SIFT Meets CNN:
A Decade Survey of Instance Retrieval

Liang Zheng, Yi Yang, and Qi Tian, *Fellow, IEEE*

Abstract—In the early days, content-based image retrieval (CBIR) was studied with global features. Since 2003, image retrieval based on local descriptors (*de facto* SIFT) has been extensively studied for over a decade due to the advantage of SIFT in dealing with image transformations. Recently, image representations based on the convolutional neural network (CNN) have attracted increasing interest in the community and demonstrated impressive performance. Given this time of rapid evolution, this article provides a comprehensive survey of instance retrieval over the last decade. Two broad categories, SIFT-based and CNN-based methods, are presented. For the former, according to the codebook size, we organize the literature into using large/medium-sized/small codebooks. For the latter, we discuss three lines of methods, *i.e.*, using pre-trained or fine-tuned CNN models, and hybrid methods. The first two perform a single-pass of an image to the network, while the last category employs a patch-based feature extraction scheme. This survey presents milestones in modern instance retrieval, reviews a broad selection of previous works in different categories, and provides insights on the connection between SIFT and CNN-based methods. After analyzing and comparing retrieval performance of different categories on several datasets, we discuss promising directions towards generic and specialized instance retrieval.

Index Terms—Instance retrieval, SIFT, convolutional neural network, literature survey.

✦

1 INTRODUCTION

CONTENT-based image retrieval (CBIR) has been a long-standing research topic in the computer vision society. In the early 1990s, the study of CBIR truly started. Images were indexed by the visual cues, such as texture and color, and a myriad of algorithms and image retrieval systems have been proposed. A straightforward strategy is to extract global descriptors. This idea dominated the image retrieval community in the 1990s and early 2000s. Yet, a well-known problem is that global signatures may fail the invariance expectation to image changes such as illumination, translation, occlusion and truncation. These variances compromise the retrieval accuracy and limit the application scope of global descriptors. This problem has given rise to local feature based image retrieval.

The focus of this survey is instance-level image retrieval. In this task, given a query image depicting a particular object/scene/architecture, the aim is to retrieve images containing the same object/scene/architecture that may be captured under different views, illumination, or with occlusions. Instance retrieval departs from class retrieval [1] in that the latter aims at retrieving images of the same class with the query. In the following, if not specified, we use "image retrieval" and "instance retrieval" interchangeably.

The milestones of instance retrieval in the past years are presented in Fig. 1, in which the times of the SIFT-based and CNN-based methods are highlighted. The majority of traditional methods can be considered to end in 2000 when Smeulders *et al.* [2] presented a comprehensive survey of CBIR "at the end of the early years". Three years later (2003)

- *L. Zheng and Y. Yang (corresponding author) are with the Centre for AI, University of Technology at Sydney, NSW, Australia.*
  E-mail: liang.zheng@uts.edu.au, yi.yang@uts.edu.au
- *Q. Tian is with the Department of Computer Science, University of Texas at San Antonio, TX, 78256 USA. E-mail: qitian@cs.utsa.edu*

the Bag-of-Words (BoW) model was introduced to the image retrieval community [3], and in 2004 was applied to image classification [4], both relying on the SIFT descriptor [5]. The retrieval community has since witnessed the prominence of the BoW model for over a decade during which many improvements were proposed. In 2012, Krizhevsky *et al.* [6] with the AlexNet achieved the state-of-the-art recognition accuracy in ILSRVC 2012, exceeding previous best results by a large margin. Since then, research focus has begun to transfer to deep learning based methods [7], [8], [9], [10], especially the convolutional neural network (CNN).

The SIFT-based methods mostly rely on the BoW model. BoW was originally proposed for modeling documents because the text is naturally parsed into words. It builds a word histogram for a document by accumulating word responses into a global vector. In the image domain, the introduction of the scale-invariant feature transform (SIFT) [5] makes the BoW model feasible [3]. Originally, SIFT is comprised of a detector and descriptor, but which are used in isolation now; in this survey, if not specified, SIFT usually refers to the 128-dim descriptor, a common practice in the community. With a pre-trained codebook (vocabulary), local features are quantized to visual words. An image can thus be represented in a similar form to a document, and classic weighting and indexing schemes can be leveraged.

In recent years, the popularity of SIFT-based models seems to be overtaken by the convolutional neural network (CNN), a hierarchical structure that has been shown to outperform hand-crafted features in many vision tasks. In retrieval, competitive performance compared to the BoW models has been reported, even with short CNN vectors [10], [16], [17]. The CNN-based retrieval models usually compute compact representations and employ the Euclidean distance or some approximate nearest neighbor (ANN) search methods for retrieval. Current literature may directly
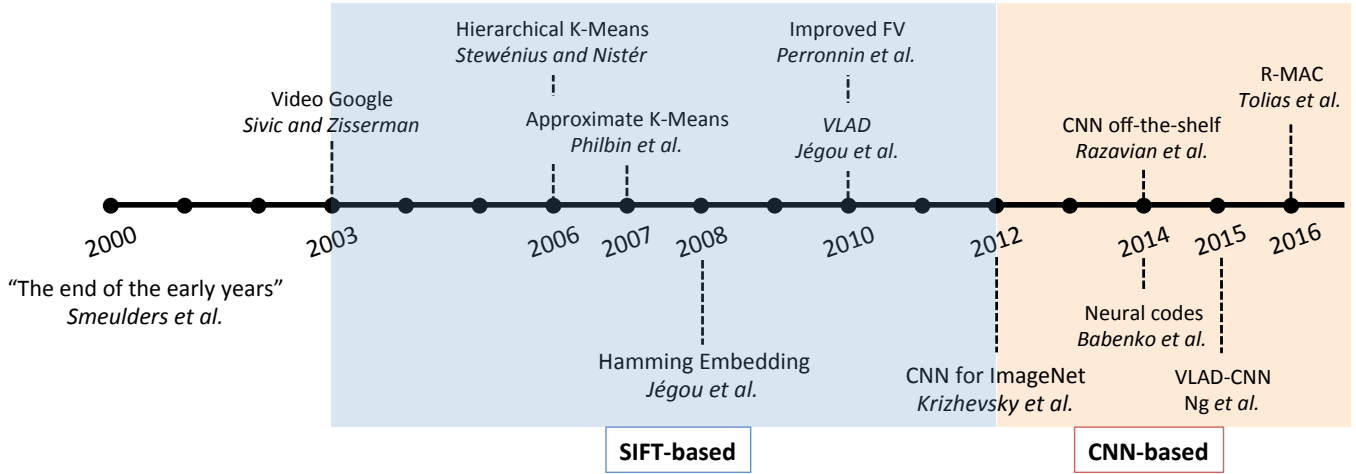
Fig. 1: Milestones of instance retrieval. After a survey of methods before the year 2000 by Smeulders *et al* [2], Sivic and Zisserman [3] proposed Video Google in 2003, marking the beginning of the BoW model. Then, the hierarchical k-means and approximate k-means were proposed by Stewénius and Nistér [11] and Philbin *et al.* [12], respectively, marking the use of large codebooks in retrieval. In 2008, Jégou *et al.* [13] proposed Hamming Embedding, a milestone in using medium-sized codebooks. Then, compact visual representations for retrieval were proposed by Perronnin *et al.* [14] and Jégou *et al.* [15] in 2010. Although SIFT-based methods were still moving forward, CNN-based methods began to gradually take over, following the pioneering work of Krizhevsky *et al.* [6]. In 2014, Razavian *et al.* [7] proposed a hybrid method extracting multiple CNN features from an image. Babenko *et al.* [8] were the first to fine-tune a CNN model for generic instance retrieval. Both [9], [10] employ the column features from pre-trained CNN models, and [10] inspires later state-of-the-art methods. These milestones are the representative works of the categorization scheme in this survey.

employ the pre-trained CNN models or perform fine-tuning for specific retrieval tasks. A majority of these methods feed the image into the network only once to obtain the descriptor. Some are based on patches which are passed to the network multiple times, a similar manner to SIFT; we classify them into hybrid methods in this survey.

1.1 Organization of This Paper

Upon the time of change, this paper provides a comprehensive literature survey of both the SIFT-based and CNN-based instance retrieval methods. We first present the categorization methodology in Section 2. We then describe the two major method types in Section 3 and Section 4, respectively. On several benchmark datasets, Section 5 summarizes the comparisons between SIFT- and CNN-based methods. In Section 6, we point out two possible future directions. This survey will be concluded in Section 7.

2 CATEGORIZATION METHODOLOGY

According to the different visual representations, this survey categorizes the retrieval literature into two broad types: SIFT-based and CNN-based. The SIFT-based methods are further organized into three classes: using large, medium-sized or small codebooks. We note that the codebook size is closely related to the choice of encoding methods. The CNN-based methods are categorized into using pre-trained or fine-tuned CNN models, as well as hybrid methods. Their similarities and differences are summarized in Table 1.

The SIFT-based methods had been predominantly studied before 2012 [6] (good works also appear in recent years [18], [19]). This line of methods usually use one type of detector, *e.g.*, Hessian-Affine, and one type of descriptor, *e.g.*,

SIFT. Encoding maps a local feature into a vector. Based on the size of the codebook used during encoding, we classify SIFT-based methods into three categories as below. leftmargin=0pt

- **Using small codebooks.** The visual words are fewer than several thousand. Compact vectors are generated [14], [15] before dimension reduction and coding.
- **Using medium-sized codebooks.** Given the sparsity of BoW and the low discriminative ability of visual words, the inverted index and binary signatures are used [13]. The trade-off between accuracy and efficiency is a major influencing factor [20].
- **Using large codebooks.** Given the sparse BoW histograms and the high discriminative ability of visual words, the inverted index and memory-friendly signatures are used [21]. Approximate methods are used in codebook generation and encoding [11], [12].

The CNN-based methods extract features using CNN models. Compact (fixed-length) representations are usually built. There are three classes:

- **Hybrid methods.** Image patches are fed into CNN multiple times for feature extraction [7]. Encoding and indexing are similar to SIFT-based methods [22].
- **Using pre-trained CNN models.** Features are extracted in a single pass using CNN pre-trained on some large-scale datasets like ImageNet [23]. Compact Encoding/pooling techniques are used [9], [10].
- **Using fine-tuned CNN models.** The CNN model (*e.g.*, pre-trained on ImageNet) is fine-tuned on a training set in which the images share similar distributions with the target database [8]. CNN features can be extracted in an end-to-end manner through a single
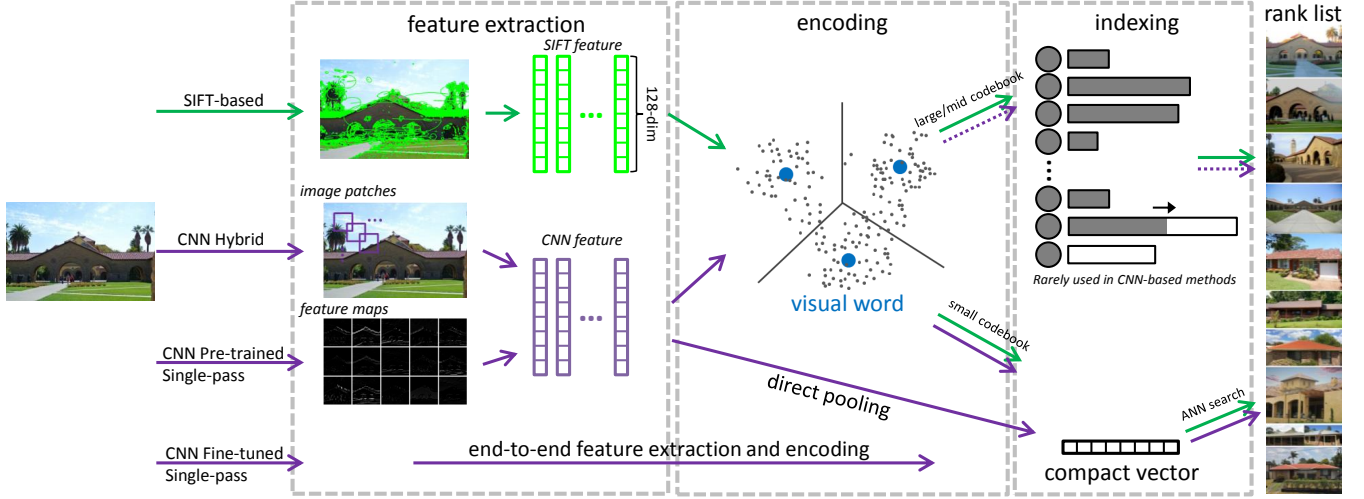
Fig. 2: A general pipeline of SIFT- and CNN-based retrieval models. Features are computed from hand-crafted detectors for SIFT, and densely applied filters or image patches for CNN. In both methods, under small codebooks, encoding/pooling is employed to produce compact vectors. In SIFT-based methods, the inverted index is necessary under large/medium-sized codebooks. The CNN features can also be computed in an end-to-end way using fine-tuned CNN models.

	method type	detector	descriptor	encoding	dim.	indexing
SIFT-based	Large voc.	DoG, Hessian-Affine, dense patches, *etc.*	Local invariant descriptors such as SIFT	Hard, soft	High	Inverted index
	Mid voc.			Hard, soft, HE	Medium	Inverted index
	Small voc.			VLAD, FV	Low	ANN methods
CNN-based	Hybrid	Image patches	CNN features	VLAD, FV, pooling	Varies	ANN methods
	Pre-trained, single-pass	Column feat. or FC of pre-trained CNN models.		VLAD, FV, pooling	Low	ANN methods
	Fine-tuned, single-pass	A global feat. is end-to-end extracted from fine-tuned CNN models.			Low	ANN methods

TABLE 1: Major differences between various types of instance retrieval models. For SIFT-based methods, hand-crafted local invariant features are extracted, and according to the codebook sizes, different encoding and indexing strategies are leveraged. For CNN-based methods, pre-trained, fine-tuned CNN models and hybrid methods are the primary types; fixed-length compact vectors are usually produced, combined with approximate nearest neighbor (ANN) methods.

pass to the CNN model. The visual representations exhibit improved discriminative ability [17], [24].

3 SIFT-BASED IMAGE RETRIEVAL

3.1 Pipeline

The pipeline of SIFT-based retrieval is introduced in Fig. 2.

Local feature extraction. Suppose we have a gallery $\mathcal{G}$ consisting of N images. Given a feature detector, we extract local descriptors from the regions around the sparse interest points or dense patches. We denote the local descriptors of D detected regions in an image as $\{f_i\}_{i=i}^{D}, f_i \in \mathbb{R}^p$.

Codebook training. SIFT-based methods train a codebook offline. Each visual word in the codebook lies in the center of a subspace, called the "Voronoi cell". A larger codebook corresponds to a finer partitioning, resulting in more discriminative visual words and vice versa. Suppose that a pool of local descriptors $\mathcal{F} \equiv \{f_i\}_{i=1}^{M}$ are computed from an unlabeled training set. The baseline approach, *i.e.,* k-means, partitions the M points into K clusters; the K visual words thus constitute a codebook of size K.

Feature encoding. A local descriptor $f_i \in \mathbb{R}^p$ is mapped into a feature embedding $g_i \in \mathbb{R}^l$ through the feature encoding process, $f_i \to g_i$. When k-means clustering is used, f_i can be encoded according to its distances to the visual words. For large codebooks, hard [11], [12] and soft quantization [25] are

good choices. In the former, the resulting embedding g_i has only one non-zero entry; in the latter, f_i can be quantized to a small number of visual words. A global signature is produced after a sum-pooling of all the embeddings of local features. For medium-sized codebooks, additional binary signatures can be generated to preserve the original information. When using small codebooks, popular encoding schemes include vector of locally aggregated descriptors (VLAD) [15], Fisher vector (FV) [14], *etc.*

3.2 Local Feature Extraction

Local invariant features aim at accurate matching of local structures between images [26]. SIFT-based methods usually share a similar feature extraction step composed of a feature detector and a descriptor.

Local detector. The **interest point detectors** aim to reliably localize a set of stable local regions under various imaging conditions. In the retrieval community, finding affine-covariant regions has been preferred. It is called "covariant" because the shapes of the detected regions change with the affine transformations, so that the region content (descriptors) can be invariant. This kind of detectors are different from keypoint-centric detectors such as the Hessian detector [27], and from those focusing on scale-invariant regions such as the difference of Gaussians (DoG) [5] detector. Elliptical regions which are adapted to the local intensity

patterns are produced by affine detectors. This ensures that the same local structure is covered under deformations caused by viewpoint variances, a problem often encountered in instance retrieval. In the milestone work [3], the Maximally Stable Extremal Region (MSER) detector [28] and the affine extended Harris-Laplace detector are employed, both of which are affine-invariant region detectors. MSER is used in several later works [11], [29]. Starting from [12], the Hessian-affine detector [30] has been widely adopted in retrieval. It has been shown to be superior to the difference of Gaussians (DoG) detector [13], [31], due to its advantage in reliably detecting local structures under large viewpoint changes. To fix the orientation ambiguity of these affine-covariant regions, the gravity assumption is made [32]. The practice which dismisses the orientation estimation is employed by later works [33], [34] and demonstrates consistent improvement on architecture datasets where the objects are usually upright. Other non-affine detectors have also been tested in retrieval, such as the Laplacian of Gaussian (LOG) and Harris detectors used in [35]. For objects with smooth surfaces [36], few interest points can be detected, so the object boundaries are good candidates for local description.

On the other hand, some employ the **dense region detectors**. In the comparison between densely sampled image patches and the detected patches, Sicre *et al.* [37] report the superiority of the former. To recover the rotation invariance of dense sampling, the dominant angle of patches is estimated in [38]. A comprehensive comparison of various dense sampling strategies, the interest point detectors, and those in between can be accessed in [39].

Local Descriptor. With a set of detected regions, descriptors encode the local content. SIFT [5] has been used as the default descriptor. The 128-dim vector has been shown to outperform competing descriptors in matching accuracy [40]. In an extension, PCA-SIFT [41] reduces the dimension from 128 to 36 to speed up the matching process at the cost of more time in feature computation and loss of distinctiveness. Another improvement is RootSIFT [33], calculated by two steps: 1) ℓ_1 normalize the SIFT descriptor, 2) square root each element. RootSIFT is now used as a routine in SIFT-based retrieval. Apart from SIFT, SURF [42] is also widely used. It combines the Hessian-Laplace detector and a local descriptor of the local gradient histograms. The integral image is used for acceleration. SURF has a comparable matching accuracy with SIFT and is faster to compute. See [43] for comparisons between SIFT, PCA-SIFT, and SURF. To further accelerate the matching speed, binary descriptors [44] replace Euclidean distance with Hamming distance during matching.

Apart from hand-crafted descriptors, some also propose learning schemes to improve the discriminative ability of local descriptors. For example, Philbin *et al.* [45] proposes a non-linear transformation so that the projected SIFT descriptor yields smaller distances for true matches. Simoyan *et al.* [34] improve this process by learning both the pooling region and a linear descriptor projection.

3.3 Retrieval Using Small Codebooks

A small codebook has several thousand, several hundred or fewer visual words, so the computational complexity of codebook generation and encoding is moderate. Representative works include BoW [3], VLAD [15] and FV [14]. We mainly discuss VLAD and FV and refer readers to [46] for a comprehensive evaluation of the BoW compact vectors.

3.3.1 Codebook Generation

Clustering complexity depends heavily on the codebook size. In works based on VLAD [15] or FV [14], the codebook sizes are typically small, *e.g.*, 64, 128, 256. For VLAD, flat k-means is employed for codebook generation. For FV, the Gaussian mixture model (GMM), *i.e.*, $u_\lambda(x) = \sum_{i=1}^{K} w_i u_i(x)$, where K is the number of Gaussian mixtures, is trained using the maximum likelihood estimation. GMM describes the feature space with a mixture of K Gaussian distributions, and can be denoted as $\lambda = \{w_i, \mu_i, \sum_i, i = 1, ..., K\}$, where w_i, μ_i and $\sum_i$ represent the mixture weight, the mean vector and the covariance matrix of Gaussian u_i, respectively.

3.3.2 Encoding

Due to the small codebook size, relative complex and information-preserving encoding techniques can be applied. We mainly describe FV, VLAD and their improvements in this section. With a pre-trained GMM model, FV describes the averaged first and second order difference between local features and the GMM centers. Its dimension is $2pK$, where p is the dimension of the local descriptors and K is the codebook size of GMM. FV usually undergoes power normalization [47], [14] to suppress the burstiness problem (to be described in Section 3.4.3). In this step, each component of FV undergoes non-linear transformation featured by parameter α, $x_i := \text{sign}(x_i)\|x_i\|^\alpha$. Then ℓ_2 normalization is employed. Later, FV is improved from different aspects. For example, Koniusz *et al.* [48] augment each descriptor with its spatial coordinates and associated tunable weights. In [49], larger codebooks (up to 4,096) are generated and demonstrate superior classification accuracy to smaller codebooks, at the cost of computational efficiency. To correct the assumption that local regions are identically and independently distributed (iid), Cinbis *et al.* [50] propose non-iid models that discount the burstiness effect and yield improvement over the power normalization.

The VLAD encoding scheme proposed by Jégou *et al.* [15] can be thought of as a simplified version of FV. It quantizes a local feature to its nearest visual word in the codebook and records the difference between them. Nearest neighbor search is performed because of the small codebook size. The residual vectors are then aggregated by sum pooling followed by normalizations. The dimension of VLAD is pK. Comparisons of some important encoding techniques are presented in [51], [52]. Again, the improvement of VLAD comes from multiple aspects. In [53], Jégou and Chum suggest the usage of PCA and whitening (denoted as PCA$_w$ in Table 5) to de-correlate visual word co-occurrences, and the training of multiple codebooks to reduce quantization loss. In [54], Arandjelović *et al.* extend VLAD in three aspects: 1) normalize the residual sum within each coarse cluster, called intra-normalization, 2) vocabulary adaptation to address the dataset transfer problem and 3) multi-VLAD for small object discovery. Concurrent to [54], Delhumeau *et al.* [55] propose to normalize each residual vector instead of the residual sums; they also advocate for local PCA within each Voronoi cell which does not perform dimension reduction as [52]. A recent work [56] employs soft assignment and empirically

learns optimal weights for each rank to improve over the hard quantization.

Note that some general techniques benefit various embedding methods, such as VLAD, FV, BoW, locality-constrained linear coding (LLC) [57] and monomial embeddings. To improve the discriminative ability of embeddings, Tolias *et al.* [58] propose the orientation covariant embedding to encode the dominant orientation of the SIFT regions jointly with the SIFT descriptor. It achieves a similar covariance property to weak geometric consistency (WGC) [13] by using geometric cues within regions of interest so that matching points with similar dominant orientations are up-weighted and vice versa. The triangulation embedding [18] only considers the direction instead of the magnitude of the input vectors. Jégou *et al.* [18] also present a democratic aggregation that limits the interference between the mapped vectors. Baring a similar idea with democratic aggregation, Murray and Perronnin [59] propose the generalized max pooling (GMP) optimized by equalizing the similarity between the pooled vector and each coding representation.

The computational complexity of BoW, VLAD and FV is similar. We neglect the offline training and SIFT extraction steps. During visual word assignment, each feature should compute its distance (or soft assignment coefficient) with all the visual words (or Gaussians) for VLAD (or FV). So this step has a complexity of $\mathcal{O}(pK)$. In the other steps, complexity does not exceed $\mathcal{O}(pK)$. Considering the sum-pooling of the embeddings, the encoding process has an overall complexity of $\mathcal{O}(pKD)$, where D is the number of features in an image. Triangulation embedding [18], a variant of VLAD, has a similar complexity. The complexity of multi-VLAD [54] is $\mathcal{O}(pKD)$, too, but it has a more costly matching process. Hierarchical VLAD [60] has a complexity of $\mathcal{O}(pKK'D)$, where K' is the size of the secondary codebook. In the aggregation stage, both GMP [59] and democratic aggregation [18] have high complexity. The complexity of GMP is $\mathcal{O}(\frac{P^2}{K})$, where P is the dimension of the feature embedding, while the computational cost of democratic aggregation comes from the Sinkhorn algorithm.

3.3.3 ANN Search

Due to the high dimensionality of the VLAD/FV embeddings, efficient compression and ANN search methods have been employed [61], [62]. For example, the principle component analysis (PCA) is usually adapted to for dimension reduction, and it is shown that retrieval accuracy even increases after PCA [53]. For hashing-based ANN methods, Perronnin *et al.* [47] use standard binary encoding techniques such as locality sensitive hashing [63] and spectral hashing [64]. Nevertheless, when being tested on the SIFT and GIST feature datasets, spectral hashing is shown to be outperformed by Product Quantization (PQ) [61]. In these quantization-based ANN methods, PQ is demonstrated to be better than other popular ANN methods such as FLANN [62] as well. A detailed discussion of VLAD and PQ can be viewed in [65]. PQ has since then been improved in a number of works. In [66], Douze *et al.* propose to re-order the cluster centroids so that adjacent centroids have small Hamming distances. This method is compatible with Hamming distance based ANN search, which offers significant speedup for PQ. We refer readers to [67] for a survey of ANN approaches.
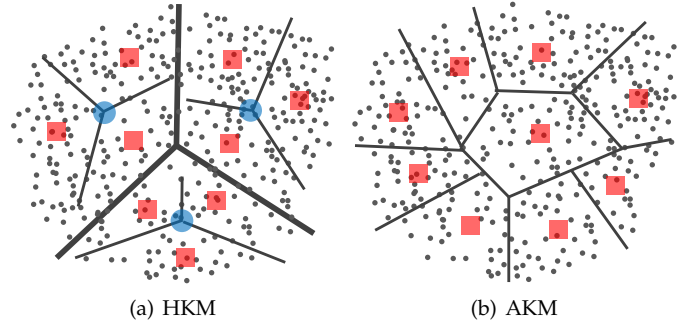

(a) HKM   (b) AKM

Fig. 3: Two milestone clustering methods (a) hierarchical k-means (HKM) [11] and (b) approximate k-means (AKM) [12] for large codebook generation. Bold borders and blue discs are the clustering boundaries and centers of the first layer of HKM. Slim borders and red squares are the final clustering results in both methods.

We also mention an emerging ANN technique, *i.e.*, group testing [68], [69], [70]. In a nutshell, the database is decomposed into groups, each represented by a group vector. Comparisons between the query and group vectors reveal how likely a group contains a true match. Since group vectors are much fewer than the database vectors, search time is reduced. Iscen *et al.* [69] propose to directly find the best group vectors summarizing the database without explicitly forming the groups, which reduces the memory consumption.

3.4 Retrieval Using Large Codebooks

A large codebook may contain 1 million [11], [12] visual words or more [71], [72]. Some major steps undergo important changes compared with using small codebooks.

3.4.1 Codebook Generation

Approximate methods are critical in assigning data into a large number of clusters. In the retrieval community, two representative works are hierarchical k-means (HKM) [11] and approximate k-means (AKM) [12], as illustrated in Fig. 1 and Fig. 3. Proposed in 2006, HKM applies standard k-means on the training features hierarchically. It first partitions the points into a few clusters (*e.g.*, $\bar{k} \ll K$) and then recursively partitions each cluster into further clusters. In every recursion, each point should be assigned to one of the $\bar{k}$ clusters, with the depth of the cluster tree being $\mathcal{O}(\log K)$, where K is the target cluster number. The computational cost of HKM is therefore $\mathcal{O}(\bar{k}M \log K)$, where M is the number of training samples. It is much smaller than the complexity of flat k-means $\mathcal{O}(MK)$ when K is large (a large codebook).

The other milestone in large codebook generation is AKM [12]. This method indexes the K cluster centers using a forest of random k-d trees so that the assignment step can be performed efficiently with ANN search. In AKM, the cost of assignment can be written as $\mathcal{O}(K \log K + vM \log K) = \mathcal{O}(vM \log K)$, where v is the number of nearest cluster candidates to be accessed in the k-d trees. So the computational complexity of AKM is on par with HKM and is significantly smaller than flat k-means when K is large. Experiments

show that AKM is superior to HKM [12] due to its lower quantization error (see Section 3.4.2). In most AKM-based methods, the default choice for ANN search is FLANN [62].

3.4.2 Feature Encoding (Quantization)

Feature encoding is interleaved with codebook clustering, because ANN search is critical in both components. The ANN techniques implied in some classic methods like AKM and HKM can be used in both clustering and encoding steps. Under a large codebook, the key trade-off is between quantization error and computational complexity. In the encoding step, information-preserving encoding methods such as FV [14], sparse coding [73] are mostly infeasible due to their computational complexity. It therefore remains a challenging problem how to reduce the quantization error while keeping the quantization process efficient.

Fro the ANN methods, the earliest solution is to quantize a local feature along the hierarchical tree structure [11]. Quantized tree nodes in different levels are assigned different weights. However, due to the highly imbalanced tree structure, this method is outperformed by k-d tree based quantization method [12]: one visual word is assigned to each local feature, using a k-d tree built from the codebook for fast ANN search. In an improvement to this hard quantization scheme, Philbin *et al.* [25] propose soft quantization by quantizing a feature into several nearest visual words. The weight of each assigned visual word relates negatively to its distance from the feature by $\exp(-\frac{d^2}{2\sigma^2})$, where d is the distance between the descriptor and the cluster center. While soft quantization is based on the Euclidean distance, Mikulik *et al.* [71] propose to find relevant visual words for each visual word through an unsupervised set of matching features. Built on a probabilistic model, these alternative words tend to contain descriptors of matching features. To reduce the memory cost of soft quantization [25] and the number of query visual words, Cai *et al.* [74] suggest that when a local feature is far away from even the nearest visual word, this feature can be discarded without a performance drop. To further accelerate quantization, scalar quantization [75] suggests that local features be quantized without an explicitly trained codebook. A floating-point vector is binarized, and the first dimensions of the resulting binary vector are directly converted to a decimal number as a visual word. In the case of large quantization error and low recall, scalar quantization uses bit-flop to generate hundreds of visual words for a local feature.

3.4.3 Feature Weighting

TF-IDF. The visual words in codebook $\mathcal{C}$ are typically assigned specific weights, called the term frequency and inverse document frequency (TF-IDF), which are integrated with the BoW encoding. TF is defined as:

$$\text{TF}(c_i^j) = o_i^j, \tag{1}$$

where o_i^j is the number of occurrences of a visual word c_i within an image j. TF is thus a local weight. IDF, on the other hand, determines the contribution of a given visual word through global statistics. The classic IDF weight of visual word c_i is calculated as:

$$\text{IDF}(c_i) = \log \frac{N}{n_i}, \text{where } n_i = \sum_{j \in \mathcal{G}} \mathbf{1}(o_i^j > 0), \tag{2}$$

where N is the number of gallery images, and n_i encodes the number of images in which word c_i appears. The TF-IDF weight for visual word c_i in image j is:

$$w(c_i^j) = \text{TF}(c_i^j)\text{IDF}(c_i). \tag{3}$$

Improvements. A major problem associated with visual word weighting is burstiness [76]. It refers to the phenomenon whereby repetitive structures appear in an image. This problem tends to dominate image similarity. Jégou *et al.* [76] propose several TF variants to deal with burstiness. An effective strategy consists in exerting a square operation on TF. Instead of grouping features with the same word index, Revaud *et al.* [77] propose detecting keypoint groups frequently happening in irrelevant images which are down-weighted in the scoring function. While the above two methods detect bursty groups after quantization, Shi *et al* [19] propose detecting them in the descriptor stage. The detected bursty descriptors undergo average pooling and are fed in the BoW architectures. From the aspect of IDF, Zheng *et al.* [78] propose the $\mathcal{L}_p$-norm IDF to tackle burstiness and Murata *et al.* [79] design the exponential IDF which is later incorporated into the BM25 formula. When most works try to suppress burstiness, Torii *et al* [80] view it as a distinguishing feature for architectures and design new similarity measurement following burstiness detection.

Another feature weighting strategy is feature augmentation on the database side [81], [33]. Both methods construct an image graph offline, with edges indicating whether two images share a same object. For [81], only features that pass the geometric verification are preserved, which reduces the memory cost. Then, the feature of the base image is augmented with all the visual words of its connecting images. This method is improved in [33] by only adding those visual words which are estimated to be visible in the augmented image, so that noisy visual words can be excluded.

3.4.4 The Inverted Index

The inverted index is designed to enable efficient storage and retrieval and is usually used under large/medium-sized codebooks. Its structure is illustrated in Fig. 4. The inverted index is a one-dimensional structure where each entry corresponds to a visual word in the codebook. An inverted list is attached to each word entry, and those indexed in the each inverted list are called indexed features or postings. The inverted index takes advantages of the sparse nature of the visual word histogram under a large codebook.

In literature, it is required that new retrieval methods be adjustable to the inverted index. In the baseline [11], [12], the image ID and term frequency (TF) are stored in a posting. When other information is integrated, they should be small in size. For example, in [82], the metadata are quantized, such as descriptor contextual weight, descriptor density, mean relative log scale and the mean orientation difference in each posting. Similarly, quantized spatial information such as the orientation can also be stored [83], [21]. In co-indexing [72], when the inverted index is enlarged with globally consistent neighbors, semantically isolated images are deleted to reduce memory consumption. In [84], the original one-dimensional inverted index is expanded to two-dimensional for ANN search, which learns a codebook for each SIFT sub-vector.
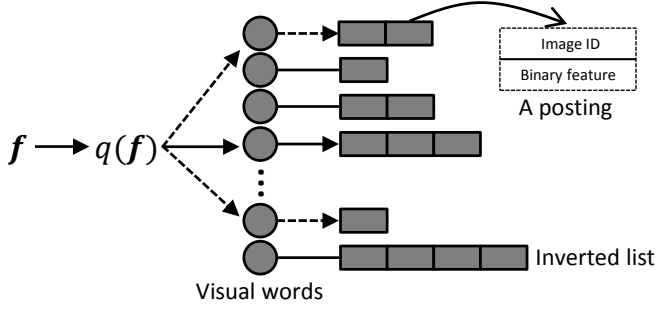
Fig. 4: The data structure of the inverted index. It physically contains K inverted lists, each consisting of some postings, which index the image ID and some binary signatures. During retrieval, a quantized feature will traverse the inverted list corresponding to its assigned visual word. Dashed line denotes soft quantization, in which multiple inverted lists are visited.

Later, it is applied to instance retrieval by [31] to fuse local color and SIFT descriptors.

3.5 Retrieval Using Medium-sized Codebooks

Medium-sized codebooks refer to those having 10-200k visual words. The visual words exhibit medium discriminative ability, and the inverted index is usually constructed.

3.5.1 Codebook Generation and Quantization

Considering the relatively small computational cost compared with large codebooks (Section 3.4.1), flat k-means can be adopted for codebook generation [85], [20]. It is also shown in [31], [86] that using AKM [12] for clustering also yields very competitive retrieval accuracy.

For quantization, nearest neighbor search can be used to find the nearest visual words in the codebook. Practice may tell that using some strict ANN algorithms produces competitive retrieval results. So comparing with the extensive study on quantization under large codebooks (Section 3.4.2) [25], [71], [74], relatively fewer works focus on the quantization problem under a medium-sized codebook.

3.5.2 Hamming Embedding and its improvements

The discriminative ability of visual words in medium-sized codebooks lies in between that of small and large codebooks. So it is important to compensate the information loss during quantization. To this end, a milestone work, *i.e.,* Hamming embedding (HE) has been dominantly employed.

Proposed by Jégou *et al.* [13], HE greatly improves the discriminative ability of visual words under medium-sized codebooks. HE first maps a SIFT descriptor $f \in \mathbb{R}^p$ from the p-dimensional space to a p_b-dimensional space:

$$x = P \cdot f = (x_1, ... x_{p_b}), \qquad (4)$$

where $P \in \mathbb{R}_b^p \times p$ is a projecting matrix, and x is a low-dimensional vector. By creating a matrix of random Gaussian values and applying a QR factorization to it, matrix P is taken as the first p_b rows of the resulting orthogonal matrix. To binarize x, Jegou *et al.* propose to compute the median vector $\overline{x_i} = (\overline{x_{1,i}}, ..., \overline{x_{p_b,i}})$ of the low-dimensional vector using descriptors falling in each Voronoi cell c_i. Given descriptor f

and its projected vector x, HE computes its visual word c_t, and the HE binary vector is computed as:

$$b_j(x) = \begin{cases} 1 & \text{if } x_j > \overline{x_{j,t}}, \\ 0 & \text{otherwise} \end{cases}, \qquad (5)$$

where $b(x) = (b_1(x), ..., b_{p_b}(x))$ is the resulting HE vector of dimension p_b. The binary feature $b(x)$ serves as a secondary check for feature matching. A pair of local features are a true match when two criteria are satisfied: 1) identical visual words and 2) small Hamming distance between their HE signatures. The extension of HE [85] estimates the matching strength between feature f_1 and f_2 reversely to the Hamming distance by an exponential function:

$$w_{\text{HE}}(f_1, f_2) = \exp(-\frac{\mathcal{H}(b(x_1), b(x_2))}{2\gamma^2}), \qquad (6)$$

where $b(x_1)$ and $b(x_2)$ are the HE binary vector of f_1 and f_2, respectively, $\mathcal{H}(\cdot, \cdot)$ computes the Hamming distance between two binary vectors, and γ is a weighting parameter. As shown in Fig. 6, HE [13] and its weighted version [85] improves accuracy considerably in 2008 and 2010.

Applications of HE include video copy detection [87], image classification [88] and re-ranking [89]. For example, in image classification, patch matching similarity is efficiently estimated by HE which is integrated into linear kernel-based SVM [88]. In image re-ranking, Tolias *et al.* [89] use lower HE thresholds to find strict correspondences which resemble those found by RANSAC, and the resulting image subset is more likely to contain true positives for query reformulation.

The improvement over HE has been observed in a number of works, especially from the view of match kernel [20]. To reduce the information loss on the query side, Jain *et al.* [90] propose a vector-to-binary distance comparison. It exploits the vector-to-hyperplane distance while retaining the efficiency of the inverted index. Further, Qin *et al.* [91] design a higher-order match kernel within a probabilistic framework and adaptively normalize the local feature distances by the distance distribution of false matches. This method is in the spirit similar to [92], in which the word-word distance, instead of the feature-feature distance [91], is normalized, according to the neighborhood distribution of each visual word. While the average distance between a word to its neighbors is regularized to be almost constant in [92], the idea of democratizing the contribution of individual embeddings has later been employed in [18]. In [20], Tolias *et al.* show that VLAD and HE share similar natures and propose a new match kernel which trades off between local feature aggregation and feature-to-feature matching, using a similar matching function to [91]. They also demonstrate that using more bits (*e.g.,* 128) in HE is superior to the original 64 bits scheme at the cost of decreased efficiency. Even more bits (256) are used in [75], but this method may be prone to relatively low recall.

3.6 Other Important Issues

3.6.1 Feature Fusion

Local-local fusion. A problem with the SIFT feature is that only local gradient description is provided. Other discriminative information encoded in an image is still not leveraged. In Fig. 5 (B), a pair of false matches cannot be
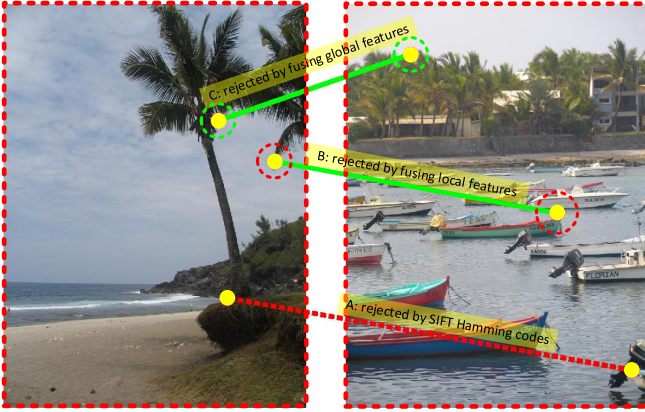
Fig. 5: False match removal by (A) HE [13], (B) local-local feature fusion, and (C) local-global feature fusion.

rejected by HE due to their similarity in the SIFT space, but the fusion of other local (or regional) features may correct this problem. A good choice for local-local fusion is to couple SIFT with color descriptors. The usage of color-SIFT descriptors can partially address the trade-off between invariance and discriminative ability. Evaluation has been conducted on several recognition benchmarks [93] of the descriptors such as HSV-SIFT [94], HueSIFT [95] and OpponentSIFT [93]. Both HSV-SIFT and HueSIFT are scale-invariant and shift-invariant. OpponentSIFT describes all the channels in the opponent color space using the SIFT descriptor and is largely robust to the light color changes. In [93], OpponentSIFT is recommended when no prior knowledge about the datasets is available. In more recent works, the binary color signatures are stored in the inverted index [96], [31]. Despite the good retrieval accuracy on some datasets, the potential problem is that intensive variation in illumination may compromise the effectiveness of colors.

Local-global fusion. Local and global features describe images from different aspects and can be complementary. In Fig. 5 (C), when local (and regional) cues are not enough to reject a false match pair, it would be effective to further incorporate visual information from a larger context scale. Early and late fusion are two possible ways. In early fusion, the image neighborhood relationship mined by global features such as FC8 in AlexNet [6] is fused in the SIFT-based inverted index [72]. In late fusion, Zhang *el al.* [97] build an offline graph for each type of feature, which is subsequently fused during the online query. In an improvement of [97], Deng *et al.* [98] add weakly supervised anchors to aid graph fusion. Both works on the rank level. For score-level fusion, automatically learned category-specific attributes are combined with pre-trained category-level information [99]. Zheng *et al.* [86] propose the query-adaptive late fusion by extracting a number of features (local or global, good or bad) and weighting them in a query-adaptive manner.

3.6.2 Geometric Matching

A frequent concern with the BoW model is the lack of geometric constraints among local features. Geometric verification can be used as a critical pre-processing step various scenarios, such as query expansion [100], [101], feature selection [81], database-side feature augmentation [81], [33], large-scale

object mining [102], *etc*. The most well-known method for global spatial verification is RANSAC [12]. It calculates affine transformations for each correspondence repeatedly which are verified by the number of inliers that fit the transformation. RANSAC is effective in re-ranking a subset of top-ranked images but has efficiency problems. As a result, how to efficiently and accurately incorporate spatial cues in the SIFT-based framework has been extensively studied.

A good choice is to discover the spatial context among local features. For example, visual phrases [103], [104], [105], [106] are generated among individual visual words to provide more strict matching criterion. Visual word co-occurrences in the entire image are estimated [107] and aggregated [108], while in [109], [110], [29] visual word clusters within local neighborhoods are discovered. Visual phrases can also be constructed from adjacent image patches [103], random spatial partitioning [106], and localized stable regions [29] such as MSER [28].

Another strategy uses voting to check geometric consistency. In the voting space, a bin with a larger value is more likely to represent the true transformation. An important work is weak geometrical consistency (WGC) [13], which focuses on the difference in scale and orientation between matched features. The space of difference is quantized into bins. Hough voting is used to locate the subset of correspondences similar in scale or orientation differences. Many later works can be viewed as extensions of WGC. For example, the method of Zhang *et al.* [21] can be viewed as WGC using x, y offsets instead of scale and orientation. This method is invariant to object translations, but may be sensitive to scale and rotation changes due to the rigid coordinate quantization. To regain the scale and the rotation variance, Shen *et al.* [111] quantize the angle and scale of the query region after applying several transformations. A drawback of [111] is that query time and memory cost are both increased. To enable efficient voting and alleviate quantization artifacts, Hough pyramid matching (HPM) [112] distributes the matches over a hierarchical partition of the transformation space. HPM trades off between flexibility and accuracy and is very efficient. Quantization artifact can also be reduced by allowing a single correspondence to vote for multiple bins [113]. HPM and [113] are much faster than RANSAC and can be viewed as extensions in the rotation and the scale invariance to the weak geometry consistency proposed along with Hamming Embedding [13]. In [114], a rough global estimate of orientation and scale changes is made by voting, which is used to verify the transformation obtained by the matched features. A recent method [115] combines the advantage of hypothesis-based methods such as RANSAC [12] and voting-based methods [21], [112], [113], [114]. Possible hypothesises are identified by voting and later verified and refined. This method inherits efficiency from voting and supports query expansion since it outputs an explicit transformation and a set of inliers.

3.6.3 Query Expansion

As a post-processing step, query expansion (QE) significantly improves the retrieval accuracy. In a nutshell, a number of top-ranked images from the original rank list are employed to issue a new query which is in turn used to obtain a new

rank list. QE allows additional discriminative features to be added to the original query, thus improving recall.

In instance retrieval, Chum *et al.* [100] are the first to exploit this idea. They propose the average query expansion (AQE) which averages features of the top-ranked images to issue the new query. Usually, spatial verification [12] is employed for re-ranking and obtaining the ROIs from which the local features undergo average pooling. AQE is used by many later works [10], [17], [24] as a standard tool. The recursive AQE and the scale-band recursive QE are effective improvement but incur more computational cost [100]. Four years later, Chum *et al.* [101] improve QE from the perspectives of learning background confusers, expanding the query region and incremental spatial verification. In [33], a linear SVM is trained online using the top-ranked and bottom-ranked images as positive and negative training samples, respectively. The learned weight vector is used to compute the average query. Other important extensions include "hello neighbor" based on reciprocal neighbors [116], QE with rank-based weighting [111], Hamming QE [89] (see Section 3.5), *etc*.

3.6.4 Small Object Retrieval

Retrieving objects that cover a small portion of images is a challenging task due to 1) the few detected local features and 2) the large amount of background noise. The Instance Search task in the TRECVID campaign [117] and the task of logo retrieval are important venues/applications for this task.

Generally speaking, both TRECVID and logo retrieval can be tackled with similar pipelines. For keypoint-based methods, the spatial context among the local features is important to discriminative target objects from others, especially in cases of rigid objects. Examples include [118], [119], [120]. Other effective methods include burstiness handling [77] (discussed in Section 3.4.3), considering the different inlier ratios between the query and target objects [121], *etc*. In the second type of methods, effective region proposals [122] or multi-scale image patches [123] can be used as object region candidates. In [123], a recent state-of-the-art method, a regional diffusion mechanism based on neighborhood graphs is proposed to further improve the recall of small objects.

4 CNN-BASED IMAGE RETRIEVAL

CNN-based retrieval methods have constantly been proposed in recent years and are gradually replacing the hand-crafted local detectors and descriptors. In this survey, CNN-based methods are classified into three categories: using pre-trained CNN models, using fine-tuned CNN models and hybrid methods. The first two categories compute the global feature with a single network pass, and the hybrid methods may require multiple network passes (see Fig. 2).

4.1 Retrieval Using Pre-trained CNN Models

This type of methods is efficient in feature computation due to the single-pass mode. Given the transfer nature, its success lies in the feature extraction and encoding steps. We will first describe some commonly used datasets and networks for pre-training, and then the feature computation process.

models	size	# layers	training Set	used in
OverFeat [132]	144M	6+3	ImageNet	[7]
AlexNet [6]			ImageNet	[22], [133]
PlacesNet [129]	60M	5+3	Places	[130], [131]
HybridNet [129]			ImageNet+Places	[130], [131]
VGGNet [124]	138M	13+3	ImageNet	[10]
GoogleNet [125]	11M	22	ImageNet	[9]
ResNet [126]	44.6M	101	ImageNet	n.a

TABLE 2: Pre-trained CNN models that can be used.

4.1.1 Pre-trained CNN Models

Popular CNN architectures. Several CNN models serve as good choices for extracting features, including AlexNet [6], VGGNet [124], GoogleNet [125] and ResNet [126], which are listed in Table 2. Briefly, CNN can be viewed as a set of non-linear functions and is composed of a number of layers such as convolution, pooling, non-linearities, *etc*. CNN has a hierarchical structure. From bottom to top layers, the image undergoes convolution with filters, and the receptive field of these image filters increases. Filters in the same layer have the same size but different parameters. AlexNet [6] was proposed the earliest among these networks, which has five convolutional layers and three fully connected (FC) layers. It has 96 filters in the first layer of sizes $11 \times 11 \times 3$ and has 256 filters of size $3 \times 3 \times 192$ in the 5th layer. Zeiler *et al.* [127] observe that the filters are sensitive to certain visual patterns and that these patterns evolve from low-level bars in bottom layers to high-level objects in top layers. For low-level and simple visual stimulus, the CNN filters act as the detectors in the local hand-crafted features, but for the high-level and complex stimulus, the CNN filters have distinct characteristics that depart from SIFT-like detectors. AlexNet has been shown to be outperformed by newer ones such as VGGNet, which has the largest number of parameters. ResNet and GoogleNet won the ILSVRC 2014 and 2015 challenges, respectively, showing that CNNs are more effective with more layers. A full review of these networks is beyond the scope of this paper, and we refer readers to [6], [128], [124] for details.

Datasets for pre-training. Several large-scale recognition datasets are used for CNN pre-training. Among them, the ImageNet dataset [23] is mostly commonly used. It contains 1.2 million images of 1000 semantic classes and is usually thought of as being generic. Another data source for pre-training is the Places-205 dataset [129] which is twice as large as ImageNet but has five times fewer classes. It is a scene-centric dataset depicting various indoor and outdoor scenes. A hybrid dataset combining the Places-205 and the ImageNet datasets has also been used for pre-training [129]. The resulting HybridNet is evaluated in [125], [126], [130], [131] for instance retrieval.

The transfer issue. Comprehensive evaluation of various CNNs on instance retrieval has been conducted in several recent works [130], [131], [134]. The transfer effect is mostly concerned. It is considered in [130] that instance retrieval, as a target task, lies farthest from the source, *i.e.,* ImageNet. Studies reveal some critical insights in the transfer process. First, during model transfer, features extracted from different layers exhibit different retrieval performance. Experiments confirm that the top layers may exhibit lower generalization ability than the layer before it. For example, for AlexNet

pre-trained on ImageNet, it is shown that FC6, FC7, and FC8 are in descending order regarding retrieval accuracy [130]. It is also shown in [10], [134] that the pool5 feature of AlexNet and VGGNet is even superior to FC6 when proper encoding techniques are employed. Second, the source training set is relevant to retrieval accuracy on different datasets. For example, Azizpour *et al.* [130] report that HybridNet yields the best performance on Holidays after PCA. They also observe that AlexNet pre-trained on ImageNet is superior to PlacesNet and HybridNet on the Ukbench dataset [11] which contains common objects instead of architectures or scenes. So the similarity of the source and target plays a critical role in instance retrieval when using a pre-trained CNN model.

4.1.2 Feature Extraction

FC descriptors. The most straightforward idea is to extract the descriptor from the fully-connected (FC) layer of the network [7], [8], [135], *e.g.,* the 4,096-dim FC6 or FC7 descriptor in AlexNet. The FC descriptor is generated after layers of convolutions with the input image, has a global receptive field, and thus can be viewed as a global feature. It yields fair retrieval accuracy under Euclidean distance and can be improved with power normalization [14].

Intermediate local features. Many recent retrieval methods [9], [10], [134] focus on local descriptors in the intermediate layers. In these methods, lower-level convolutional filters (kernels) are used to detect local visual patterns. Viewed as local detectors, these filters have a smaller receptive field and are densely applied on the entire image. Compared with the global FC feature, local detectors are more robust to image transformations such as truncation and occlusion, in ways that are similar to the local invariant detectors (Section 3.2).

Local descriptors are tightly coupled with these intermediate local detectors, *i.e.,* they are the responses of the input image to these convolution operations. In other words, after the convolutions, the resulting activation maps can be viewed as a feature ensemble, which is called the "column feature" in this survey. For example in AlexNet [6], there are $n = 96$ detectors (convolutional filters) in the 1st convolutional layer. These filters produces $n = 96$ heat maps of size 27×27 (after max pooling). Each pixel in the maps has a receptive field of 19×19 and records the response of the image *w.r.t* the corresponding filter [9], [10], [134]. The column feature is therefore of size $1 \times 1 \times 96$ (Fig. 2) and can be viewed as a description of a certain patch in the original image. Each dimension of this descriptor denotes the level of activation of the corresponding detector and resembles the SIFT descriptor to some extent. The column feature initially appears in [133], where Razavian *et al.* first do max-pooling over regularly partitioned windows on the feature maps and then concatenate them across all filter responses, yielding column-like features. In [136], column features from multiple layers of the networks are concatenated, forming the "hypercolumn" feature.

4.1.3 Feature Encoding and Pooling

When column features are extracted, an image is represented by a set of descriptors. To aggregate these descriptors into a global representation, currently two strategies are adopted: encoding and direct pooling (Fig. 2).

Encoding. A set of column features resembles a set of SIFT features. So standard encoding schemes can be directly employed. The most commonly used methods are VLAD [15] and FV [14]. A brief review of VLAD and FV can be seen in Section 3.3.2. A milestone work is [9], in which the column features are encoded into VLAD for the first time. This idea was later extended to CNN model fine-tuning [137]. The BoW encoding can also be leveraged, as the case in [138]. The column features within each layer are aggregated into a BoW vector which is then concatenated across the layers. An exception to these fix-length representations is [139], in which the column features are quantized with a codebook of size 25k and an inverted index is employed for efficiency.

Pooling. A major difference between the CNN column feature and SIFT is that the former has an explicit meaning in each dimension, *i.e.,* the response of a particular region of the input image to a filter. Therefore, apart from the encoding schemes mentioned above, direct pooling techniques can produce discriminative features as well.

A milestone work in this direction consists in the Maximum activations of convolutions (MAC) proposed by Tolias *et al.* [10]. Without distorting or cropping images, MAC computes a global descriptor with a single forward pass. Specifically, MAC calculates the maximum value of each intermediate feature map and concatenates all these values within a convolutional layer. In its multi-region version, the integral image and an approximate maximum operator are used for fast computation. The regional MAC descriptors are subsequently sum-pooled along with a series of normalization and PCA-whitening operations [53]. We also note in this survey that several other works [140], [133], [134] also employ similar ideas with [10] in employing max or average pooling on the intermediate feature maps and that Razavian *et al.* [133] are the first. It has been observed that the last convolutional layer (*e.g.,* pool5 in VGGNet), after pooling usually yields superior accuracy to the FC descriptors and the other convolutional layers [134].

Apart from direct feature pooling, it is also beneficial to assign some specific weights to the feature maps within each layer before pooling. In [140], Babenko *et al.* propose the injection of the prior knowledge that objects tend to be located toward image centers, and impose a 2-D Gaussian mask on the feature maps before sum pooling. Xie *et al.* [141] improve the MAC representation [10] by propagating the high-level semantics and spatial context to low-level neurons for improving the descriptive ability of these bottom-layer activations. With a more general weighting strategy, Kalantidis *et al.* [16] perform both feature map-wise and channel-wise weighing, which aims to highlight the highly active spatial responses while reducing burstiness effects.

4.2 Image Retrieval with Fine-Tuned CNN Models

Although pre-trained CNN models have achieved impressive retrieval performance, a hot topic consists in fine-tuning the CNN model on specific training sets. When a fine-tuned CNN model is employed, the image-level descriptor is usually generated in an end-to-end manner, *i.e.,* the network will produce a final visual representation without additional explicit encoding or pooling steps.

name	# images	# classes	content
Landmarks [8]	213,678	672	Landmark
3D Landmark [24]	163,671	713	Landmark
Tokyo TM [137]	112,623	n.a	Landmark
MV RGB-D [142]	250,000	300	House. object
Product [143]	101,945×2	n.a	Furniture

TABLE 3: Statistics of instance-level datasets having been used in fine-tuning.

4.2.1 Datasets for Fine-Tuning

The nature of the datasets used in fine-tuning is the key to learning discriminative CNN features. ImageNet [23] only provides images with class labels. So the pre-trained CNN model is competent in discriminating images of different object/scene classes, but may be less effective to tell the difference between images that fall in the same class (*e.g.,* architecture) but depict different instances (*e.g.,* "Eiffel Tower" and "Notre-Dame"). Therefore, it is important to fine-tune the CNN model on task-oriented datasets.

The datasets having been used for fine-tuning in recent years are shown in Table 3. Buildings and common objects are the focus. The milestone work on fine-tuning is [8]. It collects the **Landmarks dataset** by a semi-automated approach: automated searching for the popular landmarks in Yandex search engine, followed by a manual estimation of the proportion of relevant image among the top ranks. This dataset contains 672 classes of various architectures, and the fine-tuned network produces superior features on landmark related datasets such as Oxford5k [12] and Holidays [13], but has decreased performance on Ukbench [11] where common objects are presented. Babenko *et al.* [8] have also fine-tuned CNNs on the **Multi-view RGB-D dataset** [142] containing turntable views of 300 household objects, in order to improve performance on Ukbench. The Landmark dataset is later used by Gordo *et al.* [17] for fine-tuning, after an automatic cleaning approach based on SIFT matching. In [24], Radenović *et al.* employ the retrieval and Structure-From-Motion methods to build **3D landmark** models so that images depicting the same architecture can be grouped. Using this labeled dataset, the linear discriminative projections (denoted as L_w in Table 5) outperform the previous whitening technique [53]. Another dataset called **Tokyo Time Machine** is collected using Google Street View Time Machine which provides images depicting the same places over time [137]. While most of the above datasets focus on landmarks, Bell *et al.* [143] build a **Product dataset** consisting of furniture by developing a crowd-sourced pipeline to draw connections between in-situ objects and the corresponding products. It is also feasible to fine-tune on the query sets suggested in [144], but this method may not be adaptable to new query types.

4.2.2 Networks in Fine-Tuning

The CNN architectures used in fine-tuning mainly fall into two types: the classification-based network and the verification-based network. The classification-based network is trained to classify architectures into pre-defined categories. Since there is usually no class overlap between the training set and the query images, the learned embedding *e.g.,* FC6 or FC7 in AlexNet, is used for Euclidean distance based

retrieval. This train/test strategy is employed in [8], in which the last FC layer is modified to have 672 nodes corresponding to the number of classes in the Landmark dataset.

The verification network may either use a siamese network with pairwise loss or use a triplet loss and has been more widely employed for fine-tuning. A standard siamese network based on AlexNet and the contrastive loss is employed in [143]. In [24], Radenović *et al.* propose to replace the FC layers with a MAC layer [10]. Moreover, with the 3D architecture models built in [24], training pairs can be mined. Positive image pairs are selected based on the number of co-observed 3D points (matched SIFT features), while hard negatives are defined as those with small distances in their CNN descriptors. These image pairs are fed into the siamese network, and the contrastive loss is calculated from the ℓ_2 normalized MAC features. In a concurrent work to [24], Gordo *et al.* [17] fine-tune a triplet-loss network and a region proposal network on the Landmark dataset [8]. The superiority of [17] consists in its localization ability, which excludes the background in feature learning and extraction. In both works, the fine-tuned models exhibit state-of-the-art accuracy on landmark retrieval datasets including Oxford5k, Paris6k and Holidays, and also good generalization ability on Ukbench (Table 5). In [137], a VLAD-like layer is plugged in the network at the last convolutional layer which is amenable to training via back-propagation. Meanwhile, a new triplet loss is designed to make use of the weakly supervised Google Street View Time Machine data.

4.3 Hybrid CNN-based Methods

For the hybrid methods, multiple network passes are performed. A number of image patches are generated from an input image, which are fed into the network for feature extraction before an encoding/pooling stage. Since the manner of "detector + descriptor" is similar to SIFT-based methods, we call this method type "hybrid". It is usually less efficient than the single-pass methods.

4.3.1 Feature Extraction

In hybrid methods, the feature extraction process consists of patch detection and description steps. For the first step, the literature has seen three major types of region detectors. The first is grid image patches. For example, in [22], a two-scale sliding window strategy is employed to generate patches. In [7], the dataset images are first cropped and rotated, and then divided into patches of different scales, the union of which covers the whole image. The second type is invariant keypoint/region detectors. For instance, the difference of Gaussian feature points are used in [145]; the MSER region detector is leveraged in [146]. Third, region proposals also provide useful information on the locations of the potential objects. Mopuri *et al.* [147] employ selective search [148] to generate image patches, while EdgeBox [149] is used in [150]. In [144], the region proposal network (RPN) [151] is applied to locate the potential objects in an image.

The use of CNN as region descriptors is validated in [146], showing that CNN is superior to SIFT in image matching except on blurred images. Given the image patches, the hybrid CNN method usually employs the FC or pooled intermediate CNN features. Examples using the FC descriptors

include [7], [22], [147], [152]. In these works, the 4,096-dim FC features are extracted from the multi-scale image regions [7], [22], [152] or object proposals [147]. On the other hand, Razavian *et al.* [133] also uses the intermediate descriptors after max-pooling as region descriptors.

The above methods use pre-trained models for patch feature extraction. Based on the hand-crafted detectors, patch descriptors can also be learned through CNN in either supervised [153] or unsupervised manner [145], which improves over the previous works on SIFT descriptor learning [45], [34]. Yi *et al.* [154] further propose an end-to-end learning method integrating region detector, orientation estimator and feature descriptor in a single pipeline.

4.3.2 Feature Encoding and Indexing

The encoding/indexing procedure of hybrid methods resembles SIFT-based retrieval, *e.g.,* VLAD/FV encoding under a small codebook or the inverted index under a large codebook.

The VLAD/FV encoding, such as [22], [147], follow the standard practice in the case of SIFT features [14], [15], so we do not detail here. On the other hand, several works exploit the inverted index on the patch-based CNN features [139], [155], [156]. Again, standard techniques in SIFT-based methods such as HE are employed [156]. Apart from the above-mentioned strategies, we notice that several works [7], [133], [152] extract several region descriptors per image to do a many-to-many matching, called "spatial search" [7]. This method improves the translation and scale invariance of the retrieval system but may encounter efficiency problems. A reverse strategy to applying encoding on top of CNN activations is to build a CNN structure (mainly consisting of FC layers) on top of SIFT-based representations such as FV. By training a classification model on natural images, the intermediate FC layer can be used for retrieval [157].

4.4 Discussions

4.4.1 Relationship between SIFT- and CNN-based Methods

In this survey, we categorize current literature into six fine-grained classes. The differences and some representative works of the six categories are summarized in Table 1 and Table 5. Our observation goes below.

First, the hybrid method can be viewed as a transition zone from SIFT- to CNN-based methods. It resembles the SIFT-based methods in all the aspects except that it extracts CNN features as the local descriptor. Since the network is accessed multiple times during patch feature extraction, the efficiency of the feature extraction step may be compromised.

Second, the single-pass CNN methods tend to combine the individual steps in the SIFT-based and hybrid methods. In Table 5, the "pre-trained single-pass" category integrates the feature detection and description steps; in the "fine-tuned single-pass" methods, the image-level descriptor is usually extracted in an end-to-end mode, so that no separate encoding process is needed. In [17], a "PCA" layer is integrated for discriminative dimension reduction, making a further step towards end-to-end feature learning.

Third, fixed-length representations are gaining more popularity due to efficiency considerations. It can be obtained by aggregating local descriptors (SIFT or CNN) [15], [18], [22], [9], direct pooling [147], [10], or end-to-end feature

name	# images	# queries	content
Holidays [13]	1,491	500	scene
Ukbench [11]	10,200	10,200	common objects
Paris6k [25]	6,412	55	buildings
Oxford5k [12]	5,062	55	buildings
Flickr100k [25]	99,782	-	from Flickr's popular tags

TABLE 4: Statistics of popular instance-level datasets.

computation [8], [17]. Usually, dimension reduction methods such as PCA can employed on top of the fixed-length representations, and ANN search methods such as PQ [15] or hashing [47] can be used for fast retrieval.

4.4.2 Hashing and Instance Retrieval

Hashing is a major solution to the approximate nearest neighbor problem. It can be categorized into locality sensitive hashing (LSH) [63] and learning to hash. LSH is data-independent and is usually outperformed by learning to hash, a data-dependent hashing approach. For learning to hash, a recent survey [67] categorizes it into quantization and pairwise similarity preserving. The quantization methods are briefly discussed in Section 3.3.2. For the pairwise similarity preserving methods, some popular hand-crafted methods include Spectral hashing [64], LDA hashing [158], *etc.*

Recently, hashing has seen a major shift from hand-crafted to supervised hashing with deep neural networks. These methods take the original image as input and produce a learned feature before binarization [159], [160]. Most of these methods, however, focus on class-level image retrieval, a different task with instance retrieval discussed in this survey. For instance retrieval, when adequate training data can be collected, such as architecture and pedestrians, the deep hashing methods may be of critical importance.

5 EXPERIMENTAL COMPARISONS

5.1 Image Retrieval Datasets

Five popular instance retrieval datasets are used in this survey. Statistics of these datasets can be accessed in Table 4.

Holidays [13] is collected by Jégou *et al.* from personal holiday albums, so most of the images are of various scene types. The database has 1,491 images composed of 500 groups of similar images. Each image group has 1 query, totaling 500 query images. Most SIFT-based methods employ the original images, except [32], [71] which manually rotate the images into upright orientations. Many recent CNN-based methods [140], [137], [16] also use the rotated version of Holidays. In Table 5, results of both versions of Holidays are shown (separated by "/"). Rotating the images usually brings 2-3% mAP improvement.

Ukbench [11] consists of 10,200 images of various content, such as objects, scenes, and CD covers. All the images are divided into 2,550 groups. Each group has four images depicting the same object/scene, under various angles, illuminations, translations, *etc.* Each image in this dataset is taken as the query in turn, so there are 10,200 queries.

Oxford5k [12] is collected by crawling images from Flickr using the names of 11 different landmarks in Oxford. A total

method type		methods	detector	descriptor	encoding	indexing	voc	dim	Holidays	Ukbench	Oxford5k	+100k	Paris6k	mem/img
SIFT-based	Large Voc.	HKM [11]	MSER	SIFT	hier, soft, BoW	inv. index	1M		59.7	2.85	44.3†	26.6†	46.5†	9.8kb
		AKM [12]	hes-aff	SIFT	hard, BoW	inv. index	1M		64.1†	3.02†	49.3	34.3	50.2†	9.8kb
		Fine Voc. [71]	hes-aff	SIFT	alt. word, BoW	inv. index	16M		-/74.9 (75.8)	-	74.2 (84.9)	67.4 (79.5)	74.9 (82.4)	9.8kb
		Three Things [33]	hes-aff	rootSIFT	hard, BoW	inv. index	1M		-	-	80.9*	72.2*	76.5*	22.0kb
		Co-index [72]	DoG	SIFT, CNN	hard, BoW	co-index	1M		80.86	3.60	68.72	-	-	21.6kb
	Mid Voc.	HE+WGC [13], [85]	hes-aff	SIFT	hard, BoW, HE	inv. index	20k		81.3 (84.2)	3.42 (3.55)	61.5 (74.7)	51.6 (68.7)	-	36.8kb
		Burst [76]	hes-aff	SIFT	burst, BoW, HE	inv. index	20k		83.9 (84.8)	3.54 (3.64)	64.7 (68.5)	58§ (63§)	62.8†	36.8kb
		Q.ada [91]	hes-aff	rootSIFT	MA, BoW, HE	inv. index	20k		78.0	-	82.1	72.8	73.6	36.8kb
		ASMK [20]	hes-aff	rootSIFT	MA, BoW, agg.HE	inv. index	65k		81.0	-	80.4	75.0 (85.0)	77.0	43.2kb
		c-MI [31]	hes-aff	rootSIFT, HS	MA, BoW, HE	2D index	20k×200		84.0	3.71	58.2†	35.2†	55.1†	45.2kb
	Small Voc.	VLAD [15]	hes-aff	SIFT	VLAD	PCA, PQ	64	4,096	55.6	3.18†	37.8	27.2†	38.6†	16kb
		FV [47], [52]	hes-aff	PCA-SIFT	FV, pw.	LSH, SH	64	4,096	59.5	3.35	41.8	33.1†	43.0	16kb
		All A. VLAD [54]	hes-aff	SIFT	Improved VLAD	PCA	64	128	62.5	-	44.8	-	-	0.5kb
		NE [53]	hes-aff	rootSIFT	NE, multi.voc, VLAD	PCA$_w$	4×256	128	61.4	3.36	-	-	-	0.5kb
		Triangulation [18]	hes-aff	rootSIFT	triang+democ	PCA, pw.	16	128	61.7	3.40	43.3	35.3	-	0.5kb
CNN-based	Hybrid	Off the Shelf [7]	den. patch	OFeat 1st FC	-	-	-	-	84.3	3.64	-/68.0	-	-/79.5	4-15kb
		MSS [133]	den. patch	vgg conv5	MP	-	-	-	88.1	3.72	-/84.4	-	-/85.3	16kb
		CKN [153]	hes-aff	CKN	VLAD, power	-	256	65k	79.3	3.76	56.5	-	-	256kb
		MOP [22]	den. patch	alex FC7	multi-scale VLAD	PCA$_w$	100	2,048	80.2	-	-	-	-	8kb
		OLDFP [147]	sel. search	alex FC7	MP	ITQ [161]	-	512	88.5	3.81	60.7	-	66.2	2kb
	Pre-trained single-pass	BLCF [139]	vgg16, conv5		hard, BoW	inv. index	25k		-	-	73.9 (78.8)	59.3 (65.1)	82.0 (84.8)	0.67kb
		R-MAC [10]	vgg16, conv5		region MP, SP	PCA$_w$	-	512	85.2/86.9	-	66.9 (77.3)	61.6 (73.2)	83.0 (86.5)	1kb
		CroW [16]	vgg16, conv5		cross-dim pool.	PCA$_w$	-	512	-/85.1	-	70.8 (74.9)	65.3 (70.6)	79.7 (84.8)	2kb
		SPoC [140]	vgg16, conv5		SP, center prior	PCA$_w$	-	256	-/80.2	3.65	53.1‡/58.9	50.1‡/57.8	-	1kb
		VLAD-CNN [9]	google, various incept.		intra-norm, VLAD	PCA	100	128	83.6	-	-/55.8	-	-/58.3	0.5kb
	Fine-tuned single-pass	Faster R-CNN [144]	vgg16-reg-query, conv5		SP or MP	-	-	512	-	-	71.0‡ (78.6)	-	79.8‡ (84.2)	2kb
		Neural Codes [8]	alexnet-classification loss-Landmark, FC6			PCA	-	128	-/78.9	3.29	-/55.7	-/52.3	-	0.5kb
		NetVLAD [137]	vgg16-trip. loss-Tokyo TM, VLAD layer		VLAD layer	PCA$_w$	64	512	81.7/86.1	-	65.6/67.6	-	73.4/74.9	8kb
		SiaMAC [24]	vgg16-pair loss-3D Landmark, MAC layer		MAC layer	L$_w$	-	512	-/82.5	3.61†	77.0 (82.9)	69.2 (77.9)	83.8 (85.6)	2kb
		Deep Retrieval [17]	vgg16-trip. loss-cleaned Landmark, PCA layer used				-	512	86.7/89.1	3.62	83.1 (89.1)	78.6 (87.3)	87.1 (91.2)	2kb
		[17], [162]	ResNet101-trip. loss-cleaned Landmark, PCA layer used				-	2,048	90.3/94.8	3.84	86.1 (90.6)	82.8 (89.4)	94.5 (96.0)	8kb

TABLE 5: Performance summarization of some representative methods of the six categories on the benchmarks. "+100k" –> the addition of Flickr100k into Oxford5k. "pw." –> power low normalization [14]. "MP" –> max pooling. "SP" –> sum pooling. * and parentheses –> results are obtained with post-processing steps such as spatial verification or QE. § –> numbers are estimated from the curves. † numbers are reported by our implementation. For Holidays, results using the rotated images are presented after "/". For Oxford5k (+100k) and Paris6k, results using the full-sized queries are shown after "/". ‡ –> the full query image is fed into the network, but only the features whose centers fall into the query region of interest are aggregated. Note that in many fixed-length representations, ANN algorithms such as PQ are not used to report the results, but ANN can be readily applied after PCA during indexing.

(a) Holidays, *mAP*

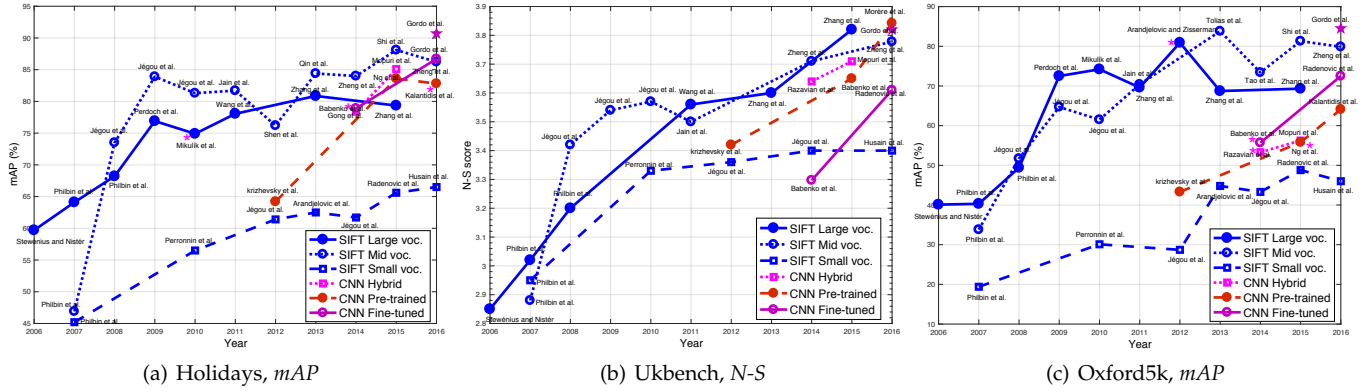(b) Ukbench, *N-S*

(c) Oxford5k, *mAP*

Fig. 6: The state of the art over the years on the (a) Holidays, (b) Ukbench, and (c) Oxford5k datasets. Six fine-grained categories are summarized (see Section 2). For each year, the best accuracy of each category is reported. For the compact representations, results of 128-bit vectors are preferentially selected. The purple star denotes the results produced by 2,048-dim vectors [17], the best performance in fine-tuned CNN methods. Methods with a pink asterisk denote using rotated images on Holidays, full-sized queries on Oxford5k, or spatial verification and QE on Oxford5k (see Table 5).

of 5,062 images form the image database. The dataset defines five queries for each landmark by hand-drawn bounding boxes, so that 55 query Regions of Interest (ROI) exist in total. Each database image is assigned one of four labels, *good*, *OK*, *junk*, or *bad*. The first two labels are true matches to the query ROIs, while "bad" denotes the distractors. In junk images, less than 25% of the objects are visible, or they undergo severe occlusion or distortion, so these images have zero impact on retrieval accuracy.

Flickr100k [25] contains 99,782 high resolution images crawled from Flickr's 145 most popular tags. In literature, this dataset is typically added to Oxford5k to test the scalability of retrieval algorithms.

Paris6k [25] is featured by 6,412 images crawled from 11 queries on specific Paris architecture. Each landmark has five queries, so there are also 55 queries with bounding boxes. The database images are annotated with the same four types of labels as Oxford5k. Two major evaluation protocols exist for Oxford5k and Paris6k. For SIFT-based methods, the cropped regions are usually used as query. For CNN-based methods, some employ the full-sized query images [8], [137]; some follow the standard cropping protocol, either by cropping the ROI and feeding it into CNN [16] or extracting CNN features using the full image and selecting those falling in the ROI [144]. Using the full image may lead to mAP improvement. These protocols are used in Table 5.

5.2 Evaluation Metrics

Precision-recall. Recall denotes the ratio of returned true matches to the total number or true matches in the database, while precision refers to the fraction of true matches in the returned images. Given a subset of n returned images, assuming there are n_p true matches among them, and a total of N_p true matches exist in the whole database, then recall@n ($r@n$) and precision@n ($p@n$) are calculated as $\frac{n_p}{N_p}$ and $\frac{n_p}{n}$, respectively. In image retrieval, given a query image and its rank list, a precision-recall curve can be drawn on the (precision, recall) points $(r@1, p@1), (r@2, p@2), ..., (r@N, p@N)$, where N is the number of images in the database.

Average precision and mean average precision. To more clearly record the retrieval performance, average precision (AP) is used, which amounts to the area under the precision-recall curve. Typically, a larger AP means a higher precision-recall curve and thus better retrieval performance. Since retrieval datasets typically have multiple query images, their respective APs are averaged to produce a final performance evaluation, *i.e.*, the mean average precision (mAP). Conventionally, we use mAP to evaluate retrieval accuracy on the Oxford5k, Paris6k, and Holidays datasets.

N-S Score. The N-S score is specifically used on the Ukbench dataset and is named after David **N**istér and Henrik **S**tewénius [11]. It is equivalent to precision@4 or recall@4 because every query in Ukbench has four true matches in the database. The N-S score is calculated as the average number of true matches in the top-4 ranks across all the rank lists.

5.3 Comparison and Analysis

5.3.1 Performance Improvement Over the Years

We present the improvement in retrieval accuracy over the past ten years in Fig. 6 and the numbers of some representative methods in Table 5. The results are computed using codebooks trained on independent datasets [13]. We can clearly observe that the field of instance retrieval has constantly been improving. The baseline approach (HKM) proposed over ten years ago only yields a retrieval accuracy of 59.7%, 2.85, 44.3%, 26.6%, and 46.5% on Holidays, Ukbench, Oxford5k, Oxford5k+Flickr100k, and Paris6k, respectively. Starting from the baseline approaches [11], [12], methods using large codebooks improve steadily when more discriminative codebooks [71], spatial constraints [21], [82], and complementary descriptors [72], [163] are introduced. For medium-sized codebooks, the most significant accuracy advance has been witnessed in the years 2008-2010 with the introduction of Hamming Embedding [13], [85] and its improvements [76], [85], [90]. From then on, major improvements come from the strength of feature fusion [31], [163], [135] with the color and CNN features, especially on the Holidays and Ukbench datasets.
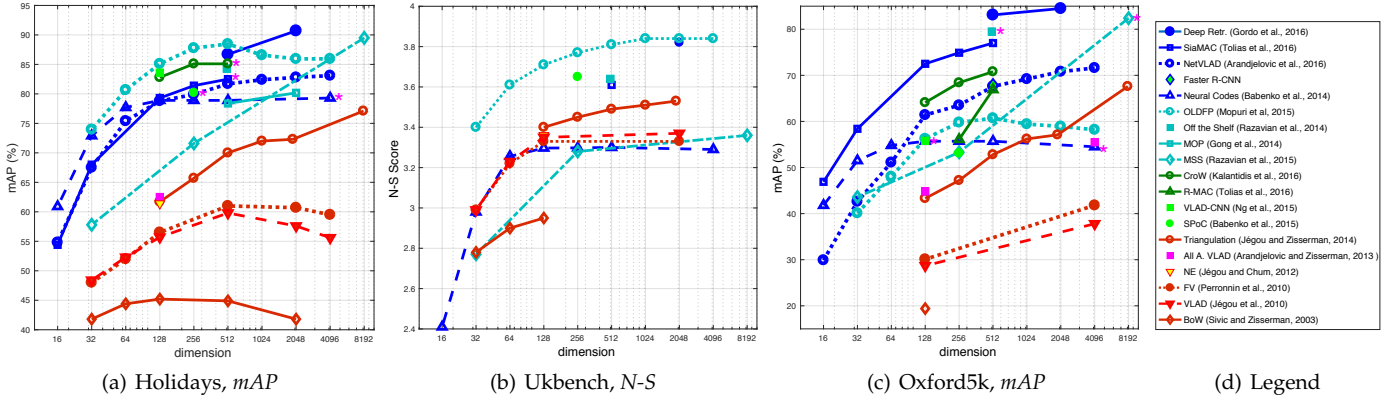
Fig. 7: The impact of feature dimension on retrieval accuracy. Compact (fixed-length) representations are shown, *i.e.,* SIFT small voc., hybrid CNN methods, pre-trained CNN methods, and fine-tuned CNN methods. Curves with a pink asterisk on the end indicates using rotated images or full-sized queries on Holidays and Oxford5k (see Table 5), *resp.*

method type	efficiency				accuracy	
	feat. ext.	retr.	mem.	train	generic	specific
SIFT large voc.	fair	high	fair	fair	high	fair
SIFT mid voc.	fair	low	low	fair	high	high
SIFT small voc.	fair	high	high	high	low	Low
CNN hybrid	low	varies	varies	varies	fair	fair
CNN pre-trained	high	high	high	high	high	fair
CNN fine-tuned	high	high	high	low	high	high

TABLE 6: A summary of efficiency and accuracy comparison between different categories. Note: feature extraction time is estimated using CPUs and GPUs (as is usually done) for SIFT and CNN, *resp.* When using GPUs for SIFT extraction, the efficiency could be high as well.

On the other hand, CNN-based retrieval models have quickly demonstrated their strengths in instance retrieval. In the year 2012 when the AlexNet [6] was introduced, the performance of the off-the-shelf FC features is still far from satisfactory compared with SIFT models during the same period. For example, the FC descriptor of AlexNet pre-trained on ImageNet yields 64.2%, 3.42, and 43.3% in mAP, N-S score, and mAP, respectively, on the Holidays, Ukbench, and Oxford5k datasets. These numbers are lower than [82] by 13.85%, 0.14 on Holidays and Ukbench, respectively, and lower than [111] by 31.9% on Oxford5k. However, with the advance in CNN architectures and fine-tuning strategies, the performance of the CNN-based methods is improving fast, being competitive on the Holidays and Ukbench datasets [17], [164], and slightly lower on Oxford5k but with much smaller memory cost [24].

5.3.2 Accuracy Comparisons

The retrieval accuracy of different categories on different datasets can be viewed in Fig. 6, Table 5 and Table 6. From these results, we arrive at three observations.

First, among the SIFT-based methods, those with medium-sized codebooks [13], [31], [19] usually lead to superior (or competitive) performance, while those based on small codebook (compact representations) [15], [18], [56] exhibit inferior accuracy. On the one hand, the visual words in the medium-sized codebooks lead to relatively high matching recall due to the large Voronoi cells. The further integration

of HE methods largely improves the discriminative ability, achieving a desirable trade-off between matching recall and precision. On the other hand, although the visual words in small codebooks have the highest matching recall, their discriminative ability is not significantly improved due to the aggregation procedure and the small dimensionality. So its performance can be compromised.

Second, among the CNN-based categories, the fine-tuned category [8], [17], [24] is advantageous in specific tasks (such as landmark/scene retrieval) which have similar data distribution with the training set. While this observation is within expectation, we find it interesting that the fine-tuned model proposed in [17] yields very competitive performance on generic retrieval (such as Ukbench) which has distinct data distribution with the training set. In fact, Babenko *et al.* [8] show that the CNN features fine-tuned on Landmarks compromise the accuracy on Ukbench. The generalization ability of [17] could be attributed to the effective training of the region proposal network. In comparison, using pre-trained models may exhibit high accuracy on Ukbench, but only yields moderate performance on landmarks. Similarly, the hybrid methods have fair performance on all the tasks, when it may still encounter efficiency problems [7], [152].

Third, comparing all the six categories, the "CNN fine-tuned" and "SIFT mid voc." categories have the best overall accuracy, while the "SIFT small voc." category has a relatively low accuracy.

5.3.3 Efficiency Comparisons

Feature computation time. For the SIFT-based methods, the dominating step is local feature extraction. Usually, it takes 1-2s for a CPU to extract the Hessian-Affine region based SIFT descriptors for a 640×480 image, depending on the complexity (texture) of the image. For the CNN-based method, it takes 0.082s and 0.347s for a single forward pass of a 224×224 and 1024×768 image through VGG16 on a TitanX card, respectively. It is reported in [17] that four images (with largest side of 724 pixels) can be processed in 1 second. The encoding (VLAD or FV) time of the pre-trained column features is very fast. For the CNN Hybrid methods, extracting CNN features out of tens of regions may take seconds. Overall speaking, the CNN pre-trained and
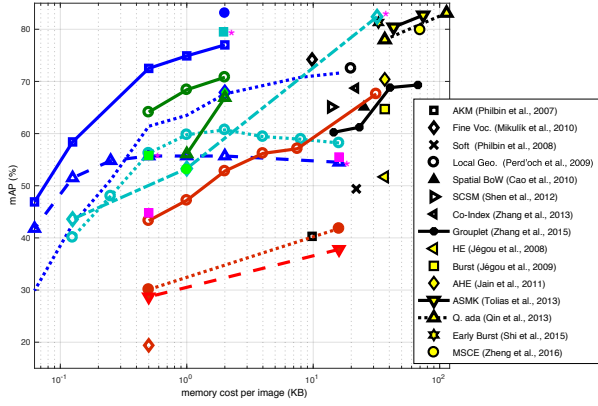
Fig. 8: Memory cost vs. retrieval accuracy on Oxford5k. In the legend, the first 8 methods are based on large codebooks, while the last 7 use medium-sized codebooks. This figure shares the same legend with Fig. 7(c) except the newly added numbers (in black).

(a) SIFT Large voc. (b) SIFT Mid voc.

Fig. 9: The impact of codebook size on SIFT-based methods using (a) large codebooks and (b) medium-sized codebooks on the Oxford5k dataset.

fine-tuned models are efficient in feature computation using GPUs. Yet it should be noted that when using GPUs for SIFT extraction, high efficiency could also be achieved.

Retrieval time. The efficiency of nearest neighbor search is high for "SIFT large voc.", "SIFT small voc.", "CNN pre-trained" and "CNN fine-tuned", because the inverted lists are short for a properly trained large codebook, and because the latter three have a compact representation to be accelerated by ANN search methods like PQ [61]. Efficiency for the medium-sized codebook is low because the inverted list contains more postings compared to a large codebook, and the filtering effect of HE methods can only correct this problem to some extent. The retrieval complexity for hybrid methods, as mentioned in Section 4.3, may suffer from the expensive many-to-many matching strategy [7], [133], [152].

Training time. Training a large or medium-sized codebook usually takes several hours with AKM or HKM. Using small codebooks reduces the codebook training time. For the fine-tuned model, Gordo *et al.* [17] report using five days on a K40 GPU for the triplet-loss model. It may take less time for the siamese [24] or the classification models [8], but should still much longer than SIFT codebook generation. Therefore, in terms of training, those using direct pooling [10], [134] or small codebooks [15], [9] are more time efficient.

Memory cost. Table 5 and Fig. 8 show that the SIFT methods with large codebooks and the compact representations are both efficient in memory cost. But the compact representations can be compressed into compact codes [53] using PQ or other competing quantization/hashing methods, so their memory consumption can be further reduced. In comparison, the methods using medium-sized codebooks are the most memory-consuming because the binary signatures should be stored in the inverted index. The hybrid methods somehow have mixed memory cost because the many-to-many strategy requires storing a number of region descriptors per image [7], [152] while some others employ efficient encoding methods [22], [147].

Spatial verification and query expansion. Spatial verification which provides refined rank lists is often used in conjunction with QE. The RANSAC verification proposed
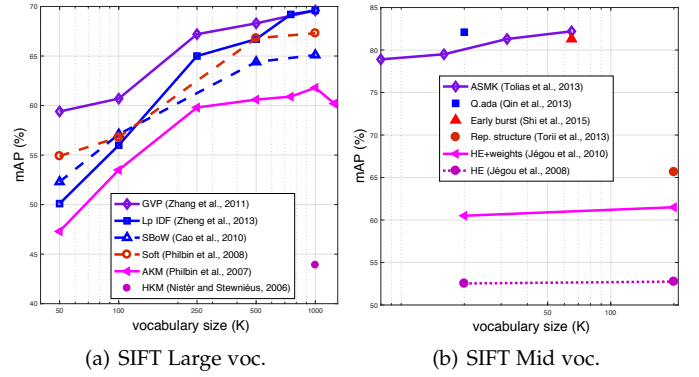
in [12] has a complexity of $\mathcal{O}(z^2)$, where z is the number of matched features. So this method is computationally expensive. The ADV approach [113] is less expensive with $\mathcal{O}(z \log z)$ complexity due to its ability to avoid unrelated Hough votes. The most efficient methods consist in [112], [115] which has a complexity of $\mathcal{O}(z)$, and [115] further outputs the transformation and inliers for QE.

From the perspective of query expansion, since new queries are issued, search efficiency is compromised. For example, AQE [100] almost doubles the search time due to the new query. For the recursive AQE and the scale-band recursive QE [100], the search time is much longer because several new searches are conducted. For other QE variants [101], [33], the proposed improvements only add marginal cost compared to performing another search, so their complexity is similar to basic QE methods.

5.3.4 Important Parameters

We summarize the impact of codebook size on SIFT methods using large/medium-sized codebooks, and the impact of dimensionality on compact representations including SIFT small codebooks and CNN-based methods.

Codebook size. The mAP results on Oxford5k are drawn in Fig. 9, and methods using large/medium-sized codebooks are compared. Two observations can be made. First, mAP usually increases with the codebook size but may reach saturation when the codebook is large enough. This is because a larger codebook improves the matching precision, but if it is too large, matching recall is lower, leading to saturated or even compromised performance [12]. Second, methods using the medium-sized codebooks have more stable performance when codebook size changes. This can be attributed to HE [13], which contributes more for a smaller codebook, compensating the lower baseline performance.

Dimensionality. The impact of dimensionality on compact vectors is presented in Fig. 7. Our finding is that the retrieval accuracy usually remains stable under larger dimensions, and drops quickly when the dimensionality is below 256 or 128. Our second finding favors the methods based on region proposals [147], [17]. These methods demonstrate very competitive performance under various feature lengths, probably due to their superior ability in object localization.

5.3.5 Discussions

We provide a brief discussion on when to use CNN over SIFT and the other way around. The above discussions provide comparisons between the two features. On the one hand, CNN-based methods with fixed-length representations have advantages in nearly all the benchmarking datasets. Specifically, in two cases, CNN-based methods can be assigned with higher priority. First, for specific object retrieval (*e.g.,* buildings, pedestrians) when sufficient training data is provided, the ability of CNN embedding learning can be fully utilized. Second, for common object retrieval or class retrieval, the pre-trained CNN models are competitive.

On the other hand, despite the usual advantages of CNN-based methods, we envision that the SIFT feature still has merits in some cases. For example, when the query or some target images are gray-scale, CNN may be less effective than SIFT because SIFT is computed on gray-scale images without resorting to color information. A similar situation involves when object color change is highly intense. In another example, for small object retrieval or when the queried object undergoes severe occlusions, the usage of local features like SIFT is favored. In applications like book/CD cover retrieval, we can also expect good performance out of SIFT due to the rich textures.

6 FUTURE RESEARCH DIRECTIONS

6.1 Towards Generic Instance Retrieval

A critical direction is to make the search engine applicable to generic search purpose. Towards this goal, two important issues should be addressed. First, large-scale instance-level datasets are to be introduced. While several instance datasets have been released as shown in Table 3, these datasets usually contain a particular type of instances such as landmarks or indoor objects. Although the RPN structure used by Gordo *et al.* [17] has proven competitive on Ukbench in addition to the building datasets, it remains unknown if training CNNs on more generic datasets will bring further improvement. Therefore, the community is in great need of large-scale instance-level datasets or efficient methods for generating such a dataset in either a supervised or unsupervised manner.

Second, designing new CNN architectures and learning methods are important in fully exploiting the training data. Previous works employ standard classification [8], pairwise-loss [24] or Triplet-loss [165], [17] CNN models for fine-tuning. The introduction of Faster R-CNN to instance retrieval is a promising starting point towards more accurate object localization [17]. Moreover, transfer learning methods are also important when adopting a fine-tuned model in another retrieval task [166].

6.2 Towards Specialized Instance Retrieval

To the other end, there are also increasing interests in specialized instance retrieval. Examples include place retrieval [167], pedestrian retrieval [168], vehicle retrieval [169], logo retrieval [77], *etc*. Images in these tasks have specific prior knowledge that can be made use of. For example in pedestrian retrieval, the recurrent neural network (RNN) can be employed to pool the body part or patch descriptors. In vehicle retrieval, the view information can be inferred during feature learning, and the license plate can also provide critical information when being captured within a short distance.

Meanwhile, the process of training data collection can be further explored. For example, training images of different places can be collected via Google Street View [137]. Vehicle images can be accessed either through surveillance videos or internet images. Exploring new learning strategies in these specialized datasets and studying the transfer effect would be interesting. Finally, compact vectors or short codes will also become important in realistic retrieval settings.

7 CONCLUDING REMARKS

This survey reviews instance retrieval approaches based on the SIFT and CNN features. According to the codebook size, we classify the SIFT-based methods into three classes: using large, medium-sized, and small codebook. According to the feature extraction process, the CNN-based methods are categorized into three classes, too: using pre-trained models, fine-tuned models, and hybrid methods. A comprehensive survey of the previous approaches is conducted under each of the defined categories. The category evolution suggests that the hybrid methods are in the transition position between SIFT and CNN-based methods, that compact representations are getting popular, and that instance retrieval is working towards end-to-end feature learning and extraction.

Through the collected experimental results on several benchmark datasets, comparisons are made between the six method categories. Our findings favor the usage of CNN fine-tuning strategy, which yields competitive accuracy on various retrieval tasks and has advantages in efficiency. Future research may focus on learning more generic feature representations or more specialized retrieval tasks.

REFERENCES

[1] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *ECCV*, 2010.

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

[3] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.

[4] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV Workshops*, 2004.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[7] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *CVPR Workshops*, 2014.

[8] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *ECCV*, 2014.

[9] J. Ng, F. Yang, and L. Davis, "Exploiting local features from deep networks for image retrieval," in *CVPR Workshops*, 2015.

[10] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," in *ICLR*, 2016.

[11] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *CVPR*, 2006.

[12] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007.

[13] H. Jégou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV*, 2008.

[14] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010.

[15] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010.

[16] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *ECCV*, 2016.

[17] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *ECCV*, 2016.

[18] H. Jégou and A. Zisserman, "Triangulation embedding and democratic aggregation for image search," in *CVPR*, 2014.

[19] M. Shi, Y. Avrithis, and H. Jégou, "Early burst detection for memory-efficient image retrieval," in *CVPR*, 2015.

[20] G. Tolias, Y. Avrithis, and H. Jégou, "To aggregate or not to aggregate: Selective match kernels for image search," in *ICCV*, 2013.

[21] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *CVPR*, 2011.

[22] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *ECCV*, 2014.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[24] F. Radenović, G. Tolias, and O. Chum, "Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples," in *ECCV*, 2016.

[25] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *CVPR*, 2008.

[26] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[27] P. R. Beaudet, "Rotationally invariant image operators," in *Proc. 4th Int. Joint Conf. Pattern Recog, Tokyo, Japan, 1978*, 1978.

[28] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.

[29] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *CVPR*, 2009.

[30] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *IJCV*, vol. 65, no. 1-2, pp. 43–72, 2005.

[31] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Packing and padding: Coupled multi-index for accurate image retrieval," in *CVPR*, 2014.

[32] M. Perdóch, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval," in *CVPR*, 2009.

[33] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*, 2012.

[34] K. Simonyan, A. Vedaldi, and A. Zisserman, "Learning local feature descriptors using convex optimisation," *TPAMI*, vol. 36, no. 8, pp. 1573–1585, 2014.

[35] T. Ge, Q. Ke, and J. Sun, "Sparse-coded features for image retrieval," in *BMVC*, 2013.

[36] R. Arandjelović and A. Zisserman, "Smooth object retrieval using a bag of boundaries," in *International Conference on Computer Vision*, 2011, pp. 375–382.

[37] R. Sicre and T. Gevers, "Dense sampling of features for image retrieval," in *ICIP*, 2014.

[38] W.-L. Zhao, H. Jégou, and G. Gravier, "Oriented pooling for dense and non-dense rotation-invariant features," in *BMVC-24th British machine vision conference*, 2013.

[39] A. Iscen, G. Tolias, P.-H. Gosselin, and H. Jégou, "A comparison of dense region detectors for image search and fine-grained classification," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2369–2381, 2015.

[40] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *TPAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

[41] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *CVPR*, 2004.

[42] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *ECCV*, 2006, pp. 404–417.

[43] L. Juan and O. Gwun, "A comparison of sift, pca-sift and surf," *International Journal of Image Processing*, vol. 3, no. 4, pp. 143–152, 2009.

[44] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*, 2011.

[45] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Descriptor learning for efficient retrieval," in *ECCV*, 2010.

[46] F. Radenović, H. Jégou, and O. Chum, "Multiple measurements and joint dimensionality reduction for large scale image search with short vectors," in *ICMR*, 2015.

[47] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *CVPR*, 2010.

[48] P. Koniusz, F. Yan, and K. Mikolajczyk, "Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection," *CVIU*, vol. 117, no. 5, pp. 479–492, 2013.

[49] P.-H. Gosselin, N. Murray, H. Jégou, and F. Perronnin, "Revisiting the fisher vector for fine-grained classification," *Pattern Recognition Letters*, vol. 49, pp. 92–98, 2014.

[50] R. G. Cinbis, J. Verbeek, and C. Schmid, "Approximate fisher kernels of non-iid image models for image categorization," *TPAMI*, vol. 38, no. 6, pp. 1084–1098, 2015.

[51] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods." in *BMVC*, 2011.

[52] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *TPAMI*, vol. 34, no. 9, pp. 1704–1716, 2012.

[53] H. Jégou and O. Chum, "Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening," in *ECCV*, 2012.

[54] R. Arandjelovic and A. Zisserman, "All about vlad," in *CVPR*, 2013.

[55] J. Delhumeau, P.-H. Gosselin, H. Jégou, and P. Pérez, "Revisiting the vlad image representation," in *ACM MM*, 2013.

[56] S. S. Husain and M. Bober, "Improving large-scale image retrieval through robust aggregation of local descriptors," *TPAMI*, 2016.

[57] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[58] G. Tolias, T. Furon, and H. Jégou, "Orientation covariant aggregation of local descriptors with embeddings," in *ECCV*, 2014.

[59] N. Murray and F. Perronnin, "Generalized max pooling," in *CVPR*, 2014.

[60] Z. Liu, S. Wang, and Q. Tian, "Fine-residual vlad for image retrieval," *Neurocomputing*, vol. 173, pp. 1183–1191, 2016.

[61] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *TPAMI*, vol. 33, no. 1, pp. 117–128, 2011.

[62] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *TPAMI*, vol. 36, no. 11, pp. 2227–2240, 2014.

[63] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the annual ACM symposium on Theory of computing*, 1998.

[64] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *NIPS*, 2009, pp. 1753–1760.

[65] E. Spyromitros-Xioufis, S. Papadopoulos, I. Y. Kompatsiaris, G. Tsoumakas, and I. Vlahavas, "A comprehensive study over vlad and product quantization in large-scale image retrieval," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1713–1728, 2014.

[66] M. Douze, H. Jégou, and F. Perronnin, "Polysemous codes," in *ECCV*, 2016.

[67] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *TPAMI*, 2017.

[68] M. Shi, T. Furon, and H. Jégou, "A group testing framework for similarity search in high-dimensional spaces," in *ACM MM*, 2014.

[69] A. Iscen, M. Rabbat, and T. Furon, "Efficient large-scale similarity search using matrix factorization," in *CVPR*, 2016.

[70] A. Iscen, T. Furon, V. Gripon, M. Rabbat, and H. Jégou, "Memory vectors for similarity search in high-dimensional spaces," *IEEE Transactions on Big Data*, 2017.

[71] A. Mikulík, M. Perdoch, O. Chum, and J. Matas, "Learning a fine vocabulary," in *ECCV*, 2010.

[72] S. Zhang, M. Yang, X. Wang, Y. Lin, and Q. Tian, "Semantic-aware co-indexing for image retrieval," in *ICCV*, 2013.

[73] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR*, 2009.

[74] Y. Cai, W. Tong, L. Yang, and A. G. Hauptmann, "Constrained keypoint quantization: towards better bag-of-words model for large-scale multimedia retrieval," in *ICMR*, 2012.

[75] W. Zhou, Y. Lu, H. Li, and Q. Tian, "Scalar quantization for large scale image search," in *ACM MM*, 2012.

[76] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *CVPR*, 2009.

[77] J. Revaud, M. Douze, and C. Schmid, "Correlation-based bursti-ness for logo retrieval," in *ACM MM*, 2012.

[78] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Lp-norm idf for large scale image search," in *CVPR*, 2013.

[79] M. Murata, H. Nagano, R. Mukai, K. Kashino, and S. Satoh, "Bm25 with exponential idf for instance search," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1690–1699, 2014.

[80] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *CVPR*, 2013.

[81] P. Turcot and D. G. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *ICCV Workshops*, 2009.

[82] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han, "Contextual weighting for vocabulary tree based image retrieval," in *ICCV*, 2011.

[83] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1794–1801.

[84] A. Babenko and V. Lempitsky, "The inverted multi-index," in *CVPR*, 2012.

[85] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *IJCV*, vol. 87, no. 3, pp. 316–336, 2010.

[86] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *CVPR*, 2015.

[87] M. Douze, H. Jégou, and C. Schmid, "An image-based approach to video copy detection with spatio-temporal post-filtering," *IEEE Transactions on Multimedia*, vol. 12, no. 4, pp. 257–266, 2010.

[88] M. Jain, R. Benmokhtar, H. Jégou, and P. Gros, "Hamming embedding similarity-based image classification," in *ICMR*, 2012.

[89] G. Tolias and H. Jégou, "Visual query expansion with or without geometry: refining local descriptors by feature aggregation," *Pattern Recognition*, vol. 47, no. 10, pp. 3466–3476, 2014.

[90] M. Jain, H. Jégou, and P. Gros, "Asymmetric hamming embedding: taking the best of our bits for large scale image search," in *ACM MM*, 2011.

[91] D. Qin, C. Wengert, and L. Gool, "Query adaptive similarity for large scale object retrieval," in *CVPR*, 2013.

[92] H. Jegou, C. Schmid, H. Harzallah, and J. Verbeek, "Accurate image search using the contextual dissimilarity measure," *TPAMI*, vol. 32, no. 1, pp. 2–11, 2010.

[93] K. Van De Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *TPAMI*, vol. 32, no. 9, pp. 1582–1596, 2010.

[94] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *TPAMI*, vol. 30, no. 4, pp. 712–727, 2008.

[95] J. Van de Weijer, T. Gevers, and A. D. Bagdanov, "Boosting color saliency in image feature detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 1, pp. 150–156, 2006.

[96] C. Wengert, M. Douze, and H. Jégou, "Bag-of-colors for improved image search," in *ACM MM*, 2011.

[97] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific fusion for image retrieval," in *ECCV*, 2012.

[98] C. Deng, R. Ji, W. Liu, D. Tao, and X. Gao, "Visual reranking through weakly supervised multi-graph learning," in *ICCV*, 2013.

[99] R. Tao, A. W. Smeulders, and S.-F. Chang, "Attributes and categories for generic instance search from one example," in *CVPR*, 2015.

[100] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *ICCV*, 2007.

[101] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall ii: Query expansion revisited," in *CVPR*, 2011.

[102] J. Philbin, J. Sivic, and A. Zisserman, "Geometric latent dirichlet allocation on a matching graph for large-scale image datasets," *IJCV*, vol. 95, no. 2, pp. 138–153, 2011.

[103] Q.-F. Zheng, W.-Q. Wang, and W. Gao, "Effective and efficient object-based image retrieval using visual phrases," in *ACM MM*, 2006.

[104] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," in *ACM MM*, 2009.

[105] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *CVPR*, 2011.

[106] Y. Jiang, J. Meng, and J. Yuan, "Randomized visual phrases for object search," in *CVPR*, 2012.

[107] L. Torresani, M. Szummer, and A. Fitzgibbon, "Learning query-dependent prefilters for scalable image retrieval," in *CVPR*, 2009.

[108] P. Koniusz, F. Yan, P.-H. Gosselin, and K. Mikolajczyk, "Higher-order occurrence pooling for bags-of-words: Visual concept detection," *TPAMI*, 2016.

[109] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: from visual words to visual phrases," in *CVPR*, 2007.

[110] C. L. Zitnick, J. Sun, R. Szeliski, and S. Winder, "Object instance recognition using triplets of feature symbols," *Technical Report MSR-TR-200753, Microsoft Research*, 2007.

[111] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu, "Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking," in *CVPR*, 2012.

[112] Y. Avrithis and G. Tolias, "Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval," *IJCV*, vol. 107, no. 1, pp. 1–19, 2014.

[113] X. Wu and K. Kashino, "Adaptive dither voting for robust spatial verification," in *ICCV*, 2015.

[114] X. Li, M. Larson, and A. Hanjalic, "Pairwise geometric matching for large-scale object retrieval," in *CVPR*, 2015.

[115] J. L. Schönberger, T. Price, T. Sattler, J.-M. Frahm, and M. Pollefeys, "A vote-and-verify strategy for fast spatial verification in image retrieval," *ACCV*, 2016.

[116] D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. Van Gool, "Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors," in *CVPR*, 2011.

[117] G. Awad, J. Fiscus, M. Michel, D. Joy, W. Kraaij, A. F. Smeaton, G. Quénot, M. Eskevich, R. Aly, and R. Ordelman, "Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking," in *Proceedings of TRECVID*, 2016.

[118] O. Chum, M. Perdoch, and J. Matas, "Geometric min-hashing: Finding a (thick) needle in a haystack," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 17–24.

[119] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis, "Scalable triangulation-based logo recognition," in *ICMR*, 2011.

[120] S. Romberg and R. Lienhart, "Bundle min-hashing for logo recognition," in *ICMR*, 2013.

[121] C.-Z. Zhu, H. Jégou, and S. Satoh, "Query-adaptive asymmetrical dissimilarities for visual object retrieval," in *ICCV*, 2013.

[122] Z. Chen, W. Zhang, B. Hu, X. Can, S. Liu, and D. Meng, "Retrieving objects by partitioning," *IEEE Transactions on Big Data*, 2017.

[123] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum, "Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations," in *CVPR*, 2017.

[124] K. Simonyan and A. Zisserman, "Very deep convolutional net-works for large-scale image recognition," in *ICLR*, 2015.

[125] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[126] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[127] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

[128] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.

[129] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014.

[130] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors of transferability for a generic convnet representation," *TPAMI*, vol. 38, no. 9, pp. 1790–1802, 2016.

[131] V. Chandrasekhar, J. Lin, O. Morère, H. Goh, and A. Veillard, "A practical guide to cnns and fisher vectors for image instance retrieval," *Signal Processing*, vol. 128, pp. 426–439, 2016.

[132] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *ICLR*, 2014.

[133] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual instance retrieval with deep convolutional networks," in *ICLR workshops*, 2015.

[134] L. Zheng, Y. Zhao, S. Wang, J. Wang, and Q. Tian, "Good practice in cnn feature transfer," *arXiv:1604.00133*, 2016.

[135] L. Zheng, S. Wang, J. Wang, and Q. Tian, "Accurate image search with multi-scale contextual evidences," *IJCV*, 2016.

[136] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *CVPR*, 2015.

[137] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *CVPR*, 2016.

[138] P. Kulkarni, J. Zepeda, F. Jurie, P. Perez, and L. Chevallier, "Hybrid multi-layer deep cnn/aggregator feature for image classification," in *ICASSP*, 2015.

[139] E. Mohedano, K. McGuinness, N. E. O'Connor, A. Salvador, F. Marqués, and X. Giró-i Nieto, "Bags of local convolutional features for scalable instance search," in *ACM MM*, 2016.

[140] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *ICCV*, 2015.

[141] L. Xie, L. Zheng, J. Wang, A. Yuille, and Q. Tian, "Interactive: Inter-layer activeness propagation," in *CVPR*, 2016.

[142] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *ICRA*, 2011.

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

[144] A. Salvador, X. Giró-i Nieto, F. Marqués, and S. Satoh, "Faster r-cnn features for instance search," in *CVPR Workshops*, 2016.

[145] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *CVPR*, 2015.

[146] P. Fischer, A. Dosovitskiy, and T. Brox, "Descriptor matching with convolutional neural networks: a comparison to sift," *arXiv:1405.5769*, 2014.

[147] K. Mopuri and R. Babu, "Object level deep feature pooling for compact image representation," in *CVPR Workshops*, 2015.

[148] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *IJCV*, 2013.

[149] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014.

[150] T. Uricchio, M. Bertini, L. Seidenari, and A. Bimbo, "Fisher encoded convolutional bag-of-windows for efficient image retrieval and social image tagging," in *ICCV Workshops*, 2015.

[151] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[152] L. Xie, R. Hong, B. Zhang, and Q. Tian, "Image classification and retrieval are one," in *ICMR*, 2015.

[153] M. Paulin, M. Douze, Z. Harchaoui, J. Mairal, F. Perronin, and C. Schmid, "Local convolutional features with unsupervised training for image retrieval," in *ICCV*, 2015.

[154] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in *ECCV*, 2016.

[155] Y. Liu, Y. Guo, S. Wu, and M. S. Lew, "Deepindex for accurate and efficient image retrieval," in *ICMR*, 2015.

[156] Y. Li, X. Kong, L. Zheng, and Q. Tian, "Exploiting hierarchical activations of neural network for image retrieval," in *ACM MM*, 2016.

[157] F. Perronnin and D. Larlus, "Fisher vectors meet neural networks: A hybrid classification architecture," in *CVPR*, 2015.

[158] C. Strecha, A. Bronstein, M. Bronstein, and P. Fua, "Ldahash: Improved matching with smaller descriptors," *TPAMI*, 2012.

[159] K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen, "Deep learning of binary hash codes for fast image retrieval," in *CVPR Workshops*, 2015.

[160] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *CVPR*, 2015.

[161] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 817–824.

[162] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *arXiv:1610.07940*, 2016.

[163] S. Zhang, X. Wang, Y. Lin, and Q. Tian, "Cross indexing with grouplets," *IEEE Transactions on Multimedia*, 2015.

[164] O. Morère, A. Veillard, J. Lin, J. Petta, V. Chandrasekhar, and T. Poggio, "Group invariant deep representations for image instance retrieval," in *Proceedings of AAAI Symposium on Science of Intelligence*, 2017.

[165] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *CVPR*, 2014.

[166] Y. Bengio, I. J. Goodfellow, and A. Courville, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[167] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *CVPR*, 2015.

[168] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[169] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *ICME*, 2016.

Liang Zheng received the Ph.D degree in Electronic Engineering from Tsinghua University, China, in 2015, and the B.E. degree in Life Science from Tsinghua University, China, in 2010. He was a postdoc researcher in University of Texas at San Antonio, USA. He is now a postdoc researcher in the Center of Artificial Intelligence, University of Technology Sydney, Australia. His research interests are image retrieval, person re-identification and deep learning.

Yi Yang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He is currently a Professor with University of Technology Sydney, Australia. He was a Post-Doctoral Research with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. His current research interest includes machine learning and its applications to multimedia content analysis and computer vision.

Qi Tian (M'96-SM'03-F'16) received the B.E. degree in electronic engineering from Tsinghua University, China, the M.S. degree in electrical and computer engineering from Drexel University, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, in 1992, 1996, and 2002, respectively. He is currently a Professor with the Department of Computer Science, University of Texas at San Antonio (UTSA). His research interests include multimedia information retrieval and computer vision.