A Decade Survey of Content Based Image Retrieval Using Deep Learning

Shiv Ram Dubey, *Member, IEEE*

***Abstract*— The content based image retrieval aims to find the similar images from a large scale dataset against a query image. Generally, the similarity between the representative features of the query image and dataset images is used to rank the images for retrieval. In early days, various hand designed feature descriptors have been investigated based on the visual cues such as color, texture, shape, etc. that represent the images. However, the deep learning has emerged as a dominating alternative of hand-designed feature engineering from a decade. It learns the features automatically from the data. This paper presents a comprehensive survey of deep learning based developments in the past decade for content based image retrieval. The categorization of existing state-of-the-art methods from different perspectives is also performed for greater understanding of the progress. The taxonomy used in this survey covers different supervision, different networks, different descriptor type and different retrieval type. A performance analysis is also performed using the state-of-the-art methods. The insights are also presented for the benefit of the researchers to observe the progress and to make the best choices. The survey presented in this paper will help in further research progress in image retrieval using deep learning.**

***Index Terms*— Content based image retrieval, deep learning, CNNs, survey, supervised and unsupervised learning.**

I. Introduction

IMAGE retrieval is a well studied problem of image matching where the similar images are retrieved from a database w.r.t. a given query image [1], [2]. Basically, the similarity between the query image and the database images is used to rank the database images in decreasing order of similarity [3]. Thus, the performance of any image retrieval method depends upon the similarity computation between images. Ideally, the similarity score computation method between two images should be discriminative, robust and efficient.

A. Hand-Crafted Descriptor Based Image Retrieval

In order to make the retrieval robust to geometric and photometric changes, the similarity between images is computed based on the content of images. Basically, the content of the images (i.e., the visual appearance) in terms of the color, texture, shape, gradient, etc. are represented in the form of a feature descriptor [4]. The similarity between the feature vectors of the corresponding images is treated as the similarity between the images. Thus, the performance of any content based image retrieval (CBIR) method heavily depends upon the feature descriptor representation of the image. Any feature descriptor representation method is expected to have the discriminating ability, robustness and low dimensionality. Various feature descriptor representation methods have been investigated to compute the similarity between the two images for content based image retrieval. The feature descriptor representation utilizes the visual cues of the images selected manually based on the need [5]–[16]. These approaches are also termed as the hand-designed or hand-engineered feature description. Moreover, generally these methods are unsupervised as they do not need the data to design the feature representation method. Various survey has been also conducted time to time to present the progress in content based image retrieval, including [17] in 2008, [18] in 2014 and [19] in 2017. The hand-engineering feature for image retrieval was a very active research area. However, its performance was limited as the hand-engineered features are not able to represent the image characteristics in an accurate manner.

B. Distance Metric Learning Based Image Retrieval

The distance metric learning has been also used very extensively for feature vectors representation [20]. It is also explored well for image retrieval [21]. Some notable deep metric learning based image retrieval approaches include Contextual constraints distance metric learning [22], Kernel-based distance metric learning [23], [24], Visuality-preserving distance metric learning [25], Rank-based distance metric learning [26], Semi-supervised distance metric learning [27], Hamming distance metric learning [28], [29], and Rank based metric learning [30], [31]. Generally, the deep metric learning based approaches have shown the promising retrieval performance compared to hand-crafted approaches. However, most of the existing deep metric learning based methods rely on the linear distance functions which limits its discriminative ability and robustness to represent the non-linear data for image retrieval. Moreover, it is also not able to handle the multi-modal retrieval effectively.

C. Deep Learning Based Image Retrieval

From a decade, a shift has been observed in feature representation from hand-engineering to learning-based after the emergence of deep learning [32], [33]. This transition is depicted in Fig. 1 where the convolutional neural networks

Manuscript received December 20, 2020; revised April 18, 2021; accepted May 7, 2021. Date of publication May 17, 2021; date of current version May 5, 2022. This work was supported by the Global Innovation & Technology Alliance (GITA) on behalf of the Department of Science and Technology (DST), Government of India under Project GITA/DST/TWN/P-83/2019. This article was recommended by Associate Editor W. Liu.

The author is with the Computer Vision Group, Indian Institute of Information Technology, Chittoor 517646, India (e-mail: shivram1987@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3080920.

Digital Object Identifier 10.1109/TCSVT.2021.3080920

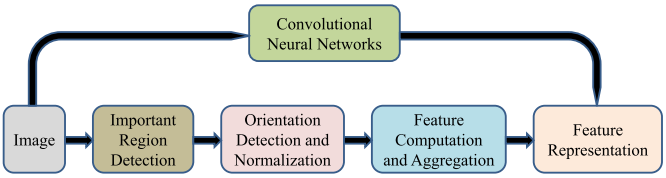

Fig. 1. The pipeline of state-of-the-art feature representation is replaced by the CNN based feature representation.

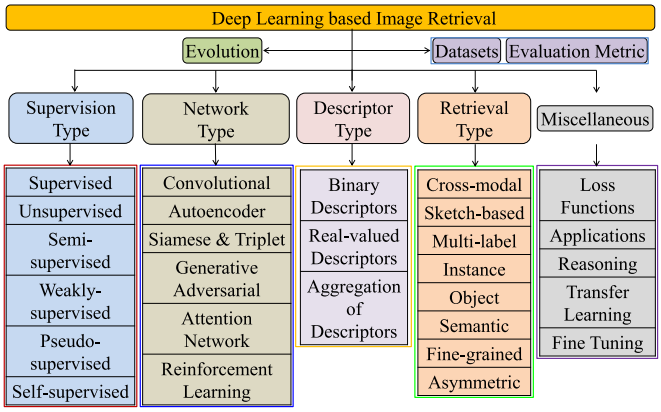

Fig. 2. Taxonomy used in this survey to categorize the existing deep learning based image retrieval approaches.

based feature learning replaces the state-of-the-art pipeline of traditional hand-engineered feature representation. The deep learning is a hierarchical feature representation technique to learn the abstract features from data which are important for that dataset and application [34]. Based on the type of data to be processed, different architectures came into existence such as Artificial Neural Network (ANN)/ Multilayer Perceptron (MLP) for 1-D data [35], [36], Convolutional Neural Networks (CNN) for image data [37], [38], and Recurrent Neural Networks (RNN) for time-series data [39], [40]. A huge progress has been made in this decade to utilize the power of deep learning for content based image retrieval [32], [41]–[44]. Thus, this survey mainly focuses over the progress in state-of-the-art deep learning based models and features for content based image retrieval from its inception. A taxonomy for the same is portrayed in Fig. 2. The major contributions of this survey can be outlined as follows:

1) This survey covers the deep learning based image retrieval approaches very comprehensively in terms of evolution of image retrieval using deep learning, different supervision type, network type, descriptor type, retrieval type and other aspects.
2) In contrast to the recent reviews [21], [42], [43], this survey specifically covers the progress in image retrieval using deep learning in 2011-2020 decade. An informative taxonomy is provided with wide coverage of existing deep learning based image retrieval approaches as compared to the recent survey [44].
3) This survey enriches the reader with the state-of-the-art image retrieval using deep learning methods with analysis from various perspectives.

TABLE I
THE SUMMARY OF LARGE-SCALE DATASETS FOR DEEP LEARNING BASED IMAGE RETRIEVAL

Dataset	Year	#Classes	Training	Test	Image Type
CIFAR-10 [45]	2009	10	50,000	10,000	Object Category Images
NUS-WIDE [46]	2009	21	97,214	65,075	Scene Images
MNIST [47]	1998	10	60,000	10,000	Handwritten Digit Images
SVHN [48]	2011	10	73,257	26,032	House Number Images
SUN397 [49]	2010	397	100,754	8,000	Scene Images
UT-ZAP50K [50]	2014	4	42,025	8,000	Shoes Images
Yahoo-1M [51]	2015	116	1,011,723	112,363	Clothing Images
ILSVRC2012 [52]	2012	1,000	∼1.2 M	50,000	Object Category Images
MS COCO [53]	2015	80	82,783	40,504	Common Object Images
MIRFlicker-1M [54]	2010	-	1 M	-	Scene Images
Google Landmarks [55]	2017	15 K	∼1 M	-	Landmark Images
Google Landmarks v2 [56]	2020	200 K	5 M	-	Landmark Images
Clickture [57]	2013	73.6 M	40 M	-	Search Log

4) This paper also presents the brief highlights and important discussions along with the comprehensive comparisons on benchmark datasets using the state-of-the-art deep learning based image retrieval approaches.

This survey is organized as follows: the background is presented in Section II the evolution of deep learning based image retrieval is compiled in Section III; the categorization of existing approaches based on the supervision type, network type, descriptor type, and retrieval type are discussed in Section IV, V, VI, and VII, respectively; some other aspects are highlighted in Section VIII; the performance comparison of the popular methods is performed in Section IX; conclusions and future directions are presented in Section X.

II. BACKGROUND

In this section the background is presented in terms of the commonly used evaluation metrics and benchmark datasets.

A. Retrieval Evaluation Measures

In order to judge the performance of image retrieval approaches, precision, recall and f-score are the common evaluation metrics. The mean average precision (mAP) is very commonly used in the literature. The precision is defined as the percentage of correctly retrieved images out of the total number of retrieved images. The recall is another performance measure being used for image retrieval by computing the percentage of correctly retrieved images out of the total number of relevant images present in the dataset. The f-score is computed from the harmonic mean of precision and recall.

B. Datasets

With the inception of deep learning models, various large-scale datasets have been created to facilitate the research in image recognition and retrieval. The details of large-scale datasets are summarized in Table I. Datasets having various types of images are available to test the deep learning based approaches such as object category datasets [45], [52], [53], scene datasets [46], [49], [90], digit datasets [47], [48], apparel datasets [50], [51], landmark datasets [55], [56], etc.

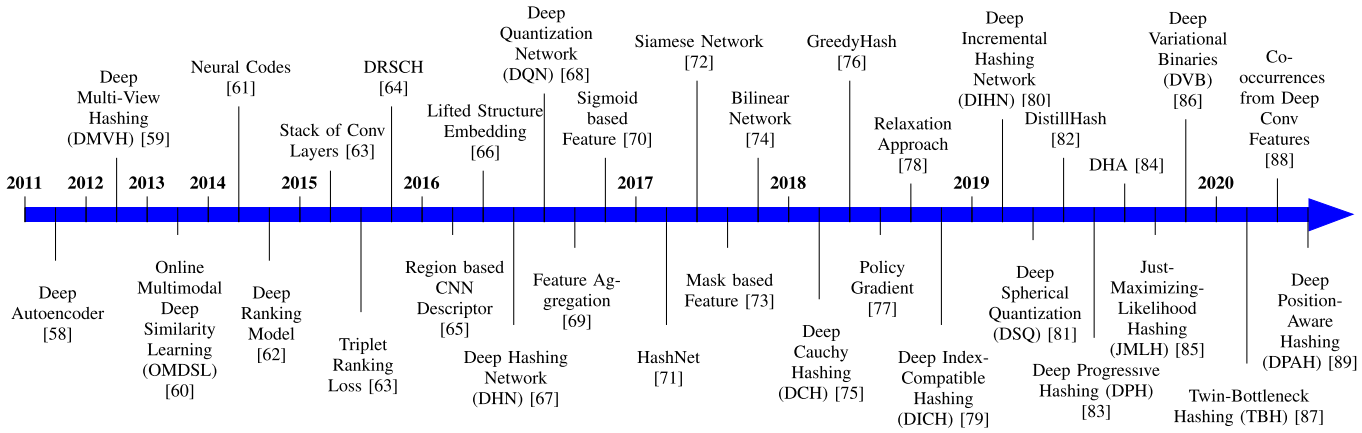

Fig. 3. A chronological view of deep learning based image retrieval methods depicting its evolution from 2011 to 2020.

The CIFAR-10 dataset is very widely used object category dataset [45]. The ImageNet (ILSVRC2012), a large-scale dataset, is also an object category dataset with more than a million number of images [52]. The MS COCO dataset [53] created for common object detection is also utilized for image retrieval purpose. Among scene image datasets commonly used for retrieval purpose, the NUS-WIDE dataset is from National University of Singapore [46]; the Sun397 is a scene understanding dataset from 397 categories with more than one lakh images [49], [91]; and the MIRFlicker-1M [90] dataset consists of a million images downloaded from the social photography site Flickr. The MNIST dataset is one of the old and large-scale digit image datasets [47] consisting of optical characters. The SVHN is another digit dataset [48] from the street view house number images which is more complex than MNIST dataset. The shoes apparel dataset, namely UT-ZAP50K [50], consists of roughly 50K images. The Yahoo-1M is another apparel large-scale dataset used in [51] for image retrieval. The Google landmarks dataset is having around a million landmark images [55]. The extended version of Google landmarks (i.e., v2) [56] contains around 5 million landmark images. There are more datasets used for retrieval in the literature, such as Corel, Oxford, Paris, etc., however, these are not the large-scale datasets. The CIFAR-10, MNIST, SVHN and ImageNet are the widely used datasets in majority of the research. Clickture is a common dataset for search log based on the queries of users [57]. The click property has been utilized for different applications, such as cross-view learning for image search [92], distance metric learning for image ranking [93] and deep structure-preserving embeddings with visual attention [94].

Note that only CIFAR-10 and MNIST datasets contain the same number of samples in each category. Other datasets are created generally in unconstrained environment with huge number of samples, thus the classes are not well balanced. The choice of dataset can be dependent upon the scenario where image retrieval models need to be used, such as object category and scene datasets for unconstrained environment, apparel datasets for e-commerce applications, and landmark datasets for driving applications.

III. Evolution of Deep Learning for Content Based Image Retrieval (CBIR)

The deep learning based generation of descriptors or hash codes is the recent trends large-scale content based image retrieval, due to its computational efficiency and retrieval quality [21]. In this section, a journey of deep learning models for image retrieval from 2011 to 2020 is presented as a chronological overview in Fig. 3.

1) 2011-2013: Among the initial attempts, in 2011, Krizhevsky and Hinton have used a deep autoencoder to map the images to short binary codes for content based image retrieval (CBIR) [58]. Kang *et al.* (2012) have proposed a deep multi-view hashing to generate the code for CBIR from multiple views of data by modeling the layers with view-specific and shared hidden nodes [59]. In 2013, Wu *et al.* have considered the multiple pretrained stacked denoising autoencoders over low features of the images [60]. They also fine tune the multiple deep networks on the output of the pretrained autoencoders.

2) 2014: In an outstanding work, the activations of the top layers of a large convolutional neural network (CNN) are utilized as the descriptors (neural codes) for image retrieval [61] as depicted in Fig. 4. A very promising performance has been recorded using the neural codes for image retrieval even if the model is trained on un-related data. The neural code is compressed using principal component analysis (PCA) to generate the compact descriptor. In 2014, deep ranking model is investigated by learning the similarity metric directly from images [62]. Basically, the triplets are employed to capture the inter-class and intra-class image differences.

3) 2015: In 2015, a deep architecture is developed which consists of a stack of convolution layers to produce the intermediate image features [63] which are used to generate the hash bits. The triplet ranking loss is also utilized to incorporate the inter-class and intra-class differences in [63] for image retrieval. Zhang *et al.* (2015) have developed a deep regularized similarity comparison hashing (DRSCH) by training a deep CNN model to simultaneously optimize the discriminative image features and hash functions [64].

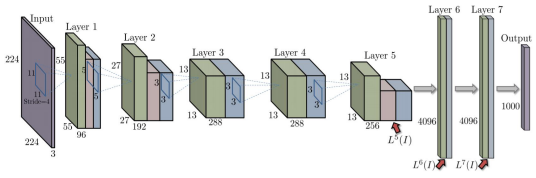

Fig. 4. The illustration of the neural code generation from a convolutional neural network (CNN) [61].

4) 2016: In 2016, Gordo *et al.* have pooled the relevant regions to form the descriptor with the help of a region proposal network to prioritize the important object regions [65]. Song *et al.* (2016) have computed the lifted structure loss between the CNN and the original features [66]. Supervised deep hashing network (DHN) learns the important image representation by controlling the quantization error [67]. At the same time, Cao *et al.* have introduced a deep quantization network (DQN) which is very similar to the DHN model [68]. The CNN based features are aggregated in [69] with the help of rank-aware multi-assignment and direction based combination. A sigmoid layer is added before the loss layer of a CNN to learn the binary code for CBIR [70].

5) 2017: In 2017, Cao *et al.* have proposed HashNet deep architecture to generate the hash code by a continuation method [71]. It learns the non-smooth binary activations using the continuation method to generate the binary hash codes from imbalanced similarity data. Gordo *et al.* (2017) have shown that the noisy training data, inappropriate deep architecture and suboptimal training procedure are the main hurdle to utilize the deep learning for image retrieval [72]. Different masking schemes are used in [73] to select the prominent CNN features for image retrieval. A bilinear network with two parallel CNNs is also used as a feature extractors [74].

6) 2018: In 2018, Cao *et al.* have investigated a deep cauchy hashing (DCH) model for binary hash code with the help of a pairwise cross-entropy loss based on Cauchy distribution [75]. Su *et al.* have employed the greedy hash by transmitting the gradient as intact during the backpropagation for hash coding layer which uses the sign function in forward propagation [76]. Different approaches such as policy gradient [77] and series expansion [78] are also utilized to train the models. Deep index-compatible hashing (DICH) method [79] is investigated by minimizing the number of similar bits between the binary codes of inter-class images.

7) 2019: In 2019, a deep incremental hashing network (DIHN) is proposed in [80] to directly learn the hash codes corresponding to the new class coming images, while retaining the hash codes of existing class images. A supervised quantization based points representation on a unit hypersphere is used in deep spherical quantization (DSQ) model [81]. DistillHash [82] distills data pairs and learns deep hash functions from the distilled data set by employing the Bayesian learning framework. A deep progressive hashing (DPH) model is developed to generate a sequence of binary codes by utilizing the progressively expanded salient regions [83]. Adaptive loss function based deep hashing [84], just-maximizing-likelihood hashing (JMLH) [85] and deep variational binaries (DVB) [86] are other approaches discovered in 2019.

8) 2020: Recently, in 2020, Shen *et al.* have come up with a twin-bottleneck hashing (TBH) model between encoder and decoder networks [87]. They have employed the binary and continuous bottlenecks as the latent variables in a collaborative manner. Forcen *et al.* (2020) have utilized the last convolution layer of CNN representation by modeling the co-occurrences from deep convolutional features [88]. A deep position-aware hashing (DPAH) model is proposed in 2020 [89] which constraints the distance between data samples and class centers.

Most of the methods developed between 2011 and 2015 use the features learnt by the autoencoders and convolutional neural networks. However, these methods face the issues in terms of the less discriminative ability as the models are generally trained for the classification problem and the information loss due to the quantization of features. Image retrieval using deep learning has witnessed a huge growth in between 2016 and 2020. As the image retrieval application needs feature learning for matching, different types of networks have been utilized to do so. The recent methods have designed the several objective functions which lead to the high inter class separation and high intra class condensation in feature space. Moreover, the development in different network architectures has also led to the growth in the image retrieval area. The key issue being addressed by deep learning methods is to learn very discriminative, robust and compact features for image retrieval.

IV. Different Supervision Categorization

This section covers the image retrieval methods in terms of the different supervision types. Basically, supervised, unsupervised, semi-supervised, weakly-supervised, pseudo-supervised and self-supervised approaches are included.

A. Supervised Approaches

The supervised deep learning models are used by researchers very heavily to learn the class specific and discriminative features for image retrieval. In 2014, Xia *et al.* have used a CNN to learn the representation of images which is used to generate a hash code and class labels [95]. The promising performance is reported over MNIST, CIFAR-10 and NUS-WIDE datasets. Shen *et al.* [96] have proposed the supervised discrete hashing (SDH) based generation of image description with the help of the discrete cyclic coordinate descent for retrieval. Liu *et al.* (2016) have done the revolutionary work by introducing a deep supervised hashing (DSH) method to learn the binary codes from the similar/dissimilar pairs of images [97]. A similar work is also presented in deep pairwise-supervised hashing (DPSH) method for image retrieval [98]. The pair-wise labels are extended to the triplet labels (i.e., query, positive and negative images) to train a shared deep CNN model for feature learning [99]. An independent layer-wise local updates are performed in [100] to efficiently train a very deep supervised hashing (VDSH) model.

In 2017, Li *et al.* have used the classification information and the pairwise label information in a single framework for the learning of the deep supervised discrete hashing (DSDH)

codes [101]. The supervised semantics-preserving deep hashing (SSDH) model integrates the retrieval and classification characteristics in feature learning [102]. The scalable image search is performed in [103] by introducing the following three characteristics: 1) minimizing the loss between the real-valued code and equivalent converted binary code, 2) ensuring the even distribution among each bit in the binary codes, and 3) decreasing the redundancy of a bit in the binary code.

The supervised training has been also the choice in asymmetric hashing [104]. A deep product quantization (DPQ) model is followed in supervised learning mode for image search and retrieval [105]. The supervised deep feature embedding is also used with the hand crafted features [106]. A very recently, a multi-Level hashing of deep features is performed by Ng *et al.* [107]. An angular hashing loss function is used to train the network in the supervised fashion [108]. A supervised hashing is also used for the multi-deep ranking [109] to improve the retrieval efficiency. Other supervised approaches are deep binary hash codes [51], deep hashing network [67], deep spherical quantization [81], and adaptive loss based supervised deep learning to hash [84].

B. Unsupervised Approaches

Though the supervised models have shown promising performance for retrieval, it is difficult to get the labelled large-scale data always. Thus, several unsupervised models have been also investigated which do not require the class labels. The unsupervised models generally enforce the constraints on hash code and/or generated output to learn the features.

Erin *et al.* [110] have used the deep networks in an unsupervised manner to learn the hash code with the help of the constraints like quantization loss, balanced bits and independent bits. Huang *et al.* [111] have utilized the CNN coupled with unsupervised discriminative clustering. In an outstanding work, DeepBit utilizes the constraints like minimal quantization loss, evenly distributed codes and uncorrelated bits for unsupervised image retrieval [112], [113]. In order to improve the robustness of DeepBit, a rotation data augmentation based fine tuning is also performed. However, the DeepBit model suffers with the severe quantization loss due to the rigid binarization of data using sign function without considering its distribution property. Deep binary descriptor with multiquantization (DBD-MQ) [114] tackles the quantization problem of DeepBit by jointly learning the parameters and the binarization functions using a K-AutoEncoders (KAEs).

It is observed in [115] that unsupervised CNN can learn more distinctive features if fine tuned with hard positive and hard negative examples. The patch representation using a patch convolutional kernel network is also adapted for patch retrieval [116]. An anchor image, a rotated image and a random image based triplets are used in unsupervised triplet hashing (UTH) to learn the binary codes for image retrieval [117]. The UTH objective function uses the combination of discriminative loss, quantization loss and entropy loss. An unsupervised similarity-adaptive deep hashing (SADH) is proposed in [118] by updating a similarity graph and optimizing the binary codes. Xu *et al.* [119] have proposed a semantic-aware part weighted aggregation using part-based detectors for CBIR systems. Unsupervised generative adversarial networks [120]–[122] are also investigated for image retrieval. The distill data pairs [82] and deep variational networks [86] are also used for unsupervised image retrieval. The pseudo triplets based unsupervised deep triplet hashing (UDTH) technique [123] is introduced for scalable image retrieval. Very recently unsupervised deep transfer learning has been exploited by Liu *et al.* [124] for retrieval in remote sensing images.

Though the unsupervised models do not need labelled data, its performance is generally lower than the supervised approaches. Thus, researchers have explored the models between supervised and unsupervised, such as semi-supervised, weakly-supervised, pseudo-supervised and self-supervised.

C. Semi, Weakly, Pseudo and Self -Supervised Approaches

The semi-supervised approaches generally use a combination of labelled and unlabelled data for feature learning [168], [169]. Semi-supervised deep hashing (SSDH) [141] uses labelled data for the empirical error minimization and both labelled and unlabelled data for embedding error minimization. The generative adversarial learning has been also utilized extensively in semi-supervised image retrieval [147], [161], [170]. A teacher-student based semi-supervised image retrieval [171] uses the pairwise information learnt by the teacher network as the guidance to train the student network.

Weakly-supervised approaches have been also explored for the image retrieval. Tang *et al.* (2017) have put forward a weakly-supervised multimodal hashing (WMH) by utilizing the local discriminative and geometric structures in the visual space [172]. Guan *et al.* [173] have performed the pre-training in weakly-supervised mode and fine-tuning in supervised mode. A weakly supervised deep hashing using tag embeddings (WDHT) [174] utilizes the word2vec semantic embeddings. A semantic guided hashing (SGH) [175] is used for image retrieval by simultaneously employing the weakly-supervised tag information and the inherent data relations.

The pseudo supervised networks have been also developed for image retrieval. The pseudo triplets are utilized in [123] for unsupervised image retrieval. K-means clustering based pseudo labels are generated and used for the training of a deep hashing network [176], [177]. An appealing performance has been observed using pseudo labels over CIFAR-10 and Flickr datasets for image retrieval.

The self-supervision is another way of supervision used in some research works for image retrieval. For example, Li *et al.* [146] have used the adversarial networks in self-supervision mode by utilizing the multi-label annotations. Zhang *et al.* [178] have introduced a self-supervised temporal hashing (SSTH) for video retrieval.

D. Summary

Following are the take always from the above discussion on deep learning based models from the supervision perspective:

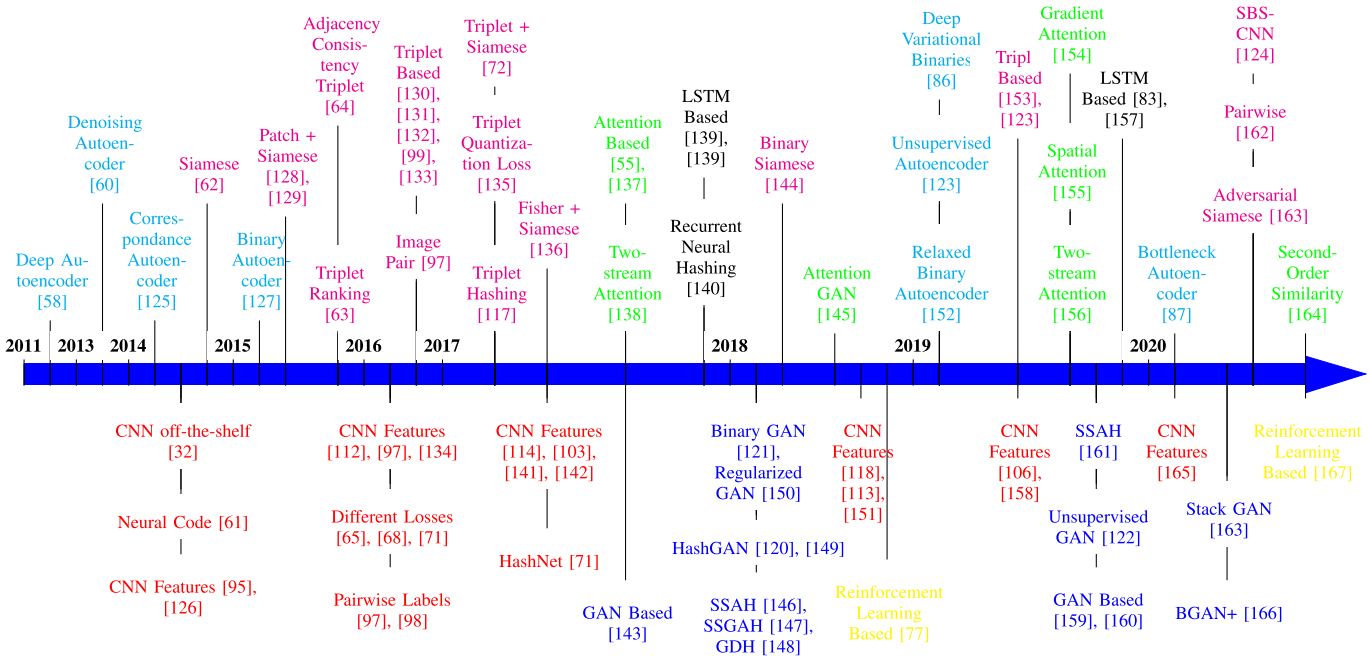

Fig. 5. A chronological view of deep learning based image retrieval methods depicting the different type of neural networks used from 2011 to 2020. The convolutional neural network, autoencoder network, siamese & triplet network, recurrent neural network, generative adversarial network, attention network and reinforcement learning network based deep learning approaches for image retrieval are depicted in Red, Cyan, Magenta, Black, Blue, Green, and Yellow colors, respectively.

- The supervised approaches utilize the class-specific semantic information through the classification error apart from the other objectives related to the hash code generation. Generally, the performance of supervised models is better than other models due to learning of the fine-grained and class specific information.
- The unsupervised models make use of the unsupervised constraints on hash code (i.e., quantization loss, independent bits, etc.) and/or data reconstruction (i.e., using an autoencoder type of networks) to learn the features.
- The semi-supervised approaches exploit the labelled and un-labelled data for the feature learning using deep networks. These approaches generally utilize the information from different modalities using different networks.
- The pseudo-supervised approaches generate the pseudo labels using some other methods to facilitate the training using generated labels. The self-supervised methods generate the temporal or generative information to learn the models over the training epochs.
- The minimal quantization error, independent bits, low dimensional feature, and discriminative code are the common objectives for most of the retrieval methods.

V. Network Types for Image Retrieval

In this section, deep learning based image retrieval approaches are presented in terms of the different architectures. A chronological overview from 2011 to 2020 is illustrated in Fig. 5 for different type of networks for image retrieval.

A. Convolutional Neural Networks for Image Retrieval

Convolutional neural networks (CNN) based feature learning has been utilized extensively for image retrieval as shown in Fig. 4. In 2014, CNN features off-the-shelf have shown a tremendous performance gain for image recognition and retrieval as compared to the hand-crafted features [32]. At the same time the activations of trained CNN has been also explored as the neural code for retrieval [61]. An image representation learning has been also performed using the CNN model to generate the descriptor for image retrieval [95]. In 2016, pairwise labels are exploited to learn the CNN feature for image retrieval [97], [98]. The CNN activations are heavily used to generate the hash codes for efficient image retrieval by employing the different losses [65], [68], [71]. The abstract features of CNN are learnt for the image retrieval in different modes, such as unsupervised image retrieval [112]–[114], [118], supervised image retrieval [95], [97], [103], [106], semi-supervised image retrieval [141], cross-modal retrieval [134], [142], sketch based image retrieval [151], [158], and object retrieval [126], [165].

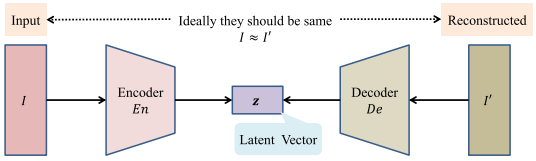

Fig. 6. A typical Autoencoder network consisting of an Encoder and a Decoder network. Generally, the encoder is a CNN and the decoder is an up-CNN. The output of the encoder is a latent space which is used to generate the hash codes.

B. Autoencoder Networks Based Image Retrieval

Autoencoder (AE) is a type of unsupervised neural network which can be used to reconstruct the input image from the latent space as portrayed in Fig. 6. Basically, it consists

of two networks, namely encoder (En) and decoder (De). The encoder network transforms the input (I) into latent feature space (z) as $En : I \rightarrow z$. Whereas, the decoder network tries to reconstruct the original image (I') from latent feature space as $De : z \rightarrow I'$. The model is trained by minimizing the reconstruction error between original image (I) and reconstructed image (I') using L_1 or L_2 loss function.

The autoencoders have been very intensively used to learn the features as the latent space for image retrieval. In the initial attempts, the deep autoencoder was used for image retrieval in 2011 [58]. A stacked denoising autoencoder is used to train the multiple deep neural networks for retrieval task [60]. Feng *et al.* (2014) have utilized the correspondence autoencoder (Corr-AE) for cross-modal retrieval [125]. A binary autoencoder is used to learn the binary code for fast image retrieval by reconstructing the image from that binary code function [127]. The use of autoencoder in image retrieval has witnessed a huge progressed in recent years, such as Deep variational binaries (DVB) using the variational Bayesian networks [86]; Autoencoder over the triplet [123]; and Relaxed binary autoencoder (RBA) [152] are investigated in 2019. In a recent work, double latent bottlenecks is used in autoencoder [87]. It includes binary latent variable and continuous latent variable. The latent variable bottleneck exchanges crucial information collaboratively and the binary codes bottleneck uses a code-driven graph to capture the intrinsic data structure.

C. Siamese and Triplet Networks for Image Retrieval

1) Siamese Network: The siamese is type of neural network that exploits the distance between features of image pairs as depicted in Fig. 7(a). The siamese network based learnt features have shown very promising performance for fine-grained image retrieval [62]. A pair of similar or dissimilar images is jointly processed by Liu *et al.* [97] to produce 1 or -1 output by CNN to learn the feature for image retrieval. Ong *et al.* (2017) have used the fisher vector computed on top of the CNN feature in autoencoder network to generate the discriminating feature descriptor for image retrieval [136]. The siamese network is also used to develop the light weight models for efficient image retrieval [124], [144]. A pairwise similarity-preserving quantization loss is employed in [162]. The siamese network is used with the stacked adversarial network in [163]. The siamese network is also used for patch based image matching [128], [129].

2) Triplet Network: A triplet network is a variation of siamese network which utilizes a triplet of images, including an anchor, a positive and a negative image as shown in Fig. 7(b). The triplet network minimizes the distance between the features of anchor and positive image and maximizes the distance between the features of anchor and negative image, simultaneously. In 2015, a triplet ranking loss is utilized on top of the shared CNN features to learn the network for computation of binary descriptors for image retrieval [63]. An adjacency consistency based regularization term is introduced in the triplet network to enforce the discriminative ability of the CNN feature description [64]. Zhuang *et al.* [130] have used triplet to learn the hash code by employing the relation weights matrix and graph cuts optimization. The triplet ranking loss, orthogonality constraint and softmax loss are minimized jointly in [131]. Triplet based siamese networks are also used for image retrieval [72], [132]. Triplet quantization based objective function minimizes the information loss [135], [179]. The triplet based feature learning has been also exploited for sketch based image retrieval [153]. Triplets are also exploited for supervised hashing [99] and unsupervised hashing [117], [123] for image retrieval.

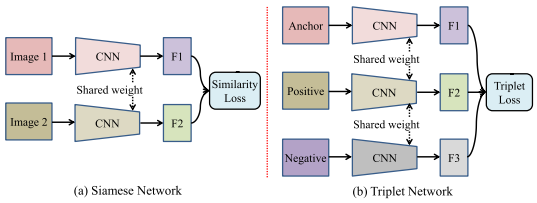

Fig. 7. (a) Siamese network computes the similarity between image pairs. (b) Triplet network minimizes the distance between the anchor and positive and maximizes the distance between the anchor and negative in feature space.

D. Generative Adversarial Networks Based Retrieval

The generative adversarial network (GAN) uses two networks, i.e., generator and discriminator. The generator network generates the new samples in the training set from the random vector. Whereas, the discriminator network distinguishes between generated image and original image. In 2018, Song *et al.* have introduced a binary generative adversarial network (BGAN) for generating the representational binary codes for image retrieval [121]. At the same time, a regularized GAN is used to introduce the BinGAN model [150] to learn the compact binary patterns. The BinGAN uses two regularizers, including a distance matching regularizer and a binarization representation entropy (BRE) regularizer. In 2018, the generative networks are also utilized in [120] to develop HashGAN in an unsupervised manner to generate the hash code for image retrieval. At the same time another HashGAN is developed by employing the paired conditional Wasserstein GAN for image retrieval [149]. GAN has been also used for cross-modal retrieval [143], [145], [146], [160], semi-supervised hashing [147], [161], sketch based image retrieval [148], [159], [163] and unsupervised adversarial hashing [122]. In 2020, binary generative adversarial networks based unified BGAN+ framework [166] is developed for image retrieval.

E. Attention Networks for Image Retrieval

The attention has been observed as a very effective way of modelling the saliency information into the feature space to avoid the effect of background. In 2017, Noh *et al.* have used the attention-based keypoints to select the important deep local features [55]. Yang *et al.* (2017) have introduced a two-stream attentive CNNs by fusing a Main and an Auxiliary CNN (MAC) for image retrieval [138]. The main CNN focuses over the discriminative visual features for semantic information, whereas the auxiliary CNN focuses

over the part of features for attentive information. Similarly, two sub-networks are employed in [155] for spatial attention and global features, respectively. Recently, Ng *et al.* [164] have computed the second-order similarity (SOS) loss over the attention based selected regions of the input image for image retrieval. The attention based models are developed for cross-modal retrieval [145] and fine-grained sketch-based image retrieval [137]. The gradient attention network based deep hashing [154] enforces the CNN binary features of a pair to minimize the distances between them, irrespective of their signs or directions. In order to localize the important image region for the feature description, an attentional heterogeneous bilinear network is employed in [180] for fashion image retrieval.

F. Recurrent Neural Networks for Image Retrieval

In 2018, Lu *et al.* have utilized the recurrent neural network (RNN) concept to perform a hierarchical recurrent neural hashing (HRNH) to produce the effective hash codes for image retrieval [140]. In 2017, Shen *et al.* have used the region-based convolutional networks with long short-term memory (LSTM) modules for textual-visual cross retrieval [139]. Bai *et al.* (2019) have also employed the LSTM based recurrent deep network in the triplet hashing framework to naturally inherit the useful information for image retrieval [83].

G. Reinforcement Learning Networks Based Retrieval

In 2018, Yuan *et al.* have exploited the reinforcement learning for image retrieval [77]. They have used a relaxation free method through policy gradient to generate the hash codes for image retrieval. The similarity preservation via the generated binary codes is used as the reward function. In 2020, Yang *et al.* [167] have utilized the deep reinforcement learning to perform the de-redundancy in hash bits to get rid of redundant and/or harmful bits, which reduces the ambiguity in the similarity computation for image retrieval.

H. Summary

The summary of the different network driven deep learning based image retrieval approaches is as follows:

- The convolutional neural network features are exploited for the hash code and descriptor learning by employing the various constraints like classification error, quantization error, independent bits, etc.
- In order to make the features more representative of the image, the autoencoder networks are used which enforces the learning based on the reconstruction loss.
- The discriminative power of descriptive hash code is enhanced by exploiting the siamese and triplet networks. Different constraints are used on the hash code to make it discriminative and compact.
- The generative adversarial network based approaches have been highly utilized to improve the discriminative ability and robustness of the learnt features by encoder network guided through the discriminator network.
- The automatic important feature selection is performed using attention module to control the redundancy in the feature space. The recurrent neural network and reinforcement learning network have been also shown very effective for the image retrieval.

VI. Type of Descriptors for Image Retrieval

This section covers the binary hash codes for efficient image retrieval, real-valued descriptors and feature aggregation for discriminative image retrieval as depicted in Fig. 8.

A. Binary Descriptors

Different types of networks are used to learn the binary description such as deep neural networks [110], convolutional neural networks [51], autoencoder networks [127], siamese networks [144], triplet networks [123], generative adversarial networks, [166], and variational networks [86]. In 2015, Liong *et al.* [110] have introduced a supervised deep hashing (SDH). The SDH method uses the quantization loss, balanced bits and independent bits constraints. Binary hash code is also learnt through a latent layer in a supervised manner in [51]. A binary autoencoder [58], [127] and a siamese network [144] are used to learn the binary features for efficient image retrieval. A binary deep neural network (BDNN) is proposed by converting a hidden layer output to binary code [186], [194]. The binary code is jointly learnt with feature aggregation in [152]. A masking technique over the convolutional features is used to generate the binary description for image retrieval [195]. A ranking optimization discrete hashing (RODH) approach is used in [196] by generating the discrete hash codes ($+1$ or -1) by employing the ranking information. A cauchy quantization loss is used in [75] to improve the discriminative power of binary descriptors. An iterative quantization approach is used to convert the features into binary codes to avoid the quantization loss [76]. Binary hash code is also used for clothing image retrieval [181]. The binary description is learnt through the supervised [96], [97], [101], unsupervised [112], [113], [120] and self-supervised [178] deep learning techniques. Among the generative approaches, a binary generative adversarial network (BGAN) is used to learn the binary code [121]. At the same time a regularized GAN is used by maximizing the entropy of binarized layer for image retrieval [150]. The GAN is trained in unsupervised mode [120] to learn the binary codes for image retrieval. In 2020, the binary GAN [166] is used for image retrieval and compression jointly.

B. Real-Valued Descriptors

The binary hashing approaches have the obvious shortcomings. First, it is difficult to represent the fine-grained similarity using binary code. Second, the generation of similar binary codes is common even for different images. Thus, researchers have also used the real-valued features to represent the images for the retrieval. The siamese networks have been extensively used to learn the real-valued feature descriptor for image retrieval [72], [132], [136]. In 2018, part-based CNN

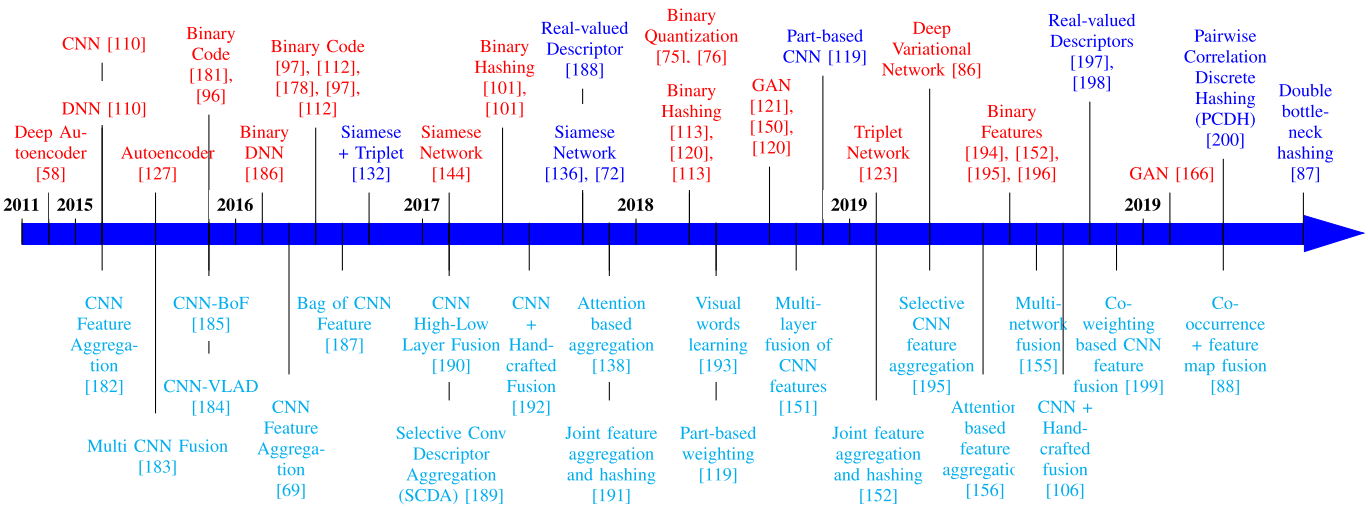

Fig. 8. A chronological view of deep learning based image retrieval methods depicting the different type of descriptors. The binary and real-valued feature vector based models are presented in Red and Blue colors, respectively. The feature aggregation based models are presented in Cyan color.

features are utilized to extract a non-binary hash code [119]. The real-valued descriptors generated using CNNs are used for medical image retrieval [188], [197] and cross-modal retrieval [198]. Chen *et al.* (2020) [200] have developed a pairwise correlation discrete hashing (PCDH) by exploiting the pairwise correlation of deep features for image retrieval. Shen *et al.* [87] have also used the real-valued descriptors with the help of double bottleneck hashing approach for image retrieval.

C. Aggregation of Descriptors

Several researchers have also tried to combine/fuse the feature at different stages of the network or multiple networks to generate the aggregation of descriptors for image retrieval [69], [182], [185]. Different strategies have been exercised for aggregation of features, such as vector locally aggregated descriptors (VLAD) [184] on the features extracted from different layers; bag of local convolutional features [187]; selective convolutional descriptor aggregation [189], [195]; fusion of multi-layer features [151], [190]; part-based weighting aggregation [119]; joint training of feature aggregation and hashing [191]; learning of feature aggregation and hash function in a joint manner [152]; and co-weighting based CNN feature fusion [199]. The features from different CNNs are also integrated for image retrieval [138], [156], [183]. One main sub-network and other attention-based sub-network are also fused at the last fully connected layer in [155]. The hand-designed features are fused with CNNs [106], [192]. Recently, Forcen *et al.* [88] have generated the image representation by combining a co-occurrence map with the feature map for image retrieval.

D. Summary

The followings are the summary of deep learning based approaches from the perspective of the type of feature descriptor:

- In order to facilitate the large-scale image retrieval, the compact and binary hash codes are generated using different networks. Different methods try to improve the discriminative ability, lower redundancy among bits, generalization of the binary hash code, etc. in different supervision modes.
- The real-valued descriptors concentrate over the discriminative ability of the learnt features for image retrieval at the cost of increased computational complexity for feature matching. Such methods try to increase the robustness and reduce the dimensionality of the descriptors.
- Feature aggregation approaches try to utilize the complementary information between the features of different networks, the features of different sub-network, and the features of different layers of same network to improve the image retrieval performance.

VII. Retrieval Type

Various retrieval types have been explored using deep learning approaches based on the nature of the problem and data as discussed in this section.

A. Cross-Modal Retrieval

The cross-modal retrieval refers to the image retrieval involving more than one modality by measuring the similarity between heterogeneous data objects. Feng *et al.* (2014) have introduced a correspondence autoencoder (Corr-AE) network for cross-modal retrieval [125]. In 2016 [201], a deep visual-semantic hashing (DVSH) network is developed for sentence and image based cross-modal retrieval by jointly learning the embeddings for images and sentences. Textual-visual deep binaries (TVDB) model represents the long descriptive sentences along with its corresponding informative images [139]. The CNN visual features have been also exploited for cross-modal retrieval, such as CNN off-the-shelf features for labelled annotation [134], CNN features with bi-directional hinge loss [202], and pairwise constraints based deep hashing network [142]. The adversarial neural network is also employed for cross-modal retrieval, such as adversarial cross-modal retrieval (ACMR) [143], self-supervised

adversarial hashing (SSAH) [146], attention-aware deep adversarial hashing (ADAH) [145], adversary guided asymmetric hashing (AGAH) [160], deep multi-level semantic hashing (DMSH) [198], and teacher-student learning [203].

B. Sketch Based Image Retrieval

Sketch based image retrieval (SBIR) is a special case of cross-modal retrieval where the query image is in the sketch domain the retrieval has to be performed in the image domain [204]. In 2017, a fine-grained SBIR (FG-SBIR) [137] is explored with the help of attention module and higher-order learnable energy function loss. Liu *et al.* [205] have introduced a semi-heterogeneous deep sketch hashing (DSH) model for SBIR by utilizing the representation of free-hand sketches. The sketches and natural photos are mapped in multiple layers in a deep CNN framework in [151] for SBIR. A zero-shot SBIR (ZS-SBIR) is proposed for retrieval of photos from unseen categories [153]. Wang *et al.* [158] have proposed a CNN based SBIR re-ranking approach to refine the retrieval results. The generative adversarial networks have been also exploited extensively for SBIR, such as generative domain-migration hashing (GDH) using cycle consistency loss [148], class sketch conditioned generative model [159], semantically aligned paired cycle-consistent generative model [206], and stacked adversarial network [163].

C. Multi-Label Image Retrieval

Multi-label retrieval involves multiple categorical labels while generating the image representations for image retrieval. Several deep learning approaches have been investigated for multi-label image retrieval using different strategies, such as multilevel similarity information [207], multilevel semantic similarity preserving hashing [208], multi-label annotations [146], category-aware object based hashing [209], [210], and fine-grained features for multilevel similarity hashing [211]. Readers may refer to the survey of multi-label image retrieval [43] published in 2020 for wider aspects and developments.

D. Instance Retrieval

In 2015, Razavian *et al.* have developed a baseline for deep CNN based visual instance retrieval [212]. An instance-aware image representations for multi-label image data by modeling the features of one category in a group is proposed in [209]. Other approaches for image instance retrieval includes bags of local convolutional features [187], learning global representations [65], and group invariant deep representation [213]. In 2020, Chen *et al.* have proposed a deep multiple-instance ranking based hashing (DMIRH) model for multi-label image retrieval by employing the category-aware bag of feature [210]. More details about image instance retrieval can be found in the survey compiled in [42].

E. Object Retrieval

The object retrieval aims to perform the retrieval based on the features derived from the specific objects in the image. In 2014, Sun *et al.* have extracted the CNN features from the region of interest detected through object detection technique for object based retrieval [126]. Several deep learning models have been investigated for object retrieval, such as integral image driven max-pooling on CNN activations [214], pooling of the relevant features based on the region proposal network [65], replicator equation based simultaneous selection and weighting of the primitive deep CNN features [215], co-weighting based aggregation of the semantic CNN features [199], and consideration of spatial and channel contribution to improve the region detection [216]. Gao *et al.* [165] have performed the 3D object retrieval with the help of a multi-view discrimination and pairwise CNN (MDPCNN) network.

F. Semantic Retrieval

In 2016, Yao *et al.* [131] have introduced a deep semantic preserving and ranking-based hashing (DSRH) method by exploiting the hash and classification losses. Similar losses are also used in [217]. A deep visual-semantic quantization (DVSQ) [218] is used by jointly learning the visual-semantic embeddings and quantizers. An adaptive Gaussian filter based aggregation of CNN features is used in [199] to exploit the semantic information. Semantic hashing has been also extensively performed for sketch based image retrieval [137], [153], [158], [206], cross-modal retrieval [139], [198], [201]. Other notable deep learning based works that model the semantic information include Multi-label retrieval [207], unsupervised image retrieval [219], supervised image retrieval [102], and semi-supervised image retrieval [141]. Semantic similarity in Hamming space based deep position-aware hashing (DPAH) [89] and semantic affinity deep semantic reconstruction hashing (DSRH) [162] are the recent methods for semantic retrieval.

G. Fine-Grained Image Retrieval

In order to increase the discriminative ability of the deep learnt descriptors, many researchers have utilized the fine-grained constraints in deep networks. Different works have incorporated the fine-grained property using different approaches, such as capturing the inter-class and intra-class image similarities using a siamese network [62], attention modules based incorporation of the spatial-semantic information [137], using selective CNN features [189], fine-grained ranking using the weighted Hamming distance [220], using the multilevel semantic similarity between multi-label image pairs [211], and using a piecewise cross entropy loss [221].

H. Asymmetric Quantization Based Retrieval

In 2017, Wu *et al.* have performed the online asymmetric similarity learning to preserve the similarity between heterogeneous data [202]. An asymmetric deep supervised hashing (ADSH) is used by learning the deep hash function only for query images, while the hash codes for gallery images are directly learned [104]. In 2019, Yang *et al.* have investigated an asymmetric deep semantic quantization (ADSQ) using three stream networks to model the heterogeneous

data [222]. A similarity preserving deep asymmetric quantization (SPDAQ) is proposed by exploiting the image subset and the label information of all the database items [223]. An adversary guided asymmetric hashing (AGAH) is introduced in [160] with the help of adversarial learning guided multi-label attention module for cross-modal image retrieval.

I. Summary

Based on the progress in image retrieval using deep learning methods for different retrieval types, following are the outlines drawn from this section:

- The cross-modal retrieval approaches learn the joint features for multiple modality using different networks. The recent methods utilize of the adversarial network for cross-modal retrieval. The similar observation and trend has been also witnessed for sketch based image retrieval.
- The multi-label and instance retrieval approaches are generally useful where more than one type of visual scenarios is present in the image. The deep learning based approaches are able to handle such retrieval by facilitating the feature learning through different type of networks.
- The region proposal network based feature selection has been employed by the existing deep learning methods for the object retrieval.
- The semantic information of the image has been used by different networks through abstract features to enhance the semantic image retrieval. The reconstruction based network is more suitable for semantic preserving hashing.
- Different feature selection and aggregation based networks have been utilized for fine-grained image retrieval.
- The asymmetric hashing has also shown the suitability of deep learning models by processing the query and gallery images with different networks.

VIII. Miscellaneous

This section covers the deep learning models for retrieval in terms of the different losses, applications and other aspects.

A. Progress in Retrieval Loss

A siamese based loss function is used in [132] by Kumar *et al.* (2016) for minimizing the global loss leading to discriminative feature learning. Zhou *et al.* (2017) have used the triplet quantization loss for deep hashing, which is based on the similarity between the anchor-positive pairs and anchor-negative pairs [135]. A listwise loss has been employed by Revaud *et al.* in 2019 [224] to directly optimize the global mean average precision in end-to-end deep learning. In 2020, a piecewise cross entropy loss function is used in [221] for fine-grained image retrieval. Several innovative losses have been used by the different feature learning approaches such as a lifted structured loss [66] and ranking loss [147].

B. Applications

The deep learning based approaches have been utilized for image retrieval pertaining to different applications such as cloth retrieval [181], biomedical image retrieval [197], face retrieval [226], [227], remote sensing image retrieval [124], landmark retrieval [228], social image retrieval [229], and video retrieval [178].

TABLE II

Mean Average Precision (mAP) With 5000 Retrieved Images (mAP@5000) in % for Different Deep Learning Based Image Retrieval Approaches Over NUS-WIDE, MS COCO and CIFAR-10 Datasets. Note That 2^{nd} Column List the Reference From Where the Results of Corresponding Approach Are Considered. Followings Are the Used Acronyms for Different Network Types in the Results: DNN - Deep Neural Network, CNN - Convolutional Neural Network, SN - Siamese Network, TN - Triplet Network, GAN - Generative Adversarial Network, DQN - Deep Q Network, PTN - Parametric Transformation Network, DVN - Deep Variational Networks, and AE - Autoencoder

Method Name	Net. Type	Result Source	NUS-WIDE 16 Bits	NUS-WIDE 32 Bits	NUS-WIDE 64 Bits	MS COCO 16 Bits	MS COCO 32 Bits	MS COCO 64 Bits
CNNH'14 [95]	CNN	[71]	57.0	58.3	60.0	56.4	57.4	56.7
SDH'15 [96]	PTN	[71]	47.6	55.5	58.1	55.5	56.4	58.0
DNNH'15 [63]	DNN	[71]	59.8	61.6	63.9	59.3	60.3	61.0
DHN'16 [67]	CNN	[71]	63.7	66.4	67.1	67.7	70.1	69.4
HashNet'17 [71]	CNN	[71]	66.2	69.9	71.6	68.7	71.8	73.6
DeepBit'16 [112]	CNN	[87]	39.2	40.3	42.9	40.7	41.9	43.0
BGAN'18 [121]	GAN	[87]	68.4	71.4	73.0	64.5	68.2	70.7
GreedyHash'18 [76]	CNN	[87]	63.3	69.1	73.1	58.2	66.8	71.0
BinGAN'18 [150]	GAN	[87]	65.4	70.9	71.3	65.1	67.3	69.6
DVB'19 [86]	DVN	[87]	60.4	63.2	66.5	57.0	62.9	62.3
DistillHash'19 [82]	SN	[87]	66.7	67.5	67.7	-	-	-
TBH'20 [87]	AE	[87]	71.7	72.5	73.5	70.6	73.5	72.2
CNNH'14 [95]	CNN	[84]	57.0	58.3	60.0	56.4	57.4	56.7
DNNH'15 [63]	DNN	[84]	59.8	61.6	63.9	59.3	60.3	61.0
DHN'16 [67]	CNN	[84]	63.7	66.4	67.1	67.7	70.1	69.4
HashNet'17 [71]	CNN	[84]	66.3	69.9	71.6	68.7	71.8	73.6
DHA'19 [84]	CNN	[84]	66.9	70.6	72.7	70.8	73.1	75.2
HashGAN'18 [120]	GAN	[120]	71.5	73.7	74.8	69.7	72.5	74.4
UH-BDNN'16 [186]	DNN	[123]	59.2	59.0	61.0	-	-	-
UTH'17 [117]	TN	[123]	54.3	53.7	54.7	-	-	-
UDTH'19 [123]	TN	[123]	64.4	67.7	69.6	-	-	-
SSDH'17 [102]	CNN	[89]	-	-	-	69.7	72.5	74.4
DPAH'20 [89]	PTN	[89]	-	-	-	73.3	76.8	**78.2**
DRDH'20 [167]	DQN	[167]	80.5	81.7	**81.8**	71.5	74.8	76.1
DVSQ'17 [218]	CNN	[223]	79.0	79.7	-	71.2	72.0	-
DTQ'18 [179]	TN	[223]	79.8	80.1	-	76.0	76.7	-
SPDAQ'19 [223]	CNN	[223]	**84.2**	85.1	-	**84.4**	**84.7**	-
DSQ'19 [81]	CNN	[81]	77.9	79.0	79.9	-	-	-
			CIFAR-10 Dataset					
			-	12 Bits	24 Bits	32 Bits	48 Bits	-
SDH'15 [96]	PTN	[225]	-	45.4	63.3	65.1	66.0	-
DSH'16 [97]	CNN	[225]	-	64.4	74.2	77.0	79.9	-
DHN'16 [67]	CNN	[225]	-	68.1	72.1	72.3	73.3	-
DPSH'16 [98]	SN	[225]	-	68.2	72.0	73.4	74.6	-
DQN'16 [68]	CNN	[225]	-	55.4	55.8	56.4	58.0	-
DSDH'17 [101]	CNN	[225]	-	74.0	78.6	80.1	82.0	-
ADSH'18 [104]	CNN	[225]	-	89.0	92.8	93.1	93.9	-
DIHN2+ADSH'19 [80]	CNN	[225]	-	89.8	92.9	92.9	93.9	-
DTH'20 [225]	CNN	[225]	-	**92.1**	**93.3**	**93.7**	**94.9**	-

C. Others

The hashing difficulty is also increased by generating the harder samples in a self-paced manner [161] to make the network training as reasoning oriented. In the initial work, the pre-trained CNN features have also very promising retrieval performance [61]. Recently, the transfer learning has been also utilized in [225] for deep transfer hashing.

TABLE III

MEAN AVERAGE PRECISION (mAP) WITH 1000 RETRIEVED IMAGES (mAP@1000) IN % FOR DIFFERENT DEEP LEARNING BASED IMAGE RETRIEVAL METHODS OVER IMAGENET, CIFAR-10 AND MNIST DATASETS

Method	Network Type	Result Source	16 Bits	32 Bits	48 Bits	64 Bits
			ImageNet Dataset			
CNNH'14 [95]	CNN	[71]	28.1	45.0	52.5	55.4
SDH'15 [96]	PTN	[71]	29.9	45.5	55.5	58.5
DNNH'15 [63]	DNN	[71]	29.0	46.1	53.0	56.5
DHN'16 [67]	CNN	[71]	31.1	47.2	54.2	57.3
HashNet'17 [71]	CNN	[71]	50.6	63.1	66.3	68.4
SSDH'17 [102]	CNN	[89]	63.4	69.2	70.1	70.7
DSQ'19 [81]	CNN	[81]	57.8	65.4	68.0	69.4
DPAH'20 [89]	PTN	[89]	**65.2**	**70.0**	**71.5**	**71.4**
			CIFAR-10 Dataset			
BGAN'18 [121]	GAN	[87]	52.5	53.1	-	56.2
GreedyHash'18 [76]	CNN	[87]	44.8	47.3	-	50.1
BinGAN'18 [150]	GAN	[87]	47.6	51.2	-	52.0
HashGAN'18 [120]	GAN	[87]	44.7	46.3	-	48.1
DVB'19 [86]	DVN	[87]	40.3	42.2	-	44.6
DistillHash'19 [82]	SN	[87]	28.4	28.5	-	28.8
TBH'20 [87]	AE	[87]	**53.2**	**57.3**	-	**57.8**
SDH'15 [110]	PTN	[110]	18.8	20.8	-	22.5
DAR'16 [111]	TN	[111]	16.8	17.0	-	17.2
DH'15 [110]	DNN	[117]	16.2	16.6	-	17.0
DeepBit'16 [112]	CNN	[117]	19.4	24.9	-	27.7
UTH'17 [117]	TN	[117]	28.7	30.7	-	32.4
DBD-MQ'17 [114]	CNN	[114]	21.5	26.5	-	31.9
UCBD'18 [113]	CNN	[113]	26.4	27.9	-	34.1
UH-BDNN'16 [186]	DNN	[123]	30.1	30.9	-	31.2
UDTH'19 [123]	TN	[123]	46.1	50.4	-	54.3
			MNIST Dataset			
SDH'15 [110]	PTN	[110]	**46.8**	**51.0**	-	**52.5**
DH'15 [110]	DNN	[117]	43.1	45.0	-	46.7
DeepBit'16 [112]	CNN	[117]	28.2	32.0	-	44.5
UTH'17 [117]	TN	[117]	43.2	46.6	-	49.9

TABLE IV

mAP@54000 AND mAP@ALL IN % FOR STATE-OF-THE-ART AND RECENT IMAGE RETRIEVAL METHODS OVER THE CIFAR-10 DATASET

Method Name	Network Type	Result Source	CIFAR-10 Dataset 16 Bits	24 Bits	32 Bits	48 Bits	64 Bits
			mAP@54000				
CNNH'14 [95]	CNN	[84]	47.6	-	47.2	48.9	50.1
DNNH'15 [63]	TN	[84]	55.9	-	55.8	58.1	58.3
SDH'15 [96]	PTN	[84]	46.1	-	52.0	55.3	56.8
DHN'16 [67]	CNN	[84]	56.8		60.3	62.1	63.5
HashNet'17 [71]	CNN	[84]	64.3	-	66.7	67.5	68.7
DHA'14 [84]	CNN	[84]	65.2		68.1	69.0	69.9
HashGAN'18 [120]	GAN	[120]	66.8	-	73.1	73.5	74.9
DTQ'18 [179]	TN	[179]	**78.9**	-	79.2	-	-
DRDH'20 [167]	DQN	[167]	78.7	-	**80.5**	**80.6**	**80.3**
			mAP@All				
DQN'16 [68]	CNN	[223]	-	55.8	56.4	58.0	-
DPSH'16 [98]	SN	[223]	-	72.7	74.4	75.7	-
DSDH'17 [101]	CNN	[223]	-	78.6	80.1	82.0	-
DTQ'18 [179]	DQN	[223]	-	79.0	79.2	-	-
DVSQ'17 [218]	CNN	[223]	-	80.3	80.8	81.1	-
SPDAQ'19 [223]	CNN	[223]	-	**88.4**	**89.1**	**89.3**	-
SSAH'19 [161]	GAN	[161]	-	87.8	-	88.6	-
DeepBit'19 [112]	CNN	[122]	22.0	-	24.1	-	29.0
BGAN'19 [121]	GAN	[122]	49.7	-	47.0	-	50.7
UADH'19 [122]	GAN	[122]	**67.7**	-	68.9	-	**69.6**
DSAH'19 [155]	CNN	[155]	-	84.1	84.5	84.9	-

D. Summary

Researchers have come up with various loss functions to facilitate the discriminative learning of features by the networks for image retrieval. The losses constraint and guide the training of the deep learning models. The image retrieval has shown a great utilization with its application to solve the real-life problems. Researchers have also tried to understand what works and what not for deep learning based image retrieval. The transfer learning has been also utilized for retrieval.

IX. PERFORMANCE COMPARISON

This survey also presents a performance analysis for the state-of-the-art deep learning based image retrieval approaches. The Mean Average Precision (mAP) reported for the different image retrieval approaches is summarized in Table II, III, and IV. The mAP@5000 (i.e., 5000 retrieved images) using various existing deep learning approaches is summarized in Table II over CIFAR-10, NUS-WIDE and MS COCO datasets. The results over CIFAR-10, ImageNet and MNIST datasets using different state-of-the-art deep learning based image retrieval methods are compiled in Table III in terms of the mAP@1000. The mAP@54000 using few methods is reported in Table IV over the CIFAR-10 dataset. The standard mAP is also depicted in Table IV by considering all the retrieved images for CIFAR-10 dataset using some of the available literature. Note that 2nd column in Table II, III, and IV list the source reference of the corresponding method reported results. Following are the observations out of these results by deep learning methods:

- Recently proposed Deep Transfer Hashing (DTH) by Zhai *et al.* [225] have shown outstanding performance over CIFAR-10 and NUS-WIDE datasets in terms of the mAP@5000. Other promising methods include Deep Spatial Attention Hashing (DSAH) by Ge *et al.* [155], Similarity Preserving Deep Asymmetric Quantization (SPDAQ) by Chen *et al.* [223], Deep Position-Aware Hashing (DPAH) by Wang *et al.* [89] and Deep Reinforcement De-Redundancy Hashing (DRDH) by Yang *et al.* [167].
- The Twin-Bottleneck Hashing (TBH) introduced by Shen *et al.* [87] is also observed as an appealing method using autoencoder having a double bottleneck over the CIFAR-10 dataset in terms of the mAP@1000. However, the Deep Position-Aware Hashing (DPAH) investigated by Wang *et al.* [89] have outperformed the other approaches over ImageNet dataset. Supervised Deep Hashing (SDH) by Erin *et al.* [110] has depicted appealing performance over the MNIST dataset.
- The deep reinforcement learning based image retrieval model, namely Deep Reinforcement De-Redundancy Hashing (DRDH) by Yang *et al.* [167], is one of recent breakthrough as supported by superlative mAP@54000 over the CIFAR-10 dataset. The Deep Triplet Quantization [179] is also one of the favourable model for feature learning.

- The Similarity Preserving Deep Asymmetric Quantization (SPDAQ) by Chen *et al.* [223] and Unsupervised ADversarial Hashing (UADH) by Deng *et al.* [122] methods have been also identified as very encouraging based on the mAP by considering all the retrieved images over the CIFAR-10 dataset.

X. Conclusion and Future Directives

A. Conclusion and Trend

This paper presents a comprehensive survey of deep learning methods for content based image retrieval. As most of the deep learning based developments are recent, this survey majorly focuses over the image retrieval methods using deep learning in a decade from 2011 to 2020. A detailed taxonomy is presented in terms of different supervision type, different networks used, different data type of descriptors, different retrieval type and other aspects. The detailed discussion under each section is also presented with the further categorization. A chronological summarization is presented to show the evolution of the deep learning models for image retrieval. Moreover, the chronological overview is also portrayed under each category to showcase the growth of image retrieval approaches. A summary of large-scale common datasets used for image retrieval is also compiled in this survey. A performance analysis of the state-of-the-art deep learning based image retrieval methods is also conducted in terms of the mean average precision for different no. of retrieved images.

The research trend in image retrieval suggests that the deep learning based models are driving the progress. The recently developed models such as generative adversarial networks, autoencoder networks and reinforcement learning networks have shown the superior performance for image retrieval. The discovery of better objective functions has been also the trend in order to constrain the learning of the hash code for discriminative, robust and efficient image retrieval. The semantic preserving class-specific feature learning using different networks and different quantization techniques is also the recent trend for image retrieval. Other trends include utilization of attention module, transfer learning, etc.

B. Future Directions

The future work in image retrieval using deep learning can include exploration of improved deep learning models, more relevant objective functions, minimum loss based quantization techniques, semantic preserving feature learning, and attention focused feature learning. The future direction in the image retrieval might be driven from the basic goal of the expected solution. There are three important aspects of any retrieval system, which include the discriminative ability, robustness capability and fast image search. In order to achieve the discriminative ability, the features corresponding to the samples of different class should be as far as possible. Thus, different approaches such as triplet based objective function, consideration of class distribution, incorporation of distance between class centroids, etc. can be exploited. In order to maintain the robustness property, various data augmentation, layer manipulation, siamese loss based objective functions, feature normalization, incorporation of class distribution and majority voting in the feature representation, etc. can be explored. In order to perform the faster image search, the learnt feature or hash code should be as low dimensional and compact as possible. Thus, better strategy for feature quantization and maximizing the relevant information into feature space in a compact way can be seen as one of the future directions. The self-supervised learning has shown very promising performance for different down-stream tasks and has potential to learn the important features in compact form. Thus, in future, the self-supervised learning can boost the performance of image retrieval models significantly.

References

[1] M. Flickner *et al.*, "Query by image and video content: The QBIC system," *Computer*, vol. 28, no. 9, pp. 23–32, 1995.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[3] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: Overview and proposals," *Pattern Recognit. Lett.*, vol. 22, no. 5, pp. 593–601, Apr. 2001.

[4] T. Deselaers, D. Keysers, and H. Ney, "Features for image retrieval: An experimental comparison," *Inf. Retr.*, vol. 11, no. 2, pp. 77–107, Apr. 2008.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[6] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[7] S. R. Dubey, S. K. Singh, and R. K. Singh, "Rotation and illumination invariant interleaved intensity order-based local descriptor," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5323–5333, Dec. 2014.

[8] I. J. Jacob, K. G. Srinivasagan, and K. Jayapriya, "Local oppugnant color texture pattern for image retrieval system," *Pattern Recognit. Lett.*, vol. 42, pp. 72–78, Jun. 2014.

[9] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 174–184, Jan. 2010.

[10] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

[11] S. Murala, R. P. Maheshwari, and R. Balasubramanian, "Local tetra patterns: A new feature descriptor for content-based image retrieval," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2874–2886, May 2012.

[12] S. R. Dubey, S. K. Singh, and R. K. Singh, "Local wavelet pattern: A new feature descriptor for image retrieval in medical CT databases," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5892–5903, Dec. 2015.

[13] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, Dec. 2013.

[14] K.-C. Fan and T.-Y. Hung, "A novel local pattern descriptor—Local vector pattern in high-order derivative space for face recognition," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2877–2891, Jul. 2014.

[15] S. R. Dubey, S. K. Singh, and R. K. Singh, "Multichannel decoded local binary patterns for content-based image retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4018–4032, Sep. 2016.

[16] S. Chakraborty, S. K. Singh, and P. Chakraborty, "Local gradient hexa pattern: A descriptor for face recognition and retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 1, pp. 171–180, Jan. 2018.

[17] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surveys*, vol. 40, no. 2, pp. 1–60, Apr. 2008.

[18] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," *ACM Comput. Surveys*, vol. 46, no. 3, pp. 1–38, Jan. 2014.

[19] W. Zhou, H. Li, and Q. Tian, "Recent advance in content-based image retrieval: A literature survey," 2017, *arXiv:1706.06064*. [Online]. Available: http://arxiv.org/abs/1706.06064

[20] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," 2013, *arXiv:1306.6709*. [Online]. Available: http://arxiv.org/abs/1306.6709

[21] J. Wang, T. Zhang, J. song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 769–790, Apr. 2018.

[22] S. C. H. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma, "Learning distance metrics with contextual constraints for image retrieval," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 2072–2078.

[23] H. Chang and D.-Y. Yeung, "Kernel-based distance metric learning for content-based image retrieval," *Image Vis. Comput.*, vol. 25, no. 5, pp. 695–703, May 2007.

[24] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang, "Supervised hashing with kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2074–2081.

[25] L. Yang *et al.*, "A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 30–44, Jan. 2010.

[26] J.-E. Lee, R. Jin, and A. K. Jain, "Rank-based distance metric learning: An application to image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[27] S. C. H. Hoi, W. Liu, and S.-F. Chang, "Semi-supervised distance metric learning for collaborative image retrieval and clustering," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 6, no. 3, pp. 1–26, Aug. 2010.

[28] M. Norouzi, D. J. Fleet, and R. R. Salakhutdinov, "Hamming distance metric learning," in *Proc. NIPS*, 2012, pp. 1061–1069.

[29] D. Song, W. Liu, and D. A. Meyer, "Fast structural binary coding," in *Proc. IJCAI*, 2016, pp. 2018–2024.

[30] D. Song, W. Liu, D. A. Meyer, D. Tao, and R. Ji, "Rank preserving hashing for rapid image search," in *Proc. Data Compress. Conf.*, Apr. 2015, pp. 353–362.

[31] D. Song, W. Liu, R. Ji, D. A. Meyer, and J. R. Smith, "Top rank supervised binary coding for visual search," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1922–1930.

[32] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 806–813.

[33] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 4, 2020, doi: 10.1109/TPAMI.2020.2992393.

[34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[35] C. H. Dagli, *Artificial Neural Networks for Intelligent Manufacturing*. Springer, 2012. [Online]. Available: https://www.springer.com/gp/book/9780412480508

[36] F. Amato, A. López, E. M. Peña-Méndez, P. Vanhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," *J. Appl. Biomed.*, vol. 11, no. 2, pp. 47–58, 2013.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[39] M. Sundermeyer, R. Schlüter, and H. Ney, "Lstm neural networks for language modeling," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 194–197.

[40] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *Proc. ICML*, 2015, pp. 2067–2075.

[41] J. Wan *et al.*, "Deep learning for content-based image retrieval: A comprehensive study," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 157–166.

[42] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May 2018.

[43] J. Rodrigues, M. Cristo, and J. G. Colonna, "Deep hashing for multi-label image retrieval: A survey," *Artif. Intel. Review*, vol. 53, no. 7, pp. 1–47, 2020.

[44] X. Luo *et al.*, "A survey on deep hashing methods," 2020, *arXiv:2003.03369*. [Online]. Available: http://arxiv.org/abs/2003.03369

[45] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2009.

[46] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world Web image database from national university of singapore," in *Proc. ACM Int. Conf. Image Video Retr. (CIVR)*, 2009, pp. 1–9.

[47] 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.

[48] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. NIPS*, 2011.

[49] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3485–3492.

[50] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 192–199.

[51] K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen, "Deep learning of binary hash codes for fast image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 27–35.

[52] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[53] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[54] M. J. Huiskes, B. Thomee, and M. S. Lew, "New trends and ideas in visual concept detection: The MIR flickr retrieval evaluation initiative," in *Proc. Int. Conf. Multimedia Inf. Retr. (MIR)*, 2010, pp. 527–536.

[55] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3456–3465.

[56] T. Weyand, A. Araujo, B. Cao, and J. Sim, "Google landmarks dataset v2—A large-scale benchmark for instance-level recognition and retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2575–2584.

[57] X.-S. Hua *et al.*, "Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines," in *Proc. 21st ACM Int. Conf. Multimedia (MM)*, 2013, pp. 243–252.

[58] A. Krizhevsky and G. E. Hinton, "Using very deep autoencoders for content-based image retrieval," in *Proc. ESANN*, vol. 1, 2011, p. 2.

[59] Y. Kang, S. Kim, and S. Choi, "Deep learning to hash with multiple representations," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 930–935.

[60] 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. 21st ACM Int. Conf. Multimedia (MM)*, 2013, pp. 153–162.

[61] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. ECCV*, 2014, pp. 584–599.

[62] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1386–1393.

[63] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3270–3278.

[64] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.

[65] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *Proc. ECCV*, 2016, pp. 241–257.

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

[67] H. Zhu, M. Long, J. Wang, and Y. Cao, "Deep hashing network for efficient similarity retrieval," in *Proc. AAAI*, 2016, pp. 2415–2421.

[68] Y. Cao, M. Long, J. Wang, H. Zhu, and Q. Wen, "Deep quantization network for efficient image retrieval," in *Proc. AAAI*, 2016, pp. 3457–3463.
[69] S. S. Husain and M. Bober, "Improving large-scale image retrieval through robust aggregation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1783–1796, Sep. 2017.
[70] G. Zhong, H. Xu, P. Yang, S. Wang, and J. Dong, "Deep hashing learning networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 2236–2243.
[71] Z. Cao, M. Long, J. Wang, and P. S. Yu, "HashNet: Deep learning to hash by continuation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5608–5617.
[72] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *Int. J. Comput. Vis.*, vol. 124, no. 2, pp. 237–254, Sep. 2017.
[73] T. Hoang, T.-T. Do, D.-K. Le Tan, and N.-M. Cheung, "Selective deep convolutional features for image retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1600–1608.
[74] A. Alzu'bi, A. Amira, and N. Ramzan, "Content-based image retrieval with compact deep convolutional features," *Neurocomputing*, vol. 249, pp. 95–105, Aug. 2017.
[75] Y. Cao, M. Long, B. Liu, and J. Wang, "Deep cauchy hashing for Hamming space retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1229–1237.
[76] S. Su, C. Zhang, K. Han, and Y. Tian, "Greedy hash: Towards fast optimization for accurate hash coding in CNN," in *Proc. NIPS*, 2018, pp. 798–807.
[77] X. Yuan, L. Ren, J. Lu, and J. Zhou, "Relaxation-free deep hashing via policy gradient," in *Proc. ECCV*, 2018, pp. 134–150.
[78] Z. Chen, X. Yuan, J. Lu, Q. Tian, and J. Zhou, "Deep hashing via discrepancy minimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6838–6847.
[79] D. Wu, J. Liu, B. Li, and W. Wang, "Deep index-compatible hashing for fast image retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.
[80] D. Wu, Q. Dai, J. Liu, B. Li, and W. Wang, "Deep incremental hashing network for efficient image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9069–9077.
[81] S. Eghbali and L. Tahvildari, "Deep spherical quantization for image search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11690–11699.
[82] E. Yang, T. Liu, C. Deng, W. Liu, and D. Tao, "DistillHash: Unsupervised deep hashing by distilling data pairs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2946–2955.
[83] J. Bai *et al.*, "Deep progressive hashing for image retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3178–3193, Dec. 2019.
[84] J. Xu, C. Guo, Q. Liu, J. Qin, Y. Wang, and L. Liu, "DHA: Supervised deep learning to hash with an adaptive loss function," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3054–3062.
[85] Y. Shen, J. Qin, J. Chen, L. Liu, F. Zhu, and Z. Shen, "Embarrassingly simple binary representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2883–2892.
[86] Y. Shen, L. Liu, and L. Shao, "Unsupervised binary representation learning with deep variational networks," *Int. J. Comput. Vis.*, vol. 127, nos. 11–12, pp. 1614–1628, Dec. 2019.
[87] Y. Shen *et al.*, "Auto-encoding twin-bottleneck hashing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2818–2827.
[88] J. I. Forcen, M. Pagola, E. Barrenechea, and H. Bustince, "Co-occurrence of deep convolutional features for image search," *Image Vis. Comput.*, vol. 97, May 2020, Art. no. 103909.
[89] R. Wang, R. Wang, S. Qiao, S. Shan, and X. Chen, "Deep position-aware hashing for semantic continuous image retrieval," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2493–2502.
[90] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. ACMMM Inf. Retr.*, 2008, pp. 39–43.
[91] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva, "SUN database: Exploring a large collection of scene categories," *Int. J. Comput. Vis.*, vol. 119, no. 1, pp. 3–22, Aug. 2016.
[92] Y. Pan, T. Yao, T. Mei, H. Li, C.-W. Ngo, and Y. Rui, "Click-through-based cross-view learning for image search," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2014, pp. 717–726.
[93] J. Yu, X. Yang, F. Gao, and D. Tao, "Deep multimodal distance metric learning using click constraints for image ranking," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4014–4024, Dec. 2017.
[94] Y. Li, Y. Pan, T. Yao, H. Chao, Y. Rui, and T. Mei, "Learning click-based deep structure-preserving embeddings with visual attention," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 3, pp. 1–19, Sep. 2019.
[95] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," in *Proc. AAAI*, 2014, pp. 2156–2162.
[96] F. Shen, C. Shen, W. Liu, and H. T. Shen, "Supervised discrete hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 37–45.
[97] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2064–2072.
[98] W.-J. Li, S. Wang, and W.-C. Kang, "Feature learning based deep supervised hashing with pairwise labels," in *Proc. IJCAI*, 2016, pp. 1711–1717.
[99] X. Wang, Y. Shi, and K. M. Kitani, "Deep supervised hashing with triplet labels," in *Proc. ACCV*, 2016, pp. 70–84.
[100] Z. Zhang, Y. Chen, and V. Saligrama, "Efficient training of very deep neural networks for supervised hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1487–1495.
[101] Q. Li, Z. Sun, R. He, and T. Tan, "Deep supervised discrete hashing," in *Proc. NIPS*, 2017, pp. 2482–2491.
[102] H.-F. Yang, K. Lin, and C.-S. Chen, "Supervised learning of semantics-preserving hash via deep convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 437–451, Feb. 2018.
[103] J. Lu, V. E. Liong, and J. Zhou, "Deep hashing for scalable image search," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2352–2367, May 2017.
[104] Q.-Y. Jiang and W.-J. Li, "Asymmetric deep supervised hashing," in *Proc. AAAI*, 2018, pp. 763–771.
[105] B. Klein and L. Wolf, "End-to-end supervised product quantization for image search and retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5041–5050.
[106] S. Kan, Y. Cen, Z. He, Z. Zhang, L. Zhang, and Y. Wang, "Supervised deep feature embedding with handcrafted feature," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5809–5823, Dec. 2019.
[107] W. W. Y. Ng, J. Li, X. Tian, H. Wang, S. Kwong, and J. Wallace, "Multi-level supervised hashing with deep features for efficient image retrieval," *Neurocomputing*, vol. 399, pp. 171–182, Jul. 2020.
[108] C. Zhou *et al.*, "Angular deep supervised hashing for image retrieval," *IEEE Access*, vol. 7, pp. 127521–127532, 2019.
[109] J. Li, W. W. Ng, X. Tian, S. Kwong, and H. Wang, "Weighted multi-deep ranking supervised hashing for efficient image retrieval," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 4, pp. 1–15, 2019.
[110] V. E. Liong, J. Lu, G. Wang, P. Moulin, and J. Zhou, "Deep hashing for compact binary codes learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2475–2483.
[111] C. Huang, C. C. Loy, and X. Tang, "Unsupervised learning of discriminative attributes and visual representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5175–5184.
[112] K. Lin, J. Lu, C.-S. Chen, and J. Zhou, "Learning compact binary descriptors with unsupervised deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1183–1192.
[113] K. Lin, J. Lu, C.-S. Chen, J. Zhou, and M.-T. Sun, "Unsupervised deep learning of compact binary descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1501–1514, Jun. 2019.
[114] Y. Duan, J. Lu, Z. Wang, J. Feng, and J. Zhou, "Learning deep binary descriptor with multi-quantization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1183–1192.
[115] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," in *Proc. ECCV*, 2016, pp. 3–20.
[116] M. Paulin, J. Mairal, M. Douze, Z. Harchaoui, F. Perronnin, and C. Schmid, "Convolutional patch representations for image retrieval: An unsupervised approach," *Int. J. Comput. Vis.*, vol. 121, no. 1, pp. 149–168, Jan. 2017.
[117] S. Huang, Y. Xiong, Y. Zhang, and J. Wang, "Unsupervised triplet hashing for fast image retrieval," in *Proc. Thematic Workshops ACM Multimedia Thematic Workshops*, 2017, pp. 84–92.
[118] F. Shen, Y. Xu, L. Liu, Y. Yang, Z. Huang, and H. T. Shen, "Unsupervised deep hashing with similarity-adaptive and discrete optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3034–3044, Dec. 2018.

[119] J. Xu, C. Shi, C. Qi, C. Wang, and B. Xiao, "Unsupervised part-based weighting aggregation of deep convolutional features for image retrieval," in *Proc. AAAI*, 2018, pp. 7436–7443.

[120] K. G. Dizaji, F. Zheng, N. S. Nourabadi, Y. Yang, C. Deng, and H. Huang, "Unsupervised deep generative adversarial hashing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3664–3673.

[121] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, and H. T. Shen, "Binary generative adversarial networks for image retrieval," in *Proc. AAAI*, 2018, pp. 394–401.

[122] C. Deng, E. Yang, T. Liu, J. Li, W. Liu, and D. Tao, "Unsupervised semantic-preserving adversarial hashing for image search," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4032–4044, Aug. 2019.

[123] Y. Gu, H. Zhang, Z. Zhang, and Q. Ye, "Unsupervised deep triplet hashing with pseudo triplets for scalable image retrieval," *Multimedia Tools Appl.*, vol. 79, pp. 35253–35274, May 2019.

[124] Y. Liu, L. Ding, C. Chen, and Y. Liu, "Similarity-based unsupervised deep transfer learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7872–7889, Nov. 2020.

[125] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 7–16.

[126] S. Sun, W. Zhou, H. Li, and Q. Tian, "Search by detection: Object-level feature for image retrieval," in *Proc. Int. Conf. Internet Multimedia Comput. Service (ICIMCS)*, 2014, pp. 46–49.

[127] M. A. Carreira-Perpinán and R. Raziperchikolaei, "Hashing with binary autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 557–566.

[128] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4353–4361.

[129] 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. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3279–3286.

[130] B. Zhuang, G. Lin, C. Shen, and I. Reid, "Fast training of triplet-based deep binary embedding networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5955–5964.

[131] T. Yao, F. Long, T. Mei, and Y. Rui, "Deep semantic-preserving and ranking-based hashing for image retrieval," in *IJCAI*, vol. 1, 2016, p. 4.

[132] B. G. Vijay Kumar, G. Carneiro, and I. Reid, "Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5385–5394.

[133] J. Lin, O. Morere, J. Petta, V. Chandrasekhar, and A. Veillard, "Tiny descriptors for image retrieval with unsupervised triplet hashing," in *Proc. DCC*, 2016, pp. 397–406.

[134] Y. Wei *et al.*, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

[135] Y. Zhou, S. Huang, Y. Zhang, and Y. Wang, "Deep hashing with triplet quantization loss," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[136] E.-J. Ong, S. Husain, and M. Bober, "Siamese network of deep Fisher-vector descriptors for image retrieval," 2017, *arXiv:1702.00338*. [Online]. Available: http://arxiv.org/abs/1702.00338

[137] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5551–5560.

[138] F. Yang, J. Li, S. Wei, Q. Zheng, T. Liu, and Y. Zhao, "Two-stream attentive CNNs for image retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1513–1521.

[139] Y. Shen, L. Liu, L. Shao, and J. Song, "Deep binaries: Encoding semantic-rich cues for efficient textual-visual cross retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4097–4106.

[140] X. Lu, Y. Chen, and X. Li, "Hierarchical recurrent neural hashing for image retrieval with hierarchical convolutional features," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 106–120, Jan. 2018.

[141] J. Zhang and Y. Peng, "SSDH: Semi-supervised deep hashing for large scale image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 212–225, Jan. 2019.

[142] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *Proc. AAAI*, 2017, pp. 1618–1625.

[143] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 154–162.

[144] A. Jose, S. Yan, and I. Heisterklaus, "Binary hashing using siamese neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2916–2920.

[145] X. Zhang, H. Lai, and J. Feng, "Attention-aware deep adversarial hashing for cross-modal retrieval," in *Proc. ECCV*, 2018, pp. 591–606.

[146] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4242–4251.

[147] G. Wang, Q. Hu, J. Cheng, and Z. Hou, "Semi-supervised generative adversarial hashing for image retrieval," in *Proc. ECCV*, 2018, pp. 469–485.

[148] J. Zhang *et al.*, "Generative domain-migration hashing for sketch-to-image retrieval," in *Proc. ECCV*, 2018, pp. 297–314.

[149] Y. Cao, B. Liu, M. Long, and J. Wang, "HashGAN: Deep learning to hash with pair conditional wasserstein GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1287–1296.

[150] M. Zieba, P. Semberecki, T. El-Gaaly, and T. Trzcinski, "BinGAN: Learning compact binary descriptors with a regularized GAN," in *Proc. NIPS*, 2018, pp. 3608–3618.

[151] D. Yu, Y. Liu, Y. Pang, Z. Li, and H. Li, "A multi-layer deep fusion convolutional neural network for sketch based image retrieval," *Neurocomputing*, vol. 296, pp. 23–32, Jun. 2018.

[152] T.-T. Do, K. Le, T. Hoang, H. Le, T. V. Nguyen, and N.-M. Cheung, "Simultaneous feature aggregating and hashing for compact binary code learning," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4954–4969, Oct. 2019.

[153] S. Dey, P. Riba, A. Dutta, J. L. Llados, and Y.-Z. Song, "Doodle to search: Practical zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2179–2188.

[154] L.-K. Huang, J. Chen, and S. Pan, "Accelerate learning of deep hashing with gradient attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5271–5280.

[155] L.-W. Ge, J. Zhang, Y. Xia, P. Chen, B. Wang, and C.-H. Zheng, "Deep spatial attention hashing network for image retrieval," *J. Vis. Commun. Image Represent.*, vol. 63, Aug. 2019, Art. no. 102577.

[156] S. Wei, L. Liao, J. Li, Q. Zheng, F. Yang, and Y. Zhao, "Saliency inside: Learning attentive CNNs for content-based image retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4580–4593, Sep. 2019.

[157] Z. Chen, J. Lin, Z. Wang, V. Chandrasekhar, and W. Lin, "Beyond ranking loss: Deep holographic networks for multi-label video search," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 879–883.

[158] L. Wang, X. Qian, Y. Zhang, J. Shen, and X. Cao, "Enhancing sketch-based image retrieval by CNN semantic re-ranking," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3330–3342, Jul. 2020.

[159] V. K. Verma, A. Mishra, A. Mishra, and P. Rai, "Generative model for zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 704–713.

[160] W. Gu, X. Gu, J. Gu, B. Li, Z. Xiong, and W. Wang, "Adversary guided asymmetric hashing for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2019, pp. 159–167.

[161] S. Jin *et al.*, "SSAH: Semi-supervised adversarial deep hashing with self-paced hard sample generation," 2019, *arXiv:1911.08688*. [Online]. Available: http://arxiv.org/abs/1911.08688

[162] Y. Wang, X. Ou, J. Liang, and Z. Sun, "Deep semantic reconstruction hashing for similarity retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 387–400, Jan. 2021.

[163] A. Pandey, A. Mishra, V. K. Verma, A. Mittal, and H. A. Murthy, "Stacked adversarial network for zero-shot sketch based image retrieval," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2540–2549.

[164] T. Ng, V. Balntas, Y. Tian, and K. Mikolajczyk, "SOLAR: Second-order loss and attention for image retrieval," 2020, *arXiv:2001.08972*. [Online]. Available: http://arxiv.org/abs/2001.08972

[165] Z. Gao, H. Xue, and S. Wan, "Multiple discrimination and pairwise CNN for view-based 3D object retrieval," *Neural Netw.*, vol. 125, pp. 290–302, May 2020.

[166] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, and H. T. Shen, "Unified binary generative adversarial network for image retrieval and compression," *Int. J. Comput. Vis.*, vol. 128, pp. 2243–2264, Feb. 2020.

[167] J. Yang, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Deep reinforcement hashing with redundancy elimination for effective image retrieval," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107116.

[168] Y. Pan, T. Yao, H. Li, C.-W. Ngo, and T. Mei, "Semi-supervised hashing with semantic confidence for large scale visual search," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 53–62.

[169] X. Yan, L. Zhang, and W.-J. Li, "Semi-supervised deep hashing with a bipartite graph," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3238–3244.

[170] Z. Qiu, Y. Pan, T. Yao, and T. Mei, "Deep semantic hashing with generative adversarial networks," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 225–234.

[171] S. Zhang, J. Li, and B. Zhang, "Pairwise teacher-student network for semi-supervised hashing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019.

[172] J. Tang and Z. Li, "Weakly supervised multimodal hashing for scalable social image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2730–2741, Oct. 2018.

[173] Z. Guan *et al.*, "Tag-based weakly-supervised hashing for image retrieval," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3776–3782.

[174] V. Gattupalli, Y. Zhuo, and B. Li, "Weakly supervised deep image hashing through tag embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10375–10384.

[175] Z. Li, J. Tang, L. Zhang, and J. Yang, "Weakly-supervised semantic guided hashing for social image retrieval," *Int. J. Comput. Vis.*, vol. 128, nos. 8–9, pp. 2265–2278, Sep. 2020.

[176] Q. Hu, J. Wu, J. Cheng, L. Wu, and H. Lu, "Pseudo label based unsupervised deep discriminative hashing for image retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1584–1590.

[177] X. Dong, L. Liu, L. Zhu, Z. Cheng, and H. Zhang, "Unsupervised deep K-means hashing for efficient image retrieval and clustering," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Nov. 4, 2020, doi: 10.1109/TCSVT.2020.3035775.

[178] H. Zhang, M. Wang, R. Hong, and T.-S. Chua, "Play and rewind: Optimizing binary representations of videos by self-supervised temporal hashing," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 781–790.

[179] B. Liu, Y. Cao, M. Long, J. Wang, and J. Wang, "Deep triplet quantization," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 755–763.

[180] H. Su, P. Wang, L. Liu, H. Li, Z. Li, and Y. Zhang, "Where to look and how to describe: Fashion image retrieval with an attentional heterogeneous bilinear network," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Oct. 30, 2020, doi: 10.1109/TCSVT.2020.3034981.

[181] K. Lin, H.-F. Yang, K.-H. Liu, J.-H. Hsiao, and C.-S. Chen, "Rapid clothing retrieval via deep learning of binary codes and hierarchical search," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 499–502.

[182] A. B. Yandex and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1269–1277.

[183] Y. Liu, Y. Guo, S. Wu, and M. S. Lew, "DeepIndex for accurate and efficient image retrieval," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 43–50.

[184] J. Y.-H. Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 53–61.

[185] T. Uricchio, M. Bertini, L. Seidenari, and A. D. Bimbo, "Fisher encoded convolutional bag-of-windows for efficient image retrieval and social image tagging," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 9–15.

[186] T.-T. Do, A.-D. Doan, and N.-M. Cheung, "Learning to hash with binary deep neural network," in *Proc. ECCV*, 2016, pp. 219–234.

[187] E. Mohedano, K. McGuinness, N. E. O'Connor, A. Salvador, F. Marques, and X. Giro-I-Nieto, "Bags of local convolutional features for scalable instance search," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2016, pp. 327–331.

[188] A. Qayyum, S. M. Anwar, M. Awais, and M. Majid, "Medical image retrieval using deep convolutional neural network," *Neurocomputing*, vol. 266, pp. 8–20, Nov. 2017.

[189] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2868–2881, Jun. 2017.

[190] W. Yu, K. Yang, H. Yao, X. Sun, and P. Xu, "Exploiting the complementary strengths of multi-layer CNN features for image retrieval," *Neurocomputing*, vol. 237, pp. 235–241, May 2017.

[191] T.-T. Do, D.-K. Le Tan, T. T. Pham, and N.-M. Cheung, "Simultaneous feature aggregating and hashing for large-scale image search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6618–6627.

[192] W. Zhou, H. Li, J. Sun, and Q. Tian, "Collaborative index embedding for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1154–1166, May 2018.

[193] X. Liu, S. Zhang, T. Huang, and Q. Tian, "E^2BoWs: An end-to-end bag-of-words model via deep convolutional neural network for image retrieval," *Neurocomputing*, vol. 395, pp. 188–198, Jun. 2020.

[194] T.-T. Do, T. Hoang, D.-K. Le Tan, A.-D. Doan, and N.-M. Cheung, "Compact hash code learning with binary deep neural network," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 992–1004, Apr. 2020.

[195] T.-T. Do, T. Hoang, D.-K.-L. Tan, H. Le, T. V. Nguyen, and N.-M. Cheung, "From selective deep convolutional features to compact binary representations for image retrieval," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 2, pp. 1–22, Jun. 2019.

[196] X. Lu, Y. Chen, and X. Li, "Discrete deep hashing with ranking optimization for image retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 6, pp. 2052–2063, Jun. 2020.

[197] S. R. Dubey, S. K. Roy, S. Chakraborty, S. Mukherjee, and B. B. Chaudhuri, "Local bit-plane decoded convolutional neural network features for biomedical image retrieval," *Neural Comput. Appl.*, vol. 32, pp. 7539–7551, Jun. 2019.

[198] Z. Ji, W. Yao, W. Wei, H. Song, and H. Pi, "Deep multi-level semantic hashing for cross-modal retrieval," *IEEE Access*, vol. 7, pp. 23667–23674, 2019.

[199] J. Zhu *et al.*, "Co-weighting semantic convolutional features for object retrieval," *J. Vis. Commun. Image Represent.*, vol. 62, pp. 368–380, Jul. 2019.

[200] Y. Chen and X. Lu, "Deep discrete hashing with pairwise correlation learning," *Neurocomputing*, vol. 385, pp. 111–121, Apr. 2020.

[201] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1445–1454.

[202] Y. Wu, S. Wang, and Q. Huang, "Online asymmetric similarity learning for cross-modal retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4269–4278.

[203] J. Liu, M. Yang, C. Li, and R. Xu, "Improving cross-modal image-text retrieval with teacher-student learning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Nov. 16, 2020, doi: 10.1109/TCSVT.2020.3037661.

[204] J. Lei, Y. Song, B. Peng, Z. Ma, L. Shao, and Y.-Z. Song, "Semi-heterogeneous three-way joint embedding network for sketch-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 3226–3237, Sep. 2020.

[205] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: Fast free-hand sketch-based image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2862–2871.

[206] A. Dutta and Z. Akata, "Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5089–5098.

[207] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1556–1564.

[208] D. Wu, Z. Lin, B. Li, M. Ye, and W. Wang, "Deep supervised hashing for multi-label and large-scale image retrieval," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2017, pp. 150–158.

[209] H. Lai, P. Yan, X. Shu, Y. Wei, and S. Yan, "Instance-aware hashing for multi-label image retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2469–2479, Jun. 2016.

[210] G. Chen, X. Cheng, S. Su, and C. Tang, "Multiple-instance ranking based deep hashing for multi-label image retrieval," *Neurocomputing*, vol. 402, pp. 89–99, Aug. 2020.

[211] Q. Qin, L. Huang, and Z. Wei, "Deep multilevel similarity hashing with fine-grained features for multi-label image retrieval," *Neurocomputing*, vol. 409, pp. 46–59, Oct. 2020.

[212] A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "A baseline for visual instance retrieval with deep convolutional networks," in *Proc. ICLR*, 2015.

[213] O. Morère, A. Veillard, L. Jie, J. Petta, V. Chandrasekhar, and T. Poggio, "Group invariant deep representations for image instance retrieval," in *Proc. AAAI Spring Symp. Ser.*, 2017, pp. 603–607.
[214] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," 2015, *arXiv:1511.05879*. [Online]. Available: http://arxiv.org/abs/1511.05879
[215] S. Pang, J. Zhu, J. Wang, V. Ordonez, and J. Xue, "Building discriminative CNN image representations for object retrieval using the replicator equation," *Pattern Recognit.*, vol. 83, pp. 150–160, Nov. 2018.
[216] X. Shi and X. Qian, "Exploring spatial and channel contribution for object based image retrieval," *Knowl.-Based Syst.*, vol. 186, Dec. 2019, Art. no. 104955.
[217] J. Guo, S. Zhang, and J. Li, "Hash learning with convolutional neural networks for semantic based image retrieval," in *Proc. KDDM*, 2016, pp. 227–238.
[218] Y. Cao, M. Long, J. Wang, and S. Liu, "Deep visual-semantic quantization for efficient image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1328–1337.
[219] Q. Qin, L. Huang, Z. Wei, K. Xie, and W. Zhang, "Unsupervised deep multi-similarity hashing with semantic structure for image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Oct. 21, 2020, doi: 10.1109/TCSVT.2020.3032402.
[220] J. Zhang and Y. Peng, "Query-adaptive image retrieval by deep-weighted hashing," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2400–2414, Sep. 2018.
[221] X. Zeng, Y. Zhang, X. Wang, K. Chen, D. Li, and W. Yang, "Fine-grained image retrieval via piecewise cross entropy loss," *Image Vis. Comput.*, vol. 93, Jan. 2020, Art. no. 103820.
[222] Z. Yang, O. I. Raymond, W. Sun, and J. Long, "Asymmetric deep semantic quantization for image retrieval," *IEEE Access*, vol. 7, pp. 72684–72695, 2019.
[223] J. Chen and W. K. Cheung, "Similarity preserving deep asymmetric quantization for image retrieval," in *Proc. AAAI*, vol. 33, 2019, pp. 8183–8190.
[224] J. Revaud, J. Almazan, R. Rezende, and C. D. Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5107–5116.
[225] H. Zhai, S. Lai, H. Jin, X. Qian, and T. Mei, "Deep transfer hashing for image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 742–753, Feb. 2021.
[226] Z. Dong, C. Jing, M. Pei, and Y. Jia, "Deep CNN based binary hash video representations for face retrieval," *Pattern Recognit.*, vol. 81, pp. 357–369, Sep. 2018.
[227] S. Ram Dubey and S. Chakraborty, "Average biased ReLU based CNN descriptor for improved face retrieval," 2018, *arXiv:1804.02051*. [Online]. Available: http://arxiv.org/abs/1804.02051
[228] T.-Y. Yang, D. K. Nguyen, H. Heijnen, and V. Balntas, "DAME Web: DynAmic MEan with whitening ensemble binarization for landmark retrieval without human annotation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2913–2922.
[229] L. Zhu, H. Cui, Z. Cheng, J. Li, and Z. Zhang, "Dual-level semantic transfer deep hashing for efficient social image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1478–1489, Apr. 2021.

Shiv Ram Dubey (Member, IEEE) received the Ph.D. degree in computer vision and image processing from the Indian Institute of Information Technology, Allahabad (IIIT Allahabad) in 2016. He has been with the Indian Institute of Information Technology (IIIT), Sri City, since June 2016, where he is currently an Assistant Professor of Computer Science and Engineering. Before that, from August 2012 to February 2013, he was a Project Officer with the Computer Science and Engineering Department, Indian Institute of Technology Madras (IIT Madras). His research interests include computer vision, deep learning, image feature description, and content based image retrieval. He was a recipient of several awards, including the Best Ph.D. Award in Ph.D. Symposium, IEEE-CICT 2017 at IIITM Gwalior, and the NVIDIA GPU Grant Award Twice from NVIDIA.