste of time and resources. For this reason, to overcome these problems, it is very convenient to use informative sample triplets, and more feasible train models with a better sample strategy could be provided instead of selecting random samples [91,92].

Hard negative samples correspond to false-positive samples that are determined by training data. Semi-hard negative mining used for the first time in [32] aims to find negative samples within the margin. Negative samples are farther from the anchor sample when compared with hard negative mining. There is also a softer transition between positive and negative samples in this approach. The negative mining relationship according to the distance among anchor, positive, and negative samples was illustrated for triplet mining [85], as seen in Figure 4. If the negative samples are too close to the anchor, we can see that the gradient has a high variance and a low signal to the noise ratio, according to [93]. Hence, distance weighted sampling was suggested to avoid noisy samples in [93]. Thanks to this method, a wider range of examples as compared with semi-hard negative mining was also offered. Negative class mining can also be found in the literature instead of negative sample mining [70]. This approach uses one of each class samples for a negative sample of the triple network. To achieve this, the authors chose multiple negative samples with a greedy search strategy.

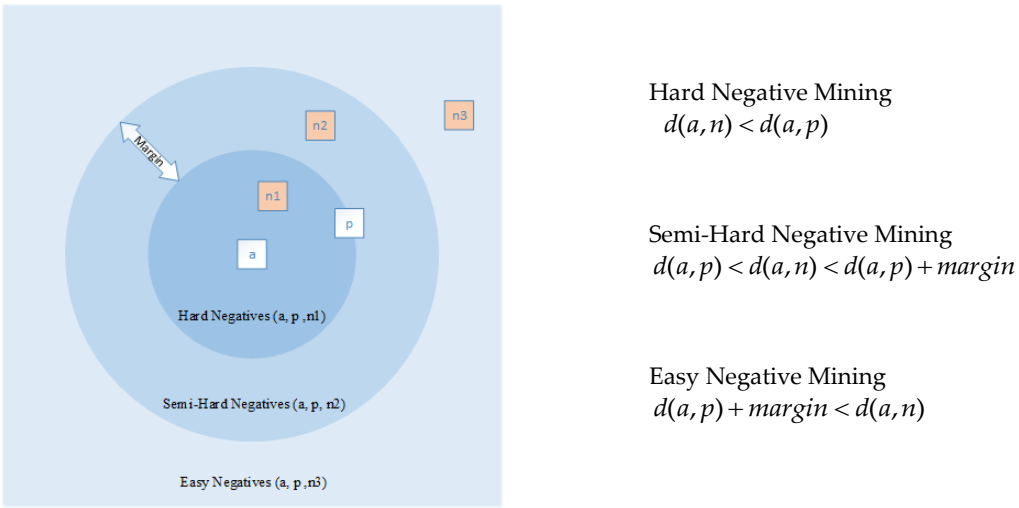


Figure 4. Negative Mining.

To summarize, even if we create good mathematical models and architectures, the learning ability of the network will be limited, depending on the discriminating power of the samples that are presented to the network. Distinguishing training examples should be presented to the network so that the network can learn better and gain better representation. For this reason, the effect of the relationship

between the samples for deep metric learning should be carefully examined. Thus, sample selection will be very useful as a preprocessing step to increase the success of network model before applying the deep metric learning model. The results in the literature point out that studies for negative mining in deep metric learning have a high impact value. When considering the benefit of choosing informative samples; on the other hand, the main one could be to avoid overfitting, because similar patterns have similar interaction when the network is being trained, because overfitting may cause slower learning or local optimal learning. Moreover, the number of all possible triplets corresponds O (n^3) time complexity when we consider a problem with two classes, which results in a waste of time and unnecessary use of resources. To overcome this problem, it is enough to only deal with valuable triple samples. Thus, significant improvements in performance can be achieved after selecting informative samples.

3.3. Loss functions for Deep Metric Learning

In this section, some loss functions that have been used to apply deep metric learning in the literature will be highlighted. The way that these functions are used will also be introduced and some details about their differences will be given. These functions provide us to increase or decrease the distance between the objects by looking at their similarity. The goal is to achieve the highest feature representation between different objects.

Initially, the Siamese network was used with neural networks for signature verification [33]. Different from [33], the Siamese network is based on learning from a discriminative learning framework for energy-based models [94]. In this approach, two identical images are taken into the Siamese network and a binary value is obtained as a result of learning from these images. The images are considered as the same class when they are "0"; if they are "1", they are considered as a different class. The Siamese network, as a metric learning approach, receives pair images, including positive and negative samples to train a network model (Figure 5) [95]. The distance between these pair images is calculated via a loss function (Equation (5)). Contrastive Loss has been benefited for the Siamese network in the literature [29,96]. This approach is a study that provides inspiration for researchers working in the field of deep metric learning. As can be seen in Figure 6a, the siamese network is a very successful model to maximize or minimize the distance between objects to improve classification performance. Shared weights that positively affect the performance of a neural network are used to obtain a meaningful pattern among images in deep metric learning, as shown in Figure 5. These weights also have important advantages in terms of time and memory. It is also possible to combine the Siamese network and Convolutional neural network, which has important advantages [97], which include similarity learning from direct image pixels, color and texture information at the same time, and its flexible structure. In the deep metric learning model [98], two Siamese Convolutional neural network and Mahalanobis metric were combined for person re-identification, where the Mahalanobis metric was used to classify the data. The deep metric learning model was created by limiting the weights to a C constant while calculating the L2 norm. The authors in [99] combined softmax loss and center loss for face recognition. While the center loss aims to find a center for deep features of each class to minimize the distances between deep features and their class center, like the contrastive loss, the softmax loss aims to find deep features that maximize the distances between different classes

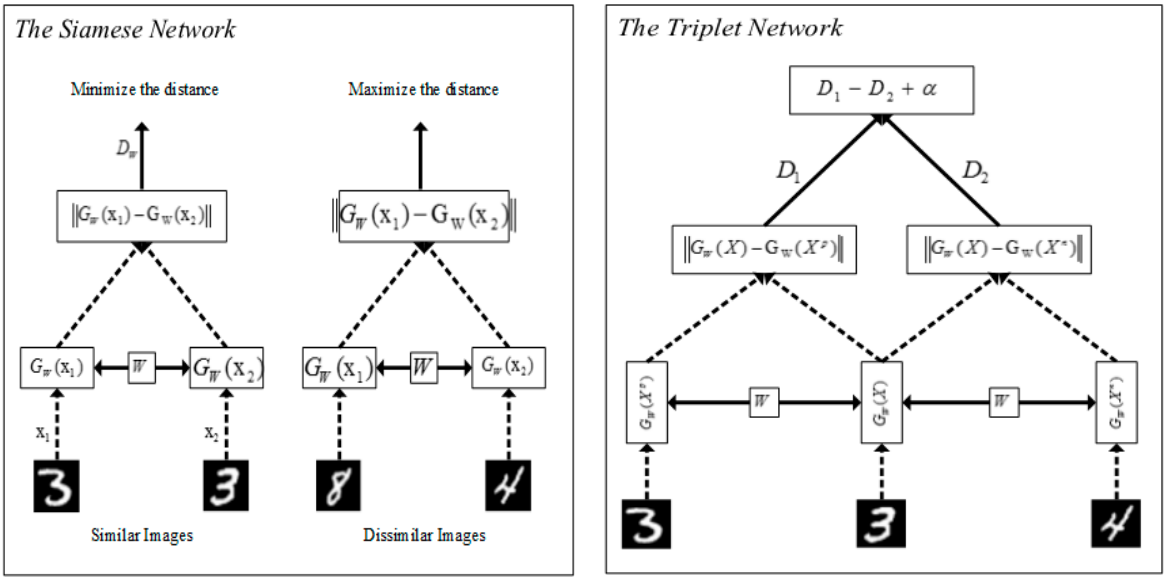


Figure 5. The Siamese network and Triplet network.

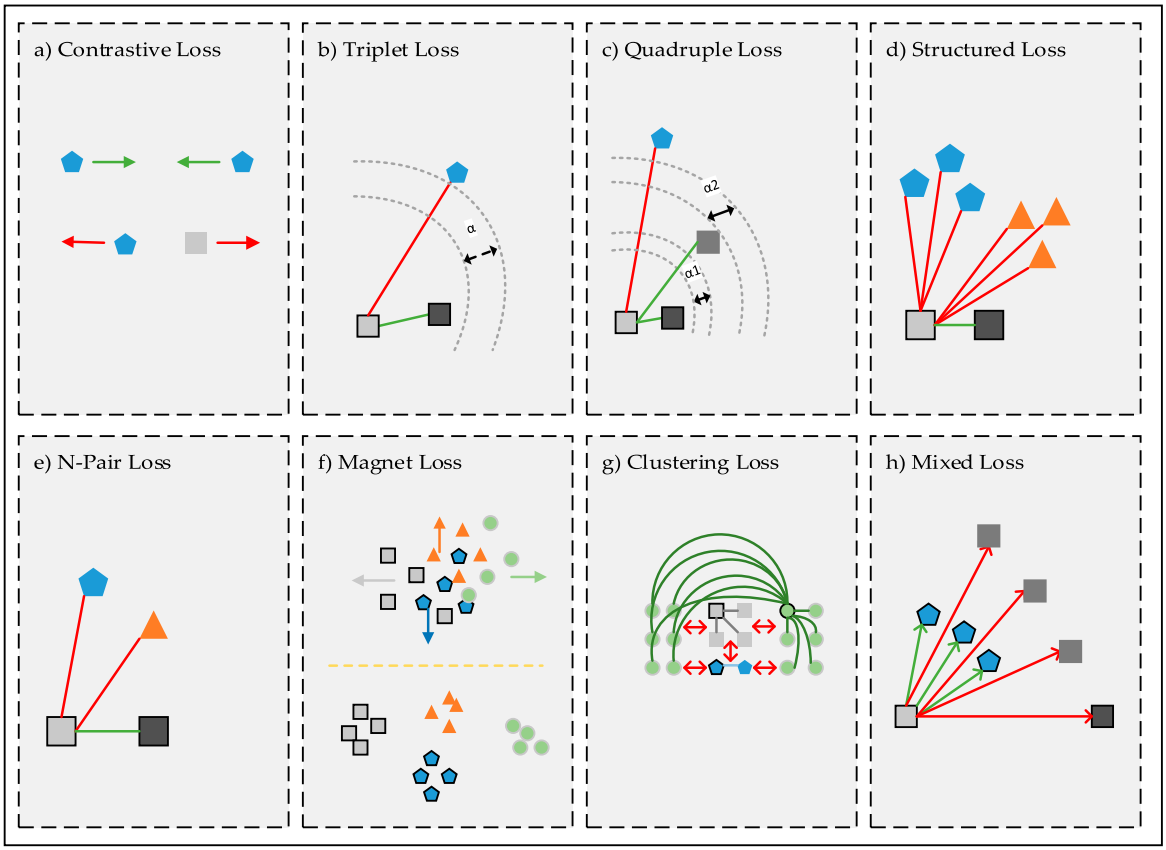


Figure 6. Metric loss functions.

Triplet network inspired by Siamese network contains three objects, which are formed positive, negative, and anchor samples [85]. Triplet networks utilize Euclidean space to compare the objects in the pattern recognition process, and this approach is directly related to metric learning. As can be seen in Equation (6), triplet loss first focuses on the similarity between the pair samples of the same and different classes using shared weights. The classification is carried out comparing the similarity of pair samples (Figure 6b). Triplet networks provide a higher discrimination power while using both

in-class and inter-class relations. The authors in [72] proposed a new metric loss that is based on triplet loss utilizing a class-level hierarchical tree. Hierarchical triplet loss improves the performance of the network with online adaptive tree update thanks to the more informative sample approach. Similarly, Wang et al. [67] suggested a novel angular loss to improve deep metric learning. Unlike Siamese and Triplet networks, angular loss focuses on angular constraint at the negative point of triplet triangles. The angular loss pushes the negative point away from the center of the positive cluster and then brings positive points closer to each other while using an angle that is a rotation and scale-invariant metric (Equation (8)). While the Triplet loss only considers positive and negative samples to compute the distance between samples, it does not exploit any similarity degree information [100]. The authors in [100] obtained a better degree of closeness between objects while using quadruple samples in each training batch. As shown in Equation (7), a new input sample similar to the sample X was added to the quadruple loss metric as an alternative approach to triplet loss. Figure 6c shows us that quadruple loss benefits from two different margin values. These margin values aim to establish more meaningful relationships among similar samples and among different samples. Histogram loss [75] uses quadruplet training samples, like quadruplet loss, for training. It benefits from histograms to calculate similarity distributions of positive and negative pairs and does not need any tuning parameters, unlike other losses. It obtains the outperforming results in experimental studies for person re-identification datasets, like CUHK03 and Market-1501, when compared to other losses. Yao et al. [76] presented part loss for person re-identification task to take advantage of learning risk minimization constraint, as in SVM. Part loss aims to use different body parts instead of focusing only on a certain point. Part loss value was calculated for each part of an image divided into five parts to evaluate each of the local patterns as a separate part

Loss functions are one of the most important parts of the deep metric learning models. Now, let us consider some loss functions while using the new representation of paired samples in embedding space. Let X_1 and X_2 be a pair of inputs in the training set. Distance for a pair of input samples $D_W(X_1, X_2)$ is,

$$D_W(X_1, X_2) = \|G_W(X_1) - G_W(X_2)\|_2 \tag{4}$$

where $G_W(X_1)$ and $G_W(X_2)$ are generated as a new representation of a pair of input samples. D_W is used to calculate the distance between the two inputs in loss functions. $L_{Contrastive}$ that is used to calculate a loss function in Siamese network model is,

$$L_{Contrastive} = (1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{\max(0, m - D_W)\}^2 \tag{5}$$

where Y is the label value. If a pair of inputs is from the same class, the value of Y is 1, otherwise its value is 0. m is margin value in $L_{Contrastive}$. Triplet network models have three inputs: anchor input X, similar to anchor input X^p, and dissimilar to anchor input X^n. Triplet loss $L_{Triplet}$ is,

$$L_{Triplet} = \max(0, \|G_W(X) - G_W(X^p)\|_2 - \|G_W(X) - G_W(X^n)\|_2 + \alpha) \tag{6}$$

where α is the margin value. Quadruple network models also have another input X^s different from Triplet network models. X^s is similar to X input like X^p input. Quadruple loss $L_{Quadruple}$ is,

$$\begin{aligned} L_{Quadruple} = \max(0, \|G_W(X) - G_W(X^p)\|_2 - \|G_W(X) - G_W(X^s)\|_2 + \alpha_1) + \\ \max(0, \|G_W(X) - G_W(X^s)\|_2 - \|G_W(X) - G_W(X^n)\|_2 + \alpha_2) \end{aligned} \tag{7}$$

Angular loss takes angle relationship into account between samples. Angular loss $L_{Angular}$ is,

$$L_{Angular} = \max\left(0, \|G_W(X) - G_W(X^p)\|_2 - 4\tan^2\alpha\|G_W(X^n) - G_W(X^c)\|_2\right) \tag{8}$$

where X^c is in the middle of X and X^p. X^c is,

$$X^c = (X + X^p)/2 \tag{9}$$

Traditional deep metric learning models, like Siamese and Triplet Networks, neglect the structural information of training samples in each training step [101]. However, it is a disadvantage for a dataset that has limited training samples. Deep structural metric learning approach proposes lifting the vector of pairwise distances within the batch to the matrix of pairwise distances through a special structured loss in the deep network [69]. Therefore, it is possible to use the full advantage of contextual information within the training batch. As can be seen in Figure 6d, it deals with the similarity relationship among many samples at the same time. The power of the proposed algorithm comes from evaluating a sufficient number of samples together. Sohn [70] proposed a multi-class N-pair loss to address the slow convergence and poor local optima of Siamese and Triplet Networks. Triplet loss deals with just a negative sample in every training stage update and it does not interact with other negative samples. Different from Triplet loss, N pair loss benefits from N-1 negative class samples to compare the anchor sample (Figure 6e). If the number of classes is limited, it can be unnecessary to use N-pair loss, because, if the number of classes decreases, the number of comparisons with an anchor sample will also decrease. The advantage of N-pair loss will be lost in this situation. Unlike all other loss metrics in the literature, the authors in [102] benefited from multi-similarity loss while using self-similarity, negative relative similarity, and positive relative similarity under the general pair weighting framework to capture the similarity between samples. This loss takes both self-similarity and relative similarities into account, which make possible the model to gather and weight informative pair samples more efficiently.

The preparation of the training data has to be done for Siamese, Triplet, and n-double networks before the training phase. This process needs more space on disk and it is time-consuming. Song et al. [71] claim that these networks also deal with the local view of the data. To overcome these problems, they proposed a new deep metric learning method while using clustering loss. As can be seen in Figure 6g, clustering loss aims to bring the samples together in a cluster while using cluster centers. It also prevents different clusters from approaching each other. The suggested method, based on structured prediction, benefits from the normalized mutual metric. Rippel et al. [103] emphasize that triplet loss evaluates a triplet sample at a time to train a dataset. This reduces the learning time of the network. However, this approach causes poor performance and insufficient training. For this reason, they suggested magnet loss, which penalizes cluster overlaps and evaluates the closest neighbors in a cluster to separate multiple clusters. The proposed method has local discrimination and it is globally consistent with the optimization procedure. Figure 6f illustrates how similar samples approach in the closest cluster while using local neighborhoods. Mixed loss [104], which is inspired by triplet loss, uses three positive samples and three negative samples to establish similarity relationship among samples, in addition to the anchor and negative sample. Figure 6h shows us that the metric loss function brings positive samples closer to anchor, while pushing negative samples away from the anchor. State-of-the-art loss metrics in the literature are summarized in Table 2 in details.

Table 2. Loss metrics.

Metric	Sample Selection	Topic	Dataset	Purpose	Year
Contrastive Loss [29]	Hard negative	Image recognition Object recognition	MNIST [105] NORB [106]	calculates a contrastive loss function that aims to obtain a higher value for pairs of dissimilar objects and aims to obtain a lower value for pairs of similar objects	2006
Triplet Loss [85]	Easy sampling	Image recognition Object recognition	MNIST [105] CIFAR10 [107] SVHN [108] STL10 [109]	calculates the distance difference between anchor-positive samples and anchor-negative samples and aims to bring similar objects closer	2014
Histogram Loss [75]	Easy sampling	Image recognition Image retrieval Person re-ID	CUB-200-2011 [68] Online Products [69] Market-1501 [74] CUHK03 [77]	aims the distributions of the similarities of less overlapping positive and negative pairs.	2016
Structured Loss [69]	Hard negative	Image retrieval	CUB-200-2011 [68] Online Products [69] CAR-196 [73]	aims a new metric learning algorithm using the lifted dense pairwise distance matrix within the batch throughout the training.	2016
N-Pair Loss [70]	Multiple negative "class"	Image retrieval Image clustering Face verification Face identification Object recognition Object verification	CUB-200-2011 [68] Online Products [69] Flower-610 [70] CAR-196 [73] LFW [80] Car-333 [110]	aims to develop triplet loss focusing on pushing a positive sample away from multiple negative samples at each training stage	2016
Magnet Loss [103]	Hard negative	Image recognition Image annotation	Stanford Dogs [111] Oxford-IIIT Pet [112] Oxford 102 Flowers [113] Object Attributes [114]	aims to retrieve a whole local neighborhood of nearest clusters and punish their overlaps	2016
Angular Loss [67]	Multiple negative	Image retrieval Image clustering	CUB-200-2011 [68] Online Products [69] CAR-196 [73]	focuses on limiting the angle in the negative sample of triplet triangles.	2017

Table 2. *Cont.*

Metric	Sample Selection	Topic	Dataset	Purpose	Year
Quadruple Loss [100]	Semi-hard negative	Patient similarity	The Ischemic Heart [100] The Cerebrovascular [100]	aims to capture the degree of similarity between patients effectively	2017
Clustering Loss [71]	Easy sampling	Image retrieval Image clustering	CUB-200-2011 [68] Online Products [69] CAR-196 [73]	aims a new metric learning approach based on the structural prediction that takes the global structure of the embedding space into account by a clustering quality metric.	2017
Hierarchical Triplet Loss [72]	Anchor-Neighbor sampling	Image retrieval Face recognition	CUB-200-2011 [68] Online Products [69] CAR-196 [73] LFW [80] In-Shop Clothes Retrieval [115]	aims to collect informative samples and capture global data context with an online class-level tree update	2018
Mixed Loss [104]	Hard-aware online exemplar mining	Image retrieval	Fashion Collocation Dataset [104]	aims to feed multiple positive and negative samples to the neural network per time	2018
Part Loss [76]	Easy sampling	Person re-ID	Market-1501 [74] CUHK03 [77] VIPeR [116]	aims to reduce empirical classification risks for training and representation learning risks for test by dividing images to K parts	2019
Multi-Similarity Loss [102]	General pair weighting	Image retrieval	CUB-200-2011 [68] Online Products [69] CAR-196 [73] In-Shop Clothes Retrieval [115]	aims to collect informative paired samples, and weights these pairs both their own and relative similarities	2019

4. Discussion

As mentioned earlier in Section 3, Deep metric learning (DML) have been used for face verification, recognition, person re-identification, and 3D shape retrieval tasks. These tasks consist of very large number of categories and limited training samples for single categories. An insufficient number of samples for any category can complicate a successful training process. However, DML algorithms can handle two, three, or more samples at the same network structure, such as Siamese network, Triplet network, and Quadruple network. These network structures allow for users to significantly increase the size of data training. Thus, it is possible to improve the training performance of the network even if the number of samples for a single category is small. As can be seen in Table 1, DML algorithms have proved to be useful with remarkable results for these tasks where both the number of categories is high and the number of samples for a single category is small.

DML, which basically consists of metric loss function, sampling strategy, and network structure, should be considered as a whole with all of the components of the network. Samples to be presented to the network and the relationship between them are related to the metric loss function. The metric loss functions such as contrastive loss [29], triplet loss [85], quadruple loss [100], n-pair loss [70], and so on allow for us to increase the data sample size (n), such as n^2 (paired samples), n^3 (triplet samples), and n^4 (quadruple samples). Inefficient paired samples or triple samples cause time consumption and too much memory space in the network training. The time complexity of network training may exponentially increase, depending on this situation. To overcome these problems, hard negative mining [89,90] and semi-hard negative mining [32,100] offers informative samples for training. The correct sampling strategy plays very important role for fast convergence [32].

Although hard mining or semi-hard mining strategies obtain the desired results in specific tasks, they still have very expensive time and memory space compared to traditional deep learning approaches [71]. Moreover, it is not always possible while using a large batch size due to the GPU memory limit. To get over these issues, clustering loss [71] offers a good metric function, which does not need any data preparation step. Although deep metric learning approaches are usually carried out on a GPU, the authors in [32] applied their mining strategy on CPU clusters to use a huge batch.

Deep metric learning approaches are highly dependent on data. The metric loss function might not provide fast convergence between the samples for some specific datasets. To overcome this problem, the weights that are obtained from pre-trained network models may ensure a fast convergence and more discriminative learning in embedding space [70].

5. Conclusions

In recent years, deep metric learning based on distance metric seems to be an important study field for researchers. At present, academic studies on this topic provide invaluable contributions to the literature. The goal of deep metric learning is to learn a similarity metric that computes the similarity or dissimilarity of two or more objects while using samples. There are notable similarity problems that are applied to face recognition, face verification, person re-Identification, 3D shape retrieval, semantic textual similarity, speaker verification, patient similarity, etc. in image, video, text and audio tasks. The aim of this article is to present a comprehensive study, in which all aspects of deep metric learning are evaluated to fill the gap in the literature.

Most of the recent studies in the literature were inspired by Siamese and Triplet networks in deep metric learning. These network structures have proved their high effectiveness on benchmark datasets and specific problem tasks. These studies contain three main parts, including informative input samples, the structure of the network model, and a metric loss function.

Firstly, If the data-dependent sampling strategy is not initially well-adjusted, the desired results may not be achieved for specific problem tasks. Additionally, poor sampling strategy might cause convergence to slow down. To address these problems, hard or semi-hard negative mining strategy provides substantial contributions to obtain informative input samples. This technique needs a margin value to separate more valuable input examples to be presented to a network. In future

studies, optimally determining the margin value and finding different sampling strategies will make a significant contribution to the success of a network. We think that one way to get informative samples is an important optimization problem. There is still an important gap in the literature to contribute to this problem. The layers, parameters, shared weights, and network model used in deep learning are another important part to be optimized. In this learning model, shared weights have an important role in obtaining a similarity relationship. Shared weights not only carry out the task of organizing in-class and inter-class relations, but they also provide more efficient memory use and less time use when training a network. The last main part is a metric loss function, which significantly determines the performance of the network. State-of-the-art studies that have been conducted in the relevant literature aim to present a distance metric, depending on the data to obtain higher discriminative features in embedding space. Although, there are notable efficient metric loss functions in the literature, a combination of these metrics or relationship between pair samples can be a considerable research topic for future studies.

There is a growing interest in deep metric learning in the literature, but studies are conducted in a limited number of areas. While this is an interesting point for researchers, there are many aspects of deep metric learning that are waiting to be explored, such as the shortcomings of existing approaches. In addition, deep metric learning studies in the literature have the flexibility of using local features, global features, and a combination of local and global features.

Author Contributions: The authors contributed equally in preparing and writing this work.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Samuel, A.L. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* **2000**, *44*, 206–226. [CrossRef]
2. Géron, A. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st ed.; O'Reilly Media: Sebastopol, CA, USA, 2017; pp. 1–9.
3. Malik, S.; Kanwal, N.; Asghar, M.N.; Sadiq, M.A.A.; Karamat, I.; Fleury, M. Data Driven Approach for Eye Disease Classification with Machine Learning. *Appl. Sci.* **2019**, *9*, 2789. [CrossRef]
4. Globerson, A.; Roweis, S. Metric learning by collapsing classes. In Proceedings of the 18th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 5–8 December 2005; pp. 451–458.
5. Wang, F.; Sun, J. Survey on distance metric learning and dimensionality reduction in data mining. *Data Min. Knowl. Discov.* **2015**, *29*, 534–564. [CrossRef]
6. Weinberger, K.Q.; Saul, L.K. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* **2009**, *10*, 207–244.
7. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
8. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
9. Tran, K.; Phan, T.; Tran, T.K.; Phan, T.T. Deep Learning Application to Ensemble Learning—The Simple, but Effective, Approach to Sentiment Classifying. *Appl. Sci.* **2019**, *9*, 2760. [CrossRef]
10. Lu, J.; Hu, J.; Zhou, J. Deep Metric Learning for Visual Understanding: An Overview of Recent Advances. *IEEE Signal Process. Mag.* **2017**, *34*, 76–84. [CrossRef]
11. Xing, E.P.; Jordan, M.I.; Russell, S.J.; Ng, A.Y. Distance metric learning with application to clustering with side-information. In Proceedings of the 15th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 9–14 December 2002; pp. 521–528.
12. Weinberger, K.Q.; Blitzer, J.; Saul, L.K. Distance metric learning for large margin nearest neighbor classification. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 10 December 2005; pp. 1473–1480.

13. Davis, J.V.; Kulis, B.; Jain, P.; Sra, S.; Dhillon, I.S. Information-theoretic metric learning. In Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, USA, 20–24 June 2007; pp. 209–216.
14. Nguyen, H.V.; Bai, L. Cosine similarity metric learning for face verification. In Proceedings of the 10th Asian Conference on Computer Vision, Queenstown, New Zealand, 8–12 November 2010; pp. 709–720.
15. Duan, Y.; Lu, J.; Feng, J.; Zhou, J. Deep Localized Metric Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 2644–2656. [CrossRef]
16. Bellet, A.; Habrard, A.; Sebban, M. A Survey on Metric Learning for Feature Vectors and Structured Data. *arXiv* **2014**, arXiv:1306.6709.
17. Nvidia Developer. Available online: https://developer.nvidia.com/deep-learning-frameworks (accessed on 16 April 2019).
18. Matusita, K. Decision Rules, Based on the Distance, for Problems of Fit, Two Samples, and Estimation. *Ann. Math. Stat.* **1955**, *26*, 631–640. [CrossRef]
19. Thacker, N.A.; Aherne, F.J.; Rockett, P.I. The Bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika* **1997**, *34*, 363–368.
20. Elgammal, A.; Duraiswami, R.; Davis, L.S. Probabilistic tracking in joint feature-spatial spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Madison, WI, USA, 16–22 June 2003; pp. 1–8.
21. Yu, J.; Yang, X.; Gao, F.; Tao, D. Deep multimodal distance metric learning using click constraints for image ranking. *IEEE Trans. Cybern.* **2016**, *47*, 4014–4024. [CrossRef] [PubMed]
22. Cai, X.; Wang, C.; Xiao, B.; Chen, X.; Zhou, J. Deep nonlinear metric learning with independent subspace analysis for face verification. In Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 29 October–2 November 2012; pp. 749–752.
23. Dai, G.; Xie, J.; Zhu, F.; Fang, Y. Deep Correlated Metric Learning for Sketch-based 3D Shape Retrieval. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4002–4008.
24. Li, Z.; Tang, J. Weakly Supervised Deep Metric Learning for Community-Contributed Image Retrieval. *IEEE Trans. Multimedia* **2015**, *17*, 1989–1999. [CrossRef]
25. Kumar, V.B.; Harwood, B.; Carneiro, G.; Reid, I.; Drummond, T. Smart Mining for Deep Metric Learning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2821–2829.
26. Gundogdu, E.; Solmaz, B.; Koç, A.; Yücesoy, V.; Alatan, A.A. Deep distance metric learning for maritime vessel identification. In Proceedings of the 25th Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey, 15–18 May 2017; pp. 1–4.
27. Liu, J.; Deng, Y.; Huang, C. Targeting Ultimate Accuracy: Face Recognition via Deep Embedding. *arXiv* **2015**, arXiv:1506.07310.
28. Hoffer, E.; Ailon, N. Semi-supervised deep learning by metric embedding. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 14–26 April 2017.
29. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality Reduction by Learning an Invariant Mapping. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, 17–22 June 2006; pp. 1735–1742.
30. Lim, I.; Gehre, A.; Kobbelt, L. Identifying Style of 3D Shapes using Deep Metric Learning. *Comput. Graph. Forum* **2016**, *35*, 207–215. [CrossRef]
31. Hu, J.; Lu, J.; Tan, Y.-P. Discriminative Deep Metric Learning for Face Verification in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1875–1882.
32. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823.
33. Bromley, J.; Bentz, J.W.; Bottou, L.; Guyon, I.; LeCun, Y.; Moore, C.; Sackinger, E.; Shah, R. Signature verification using a "siamese" time delay neural network. *Int. J. Pattern Recognit. Artif. Intell.* **1993**, *7*, 669–688. [CrossRef]

34. Lee, J.; Abu-El-Haija, S.; Varadarajan, B.; Natsev, A. Collaborative Deep Metric Learning for Video Understanding. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 481–490.
35. Huang, W.; Ding, H.; Chen, G. A novel deep multi-channel residual networks-based metric learning method for moving human localization in video surveillance. *Signal Process.* **2018**, *142*, 104–113. [CrossRef]
36. Hu, J.; Lu, J.; Tan, Y.-P. Deep Metric Learning for Visual Tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *26*, 2056–2068. [CrossRef]
37. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person Re-identification: Past, Present and Future. *arXiv* **2016**, arXiv:1610.02984.
38. Chen, M.; Ge, Y.; Feng, X.; Xu, C.; Yang, D. Person Re-Identification by Pose Invariant Deep Metric Learning with Improved Triplet Loss. *IEEE Access* **2018**, *6*, 68089–68095. [CrossRef]
39. Yang, X.; Zhou, P.; Wang, M. Person Reidentification via Structural Deep Metric Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, 1–12. [CrossRef]
40. Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. *arXiv* **2017**, arXiv:1703.07737.
41. Ahmed, E.; Jones, M.; Marks, T.K. An improved deep learning architecture for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3908–3916.
42. Ding, S.; Lin, L.; Wang, G.; Chao, H. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognit.* **2015**, *48*, 2993–3003. [CrossRef]
43. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef]
44. Annarumma, M.; Montana, G. Deep metric learning for multi-labelled radiographs. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing, Pau, France, 9–13 April 2018; pp. 34–37.
45. Wang, F.; Kang, L.; Li, Y. Sketch-based 3D shape retrieval using Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1875–1883.
46. Dai, G.; Xie, J.; Fang, Y. Deep Correlated Holistic Metric Learning for Sketch-Based 3D Shape Retrieval. *IEEE Trans. Image Process.* **2018**, *27*, 3374–3386. [CrossRef] [PubMed]
47. Xie, J.; Dai, G.; Zhu, F.; Shao, L.; Fang, Y. Deep Nonlinear Metric Learning for 3-D Shape Retrieval. *IEEE Trans. Cybern.* **2018**, *48*, 412–422. [CrossRef]
48. He, X.; Zhou, Y.; Zhou, Z.; Bai, S.; Bai, X. Triplet-Center Loss for Multi-view 3D Object Retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 18–22 June 2018; pp. 1945–1954.
49. Lu, J.; Hu, J.; Tan, Y.-P. Discriminative Deep Metric Learning for Face and Kinship Verification. *IEEE Trans. Image Process.* **2017**, *26*, 4269–4282. [CrossRef]
50. Chen, X.; He, L.; Xu, C.; Liu, J. Distance-Dependent Metric Learning. *IEEE Signal Process. Lett.* **2019**, *26*, 357–361. [CrossRef]
51. Liu, X.; Kumar, B.V.K.V.; You, J.; Jia, P. Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 522–531.
52. Liu, H.; Lu, J.; Feng, J.; Zhou, J. Label-Sensitive Deep Metric Learning for Facial Age Estimation. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 292–305. [CrossRef]
53. Mueller, J.; Thyagarajan, A. Siamese recurrent architectures for learning sentence similarity. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 2786–2792.
54. Benajiba, Y.; Sun, J.; Zhang, Y.; Jiang, L.; Weng, Z.; Biran, O. Siamese Networks for Semantic Pattern Similarity. In Proceedings of the IEEE 13th International Conference on Semantic Computing (ICSC), Newport Beach, CA, USA, 30 January–1 February 2019; pp. 191–194.
55. Zhu, W.; Yao, T.; Ni, J.; Wei, B.; Lu, Z. Dependency-based Siamese long short-term memory network for learning sentence representations. *PLoS ONE* **2018**, *13*, e0193919. [CrossRef] [PubMed]

56. Ein-Dor, L.; Mass, Y.; Halfon, A.; Venezian, E.; Shnayderman, I.; Aharonov, R.; Slonim, N. Learning Thematic Similarity Metric Using Triplet Networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), Melbourne, Australia, 15–20 July 2018; pp. 49–54.
57. Narayanaswamy, V.S.; Thiagarajan, J.J.; Song, H.; Spanias, A. Designing an Effective Metric Learning Pipeline for Speaker Diarization. *arXiv* **2018**, arXiv:1811.00183.
58. Wang, J.; Wang, K.-C.; Law, M.T.; Rudzicz, F.; Brudno, M. Centroid-based Deep Metric Learning for Speaker Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 3652–3656.
59. Snell, J.; Swersky, K.; Zemel, R.S. Prototypical Networks for Few-shot Learning. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 4077–4087.
60. Lu, R.; Wu, K.; Duan, Z.; Zhang, C. Deep ranking: Triplet MatchNet for music metric learning. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 121–125.
61. Wang, Q.; Wan, J.; Yuan, Y. Deep Metric Learning for Crowdedness Regression. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 2633–2643. [CrossRef]
62. Liu, Y.; Zhao, K.; Cong, G. Efficient Similar Region Search with Deep Metric Learning. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 1850–1859.
63. Wang, X.; Liu, M. Multi-View Deep Metric Learning for Volumetric Image Recognition. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
64. Fathi, A.; Wojna, Z.; Rathod, V.; Wang, P.; Song, H.O.; Guadarrama, S.; Murphy, K.P. Semantic Instance Segmentation via Deep Metric Learning. *arXiv* **2017**, arXiv:1703.10277.
65. Cai, S.; Huang, J.; Ding, X.; Zeug, D. Semantic edge detection based on deep metric learning. In Proceedings of the International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Xiamen, China, 6–9 November 2017; pp. 707–712.
66. Xing, Y.; Wang, M.; Yang, S.; Jiao, L. Pan-sharpening via deep metric learning. *ISPRS J. Photogramm. Remote. Sens.* **2018**, *145*, 165–183. [CrossRef]
67. Wang, J.; Zhou, F.; Wen, S.; Liu, X.; Lin, Y. Deep Metric Learning with Angular Loss. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2593–2601.
68. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. *The Caltech-UCSD Birds-200-2011 Dataset*; Computation & Neural Systems Technical Report, CNS-TR-2011-001; California Institute of Technology: Pasadena, CA, USA, 2011; pp. 1–8.
69. Song, H.O.; Savarese, S.; Xiang, Y.; Jegelka, S. Deep Metric Learning via Lifted Structured Feature Embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4004–4012.
70. Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Barcelona, Spain, 5–10 December 2016; pp. 1857–1865.
71. Song, H.O.; Jegelka, S.; Rathod, V.; Murphy, K. Deep Metric Learning via Facility Location. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2206–2214.
72. Ge, W.; Huang, W.; Dong, D.; Scott, M.R. Deep Metric Learning with Hierarchical Triplet Loss. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 272–288.
73. Krause, J.; Stark, M.; Deng, J.; Fei-Fei, L. 3D Object Representations for Fine-Grained Categorization. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), Sydney, Australia, 1–8 December 2013; pp. 554–561.
74. Zheng, L.; Shen, L.; Tian, L.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015; pp. 1116–1124.
75. Ustinova, E.; Lempitsky, V. Learning Deep Embeddings with Histogram Loss. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Barcelona, Spain, 5–10 December 2016; pp. 4170–4178.

76. Yao, H.; Zhang, S.; Hong, R.; Zhang, Y.; Xu, C.; Tian, Q. Deep Representation Learning with Part Loss for Person Re-Identification. *IEEE Trans. Image Process.* **2019**, *28*, 2860–2871. [CrossRef]
77. Li, W.; Zhao, R.; Xiao, T.; Wang, X. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 152–159.
78. Li, B.; Lu, Y.; Godil, A.; Schreck, T.; Bustos, B.; Ferreira, A.; Furuya, T.; Fonseca, M.J.; Johan, H.; Matsuda, T.; et al. A comparison of methods for sketch-based 3D shape retrieval. *Comput. Vis. Image Underst.* **2014**, *119*, 57–80. [CrossRef]
79. Li, B.; Lu, Y.; Li, C.; Godil, A.; Schreck, T.; Aono, M.; Burtscher, M.; Chen, Q.; Chowdhury, N.K.; Fang, B.; et al. A comparison of 3D shape retrieval methods based on a large-scale benchmark supporting multimodal queries. *Comput. Vis. Image Underst.* **2015**, *131*, 1–27. [CrossRef]
80. Howell, A.J.; Buxton, H. Towards unconstrained face recognition from image sequences. In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, Killington, VT, USA, 14–16 April 1996; pp. 224–229.
81. Hu, J.; Lu, J.; Tan, Y.-P. Sharable and Individual Multi-View Metric Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2281–2288. [CrossRef]
82. Wolf, L.; Hassner, T.; Maoz, I. Face recognition in unconstrained videos with matched background similarity. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 21–23 June 2011; pp. 529–534.
83. Marelli, M.; Bentivogli, L.; Baroni, M.; Bernardi, R.; Menini, S.; Zamparelli, R. SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–29 August 2014; pp. 1–8.
84. Greenberg, C.; Banse, D.; Doddington, G.; GarciaRomero, D.; Godfrey, J.; Kinnunen, T.; Martin, A.; McCree, A.; Przybocki, M.; Reynolds, D. The NIST 2014 Speaker Recognition i-Vector Machine Learning Challenge. In Proceedings of the Odyssey 2014: The Speaker and Language Recognition Workshop, Joensuu, Finland, 16–19 June 2014.
85. Hoffer, E.; Ailon, N. Deep Metric Learning Using Triplet Network. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*; Springer: Berlin/Heidelberg, Germany, 2015; Volume 9370, pp. 84–92.
86. Veaux, C.; Yamagishi, J.; MacDonald, K. *Superseded—CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit*; University of Edinburgh, The Centre for Speech Technology Research (CSTR): Edinburgh, UK, 2016.
87. Chung, J.S.; Nagrani, A.; Zisserman, A. VoxCeleb2: Deep Speaker Recognition. *Proc. Interspeech* **2018**, 1086–1090. [CrossRef]
88. Bell, S.; Bala, K. Learning visual similarity for product design with convolutional neural networks. *ACM Trans. Graph.* **2015**, *34*, 98. [CrossRef]
89. Simo-Serra, E.; Trulls, E.; Ferraz, L.; Kokkinos, I.; Fua, P.; Moreno-Noguer, F. Discriminative learning of deep convolutional feature point descriptors. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015; pp. 118–126.
90. Bucher, M.; Herbin, S.; Jurie, F. Hard Negative Mining for Metric Learning Based Zero-Shot Classification. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 524–531.
91. Lin, Y.; Cui, Y.; Zhou, F.; Belongie, S. Fine-Grained Categorization and Dataset Bootstrapping Using Deep Metric Learning with Humans in the Loop. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1153–1162.
92. Movshovitz-Attias, Y.; Toshev, A.; Leung, T.K.; Ioffe, S.; Singh, S. No Fuss Distance Metric Learning Using Proxies. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 360–368.
93. Manmatha, R.; Wu, C.-Y.; Smola, A.J.; Krähenbühl, P. Sampling matters in deep embedding learning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2840–2848.

94. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; pp. 539–546.
95. Filković, I.; Kalafatić, Z.; Hrkać, T. Deep metric learning for person Re-identification and De-identification. In Proceedings of the 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 30 May–3 June 2016; pp. 1360–1364.
96. Jeong, Y.; Lee, S.; Park, D.; Park, K.H. Accurate Age Estimation Using Multi-Task Siamese Network-Based Deep Metric Learning for Frontal Face Images. *Symmetry* **2018**, *10*, 385. [CrossRef]
97. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Deep Metric Learning for Person Re-identification. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 34–39.
98. Shi, H.; Zhu, X.; Liao, S.; Lei, Z.; Yang, Y.; Li, S.Z. Constrained deep metric learning for person re-identification. *arXiv* **2015**, arXiv:1511.07545.
99. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A Discriminative Feature Learning Approach for Deep Face Recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Volume 9911, pp. 499–515.
100. Ni, J.; Liu, J.; Zhang, C.; Ye, D.; Ma, Z. Fine-grained Patient Similarity Measuring using Deep Metric Learning. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 1189–1198.
101. Gong, Z.; Zhong, P.; Yu, Y.; Hu, W. Diversity-Promoting Deep Structural Metric Learning for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote. Sens.* **2018**, *56*, 371–390. [CrossRef]
102. Wang, X.; Han, X.; Huang, W.; Dong, D.; Scott, M.R. Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5022–5030.
103. Rippel, O.; Paluri, M.; Dollar, P.; Bourdev, L. Metric learning with adaptive density discrimination. In Proceedings of the International Conference on Learning Representations, San Juan, PR, USA, 2–4 May 2016; pp. 1–15.
104. Chen, L.; He, Y. Dress Fashionably: Learn Fashion Collocation with Deep Mixed-Category Metric Learning. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2017; pp. 2103–2110.
105. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
106. LeCun, Y.; Huang, F.J.; Bottou, L. Learning methods for generic object recognition with invariance to pose and lighting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA, 27 June–2 July 2004; pp. 97–104.
107. Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. Master's Thesis, Department of Computer Science, University of Toronto, Toronto, ON, Canada, 2009.
108. Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A.Y. Reading digits in natural images with unsupervised feature learning. In Proceedings of the Workshop on Deep Learning and Unsupervised Feature Learning, Granada, Spain, 16 December 2011; pp. 1–9.
109. Coates, A.; Ng, A.; Lee, H. An analysis of single-layer networks in unsupervised feature learning. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 215–223.
110. Xie, S.; Yang, T.; Wang, X.; Lin, Y. Hyper-class augmented and regularized deep learning for fine-grained image classification. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2645–2654.
111. Khosla, A.; Jayadevaprakash, N.; Yao, B.; Fei-Fei, L. Novel dataset for fine-grained image categorization: Stanford dogs. In Proceedings of the First Workshop on Fine-Grained Visual Categorization (FGVC) in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 21–23 June 2011; pp. 1–2.
112. Parkhi, O.M.; Vedaldi, A.; Zisserman, A.; Jawahar, C. Cats, and dogs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 18–20 June 2012; pp. 3498–3505.

113. Nilsback, M.-E.; Zisserman, A. Automated Flower Classification over a Large Number of Classes. In Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, Bhubaneswar, India, 16–19 December 2008; pp. 722–729.
114. Russakovsky, O.; Fei-Fei, L. Attribute learning in large-scale datasets. In Proceedings of the European Conference on Computer Vision (ECCV), Crete, Greece, 5–11 September 2010; pp. 1–14.
115. Tang, X.; Liu, Z.; Luo, P.; Qiu, S.; Wang, X. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1096–1104.
116. Gray, D.; Brennan, S.; Tao, H. Evaluating appearance models for recognition, reacquisition, and tracking. In Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), Rio de Janeiro, Brazil, 14 October 2007; pp. 1–7.