Information Processing and Management 60 (2023) 103119

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Adaptive multi-feature fusion via cross-entropy normalization for effective image retrieval

Wentao Ma [a], Tongqing Zhou [a], Jiaohua Qin [b,*], Xuyu Xiang [b], Yun Tan [b], Zhiping Cai [a,*]

[a] College of Computer, National University of Defense Technology, Changsha, Hunan, 410073, China
[b] College of Computer Science and Information Technology, Central South University of Forestry & Technology, Changsha, Hunan, 410000, China

ARTICLE INFO

Keywords:
Image retrieval
Cross-entropy
Feature fusion
High-level semantic features

ABSTRACT

Multi-feature fusion has achieved gratifying performance in image retrieval. However, some existing fusion mechanisms would unfortunately make the result worse than expected due to the domain and visual diversity of images. As a result, a burning problem for applying feature fusion mechanism is how to figure out and improve the complementarity of multi-level heterogeneous features. To this end, this paper proposes an adaptive multi-feature fusion method via cross-entropy normalization for effective image retrieval. First, various low-level features (e.g., SIFT) and high-level semantic features based on deep learning are extracted. Under each level of feature representation, the initial similarity scores of the query image w.r.t. the target dataset are calculated. Second, we use an independent reference dataset to approximate the tail of the attained initial similarity score ranking curve by cross-entropy normalization. Then the area under the ranking curve is calculated as the indicator of the merit of corresponding feature (i.e., a smaller area indicates a more suitable feature.). Finally, fusion weights of each feature are assigned adaptively by the statistically elaborated areas. Extensive experiments on three public benchmark datasets have demonstrated that the proposed method can achieve superior performance compared with the existing methods, improving the metrics mAP by relatively 1.04% (for Holidays), 1.22% (for Oxf5k) and the N-S by relatively 0.04 (for UKbench), respectively.

1. Introduction

Content-based image retrieval (CBIR) has been widely studied in computer vision tasks, which mainly leverages global features, local features and convolution features of images to explore retrieval task, and has achieved gratifying results. Although many efforts have been spent in this field, the burning challenging problem of CBIR, known as the 'semantic gap', still exists between low-level features captured by machines and high-level semantic features perceived by humans. Gkelios, Sophokleous, Plakias, Boutalis, and Chatzichristofis (2021), Zheng, Yang, and Tian (2018) comprehensively introduced the development of image retrieval in the past two decades. On the hand, the methods based on low-level features of SIFT mostly depend on the Bag-of-Word (BoW). As surveyed in Refs. Chen, Hu, and Shen (2009), Elsayad, Martinet, Urruty, and Djeraba (2010), Xia et al. (2018), Zhu, Jin, Zheng, and Feng (2014), Zhu et al. (2021), such methods generally use local feature descriptors such as SIFT (Xie, Tian, & Zhang, 2014; Zheng

* Corresponding authors.
E-mail addresses: wtma@nudt.edu.cn (W. Ma), zhoutongqing@nudt.edu.cn (T. Zhou), qinjiaohua@163.com (J. Qin), xyuxiang@163.com (X. Xiang), tantanyun@hotmail.com (Y. Tan), zpcai@nudt.edu.cn (Z. Cai).

https://doi.org/10.1016/j.ipm.2022.103119
Received 10 June 2022; Received in revised form 15 September 2022; Accepted 10 October 2022
Available online 28 October 2022

et al., 2018) and establish a corresponding feature codebook database. Compared with other low-level features, the BoW model can make a better scheme; however, its retrieval efficiency is greatly reduced due to the huge computational complexity. On the other hand, high-level semantic features, usually extracted by the popular Convolutional Neural Network (CNN), are widely used as feature descriptors recently (Gkelios et al., 2021; Ma, Zhou, Qin, Xiang et al., 2022; Qiao, Wu, & Jin, 2021; Zhan, Zhang, Hu, & Sheng, 2021). In general, the deep CNN model can achieve superior results in image classification, clustering, retrieval and object detection (Huang, Liu, Pleiss, Van Der Maaten, & Weinberger, 2019; Ma, Zhou, Qin, Zhou, & Cai, 2022), because it can approach human perception to some extent via a series of hidden layer operations to improve feature representations. Therefore, the high-level semantic features are believed to yield higher accuracy than their low-level counterparts. Among the typical CNN model (e.g., AlexNet, InceptionV3, VGG19), while DenseNet121 is featured with extremely high feature utilization rate, stronger feature expression ability, fewer parameters and less computation complexity (Huang et al., 2019).

Both types of features can help to attain decent performance (Zheng et al., 2018), but they are considered independently with the joint benefit neglected, leaving a large room for improvement. Given this, researchers have proposed some effective fusion mechanisms, such as feature-level fusion (Abdi, Shamsuddin, Hasan, & Piran, 2019; Douze, Ramisa, & Schmid, 2011; Li, Yang, Yang, Sun, & Xu, 2021; Liu, Guo, Wu, & Cai, 2017; Wu, Li, Xu, & Yang, 2021; Zhao, Lu, & Wang, 2017), index-level fusion (Zhang, Yang, Wang, Lin, & Tian, 2015; Zhang et al., 2019; Zheng, Wang, Liu, & Tian, 2014; Zheng, Wang, & Tian, 2014), graph-level fusion (Deng, Ji, Liu, Tao, & Gao, 2013; Lao et al., 2021; Liu et al., 2020; Zhang, Yang, Cour, Yu, & Metaxas, 2014), similarity score-level fusion (Zhang et al., 2018; Zheng et al., 2015) and rank-level fusion (Liu, Wang, Zheng, & Tian, 2017; Valem & Pedronette, 2020a, 2020b). These fusion methods have greatly promoted the development of image retrieval technology. However, it is hard to balance the relationship between weight distribution and complementarity of heterogeneous features, which results in a performance worse than expected (Zheng, Wang, Liu et al., 2014). In some cases, there is no use of high-level semantic features (Zhao et al., 2017) or a large amount of redundancy in high-level semantic features, which cannot adequately represent image information (Zhang et al., 2014). As a result, the complementarity of heterogeneous features cannot be fully utilized, wherein the inappropriate fusion mechanism is the essence.

To tackle the above issues, our work mainly considers two aspects: on the one hand, all features are treated equally, and the failure of identifying better features, which may under-utilize features' discriminative ability. On the other hand, bad features that escape being punished may lead to unexpected consequences, namely, accuracy gets even lower after fusion (Zhang et al., 2014; Zheng, Wang, Liu et al., 2014). Thus our work proposes an adaptive multi-feature late fusion via cross-entropy normalization for effective image retrieval. The main idea of our solution is to adaptively identify the effectiveness of different features based on the similarity it shares with each query, so that those 'good' features are endowed with larger weights for providing greater contributions, while the 'bad' features are punished, thus attaining differential fusion of heterogeneous features. The weight allocation of each feature fusion is adaptive for each input query image, in this way, 'good' feature for a specific query is assigned a larger weight. Specifically, first various low-level features and high-level semantic features with higher feature utilization and low redundancy are extracted, meanwhile we calculate the similarity between a query and each image in the target datasets (Holidays, UKbench and Oxf5k) under several feature representations, and obtain a similarity curve for each representation (known as target distribution). Then, we perform another matching between the query and a reference dataset (Flickr1M & Flickr343Places), which is considerably large, so that to obtain reference curves for each representation (known as reference distribution), too. Next, cross-entropy normalization is used to approximate the target distribution with the reference distribution, giving us a normalized similarity curve for each representation. By doing this, we assess the validity of a representation's curve. Namely, better matching between the target distribution and the reference distribution indicates the original curve is acceptable on a larger dataset. After that, the area under the normalized similarity curve is denoted as the merit of the representation, which is further used as its weight during fusion. In this way, adaptivity is achieved by evaluating the representation with its generalization to different (reference) datasets. The main contributions of this paper are listed as follows:

- We propose an efficient multi-feature image retrieval method, which extracts four types of low-level features as well as high-level semantic features of five pre-trained CNN models. Our attempt to adopt such a wide exploitation of multi-feature is believed to represent visual characteristics under different image retrieval contexts, thus bringing insights for leveraging ensemble knowledge into retrieval.
- We propose an adaptive multi-feature score-level fusion via cross-entropy normalization to improve the complementarity of heterogeneous features. This mechanism combines the performance of multiple features in an unsupervised and has better retrieval accuracy and generalization than single feature retrieval. To be specific, the relationship between fusion weight and feature complementarity can be realized in an adaptive manner.
- Extensive experiment results on three public benchmark datasets have demonstrated that our method is highly competitive and consistently outperforms several popular fusion methods by an obvious margin.

The rest of this paper is organized as follows. After briefly reviewing the related work in Section 2, we introduce the proposed adaptive multi-feature late fusion via cross-entropy normalization for effective image retrieval in Section 3. Experimental settings and analysis of the corresponding experimental results are given in Section 4 and Section 5, respectively. Section 6 introduces the discussions about our work and this paper will be concluded in Section 7.

2. Related work

The related work of this paper involves the following two groups: single feature retrieval and feature fusion designs.

2.1. Single feature retrieval

In the past decades, many advanced image processing methods have been investigated and applied in image classification, object detection (Huang et al., 2019; Ma, Zhou, Qin, Zhou et al., 2022), image retrieval (Zheng et al., 2018) and so on. Especially for image retrieval, (Jégou, Douze, Schmid, & Pérez, 2010) proposes the vector of locally aggregated descriptors (VLAD), joint optimization of memory usage, retrieval accuracy and efficiency. To speed up the retrieval of inverted lists, PCA transformation and product quantization coding are applied to VLAD (Jegou, Douze, & Schmid, 2010). The BoW is used to obtain a high accuracy rate, which usually requires a large codebook. Perronnin and Dance (2007), Sánchez, Perronnin, Mensink, and Verbeek (2013) adopt a mixed Gaussian model to approximate the distribution of low-level feature vectors, and further derived Fisher-Vector (FV). Xie et al. (2014) provides a local descriptor Max-SIFT that is invariant to horizontal flips. Experimental results on datasets such as scene classification and fine-grained classification show that Max-SIFT is better than SIFT, and better than the Dirichlet-based Histogram Feature Transform (DHFT) proposed in Kobayashi (2014). Researchers also employ global descriptors for image retrieval, such as HSV, GIST, and other visual attribute features. Moreover, with the development of Artificial Intelligence, CNN-based high-level semantic feature image retrieval methods are considered as the common practice in image retrieval, which almost completely replaces the traditional low-level image descriptors for image retrieval (Amato, Carrara, Falchi, Gennaro, & Vadicamo, 2020; Ge, Wei, Yu, Singh, & Xiong, 2021; Pandey, Khanna, & Yokota, 2016; Zheng et al., 2018). High-level semantic features are used to replace traditional low-level image descriptors for image retrieval and are divided into two categories: (1) Convolutional layer features and (2) Full connection layer features. This classification takes into account mainly the location of extracted features derived from CNN model.

Convolutional layer features: (Liu, Shen, & van den Hengel, 2015) proposes cross-convolutional layer pooling, which is used to encode convolutional layer features and achieved better results. Kalantidis, Mellina, and Osindero (2016) presents a cross-convolutional layer weighting to generate image descriptors by the last convolutional layer of CNN. Arandjelović, Gronat, Torii, Pajdla, and Sivic (2018) designs a novel generalized VLAD layer and leveraged an 'end-to-end' learning to generate NetVLAD descriptors. Furthermore, experiments show that this framework and training procedure are significantly better than the existing image feature descriptors. To enrich the semantic information, (Chen et al., 2019) investigates a novel CMF-based framework, namely, SCRATCH. This method utilizes Collective Matrix Factorization on the original features and semantic embedding to find a shared latent semantic space while preserving the intra- and inter-modal similarities.

Full connection layer features: (Babenko, Slesarev, Chigorin, & Lempitsky, 2014) adopts pre-trained CNN to extract image features and also provides 'fine-tuning' neural network to extract features. To increase the invariance of CNN features, (Gong, Wang, Guo, & Lazebnik, 2014) extracts the multi-scales full connection layer features and uses them to form the VLAD vector. Bao and Li (2017) leverages VLAD-pooling to aggregate the full connection layer features extracted from each slice, and the resulting descriptor is called object-based deep feature aggregation. Öztürk (2020) designs an effective hash-generating method for medical image retrieval, which leverages full-connection layer features at the output of the CNN architecture to generate hash codes, thus reducing the semantic gap between low-level features and high-level semantics. Many types of research have proved the advantages of global high-level semantic features in image retrieval, which is a promising technique to compensate for the weaknesses of local features.

2.2. Feature fusion designs

Both retrieval methods based on a single low-level feature and a single high-level semantic feature can achieve gratifying performance (Gkelios et al., 2021; Zheng et al., 2018), but these methods still have a large room for improvement. For this, some effective fusion methods have been proposed, such as graph-based fusion, index-based fusion, and score-based fusion.

2.2.1. Graph-based fusion

It is well known that graph-based fusion visual retrieval has been proven to be effective, which integrates the initial retrieval and visual consistency of images. Zhang et al. (2014) proposes a query-specific fusion based on the undirected graph. This method modeled the retrieval ranks as graphs of candidate images, where multiple graphs are merged and reranked by conducting a link analysis on a fused graph. Liu, Wang et al. (2017) believes that graph fusion is susceptible to 'Outliers', so they design an image graph that is less susceptible. Different from graph fusion, image graph adopts the features of SIFT, GIST, HSV, and CNN, and achieves better results than graph fusion in Holidays and UKbench. Inspired by the three-degree influence principle in social networks, (Liu et al., 2020) provides a reranking method (N3G) based on a single feature and a multi-graph fusion ranking (MFR) method based on social network group relation theory, which takes into account the correlation of all images in multiple neighborhood graphs. Lao et al. (2021) presents a Three Degree Binary Graph (TDBG) to eliminate outlier candidates irrelevant to the query and utilizes a set-based greedy algorithm to reduce the influence of adjacent manifolds, which further improved the retrieval performance of the system.

2.2.2. Index-based fusion

Index construction plays an essential role in image retrieval, and index-level fusion is an effective fusion mechanism. Zhang et al. (2015) proposes a Semantic Aware Co-indexing, in which both embedded image low-level local invariant features with strong robustness and semantic attributes features with high-level semantic meaning into the inverted index, thus improving the indexing differentiation. To improve the accuracy, (Liu, Guo, Wu, & Lew, 2015) includes CNN features in an index and designs the Deep-Index, but the performance is inferior to the features of fusion SIFT and CNN 'Deep-Embedding' (Zheng, Wang, He, & Tian, 2014). The

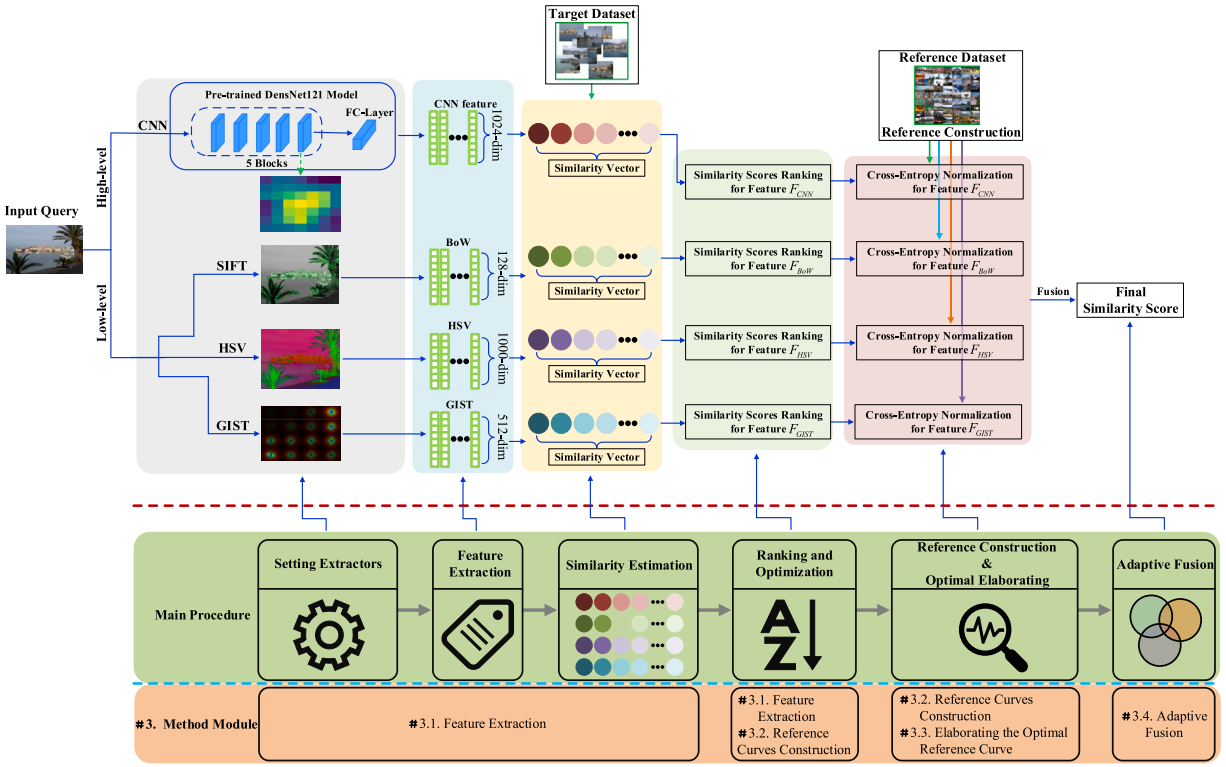


Fig. 1. An overview of our feature fusion image retrieval system.

inverted multi-index structure proposed by Babenko and Lempitsky (2014) generalizes the inverted index idea by replacing the standard quantization within inverted indices with product quantization, which achieves a much denser subdivision of the search space compared to inverted indices for very similar retrieval complexity and pre-processing time. Bhowmik, González, Gouet-Brunet, Pedrini, and Bloch (2014) presents and discusses a multi-dimensional feature fusion strategy based on inversion multi-index and similarity search, and experimental results prove that the combined use of several descriptions achieves the purpose of improving similarity search. Zhang et al. (2019) investigates a novel multi-index fusion image retrieval based on AlexNet and ResNet50 networks. This method inherits the core idea of Collaborative Indexing Embedding (CIE), which integrates the features of different visual representations at index-level.

2.2.3. Score-based fusion

As we all know, for the multi-feature fusion image retrieval technique, given a query image, one does not know which feature is valid or invalid without prior knowledge. Hence, it is significant to identify the effectiveness of features by an adaptive manner. Zhang et al. (2018) observe that the geometric relational features based on distances between joints and selected lines outperform other features, then proposes a multi-stream LSTM architecture with a new smoothed score fusion learn classification from different geometric feature streams. The final fusion results reach the most advanced performance, but the advantage of this method is only in that skeleton-based action recognition cannot be used for natural images. Then (Zheng et al., 2015) provided a similar score-level multi-feature fusion with the name of 'Score Fusion'. Zheng et al. (2015) argues that for a good feature, the similarity ranking curve should decrease rapidly at the beginning and then tend to be stable, while the similarity ranking curve of a bad feature will drop gradually. Score Fusion has two main characteristics: (1) The weights of each feature are not fixed and are not easily affected by noneffective features. (2) The effectiveness of features is estimated online by using independent datasets. Zheng et al. (2015) adopts five types of features: SIFT, HSV, CaffeNet full connective layer features, GIST, and Random features. Extensive experiments show that Score Fusion is superior to the Graph-level Fusion (Zhang et al., 2014) and the Semantic Aware Co-indexing (Zhang et al., 2015) in both retrieval performance and efficiency.

Differently, inspired by Score Fusion, we propose an adaptive multi-feature late fusion via cross-entropy normalization for effective image retrieval. This method has two main advantages: (1) Compared with CaffeNet's full connection layer features are 4096-dim, while in our work the high-level semantic features of 2048-dim and 1024-dim with higher feature utilization and smaller redundancy are adopted respectively. (2) The mechanism of Score Fusion is beneficial to improving performance to a certain extent,

Table 1
Key notations.

Notation	Description
$F = \{F_1, F_2, \dots, F_i \mid i = 1, 2, \dots, c\}$	F indicates the features set, F_i represents the ith type of feature, and c denotes the number of features to be fused.
$T = \{T_1, T_2, \dots, T_i \mid i = 1, 2, \dots, c\}$	T indicates the target set, T_i represents the F_i type target ranking curve, and c denotes the number of features to be fused.
$R = \{R_1, R_2, \dots, R_i \mid i = 1, 2, \dots, c\}$	R indicates the references set, R_i represents the F_i type reference ranking curve, and c denotes the number of features to be fused.
$S^q_{T_i}$, $S^q_{R_i}$	The similarity ranking between query image q and target dataset (Holidays, UKbench and Oxf5k), known as the target distribution $S^q_{T_i}$, and the similarity ranking between query image q and the reference dataset (Flickr1M & Flickr343Places), known as the reference distribution $S^q_{R_i}$, both obtained on feature F_i.
$(u : v)$	u and v are the parameters of the vector segment that restricts the matching region between two vectors. We require that u not be too small and that v be relatively large.
$S^q_{T_i}(u : v)$, $S^q_{R_i}(u : v)$	$S^q_{T_i}(u : v)$ and $S^q_{R_i}(u : v)$ represent the distribution of similarity ranking curve on the target and reference datasets in the $(u : v)$ vector segment respectively.
$H(S^q_{T_i}(u : v))$	The entropy of target distribution.
$D(S^q_{T_i}(u : v), S^q_{R_i}(u : v))$, $E(S^q_{T_i}(u : v), S^q_{R_i}(u : v))$	$D(S^q_{T_i}(u : v), S^q_{R_i}(u : v))$ and $E(S^q_{T_i}(u : v), S^q_{R_i}(u : v))$ represent the relative entropy and cross-entropy of distribution between the target distribution and the reference distribution, respectively.
$E(S^q_{T_i}(u : v), S^q_{R_i}(u : v))$	The cross-entropy of distribution between target distribution and reference distribution.
$L(S^q_{T_i}(u : v), S^q_{R_i}(u : v))$	The final loss function.
$S^{q^*}_{R_i}$	$S^{q^*}_{R_i}$ is the optimal reference curve to query image q for feature F_i.
A_{F_i}	A_{F_i} indicates the area under the cross-entropy normalized similarity curve of features F_i.
$W^q_{F_i}$	The weight of feature F_i for query image q.

but there is still room for improvement in the complementarity of heterogeneous features. Therefore, by the distribution of the similarity curve and reference curve, the proposed method adopts cross-entropy normalization to enhance complementarity.

3. The proposed method

Fig. 1 shows the pipeline of our proposal, and the step division of the pipeline is basically consistent with the method description module in Section 3.1∼3.4. Each module described in our method includes feature extraction, reference curve construction, elaborating the optimal reference curve, and adaptive fusion, which can be reflected in the corresponding step in the pipeline. Meanwhile, for readability and clarity, some of the notations adopted in this paper and their definitions are described in Table 1.

3.1. Feature extraction

This paper adopts four low-level features: HSV, SIFT, GIST, RAND, and five high-level semantic features. Also, similar to Liu, Wang et al. (2017), Zheng et al. (2015), the similarity of all features is normalized by l_2 paradigm.

[HSV] For each image, we calculate the HSV color histogram of 1000-dim, and the H, S, and V components are multiplied by $20 \times 10 \times 5$ bins, respectively.

[BoW] For the SIFT descriptors of each image, our work leverages the 128-bit Hamming signature to embed each SIFT descriptor into the inverted file to filter out false matches, with a hamming threshold of 52 and a weighting parameter of 26. Moreover, rootSIFT and burstiness strategy (Zheng et al., 2015) are employed on two public benchmark datasets.

[GIST] In order to calculate the GIST descriptor, we adjust the image size to 256×256 and employ four types of feature scales with the number of scale orientations (8, 8, 8, 8) respectively.

[RAND] In order to illustrate the robustness of our method to 'bad' features, and also to reveal the performance superiority of the Score Fusion (Zheng et al., 2015), we utilize a random feature, namely RAND. In effect, the RAND feature is a random transform matrix $P \in \mathbb{R}^{d \times m}$ (Zheng et al., 2015), where d is the target feature dimension (set to 1000 in our experiment), and m is the dimension of the input image (with all pixels concatenated by columns).

[CNN] Pre-trained CNN architectures on ImageNet have demonstrated their generalization for unseen data. In our paper, the pre-training models of AlexNet, InceptionV3, VGG19, ResNet50, and DenseNet121 are adopted to extract high-level semantic features

of three public benchmark datasets. It is worth noting that due to the difference in pre-training models, the dimensions of features will also be different: the features extracted from both AlexNet and VGG19 models are 4096-dim, InceptionV3 and ResNet50 models are 2048-dim, while the DenseNet121 model is 1024-dim.

3.2. Reference curve construction

The reference dataset Flickr1M (Jegou, Douze, & Schmid, 2008) and Flickr343Places (Zheng et al., 2015) both have 1M images. These two independent public datasets are adopted to construct the reference curves, which are then cross-entropy normalized to the target ranking curve. The reference curve is mainly used to eliminate the 'high tail' of bad features. To find the optimal reference curve, our work considers the relationship between the target curve distribution and the reference curve distribution. While the optimal reference curve can be regarded as approximating the end of the target curve and subtracting 'end' to highlight the top of the target curve. In particular, experimental statistics show that considering the intrinsic distribution of the target curve and reference curve can not only eliminate the 'high tail' but also enhance the complementarity of heterogeneous features. The reference curve for SIFT uses Flickr1M, while the reference curves for other features adopt dataset Flickr343Places. Q queries are randomly selected as reference curve. Then, image retrieval of GIST, HSV, SIFT, RAND, and five high-level semantic features are conducted in dataset of Flickr1M & Flickr343Place respectively. For feature F_i, there are Q reference curves, denoted as R_i.

3.3. Elaborating the optimal reference curve

Our method can not only effectively remove the 'high tail' but also improve the complementarity between heterogeneous features. Specifically, the similarity ranking between query image q and the target dataset (Holidays, UKbench and Oxf5k) is denoted as $S^q_{T_i}$ (i.e., the target distribution) and the similarity ranking between query image q and the reference dataset (Flickr1M & Flickr343Places) is denoted as $S^q_{R_i}$ (i.e., the reference distribution), which are both obtained for feature F_i. Since $S^q_{T_i}$ can be easily calculated given the query and the target dataset, our goal is to find the reference that best matches the tail of $S^q_{T_i}$ in R_i.

According to the target distribution $S^q_{T_i}$, our strategy is to eliminate the 'high tail' of bad features by matching the optimal reference curve at the minimum cost. $S^q_{T_i}(u:v)$ and $S^q_{R_i}(u:v)$ represent the distribution of similarity ranking curve on the target dataset and reference dataset in the $(u:v)$ vector segment respectively. Then entropy of target distribution $S^q_{T_i}(u:v)$ is defined as $H(S^q_{F_i}(u:v)) = -\sum_i S^q_{T_i}(u:v)\log S^q_{T_i}(u:v)$. While relative entropy $D(S^q_{T_i}(u:v), S^q_{R_i}(u:v))$ is a non-symmetric measure of the difference between two distributions. Concretely, in our work it is a measure of the heterogeneous features' complementarity when $S^q_{R_i}(u:v)$ is used to approximate $S^q_{T_i}(u:v)$. Formally, the relative entropy is defined as

$$D(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) = \sum_i S^q_{T_i}(u:v)\log\frac{S^q_{T_i}(u:v)}{S^q_{R_i}(u:v)} = \sum_i [S^q_{T_i}(u:v)\log S^q_{T_i}(u:v) - S^q_{T_i}(u:v)\log S^q_{R_i}(u:v)] \tag{1}$$

From Eq. (1), $S^q_{T_i}(u:v)$ and $S^q_{R_i}(u:v)$ are approximate to each other, so relative entropy $D(S^q_{F_i}(u:v), S^q_{R_i}(u:v))$ is smaller. Ideally, if the distribution of target curve and reference curve is the same, $D(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) = 0$. By Theorem 1 (please kindly refer to the appendix of Appendix), only if $S^q_{T_i}(u:v) = S^q_{R_i}(u:v)$, $D(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) = 0$. However, the curves come from different datasets then $S^q_{T_i}(u:v) \neq S^q_{R_i}(u:v)$, so by Eq. (1), we have $D(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) > 0$. Our goal is to minimize the cost of complementarity between heterogeneous features when $S^q_{R_i}(u:v)$ approaches $S^q_{T_i}(u:v)$. According to the above discussion, cross-entropy can be defined as

$$E(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) = \sum_i S^q_{T_i}(u:v)\log\frac{1}{S^q_{R_i}(u:v)} = -\sum_i S^q_{T_i}(u:v)\log S^q_{R_i}(u:v) \tag{2}$$

Eq. (1) can be simplified as followed:

$$D(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) = E(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) - H(S^q_{T_i}(u:v)) \tag{3}$$

From Theorem 1, $S^q_{T_i}(u:v) \neq S^q_{R_i}(u:v)$ implies that $D(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) > 0$, so $E(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) - H(S^q_{T_i}(u:v)) > 0$. Moreover, according to Eq. (3), minimizing the relative entropy $D(S^q_{T_i}(u:v), S^q_{R_i}(u:v))$ is to minimize the cross-entropy $E(S^q_{T_i}(u:v), S^q_{R_i}(u:v))$.

By Theorem 2 (please refer to the appendix of Appendix), minimizing the difference between the distribution $S^q_{T_i}(u:v)$ of target curves and the distribution $S^q_{R_i}(u:v)$ of reference curves is equivalent to minimize the relative entropy between them. So the cross-entropy loss model for the reference distribution $S^q_{R_i}(u:v)$ and target distribution $S^q_{T_i}(u:v)$ is defined as

$$\{S^{q^*}_{R_i}\} = \arg\min_i E(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) \Leftrightarrow \arg\min_i L(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) \tag{4}$$

, wherein the loss function in the last term is defined as

$$L(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) = \sum_{i=1}^{c}[S^q_{T_i}(u:v)\log S^q_{R_i}(u:v) - (1 - S^q_{T_i}(u:v))\log(1 - S^q_{R_i}(u:v))] \tag{5}$$

We can find the optimal reference curve $S^{q^*}_{R_i}$ from bow, hsv, gist, rand, and high-level semantics by minimizing cross-entropy loss. The cross-entropy loss algorithm is shown in **Algorithm** 1.

Algorithm 1 *Cross-Entropy Loss Algorithm*

Offline:

1: The target distribution between query image q and the target dataset (Holidays, UKbench and Oxf5k) d on feature F_i is denoted as $S^q_{T_i}$.
2: The reference distribution between query image q and the reference dataset (Flickr1M & Flickr343Places) on feature F_i is denoted as $S^q_{R_i}$.

Online:

1: Input: $F = \{F_1, F_2, ..., F_i | i = 1, 2, ..., c\}$ kinds of feature, target distribution $S^q_{T_i}$ and reference distribution $S^q_{R_i}$, vector segment of matching constraint region $(u : v)$.
2: Output: matching optimal reference curve $S^{q^*}_{R_i}$.
3: Initialize the target distribution $S^q_{T_i}(u : v)$ and reference distribution $S^q_{R_i}(u : v)$.
4: Calculate cross-entropy loss between $S^q_{T_i}(u : v)$ and $S^q_{R_i}(u : v)$ by Eq. (5).
5: Update $R_{F_{(i+1)}}$ (where i represents the type of feature to be fused, R_{F_i} denotes the reference curve of corresponding feature).
6: Repeat steps. 4-5 until $S^{q^*}_{R_i}$ satisfies requirement of minimizing the cross-entropy loss between the target distribution $S^q_{T_i}(u : v)$ and the reference distribution $S^q_{R_i}(u : v)$.
7: Output the best match $S^{q^*}_{R_i}$.

3.4. Adaptive fusion

We obtained the optimal reference curve $S^{q^*}_{R_i}$ by cross-entropy loss, the next step is to eliminate the 'high tail' of the target curve through the optimal reference curve $S^{q^*}_{R_i}$.

$$\widehat{S^q_{T_i}} = S^q_{T_i} - S^{q^*}_{R_i} \tag{6}$$

, wherein $S^q_{T_i}$ is the target distribution, $S^{q^*}_{R_i}$ is the optimal reference curve from reference distribution, both of which are obtained on feature F_i. Then, $\widehat{S^q_{T_i}}$ undergoes min–max normalization:

$$\tilde{S}^q_{T_i} = \frac{\widehat{S^q_{T_i}} - (\widehat{S^q_{T_i}})_{min}}{(\widehat{S^q_{T_i}})_{max} - (\widehat{S^q_{T_i}})_{min}} \tag{7}$$

After min–max normalization by Eq. (7), good query features tend to have a small area under the score curve, and vice versa. In this way, the similarity curves $\tilde{S}^q_{T_i}$ can be used to estimate the effectiveness of different feature representations based on the similarity it shares with each query, so that those 'good' features are endowed with larger weights for providing greater contributions, while the 'bad' features are punished, thus attaining differential fusion of heterogeneous features.

$$W^q_{F_i} = \frac{\frac{1}{A_{F_i}}}{\sum_{i=1}^{c} \frac{1}{A_{F_i}}}, \tag{8}$$

where A_{F_i} represents the area under the similarity curve $\tilde{S}^q_{T_i}$ normalized by Eq. (7). Next, the Eq. (8) will be applied to Eq. (9) to assign feature fusion weights adaptively. To be specific, suppose that given query image q with c features to be fused, the similarity between query image q and every image in the target dataset d on feature F_i is denoted as $S^q_{T_i}$, where $i = 1, 2, \ldots, c$. Let the weight of query image q with respect to feature F_i be $W^q_{F_i}$, and its sum be 1. Therefore, the similarity measurement of adaptive multi-feature fusion can be defined as,

$$sim(q, d) = \prod_{i=1}^{c} (S^q_{T_i})^{(W^q_{F_i})}, \ where \ \sum_{i=1}^{c} W^q_{F_i} = 1 \tag{9}$$

For each input query image, we complete the fusion by extracting features, constructing target/reference distribution, and normalization of cross-entropy loss. Also, since our method is late-processing, the above steps are used to complete the retrieval task for each new query. Hence, feature extraction of the target dataset and reference dataset is offline, while feature extraction of new query and cross-entropy normalization feature fusion is online. Meanwhile, in our work, we only consider time complexity of the online part, and Section 5.6 has a detailed time complexity analysis.

4. Experimental settings

4.1. Benchmark datasets

We briefly introduce the public datasets: Holidays (Jegou et al., 2008), UKbench (Nister & Stewenius, 2006), Oxf5k (Philbin, Chum, Isard, Sivic, & Zisserman, 2007), Flickr1M (Jegou et al., 2008), and Flickr343Places (Zheng et al., 2015).

The **Holidays** dataset is collected by Jegou et al. (2008) from personal holiday albums, so most of the images are of various scene types. The dataset consists of 1491 images from 500 categories of similar images. Each image category has one query, totaling 500 query images. To more clearly represent retrieval performance, we adopt average precision (AP) as a metric to evaluate model performance. In general, a larger AP means a higher precision–recall curve, resulting in better retrieval results. Moreover, since retrieval datasets typically have multiple query images, their respective APs are averaged to produce the final performance evaluation, namely, the mean average precision (mAP).

The **UKbench** dataset is collected by Nister and Stewenius (2006), which contains 2550 categories and 10,200 images in total. Each category has four images depicting the same scene, under various viewpoints, illuminations, etc. In the experiment, each image is taken as the query in turn so there are 10,200 queries and using an N-S index (maximum 4) as a performance evaluation indicator.

The **Oxf5k** dataset is collected by crawling for particular Oxford landmark images from Flickr. This dataset consists of 5062 images, which has a comprehensive ground truth for 11 different landmarks, each containing five possible queries. For each query, it has many similar images of the same instance, which are taken from different views. The performance is measured by mean Average Precision (mAP).

The **Flickr1M & Flickr343Places** datasets both have 1M images. Since both two datasets cover images that have similar scene categories as the target datasets (including Holidays, UKbench, and Oxf5k), so providing sufficient feature correlation. Thus, these two independent and extremely public datasets are used to construct the reference curve, which is then cross-entropy normalized to the target curve.

We select the datasets for both task variety and benchmark consistency. On one hand, the four selected datasets cover large/small collections, campus/people/city/scenery domains, dense/sparse categories, which together represent typical retrieval tasks. On the other hand, these datasets are also adopted by recent image retrieval literature.

4.2. Experiment setup

We demonstrate the experimental result of our method on three public datasets and compare it with existing baseline methods. The experiments are conducted with Windows 64, Intel i7-7800X CPU, 64.00 GB RAM, and two Nvidia GeForce GTX-1080Ti GPUs. Meanwhile, to alleviate the adverse influence of randomness, we repeat our experiment for 5 times and report the average values and the corresponding standard deviations.

4.3. Research questions

To examine the effectiveness of our proposal, we conduct extensive experiments to answer the following research questions (RQs):

RQ1: What is the retrieval baseline of 9 single features?
RQ2: What is the retrieval performance of multi-feature fusion?
RQ3: How do the involved parameter variables in our method affect the retrieval performance?
RQ4: How does the fusion of different high-level semantic features in our method affect retrieval performance?
RQ5: Is the overall performance of our method superior to the baselines?
RQ6: What are the efficiency (run-time) and memory cost of our method?

5. Experimental results and analysis

In this section, we present the results of the experiments with the corresponding analysis on public benchmark datasets for image retrieval.

5.1. Retrieval performance of single feature

To answer **RQ1**, we extract 9 features on Holidays, UKbench and Oxf5k datasets, including four low-level features and five high-level semantic features, respectively. Meanwhile, to be fair, we compare descriptions given the same parameters (single feature retrieval parameters u, v and Q are the same, the sensitivity will be evaluated in Section 5.3). The retrieval accuracy of a single feature is presented in Tables 2 and 3.

It shows that the low-level BoW has achieved 80.16% of mAP, 3.582 of N-S and 74.83% of mAP performance on Holidays, UKbench and Oxf5k datasets, respectively. However, GIST and RAND have achieved poor performance on the three public benchmark datasets. Specifically, the features of GIST and RAND are introduced mainly for comparison with Score Fusion. For the high-level semantic features, the features extracted from the DenseNet121 network model achieved 75.90% of the mAP, 3.685 of the N-S and 48.19% of the mAP on Holidays, UKbench and Oxf5k datasets, respectively. Moreover, the feature extracted from

Table 2
Image retrieval performance of four low-level features.

Dataset/Method	HSV	BoW	GIST	RAND
Holidays, mAP(%)	61.32	80.16	33.81	13.49
UKbench, N-S	3.195	3.582	1.856	1.422
Oxf5k, mAP(%)	35.32	74.83	12.94	9.38

Table 3
Image retrieval performance of five high-level semantic features.

Dataset/Method	AlexNet	VGG19	InceptionV3	ResNet50	DenseNet121
Holidays, mAP(%)	69.33	63.59	65.50	74.83	75.90
UKbench, N-S	3.397	3.319	3.388	3.638	3.685
Oxf5k, mAP(%)	44.56	41.32	43.28	47.51	48.19

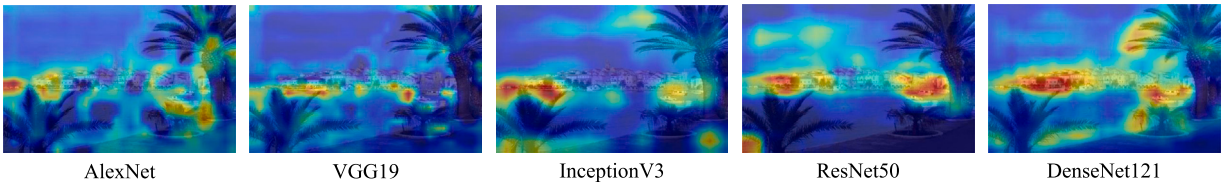


Fig. 2. Feature visualization of different CNN pre-trained models.

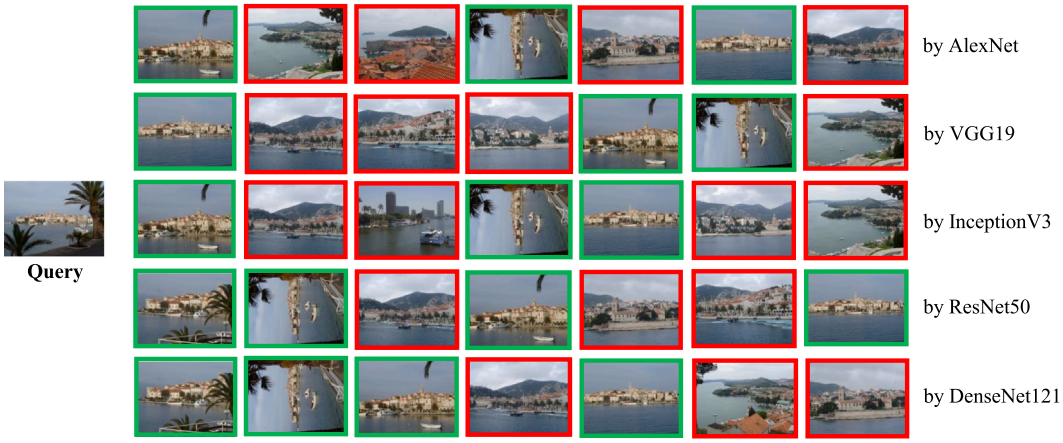
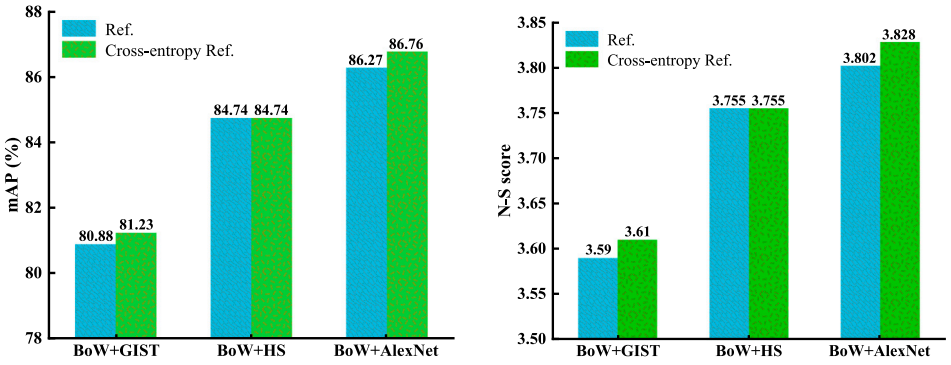


Fig. 3. A visualization example of image retrieval using the high-level semantic feature extracted by different CNN pre-trained models.

both AlexNet and VGG19 models is 4096-dim, InceptionV3 and ResNet50 models are 2048-dim, while DenseNet121 model is 1024-dim. As the dimension decreases, the redundancy of features is also decreasing and the utilization rate of features is increasing, thus achieving a better performance. This trend is mainly due to the use of dense residual blocks in DenseNet121, which enables features to be fused at multi-scales. This is consistent with the performance obtained in image classification (Huang et al., 2019). In a word, the BoW features correspond to the best performances on Holidays and Oxf5k when a single feature is adopted for retrieval. While the features extracted from the DenseNet121 network model correspond to the best performances on UKbench when single feature is leveraged for retrieval.

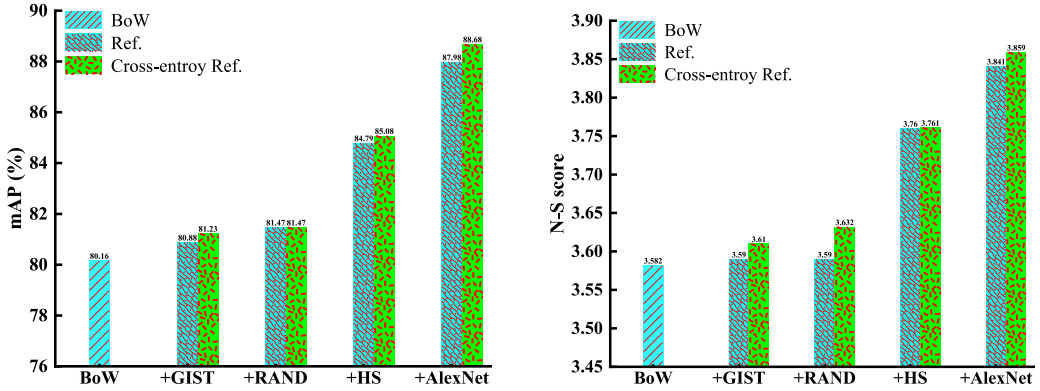
As shown in Figs. 2 and 3, we have added two qualitative experimental results, feature visualization and retrieval visualization, to better illustrate the differences in features from different pre-trained CNN models. In Fig. 2, we visualize five high-level semantic features: AlexNet and VGG19 models are 4096-dim, InceptionV3 and ResNet50 models are 2048-dim, while DenseNet121 model is 1024-dim. One can see that from AlexNet to DenseNet, the models focus more on fine-grained information while ignoring background and noise information. Namely, as the dimension decreases the redundancy of features is decreasing and the utilization rate of features is increasing. Meanwhile, we demonstrate some examples of retrieving relevant images using high-level semantic features extracted from various CNN models. True-matched images are marked with the green symbol, and false-matched ones are red. As can be seen from Fig. 3, although the same query image is used, there will be great differences in the results due to the semantic features of diversity extracted by distinct models.



(a) Holidays: Feasibility of Cross-entropy Reference.

(b) UKbench: Feasibility of Cross-entropy Reference.

Fig. 4. The feasibility of the proposed method. In without changing the features, we compare with Score Fusion by the feature fusion manner of 'BoW+GIST', 'BoW+HS' and 'BoW+AlexNet' on Holidays and UKbench. The cyan and green bars show results by 'Reference' of Score Fusion and our 'Cross-entropy Reference', respectively.



(a) Holidays: Ref. *vs* Cross-entropy Ref.

(b) UKbench: Ref. *vs* Cross-entropy Ref.

Fig. 5. In without changing the features, comparison of the performance with Score Fusion on Holidays and UKbench. The slash cyan bar represents the baseline of BoW, while the cyan and green bars show results by 'Reference' of Score Fusion and our 'Cross-entropy Reference', respectively.

5.2. Fusion multi-feature

To answer **RQ2**, we employ their (Score Fusion) released code and default parameters to conduct the cross-entropy normalized fusion with four features: BoW, HS, GIST, and AlexNet. The fusion results on the two datasets are shown in Fig. 4, it can be seen that the fusion mechanism is effective. In the datasets of Fig. 4(a) Holidays and Fig. 4(b) UKbench, three fusion manners of four features are used respectively. In Holidays, the mAP of 'BoW+GIST' and 'BoW+AlexNet' increased to 81.23% and 86.76% respectively, and the N-S of 'BoW+GIST' and 'BoW+AlexNet' increased to 3.610 and 3.828 respectively in UKbench.

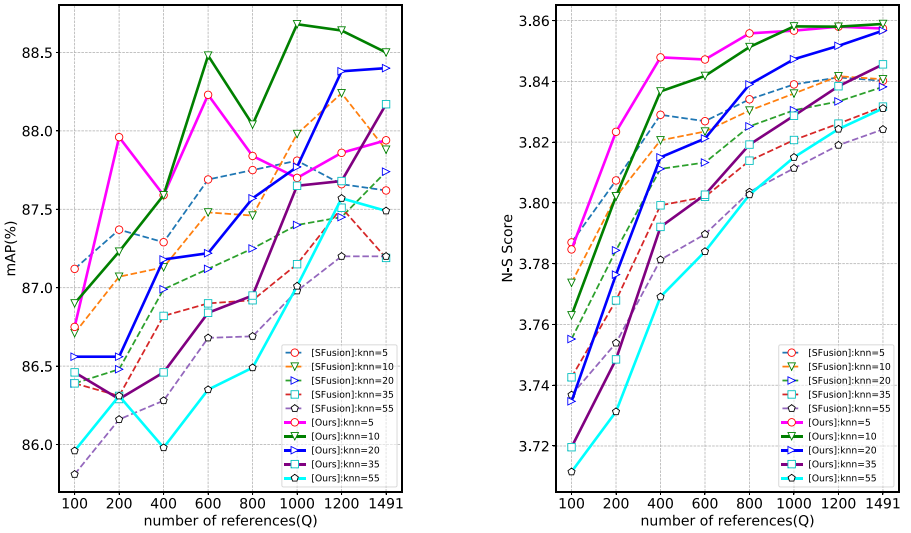
In addition, multi-feature is fused with BoW on two datasets and the result comparisons are presented in Fig. 5. It shows that on both datasets the proposed method outperforms Score Fusion. Experiments are conducted on five features fusion: 'BoW+GIST', 'BoW+GIST+RAND', 'BoW+GIST+RAND+HS' and 'BoW+GIST+RAND+HS+AlexNet'. The slash cyan bar represents the baseline result of BoW, while the cyan bar and the green bar respectively represent the Score Fusion and our method. Overall, our method is superior to Score Fusion, especially after the fusion of high-level semantic features. The detailed experimental results of multi-feature fusion on three datasets comparing two methods are shown in Table 4. When multiple features are fused, the performance is further boosted. The fusion of five features (BoW+GIST+RAND+HS+AlexNet) achieves 88.68% in mAP, 3.859 in N-S and 84.65% in mAP on Holidays, UKBench and Oxf5k, respectively. Moreover, with DenseNet121 feature, our performance are further enhanced to 89.02% in mAP, 3.8813 in N-S and 84.95% in mAP, respectively.

5.3. Ablation study of parameter tuning

To answer **RQ3**, we conduct a series of ablation studies to investigate the contributions of different parameter variables on Holidays and UKbench. In our work, the experiment mainly involves four parameter variables, which u and v are the vector segment

Table 4
Fusion results (mean±std) of different feature combinations on three public benchmarks datasets.

Feature combinations	Holidays, mAP(%)		UKbench, N-S		Oxf5k, mAP(%)	
	Score Fusion (Zheng et al., 2015)	Ours	Score Fusion (Zheng et al., 2015)	Ours	Score Fusion (Zheng et al., 2015)	Ours
BoW+GIST	80.88 ± 0.2	81.61 ± 0.2	3.590 ± 0.02	3.610 ± 0.02	79.51 ± 0.1	79.62 ± 0.1
BoW+RAND	80.91 ± 0.1	80.91 ± 0.1	3.596 ± 0.01	3.605 ± 0.01	79.23 ± 0.2	79.30 ± 0.1
BoW+GIST+RAND	81.47 ± 0.1	81.47 ± 0.1	3.590 ± 0.02	3.632 ± 0.01	79.72 ± 0.2	79.88 ± 0.1
BoW+HS	84.47 ± 0.2	84.47 ± 0.1	3.755 ± 0.02	3.755 ± 0.02	80.27 ± 0.2	80.42 ± 0.2
BoW+AlexNet	86.27 ± 0.1	86.76 ± 0.2	3.802 ± 0.03	3.828 ± 0.02	81.53 ± 0.3	82.45 ± 0.2
BoW+HS+AlexNet	87.95 ± 0.2	88.49 ± 0.2	3.840 ± 0.01	3.855 ± 0.02	82.38 ± 0.3	83.04 ± 0.2
BoW+GIST+RAND+HS+AlexNet	87.98 ± 0.1	88.68 ± 0.2	3.841 ± 0.02	3.859 ± 0.03	83.73 ± 0.3	84.65 ± 0.3
BoW+VGG19	81.39 ± 0.4	85.59 ± 0.2	3.623 ± 0.01	3.772 ± 0.02	76.89 ± 0.1	80.27 ± 0.2
BoW+HS+VGG19	84.93 ± 0.3	87.17 ± 0.2	3.770 ± 0.02	3.832 ± 0.03	79.52 ± 0.2	82.93 ± 0.3
BoW+GIST+RAND+HS+VGG19	85.31 ± 0.2	87.27 ± 0.2	3.775 ± 0.05	3.836 ± 0.03	80.28 ± 0.3	83.14 ± 0.1
BoW+InceptionV3	82.62 ± 0.4	84.98 ± 0.3	3.641 ± 0.03	3.837 ± 0.01	77.62 ± 0.2	80.84 ± 0.2
BoW+HS+InceptionV3	86.12 ± 0.1	87.87 ± 0.1	3.779 ± 0.02	3.853 ± 0.02	80.25 ± 0.1	83.37 ± 0.2
BoW+GIST+RAND+HS+InceptionV3	86.13 ± 0.2	87.94 ± 0.1	3.783 ± 0.02	3.854 ± 0.03	80.38 ± 0.2	83.79 ± 0.2
BoW+ResNet50	82.70 ± 0.6	88.02 ± 0.4	3.634 ± 0.01	3.823 ± 0.03	78.03 ± 0.3	81.24 ± 0.3
BoW+HS+ResNet50	85.48 ± 0.2	88.19 ± 0.3	3.773 ± 0.03	3.853 ± 0.02	80.74 ± 0.4	83.50 ± 0.3
BoW+GIST+RAND+HS+ResNet50	85.57 ± 0.4	88.29 ± 0.4	3.777 ± 0.02	3.855 ± 0.02	80.86 ± 0.2	83.67 ± 0.2
BoW+DenseNet121	83.23 ± 0.2	87.95 ± 0.3	3.670 ± 0.01	3.858 ± 0.02	79.84 ± 0.1	82.74 ± 0.2
BoW+HS+DenseNet121	86.06 ± 0.2	88.97 ± 0.2	3.787 ± 0.02	3.8812 ± 0.01	82.47 ± 0.2	84.81 ± 0.2
BoW+GIST+RAND+HS+DenseNet121	86.17 ± 0.2	89.02 ± 0.3	3.790 ± 0.01	3.8813 ± 0.01	82.54 ± 0.3	84.95 ± 0.2



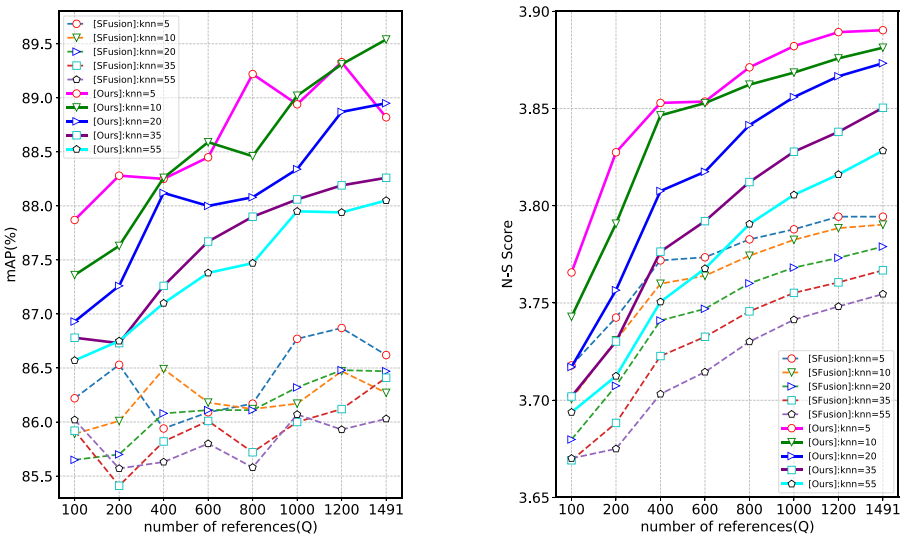
(a) Holidays: mAP *vs* Q (b) UKbench: NS *vs* Q

Fig. 6. The sensitivity of parameter Q on Holidays and UKbench. We set the ablation experiment with fine-tuning parameters of kNN=5, 10, 20, 35, 55 and Q=100, 200, 400, 600, 800, 1000, 1200, 1491, respectively. Specifically, if not explicitly stated we set Q=1000 and kNN=10 in our work.

parameters that restrict search and matching of the optimal reference curve, the number of reference curves Q, and kNN. Compared with other methods, u and v are the same as Score Fusion, which is 10 and 400 respectively. The performance sensitivity of the number of reference curves Q is evaluated and the results are shown in Fig. 6.

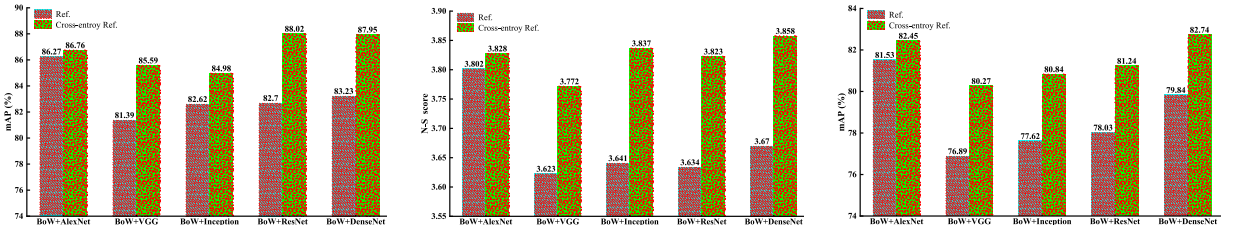
It is shown in Fig. 6 that with the number of reference curves Q increases, the accuracy is gradually improved. Our work sets kNN=5, 10, 20, 35, 55 and Q=100, 200, 400, 600, 800, 1000, 1200, 1491, respectively. This result also confirms the necessity of reference curve for performance improvement. In particular, the proposed method can find a good approximation to the tail if the best matching reference curve is searched over a larger range. Nevertheless, the computational complexity of the whole system will also increase with the increase of Q. Therefore, in order to balance efficiency and accuracy, we set Q=1000 in the experiment without explicitly stating it.

Our method uses the codes released by Score Fusion to conduct the effect of cross-entropy and non-cross-entropy normalization on fine-tuning parameters, in Fig. 6. To further illustrate the strength of our method, we employ high-level semantic features with higher



(a) Holidays: mAP *vs* Q (b) Ukbench: NS *vs* Q

Fig. 7. The ablation research of fusing different high-level semantic features and fine-tuning parameters Q on Holidays and UKbench. Similar to Fig. 6, this paper sets the kNN=5, 10, 20, 35, 55 and Q=100, 200, 400, 600, 800, 1000, 1200, 1491, respectively.



(a) (Holidays_BoW+CNNs): Ref. *vs* Cross-entropy Ref. (b) (UKbench_BoW+CNNs): Ref. *vs* Cross-entropy Ref. (c) (Oxf5k_BoW+CNNs): Ref. *vs* Cross-entropy Ref.

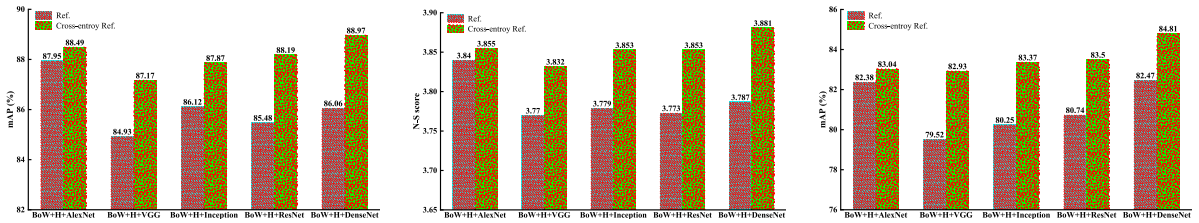
Fig. 8. Comparison of different high-level semantic feature fusion BoW. The cyan and green bars show results by 'Reference' of Score Fusion and 'Cross-entropy Reference' of our method, respectively.

feature utilization rate to conduct fine-tuning parameter experiments. As can be seen from Fig. 7, the high-level semantic features extracted by DenseNet121 with higher feature utilization rate are used. Our method not only maintains the steady improvement of fusion accuracy along with the increment of reference curve Q, but also on both datasets our method is superior to Score Fusion.

5.4. Fusion of different high-level semantic features

To answer **RQ4**, we adopt high-level semantic information features with less redundancy and a higher feature utilization rate. Five kinds of pre-training CNN models, AlexNet, InceptionV3, VGG19, ResNet50 and DenseNet121 are leveraged to extract high-level semantic features, and their feature utilization rate is gradually improved. The experimental results on the three datasets are shown in Figs. 8 and 9 that one can draw the following conclusions:

- In Fig. 8, BoW and five high-level semantic features are fused respectively on three datasets: 'BoW+AlexNet', 'BoW+VGG19', 'BoW+InceptionV3', 'BoW+ResNet50' and 'BoW+DenseNet121'. The cyan bar and the green bar represent the Score Fusion and our method, respectively. It can be seen that our performance is superior than that of Score Fusion. In particular, when combined with 'BoW+ResNet50' features, our work has a great improvement on Holidays, by 5.3 points. BoW is not well fused with VGG19 and InceptionV3 respectively, since the performance of AlexNet is better than that of VGG19 and InceptionV3 on individual feature retrieval in Table 3.
- In Fig. 9, BoW, HSV and five high-level semantic features are fused on three datasets, 'BoW+HS+AlexNet', 'BoW+HS+VGG19', 'BoW+HS+InceptionV3', 'BoW+HS+ResNet50', 'BoW+HS+DenseNet121'. The cyan bar and the green bar respectively represent the Score Fusion and our method. When combining three features, our performance is also better than Score Fusion. Especially, the 'BoW+HS+AlexNet' is better than the five features fusion of Score Fusion. With the improvement of the



(a) (Holidays_BoW+HS+CNNs): Ref. *vs* Cross-entropy Ref. (b) (UKbench_BoW+HS+CNNs): Ref. *vs* Cross-entropy Ref. (c) (Oxf5k_BoW+HS+CNNs): Ref. *vs* Cross-entropy Ref.

Fig. 9. Comparison of different high-level semantic feature fusion BoW and HSV. The cyan and green bars show results by 'Reference' of Score Fusion and 'Cross-entropy Reference' of the proposed method, respectively. In particular, the accuracy of 'BoW+HS+AlexNet' using our work is better than that of 'BoW+GIST+RAND+HS+AlexNet' using Score Fusion on Holidays.

(a) Holidays (b) Ukbench

Fig. 10. Some illustrative search examples from Holidays (left) and UKBench (right), respectively. For each query, its top-5 ranked images resulted from GIST (the first row), HSV (the second row), DenseNet121 (the third row), BoW (the fourth row) and the proposed method (the fifth row) are shown, respectively. True matched images are marked with green box, and false matched ones red.

utilization rate of high-level semantic features, our performance is also improved steadily, and the 'BoW+HS+DenseNet121' has the best fusion performance. The experimental results of five high-level semantic feature fusion are shown in Table 4.

- Furthermore, we also qualitatively illustrate the performance and the advantages of our method. Specifically, it can be seen from Fig. 10 that our method allows more query-related images to be retrieved in the top returned shortlist compared with the other competitors.

5.5. Comparison with other fusion methods

To answer **RQ5**, we compare our results with three existing fusion mechanisms: Semantic Aware Co-indexing (Zhang et al., 2015), Graph Fusion (Zhang et al., 2014) and Score Fusion (Zheng et al., 2015). In the experiment, our method adopts their codes and default parameters for Graph Fusion. The only difference of this method is that kNN parameters are different (all parameters of our method are the same as Score Fusion). Multiple features are fused with BoW. In Table 5, the experimental results show that our method is better than Graph Fusion. In Score Fusion* (Zheng et al., 2015), the author also tried fine-tuning the global weight, manually adjusting the weight at a step size of 0.1. From the experimental results, the performance of the adaptive weight adjustment fusion of two features is not as good as 'fine-tuning the global weight', but the effect of the adaptive weight adjustment is better when three or five features are fused. Our work is also adaptive weight, whether two or three or five features are fused, and its performance is almost better than the first two.

Compared with Semantic Aware Co-indexing, the BoW is adopted by Co-indexing (Zhang et al., 2015) and Score Fusion‡ (Zheng et al., 2015) is not hamming embedded, and the precision of single feature retrieval on the two public benchmark datasets is mAP=50.10% and N-S=3.112, respectively. Even in this case, the Score Fusion is improved a lot. The performance of the BoW used in our paper has been improved to mAP=80.16% and N-S=3.582 respectively on Holidays and UKbench by hamming embedded, and fusing other features is even better. According to the above experimental results, and compared with Semantic Aware Co-indexing, Graph Fusion and Score Fusion, our method can better optimize the complementarity of heterogeneous features, thus the performance is greatly improved. Furthermore, the experimental results of the proposed method are compared with the state-of-the-art in Table 6. Our results achieve mAP = 89.02%, N-S = 3.881 and mAP = 84.95% on Holidays, UKbench and Oxf5k, respectively.

Table 5
The overall performance (mean±std) comparison with Co-indexing (Zhang et al., 2015), Graph Fusion (Zhang et al., 2014) and Score Fusion (Zheng et al., 2015) on two public benchmark datasets. In particular, Score Fusion* (Zheng et al., 2015) indicates fine-tuning the global weight, the author manually adjusts the weight at a step size of 0.1. Score Fusion‡ (Zheng et al., 2015) represents that all feature descriptors are not hamming embedded.

Feature combinations	Holidays, mAP(%)				UKbench, N-S			
	Graph Fusion (Zhang et al., 2014)	Score Fusion* (Zheng et al., 2015)	Score Fusion (Zheng et al., 2015)	Ours	Co-indexing (Zhang et al., 2015)	Score Fusion‡ (Zheng et al., 2015)	Score Fusion (Zheng et al., 2015)	Ours
BoW+GIST	76.39 ± 0.2	81.54 ± 0.3	80.88 ± 0.2	81.61 ± 0.2	2.766 ± 0.03	3.177 ± 0.03	3.590 ± 0.02	3.610 ± 0.02
BoW+RAND	76.57 ± 0.2	81.18 ± 0.2	80.91 ± 0.1	80.91 ± 0.1	2.701 ± 0.02	3.210 ± 0.01	3.596 ± 0.01	3.605 ± 0.01
BoW+GIST+RAND	70.59 ± 0.1	81.65 ± 0.2	81.47 ± 0.1	81.47 ± 0.1	2.829 ± 0.01	3.263 ± 0.03	3.590 ± 0.02	3.632 ± 0.01
BoW+HS	81.58 ± 0.2	84.18 ± 0.1	84.47 ± 0.2	84.47 ± 0.1	3.504 ± 0.02	3.541 ± 0.03	3.755 ± 0.02	3.755 ± 0.02
BoW+AlexNet	83.36 ± 0.1	86.60 ± 0.3	86.27 ± 0.1	86.76 ± 0.2	3.562 ± 0.02	3.624 ± 0.03	3.802 ± 0.03	3.828 ± 0.02
BoW+HS+AlexNet	83.75 ± 0.3	87.23 ± 0.4	87.95 ± 0.2	88.49 ± 0.2	3.661 ± 0.01	3.750 ± 0.02	3.840 ± 0.01	3.855 ± 0.02
BoW+GIST+RAND+HS+AlexNet	81.04 ± 0.1	87.34 ± 0.2	87.98 ± 0.1	88.68 ± 0.2	3.608 ± 0.02	3.752 ± 0.04	3.841 ± 0.02	3.859 ± 0.03

Table 6
Performance comparison with other state-of-the-art methods. Here, "-" denotes that no experimental results with same settings are available.

Method/Dataset&Indicators	Holidays, mAP(%)	UKbench, N-S	Oxf5k, mAP(%)	Query time (s)	Memory cost (GB)
BOF (Jégou, Douze, & Schmid, 2009)	84.80	3.64	–	–	–
Co-RMGL (Deng et al., 2013)	84.70	3.75	84.3	–	–
MBF (Zheng, Wang, & Tian, 2014)	85.20	3.79	–	0.145	6.1
Index fusion (Zheng, Wang, Liu et al., 2014)	85.80	3.85	–	1.413	6.1
CroW (Kalantidis et al., 2016)	85.10	3.63	71.8	–	–
FDLCDF (Liu, Guo et al., 2017)	82.05	3.636	–	**0.075**	–
GatedSQU (Chen et al., 2018)	88.80	3.74	69.4	–	–
Graph Fusion (Zhang et al., 2014)	84.64	3.83	–	0.749	–
MMF-SIFT (Zhang et al., 2019)	84.40	3.04	–	–	–
Score Fusion (Zheng et al., 2015)	87.98	3.841	83.73	–	**0.076**
SaCoCo (Iakovidou et al., 2019)	76.10	3.33	–	–	–
RGSF (Valem & Pedronette, 2020a)	**90.51**	3.79	79.42	–	–
N3G-MFR (Liu et al., 2020)	86.45	3.88	–	1.68	–
BMSL-CSRSL (Wu et al., 2021)	84.70	3.64	72.2	–	–
SMVF (Li et al., 2021)	87.80	3.77	65.1	–	–
MFBCR (Lao et al., 2021)	–	**3.93**	–	4.667	0.211
DSFH (Liu & Yang, 2021)	74.76	3.528	62.2	–	–
Ours	**89.02**	3.8813	**84.95**	**0.214**	**0.41**

5.6. Retrieval efficiency and cost

To answer **RQ6**, we discuss the average query time and memory cost in this section. Our proposal includes both offline procedures and online procedures. Considering the experimental scenario, we neglect the model training step and feature extraction step on the offline part. For the online part, in the proposed method, most of the computation costs focus on similarity computing, including the similarity between a query and every image in the target datasets as well as the similarity between a query and every image in the reference datasets. Suppose the size of the target dataset is N and the reference dataset is M, while the fundamental computation complexity is required for the similarity measure $O(NM)$. However, in practical experiments, we perform two matches (query for target dataset and query for reference dataset) under several feature representations. As a result, given a new query q, we need to compute the similarity between q and every image in two datasets with computation complexity $O(KNM)$, where K is the number of to-be-fused feature representations in a query image.

Table 6 shows the comparison of average query time and memory cost on two datasets with other competitors. For query time, it should be noted that the feature dimensions and the number of features to-be-fused in our work are not exactly the same as other methods. Specifically, FDLCDF (Liu, Guo et al., 2017) (adopts the semantic feature of 1000-dim, which integrates three features; ours 1024-dim, combines five features, so the average query time is about 0.214 s) and Score Fusion (Zheng et al., 2015) (semantic feature 4096-dim is adopted, while ours 1024-dim). Furthermore, our method is a late-processing fusion technology that works on a given similarity score. We also compares the time of late-processing steps of our method with other late-processing counterpart considered in Table 6, such as Graph Fusion (Zhang et al., 2014) and Score Fusion (Zheng et al., 2015). However, the efficiency of the retrieval system depends on many objective factors, such as the performance of the computer, the dimension of the features, the number of fused features and so on. Still, it can broadly reflect the highly competitive result of our method.

In addition, the memory footprint is also an indicator of retrieval performance, the memory cost of the proposed method is about 0.41 GB. Our method adopts the same framework with Graph Fusion and Score Fusion, thus the memory costs of these methods are similar to ours in theory. Yet, it is about 5 times higher than the Score Fusion, mainly because the reference curve codebook used in the online evaluation requires only 0.076 GB of extra memory. In this paper, there is a process to calculate and select the optimal reference curve, so the theoretical memory footprint is about 5 times that of Score Fusion. The experimental results show that this trend is indeed consistent with the theory.

6. Discussions

In this paper, we propose an adaptive multi-feature late fusion via cross-entropy normalization for effective image retrieval, which achieves competitive results compared with existing methods. Hence, the theoretical and practical implications of our research can greatly promote the development of image retrieval technique to a certain extent. Specifically, it is summarized as follows:

- No free lunch for feature representation in retrieval tasks. The discriminability of feature representations varies from query to query, influenced by its semantic domain, content property, and even subjective characteristics. In other tasks, adopting a flexible ensembling method is also a good manner to practice improving performance.
- The knowledge in large datasets is beneficial for guiding computer vision tasks, such as its feature representation can be leveraged to evaluate the generalization of models and can also be adopted as an indicator of utility sensitivity.

Yet, this work also has some drawbacks. Specifically, it is summarized as follows:

- Our work employs the pre-training CNN model based on image classification to extract the high-level semantic features and achieve promising performance. However, image classification is different from image retrieval. Image retrieval is a more fine-grained issue, which pays more attention to the local visual information of images. The discrimination of this pattern information by this manner is an important factor affecting the retrieval performance. Therefore, in future work, we will further explore a more suitable CNN model for multi-feature fusion image retrieval.
- In addition, our method is a late-processing fusion technology that works on a given similarity score. By investigating the area under the ranking curve as an indicator of feature validity, namely, the area of good features is small, and vice versa. Nevertheless, this process requires significant computational overhead and memory footprint. Hence, in future work, we will explore another point: designing an 'early' coding method to determine the effectiveness of features.

7. Conclusions

This paper proposes an adaptive multi-feature fusion via cross-entropy normalization for effective image retrieval. On the one hand, the pre-trained deep learning network model is used to extract high-level semantic features with high feature utilization rate to improve the performance of semantic information. On the other hand, this work adopts the cross-entropy normalization to optimize the complementarity of heterogeneous features. It can not only reduce the 'high tail' of bad features, but also balance the relationship between the weight distribution and complementarity. And this proposal has proved that our method can achieve superior performance to other counterparts via extensive experiments on three public benchmark datasets.

CRediT authorship contribution statement

Wentao Ma: Conceived and designed the study, Performed the experiments, Writing – review & editing. **Tongqing Zhou:** Performed the experiments, Writing – review & editing. **Jiaohua Qin:** Conceived and designed the study, Reviewed and edited the manuscript. **Xuyu Xiang:** Writing – review & editing. **Yun Tan:** Performed the experiments, Reviewed and edited the manuscript. **Zhiping Cai:** Conceived and designed the study, Reviewed and edited the manuscript.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant 61772561, 62002392, 62072465, 62172155 and 62102425; in part by the Key Research and Development Plan of Hunan Province under Grant 2019SK2022; in part by the Postgraduate Excellent teaching team Project of Hunan Province under Grant [2019]370-133; in part by the Natural Science Foundation of Hunan Province, China under Grant 2020JJ4141 and 2020JJ4140; in part by the Science and Technology Innovation Program of Hunan Province under Grant 2021RC2071; in part by the Postgraduate Research and Innovation Project of Hunan Province under Grant CX20210080. All authors read and approved the manuscript.

Appendix

Theorem 1. *If the distribution is* $S^q_{T_i}(u:v) \neq S^q_{R_i}(u:v)$*, then* $D(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) \geq 0$.

Proof. By Jensen's inequality, we derive the objective of Eq. (1). Note that

$$\begin{aligned} D(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) &= \sum_i S^q_{T_i}(u:v)\log(\frac{S^q_{T_i}(u:v)}{S^q_{R_i}(u:v)}) \\ &= -\sum_i S^q_{T_i}(u:v)\log(\frac{S^q_{R_i}(u:v)}{S^q_{T_i}(u:v)}) \\ &\geq -\log\sum_i (S^q_{T_i}(u:v))(\frac{S^q_{R_i}(u:v)}{S^q_{T_i}(u:v)}) \\ &= -\log\sum_i S^q_{R_i}(u:v) = 0 \end{aligned} \tag{10}$$

Only if $S^q_{T_i}(u:v) = S^q_{R_i}(u:v)$, the $D(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) = 0$. However, the curves come from different datasets then $S^q_{T_i}(u:v) \neq S^q_{R_i}(u:v)$, so by Eq. (1) the $D(S^q_{T_i}(u:v), S^q_{R_i}(u:v)) > 0$. □

Theorem 2. *If $H(S^q_{T_i}(u:v))$ is a constant, then $min[D(S^q_{T_i}(u:v), S^q_{R_i}(u:v))]$ can be equivalent to $min[E(S^q_{T_i}(u:v), S^q_{R_i}(u:v))]$.*

Proof. By Eq. (3),

$$\begin{aligned} min[D(S^q_{T_i}(u:v), S^q_{R_i}(u:v))] &\Leftrightarrow min[E(S^q_{T_i}(u:v)), \\ &\quad S^q_{R_i}(u:v) - H(S^q_{T_i}(u:v))] \\ &\Leftrightarrow min[E(S^q_{T_i}(u:v), S^q_{R_i}(u:v))] \end{aligned} \tag{11}$$

To minimize the difference between the distribution $S^q_{T_i}(u:v)$ of similarity ranking and the distribution $S^q_{R_i}(u:v)$ of reference curve is equivalent to minimize the relative entropy between them. □

References

Abdi, A., Shamsuddin, S. M., Hasan, S., & Piran, J. (2019). Deep learning-based sentiment classification of evaluative text based on multi-feature fusion. *Information Processing & Management*, *56*(4), 1245–1259.

Amato, G., Carrara, F., Falchi, F., Gennaro, C., & Vadicamo, L. (2020). Large-scale instance-level image retrieval. *Information Processing & Management*, *57*(6), Article 102100.

Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2018). NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1437–1451. http://dx.doi.org/10.1109/TPAMI.2017.2711011.

Babenko, A., & Lempitsky, V. (2014). The inverted multi-index. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(6), 1247–1260.

Babenko, A., Slesarev, A., Chigorin, A., & Lempitsky, V. (2014). Neural codes for image retrieval. In *Proceedings of the european conference on computer vision* (pp. 584–599). Springer.

Bao, Y., & Li, H. (2017). Object-based aggregation of deep features for image retrieval. In *Proceedings of the international conference on multimedia modeling* (pp. 478–489). Springer.

Bhowmik, N., González, R., Gouet-Brunet, V., Pedrini, H., & Bloch, G. (2014). Efficient fusion of multidimensional descriptors for image retrieval. In *Proceedings of the IEEE international conference on image processing* (pp. 5766–5770). IEEE.

Chen, X., Hu, X., & Shen, X. (2009). Spatial weighting for bag-of-visual-words and its application in content-based image retrieval. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 867–874). Springer.

Chen, Z.-D., Li, C.-X., Luo, X., Nie, L., Zhang, W., & Xu, X.-S. (2019). SCRATCH: A scalable discrete matrix factorization hashing framework for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(7), 2262–2275.

Chen, Z., et al. (2018). Gated square-root pooling for image instance retrieval. In *IEEE international conference on image processing* (pp. 1982–1986). IEEE.

Deng, C., Ji, R., Liu, W., Tao, D., & Gao, X. (2013). Visual reranking through weakly supervised multi-graph learning. In *Proceedings of the IEEE international conference on computer vision* (pp. 2600–2607). IEEE.

Douze, M., Ramisa, A., & Schmid, C. (2011). Combining attributes and fisher vectors for efficient image retrieval. In *IEEE conference on computer vision and pattern recognition* (pp. 745–752). IEEE.

Elsayad, I., Martinet, J., Urruty, T., & Djeraba, C. (2010). A new spatial weighting scheme for bag-of-visual-words. In *International workshop on content based multimedia indexing* (pp. 1–6). IEEE.

Ge, X., Wei, G., Yu, J., Singh, A., & Xiong, N. (2021). An intelligent fuzzy phrase search scheme over encrypted network data for IoT. *IEEE Transactions on Network Science and Engineering*.

Gkelios, S., Sophokleous, A., Plakias, S., Boutalis, Y., & Chatzichristofis, S. A. (2021). Deep convolutional features for image retrieval. *Expert Systems with Applications*, *177*, Article 114940.

Gong, Y., Wang, L., Guo, R., & Lazebnik, S. (2014). Multi-scale orderless pooling of deep convolutional activation features. In *Proceedings of the european conference on computer vision* (pp. 392–407). Springer.

Huang, G., Liu, Z., Pleiss, G., Van Der Maaten, L., & Weinberger, K. (2019). Convolutional networks with dense connectivity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. http://dx.doi.org/10.1109/TPAMI.2019.2918284.

Iakovidou, C., et al. (2019). Composite description based on salient contours and color information for CBIR tasks. *IEEE Transactions on Image Processing*, *28*(6), 3115–3129.

Jegou, H., Douze, M., & Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *Prceedings of the european conference on computer vision* (pp. 304–317). Springer.

Jégou, H., Douze, M., & Schmid, C. (2009). On the burstiness of visual elements. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1169–1176). IEEE.

Jegou, H., Douze, M., & Schmid, C. (2010). Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(1), 117–128.

Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3304–3311). IEEE.
Kalantidis, Y., Mellina, C., & Osindero, S. (2016). Cross-dimensional weighting for aggregated deep convolutional features. In *Proceedings of the european conference on computer vision* (pp. 685–701). Springer.
Kobayashi, T. (2014). Dirichlet-based histogram feature transform for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3278–3285). IEEE.
Lao, G., Liu, S., Tan, C., Wang, Y., Li, G., Xu, L., et al. (2021). Three degree binary graph and shortest edge clustering for re-ranking in multi-feature image retrieval. *Journal of Visual Communication and Image Representation*, *80*, Article 103282.
Li, J., Yang, B., Yang, W., Sun, C., & Xu, J. (2021). Subspace-based multi-view fusion for instance-level image retrieval. *Visual Computer*, *37*(3), 619–633.
Liu, P., Guo, J.-M., Wu, C.-Y., & Cai, D. (2017). Fusion of deep learning and compressed domain features for content-based image retrieval. *IEEE Transactions on Image Processing*, *26*(12), 5706–5717.
Liu, Y., Guo, Y., Wu, S., & Lew, M. S. (2015). Deepindex for accurate and efficient image retrieval. In *Proceedings of the ACM on international conference on multimedia retrieval* (pp. 43–50). ACM.
Liu, L., Shen, C., & van den Hengel, A. (2015). The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4749–4757). IEEE.
Liu, S., Sun, M., Feng, L., Qiao, H., Chen, S., & Liu, Y. (2020). Social neighborhood graph and multigraph fusion ranking for multifeature image retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(3), 1389–1399.
Liu, Z., Wang, S., Zheng, L., & Tian, Q. (2017). Robust imagegraph: Rank-level feature fusion for image search. *IEEE Transactions on Image Processing*, *26*(7), 3128–3141.
Liu, G.-H., & Yang, J.-Y. (2021). Deep-seated features histogram: A novel image retrieval method. *Pattern Recognition*, *116*, Article 107926.
Ma, W., Zhou, T., Qin, J., Xiang, X., Tan, Y., & Cai, Z. (2022). A privacy-preserving content-based image retrieval method based on deep learning in cloud computing. *Expert Systems with Applications*, Article 117508.
Ma, W., Zhou, T., Qin, J., Zhou, Q., & Cai, Z. (2022). Joint-attention feature fusion network and dual-adaptive NMS for object detection. *Knowledge-Based Systems*, Article 108213.
Nister, D., & Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2161–2168). IEEE.
Öztürk, Ş. (2020). Stacked auto-encoder based tagging with deep features for content-based medical image retrieval. *Expert Systems with Applications*, *161*, Article 113693.
Pandey, S., Khanna, P., & Yokota, H. (2016). A semantics and image retrieval system for hierarchical image databases. *Information Processing & Management*, *52*(4), 571–591.
Perronnin, F., & Dance, C. (2007). Fisher kernels on visual vocabularies for image categorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–8). IEEE.
Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–8). IEEE.
Qiao, Y., Wu, K., & Jin, P. (2021). Efficient anomaly detection for high-dimensional sensing data with one-class support vector machine. *IEEE Transactions on Knowledge and Data Engineering*.
Sánchez, J., Perronnin, F., Mensink, T., & Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, *105*(3), 222–245.
Valem, L. P., & Pedronette, D. C. G. (2020a). Graph-based selective rank fusion for unsupervised image retrieval. *Pattern Recognition Letters*, *135*, 82–89.
Valem, L. P., & Pedronette, D. C. G. (2020b). Unsupervised selective rank fusion for image retrieval tasks. *Neurocomputing*, *377*, 182–199.
Wu, Z., Li, J., Xu, J., & Yang, W. (2021). Beyond ITQ: Efficient binary multi-view subspace learning for instance retrieval. *Journal of Visual Communication and Image Representation*, *79*, Article 103234.
Xia, Z., Yuan, C., Lv, R., Sun, X., Xiong, N. N., & Shi, Y.-Q. (2018). A novel weber local binary descriptor for fingerprint liveness detection. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *50*(4), 1526–1536.
Xie, L., Tian, Q., & Zhang, B. (2014). Max-sift: Flipping invariant descriptors for web logo search. In *Proceedings of the IEEE international conference on image processing* (pp. 5716–5720). IEEE.
Zhan, H., Zhang, K., Hu, C., & Sheng, V. (2021). Multi-objective privacy-preserving text representation learning. In *Proceedings of the international conference on information & knowledge management* (pp. 3612–3616).
Zhang, S., Yang, M., Cour, T., Yu, K., & Metaxas, D. N. (2014). Query specific rank fusion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(4), 803–815.
Zhang, S., Yang, M., Wang, X., Lin, Y., & Tian, Q. (2015). Semantic-aware co-indexing for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(37), 2573–2587.
Zhang, S., Yang, Y., Xiao, J., Liu, X., Yang, Y., Xie, D., et al. (2018). Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks. *IEEE Transactions on Multimedia*, *20*(9), 2330–2343.
Zhang, Z., et al. (2019). Effective image retrieval via multilinear multi-index fusion. *IEEE Transactions on Multimedia*, *21*(11), 2878–2890.
Zhao, W., Lu, H., & Wang, D. (2017). Multisensor image fusion and enhancement in spectral total variation domain. *IEEE Transactions on Multimedia*, *20*(4), 866–879.
Zheng, L., Wang, S., He, F., & Tian, Q. (2014). Seeing the big picture: Deep embedding with contextual evidences. arXiv preprint arXiv:1406.0132.
Zheng, L., Wang, S., Liu, Z., & Tian, Q. (2014). Packing and padding: Coupled multi-index for accurate image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1939–1946). IEEE.
Zheng, L., Wang, S., & Tian, Q. (2014). Coupled binary embedding for large-scale image retrieval. *IEEE Transactions on Image Processing*, *23*(8), 3368–3380.
Zheng, L., Wang, S., Tian, L., He, F., Liu, Z., & Tian, Q. (2015). Query-adaptive late fusion for image search and person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1741–1750).
Zheng, L., Yang, Y., & Tian, Q. (2018). SIFT meets CNN: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(5), 1224–1244.
Zhu, L., Jin, H., Zheng, R., & Feng, X. (2014). Weighting scheme for image retrieval based on bag-of-visual-words. *IET Image Processing*, *8*(9), 509–518.
Zhu, X., Luo, Y., Liu, A., Xiong, N. N., Dong, M., & Zhang, S. (2021). A deep reinforcement learning-based resource management game in vehicular edge computing. *IEEE Transactions on Intelligent Transportation Systems*.