Boosting Diversity in Visual Search with Pareto Non-Dominated Re-Ranking

SI-CHAO LEI, South China University of Technology, China
YUE-JIAO GONG, South China University of Technology, China and and University of Electronic Science and Technology of China, China
XIAO-LIN XIAO, South China Normal University, China
YI-CONG ZHOU, University of Macau, China
JUN ZHANG, Hanyang University ERICA, South Korea, and Nankai University, China

The field of visual search has gained significant attention recently, particularly in the context of web search engines and e-commerce product search platforms. However, the abundance of web images presents a challenge for modern image retrieval systems, as they need to find both relevant and diverse images that maximize users' satisfaction. In response to this challenge, we propose a non-dominated visual diversity re-ranking (NDVDR) method based on the concept of Pareto optimality. To begin with, we employ a fast binary hashing method as a coarse-grained retrieval procedure. This allows us to efficiently obtain a subset of candidate images for subsequent re-ranking. Fed with this initial retrieved image results, the NDVDR performs a fine-grained re-ranking procedure for boosting both relevance and visual diversity among the top-ranked images. Recognizing the inherent conflict nature between the objectives of relevance and diversity, the re-ranking procedure is simulated as the analytical stage of a multi-criteria decision-making process, seeking the optimal tradeoff between the two conflicting objectives within the initial retrieved images. In particular, a non-dominated sorting mechanism is devised that produces Pareto non-dominated hierarchies among images based on the Pareto dominance relation. Additionally, two novel measures are introduced for the effective characterization of the relevance and diversity scores among different images. We conduct experiments on three popular real-world image datasets and compare our re-ranking method with several state-of-the-art image search re-ranking methods. The experimental results validate that our re-ranking approach guarantees retrieval accuracy while simultaneously boosting diversity among the top-ranked images.

CCS Concepts: • **Information systems** → **Information retrieval diversity**; *Top-k retrieval in databases;* • **Computing methodologies** → *Visual content-based indexing and retrieval*;

Additional Key Words and Phrases: Visual image search, re-ranking, Pareto optimality, image diversity

This work was supported in part by the National Natural Science Foundation of China under grant 62276100, in part by the Guangdong Natural Science Funds for Distinguished Young Scholars under grant 2022B1515020049, in part by the Guangdong Regional Joint Fund for Basic and Applied Research under grant 2021B1515120078, in part by the TCL Young Scholars Program, and in part by the National Research Foundation of Korea under grant NRF2022H1D3A2A01093478.
Authors' addresses: S.-C. Lei and Y.-J. Gong (Corresponding author), School of Computer Science and Engineering, South China University of Technology, Guangzhou 510000, China; e-mails: cssclei@outlook.com, gongyuejiao@gmail.com; X.-L. Xiao, School of Computer Science, South China Normal University, Guangzhou 51000, China; e-mail: shellyxiaolin@gmail.com; Y.-C. Zhou, Department of Computer and Information Science, University of Macau, Macau 999078, China; e-mail: yicongzhou@um.edu.mo; J. Zhang (Corresponding author), Hanyang University ERICA, Ansan 15588, South Korea, and Nankai University, Tianjin 30071, China; e-mail: junzhang@ieee.org.

1551-6857/2023/11-ART79 $15.00
https://doi.org/10.1145/3625296

ACM Reference format:
Si-chao Lei, Yue-Jiao Gong, Xiao-Lin Xiao, Yi-Cong Zhou, and Jun Zhang. 2023. Boosting Diversity in Visual Search with Pareto Non-Dominated Re-Ranking. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 3, Article 79 (November 2023), 23 pages.
https://doi.org/10.1145/3625296

1 INTRODUCTION

As of recently, most popular search engines or e-commerce platforms, such as Google, Bing, eBay, and Alibaba, provide the function of *visual search*. Typically, visual search allows users to search the web using an image instead of text as a query to retrieve similar images, products, pages, and so on. However, the proliferation of web images has challenged the effectiveness and efficiency of visual image retrieval. The initial search results usually contain irrelevant or unexpected images. This issue has given rise to the research field of *image re-ranking* to fine-tune the initial search results. After re-ranking, the retrieval accuracy of image search can be substantially improved [9, 28, 33].

Most image re-ranking approaches are relevance-based, aiming to re-rank initial image lists to maximize the relevance of the top images to the query [7, 50]. However, in practical applications, such relevance-based re-ranking methods often fail to fully satisfy the intentions of users [52]. This is primarily due to the vast number of images being generated and shared on the web every day, leading to a significant presence of near-duplicate images. Figure 1 illustrates an example of the top 10 images provided by Google visual search, where the leftmost image is the query instance. Clearly, most of the results are nearly identical by the default visual relevance-only search. This issue, commonly known as "the lack-of-diversity problem", is frequently encountered in real-world visual image search systems [24, 56]. With the exponential growth of web images, relying solely on visual relevance for image re-ranking tends to yield redundant image results. This redundancy places unexpected burdens on users, wasting their time spent on manually viewing and selecting images, thereby diminishing their overall experience. Conversely, enhancing the diversity of top images in the retrieved results can significantly aid users in identifying preferred images. The importance of diversity is also paramount in visual image search-based product recommendations on e-commerce platforms [56]. When the majority of retrieved products are similar, users can easily become overwhelmed and quickly lose their shopping interest.

Therefore, modern image visual search systems require to return both accurate and diverse image results to provide a more comprehensive response to user queries [36, 42]. Many researchers have dedicated their efforts to making the top-ranked results diversified [4, 5, 16, 20, 34]. Generally, the ultimate goal of image search re-ranking is to refine the retrieved results toward enhancing diversity that covers as many different visual aspects as possible while ensuring retrieval accuracy simultaneously. Most existing image re-ranking methods primarily focus on text-based diversification [6], such as query explanation [31] and textual mining [6, 39]. As a complement, visual diverse search conducts visual content-based diversification that is typically concentrating on two aspects, namely image representations and search result diversification strategies. On one hand, various visual descriptors are employed to mine visual information from multiple perspectives [2, 8, 19, 35]. On the other hand, various diversification strategies are used to rank or re-rank image search results, among which representative techniques include clustering-based methods [1, 43, 47], optimization-based methods [18, 29, 45], and graph-based methods [9, 28].

As mentioned previously, the image search re-ranking task is a typical multi-criteria optimization problem. An image retrieval system may desire to explore all possible optimal solutions of relevance and diversity objectives simultaneously, known as a tradeoff analysis. In the pursuit of

(a) Results from Google (default) (b) Results after re-ranking

Fig. 1. An example of top 10 results retrieved from Google image search.

finding the optimal tradeoff between conflicting objectives, the concept of *Pareto optimality* often comes into play. Pareto optimality is a well-developed principle in the field of multi-objective optimization [12]. The concept of Pareto optimality is explained based on the Pareto dominance relation to identifying solutions that are superior to others regarding the conflicting objectives for decision making. For a multi-objective problem with m objectives to be simultaneously optimized, a solution x is said to dominate another solution y if x is better than y for at least one objective and is not worse for any other objectives. In other words, x dominates y if and only if

$$x \succ y, \quad \text{if} \begin{cases} f_i(x) > f_i(y), \exists i \\ f_j(x) \geq f_j(y), \forall j \neq i \end{cases} \tag{1}$$

where f_i $(i = 1, \ldots, m)$ denotes the objective function, the symbols '$>$' and '$\geq$' measure the betterness of two solutions, and the symbol '$\succ$' denotes the domination relation. The Pareto-optimal solutions provide the optimal possible tradeoffs between the objectives, and any solution that is not Pareto-optimal is considered to be dominated by one or more Pareto-optimal solutions. In the context of visual image search, it is crucial to ensure that each top-ranked result is not dominated by the subsequent image results in terms of both relevance and diversity. However, to the best of our knowledge, none of the existing visual image search methods explicitly achieve Pareto optimality.

This article proposes a novel approach called the Non-Dominated Visual Diversity Re-ranking (NDVDR) method for relevant and diverse visual image search. Starting with a candidate subset of images retrieved using binary hashing, which is known for its efficiency in large-scale visual image retrieval, the NDVDR method performs visual re-ranking with non-dominated sorting based on Pareto dominance relationships among images to refine the initial results and enhance both relevance and diversity among the top-ranked search results. The main contributions of this work can be summarized as follows:

- We introduce the concept of Pareto optimality into the field of visual image search re-ranking, providing a principled approach for multi-criteria decision making when returning image results.
- We define optimality conditions on image pairs to explore the Pareto domination relation and introduce relevance and diversity scoring functions to quantify the relations among different image pairs. These mechanisms enable an effective evaluation of the tradeoff between relevance and diversity in the context of image retrieval.
- The NDVDR we propose provides a generic framework that can be applied to any initial retrieval results obtained from existing methods. This flexibility allows for easy extension and integration of our re-ranking mechanism into different retrieval systems, making it applicable and adaptable to various scenarios.
- As preliminary results, we demonstrate the significant effectiveness of our proposed NDVDR method through comprehensive experimental evaluations on three real-world image

datasets. Furthermore, we integrate our re-ranking mechanism into other image retrieval systems and different visual image representations to validate its generalizability and extensibility across different retrieval settings.

The rest of the article is outlined as follows. Section 2 gives an comprehensive overview of image search re-ranking. Section 3 elaborates the proposed NDVDR framework and the re-ranking algorithm. Section 4 discusses the experimental results in details, and finally, some conclusions are made in Section 5.

2 RELATED WORK

While the majority of image search systems prioritize relevance, modern image retrieval tasks face the challenge of providing both relevant and diverse results [25, 26]. To address this challenge, image search re-ranking methods have been extensively developed to enhance both the relevance and diversity among the top-ranked image results. Originally stemming from text-based retrieval systems, image search re-ranking emerged as a solution to tackle ambiguity in textual queries [20, 39]. Additionally, with the popularity of exploiting visual information in image content description, visual image search re-ranking has also garnered attention, aiming to provide a more comprehensive feedback of image visual content [2, 35]. Our work focuses on visual image search re-ranking. A typical image search re-ranking method involves a two-step process: first, an initial ranked list of candidate images is retrieved using various image retrieval methods based on their relevance to the query; second, the initial image list is refined to include a subset of images that exhibit maximum diversity. Existing methods in image re-ranking primarily center around image representations [19, 35], image re-ranking strategies, and various combinations thereof. A plethora of image re-ranking methods exist, ranging from image clustering [16, 20], objective optimization [29, 37], and graph-based re-ranking [9, 28] to hybrid approaches [44, 53]. This section provides a comprehensive review of the aforementioned methods.

2.1 Image Representation

Various types of image features can be extracted from different information sources, including text, credibility, visual, or hybrid sources. For text sources, images shared on social media are often annotated with free tags by users to describe their content, and these textual descriptions can vary depending on different user habits [20, 34, 39]. Some approaches perform interactive learning based on human-machine or human-in-the-loop efforts to achieve high accuracy [3, 4, 40, 54] by employing user feedback regarding relevance and diversity. In terms of visual representation, various image features have been developed. Boato et al. [2] highlighted that image visual representations contain much more information than their textual descriptors. They proposed to exploit visual saliency for object-category level diversification that combines similarity measures of image background and foreground. Leuken et al. [48] introduced a dynamic weighting function for different visual features to capture image information at different scales. Deep visual features were also widely used in visual image search [35, 49]. Milbich et al. [35] employed deep metric learning that aggregates diverse visual features to capture a diverse range of image representations for better learning of visual similarity. Inspired by recent advancements in vision transformers [13, 17, 46] across various computer vision tasks, Chen et al. [13] proposed learning multi-scale feature representations using vision transformer models to produce stronger image features. Tang et al. [46] developed an augmented shortcut scheme in a vision transformer to improve the diversity of image features. Conventional hybrid approaches combine textual and visual information [10, 16, 22]. For instance, Goynuk and Altingovde [22] fused different methods of relevance and diversity evaluation based on various textual and visual features to diversify image search results.

2.2 Clustering-Based Image Re-Ranking

In this category, clustering methods are employed to group images into clusters so that similar images are placed in the same cluster, whereas dissimilar ones are separated into different clusters. Afterward, image re-ranking would iteratively select representative images from each cluster to promote the diversity of the top-ranked image results. Various clustering techniques, such as hierarchical clustering [1, 11, 16, 20, 38, 43, 49] and density-based spatial clustering (DBSCAN) [43], have been widely used for image re-ranking tasks. The determination of the number of clusters significantly impacts the quality of the clustering results. The hierarchical clustering methods merge or split image sets based on their textual or visual distance to form clusters. Vandersmissen et al. [49] proposed an adaptive hierarchical clustering method that tries to identify the optimal number of clusters by observing reference points from the plot of the number of clusters versus the inter-cluster distance. Castellanos et al. [11] utilized hierarchical agglomerative clustering to group images based on the latent topics of their textual content. Peng et al. [38] improved the relevance and visual diversity of results by re-ranking images using social metadata through hierarchical clustering. In the work of Seddati et al. [43], DBSCAN was utilized to perform clustering on the weighted textual features and deep visual features, which not only adaptively adjust the number of clusters but also resist noise in the initial results. Dang-Nguyen et al. [16] employed the balanced iterative reducing and clustering method to group similar images based on their textual and visual descriptions, then images were extracted from different clusters based on a measure of the user's credibility. Benavent et al. [1] employed both classical *k*-means and hierarchical agglomerative clustering based on the latent textual information to generate a relevant and diverse image list. More recently, Figuerêdo and Calumby [20] utilized clustering to find subtopics of textual queries to diversify image results.

2.3 Optimization-Based Image Re-Ranking

The image re-ranking task is considered as an optimization problem that takes into account relevance and diversity. The optimization-based methods address this task through utilizing different evaluation metrics and integrating relevance and diversity in their optimization frameworks [18, 29, 37, 41, 44, 45, 51]. Wang et al. [51] introduced a diverse relevance ranking algorithm that aimed to maximize a novel metric called Average Diversity Precision (ADP). Both relevance and diversity are considered in their optimization framework by incorporating the similarity among image visual features and the associated tags. Escalante and Morales-Reyes [18] employed a multi-objective evolutionary algorithm called NSGA-II to maximize diversity in consecutive positions of an image list on a tourist destination dataset. Spyromitros-Xioufis et al. [45] formulated the re-ranking process as a function optimization problem that integrates relevance and diversity into optimization, and trained ensemble classifiers to calculate relevance scores of images to the query. Rao et al. [41] combined relevance and diversity into hash functions and proposed to optimize hash objective functions to retrieve both relevant and diverse results. Karako and Manggala [29] considered fairness as an aspect of diversity and introduced a fairness-based MMR method, which tries to enhance gender fairness among image results on a stock photo dataset. Ouyang et al. [37] tried to optimize the re-ranking size of images for different queries and re-ranked the images by CNN model optimization to maximize the relevance of top-ranked images.

2.4 Graph-Based Image Re-Ranking

Graph-based methods focus on constructing graphs that represent the relationships among images and leveraging graph information to re-rank the image results [8, 9, 27, 28, 39, 53, 55]. Ji et al. [27] constructed a graph based on four visual features and then applied absorbing random walks on

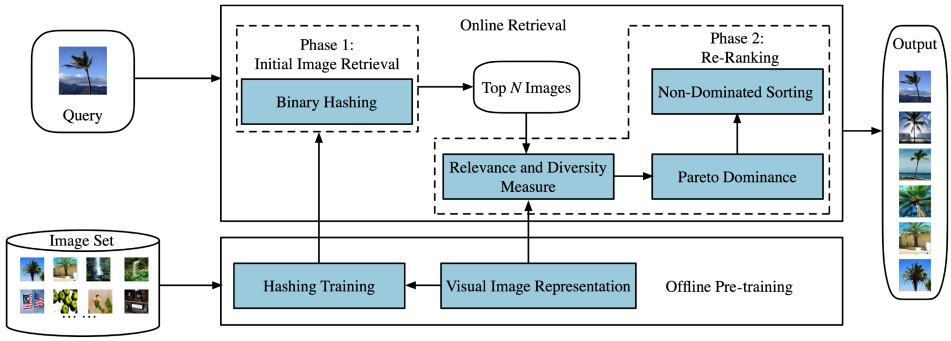

Fig. 2. The overall framework of NDVDR.

the graph to enhance both the relevance and diversity of the image results. Qian et al. [39] constructed a tag graph based on the similarity between image tags, aiming to maximize the topic coverage of the image results. Yan et al. [53] proposed a graph clustering model that first clusters images into topics and then performs Markov random walks under the constraints imposed by image cluster information. Zeng et al. [55] introduced a variational multiple instance graph that focuses on capturing diverse query semantics by learning a continuous semantic space. Additionally, the hypergraph has also been widely adopted for image search re-ranking, with most existing hypergraph-based methods focusing on refining the relevance of top-ranked images. Bouhlel et al. [8] introduced diversity-induced visual hypergraph ranking with absorbing nodes to achieve image diversification. Each ranked image node on the hypergraph is set as an absorbing node with zero weight, which helps reduce the likelihood of duplicate images receiving high ranking scores, as the remaining unranked duplicate images share the same edge with the absorbing node on the hypergraph. Jing et al. [28] built a hypergraph based on relevance and pseudo-relevance information, aiming to capture the intrinsic geometric structure of the data distribution. Bouhlel et al. [9] utilized ridge regression for hypergraph construction, alleviating the computational burden associated with hypergraph construction. Shen et al. [44] proposed a hybrid approach to optimize the similarity measure between images by few-shot learning based on the similarity graph of images.

3 METHODOLOGY

This article presents a visual re-ranking method called NDVDR, with the primary goal of refining the top-ranked images in a way that they possess high relevance to the query while also showcasing a diverse range of visual characteristics. NDVDR begins by employing fast binary hashing retrieval to efficiently search for a candidate set of images that match the query image. This initial retrieval step serves as a starting point for further re-ranking refinement. Using the initial retrieval results as inputs, NDVDR subsequently performs a non-dominated re-ranking procedure to derive a re-ranked list of images as the final outputs, which exhibit an optimal balance between being visually relevant and diverse to the query.

The overall framework of the proposed image search re-ranking method is depicted in Figure 2, which consists of online retrieval and offline pretraining.

Online Retrieval. The online retrieval comprises two successive phases of *Initial Image Retrieval* and *Re-Ranking*. In the initial image retrieval stage, the query image is processed using a fast binary hashing method. This process generates an initial ranked list of retrieval results. From the initial ranked list, only the top N images are selected for the subsequent re-ranking phase. During the re-ranking stage, the Pareto dominance relation is defined among

the selected N images, which considers both relevance and diversity factors for each image pair. By non-dominated sorting, thereafter the overall quality of the re-ranked image list is enhanced.

Offline Pre-Training. The offline pre-training consists of *Hashing Training* and *Visual Image Representation*. In visual image representation, different visual features are extracted from raw images to represent their visual content. Two types of visual features are utilized: the generalized search trees (GIST) feature and a CNN-based visual feature. Following Gou et al. [21], the GIST feature captures global information of the images and is specifically used for hashing training. The goal of hash training is to transform images into compact binary hash codes that can efficiently represent the images while preserving their visual characteristics. The CNN-based feature captures high-level visual information and is utilized during the re-ranking stage for the evaluation of relevance and diversity among images. We will also investigate the use of other advanced image representations in Section 4 of this article.

3.1 Initial Image Retrieval

Binary hashing has gained widespread popularity in current image retrieval tasks due to its low storage cost and high search efficiency. The fundamental principle of image hashing involves training a set of hash functions that project high-dimensional feature representations of images into low-dimensional binary Hamming space. In this work, we specifically opt HCSDH [21] as our chosen binary hashing method for its efficient search capabilities and relatively high precision in large-scale image retrieval tasks. An overarching advantage of HCSDH lies in its remarkable bit-level scalability, which means that the retrieval performance and training time are not dependent on the selected length of the binary hash codes.

Hashing training computes the projection matrix P that projects images to compact hash codes as

$$P = (XX^T)^{-1}XB^T, \tag{2}$$

where X is the feature representations of images and B is a binary matrix. The pre-processing step converts original GIST features into feature vectors using the Gaussian kernel transformation as

$$X = e^{\left(-\frac{\|X_{\text{GIST}} - A\|^2}{\sigma}\right)}, \tag{3}$$

where X_{GIST} consists of GIST descriptions of images and A represents a set of anchor vectors randomly sampled from X_{GIST}. Suppose the pre-defined hash code length is L, the size of the training set is M, and the number of classes in the training dataset is C; the HCSDH first computes a Hadamard matrix $H \in \{-1, 1\}^{L \times L}$. A set of matrix $[b_1^{'}, \ldots, b_i^{'}, \ldots, b_C^{'}] \in \{-1, 1\}^{L \times C}$ is derived from H based on the Hadamard matrix, where each element $b_i^{'}$ corresponds to the ith column of H. Given the available label information of training images $\{y_i\}_{i=1}^{M}$, the binary matrix B is constructed from $[b_1^{'}, \ldots, b_C^{'}]$, where $b_i = b_{y_i}^{'}$. Finally, according to Equation (2), the projection matrix can be calculated to map the images in GIST feature space into binary Hamming space, and thereby each image is then represented by a compact binary hash code.

Binary Hashing retrieval calculates the binary Hamming distance between the query image and all the images in the dataset, which is computationally efficient. The result is a ranked list of images sorted based on their Hamming distances to the query, with closer images having smaller Hamming distances. Note that the retrieval ranks all images in the dataset. However, to ensure the precision of the initially retrieved result, only the top N images are selected to participate in the subsequent re-ranking stage. This approach represents a coarse-grained retrieval procedure, helping to narrow down the scope of the re-ranking process, making NDVDR more efficient and effective in producing high-quality image re-ranking results.

3.2 Determination of the Re-Ranking Data Size

The number of preserved images, denoted as N, plays a crucial role in connecting the initial retrieval phase and the subsequent re-ranking phase. Selecting an appropriate value for N is significant to balance the effectiveness and efficiency of the re-ranking process. If the candidate image set is too large, the likelihood of including noise images increases, and it also burdens the efficiency of re-ranking. Conversely, if N is too small, it risks the probability of diversity enhancement. In this case, some diverse yet relevant images may not be included in the preserved image set and would have no chance to participate in re-ranking.

Hence, we adapt the value of N as

$$N = count(hammD(q, x) \leq R), \tag{4}$$

where $hammD(q, x)$ represents the Hamming distance between the query q and an image x in the dataset. The function $count()$ represents the number of images located in the vicinity of the query image within the pre-defined radius R. One may empirically set R as a constant. In contrast, we take a step forward to adapt R according to the hash code length L. Intuitively, a longer L can better preserve the similarity relationships among images, resulting in higher retrieval precision but a lower recall rate. Conversely, a shorter hash code increases the recall rate but also increases the probability of showing irrelevant images. Therefore, to stabilize the performance of the initial retrieval, we no longer fix the radius R but adapt it proportionally to the hash code length L, as follows:

$$R = \omega \times L, \tag{5}$$

where ω is a scaling factor. With a pre-defined Hamming threshold R, a short L results in a large N, whereas a long L results in a small N.

3.3 Non-Dominated Visual Diversity Re-Ranking

For modern image retrieval systems, it is crucial to consider both the relevance and diversity levels of the retrieved image list when given a query image. Real-world image databases often contain a significant number of redundant images. If the ranking criteria are solely based on relevance, duplicates or visually similar images tend to be ranked closely together. However, solely considering diversity can lead to an increase in irrelevant images, which reduces retrieval precision. Neither relevance-only nor diversity-only approaches fully satisfy user expectations. Therefore, the image re-ranking task is a typical multi-objective optimization problem where the conflicting objectives of relevance and diversity must be addressed simultaneously. The optimal solutions that strike a tradeoff between conflicting objectives are typically defined in terms of Pareto optimality [12]. Pareto optimality represents the concept of achieving the best possible outcomes where no single objective can be improved without compromising any others. In this section, we define the Pareto relation of images as follows.

Definition 1 (Pareto Dominance). An image x_i is said to Pareto dominate another image x_j, denoted as $x_i \succ x_j$, if the following two conditions are satisfied:

$$\begin{aligned} &\text{Condition 1: } f_{rel}(x_i) \geq f_{rel}(x_j) \text{ and } f_{div}(x_i) \geq f_{div}(x_j), \\ &\text{Condition 2: } f_{rel}(x_i) > f_{rel}(x_j) \text{ or } f_{div}(x_i) > f_{div}(x_j), \end{aligned} \tag{6}$$

where f_{rel} and f_{div} are the objective evaluations of the relevance and diversity terms for the images, respectively. The symbols '$>$' and '$\geq$' measure the betterness between the two images, and the symbol '$\succ$' denotes the domination relation between the two images.

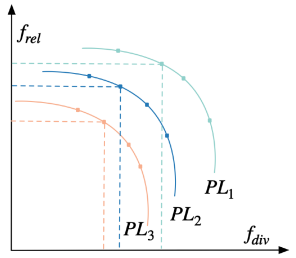

Fig. 3. The ideal Pareto hierarchy among images regarding relevance and diversity.

Definition 2 (Pareto Layer). The Pareto layer (*PL*) consists of a subset of images that are non-dominated to each other:

$$\forall x_i \in PL : \nexists x_j \in PL, x_j \succ x_i. \tag{7}$$

Definition 3 (Priority of Pareto Layers). A Pareto layer PL_a has priority over PL_b, denoted as $PL_a \succ_{\mathrm{L}} PL_b$, if

$$\forall x_i \in PL_a, \forall x_j \in PL_b : x_i \succ x_j, \tag{8}$$

where the symbol $\succ_{\mathrm{L}}$ denotes the priority relation among Pareto layers.

Definition 4 (Pareto Optimality of Image Re-Ranking). A re-ranked image list is said to be Pareto optimal if it can be divided into a few Pareto layers $\{PL_1, PL_2, \ldots, PL_u\}$ that satisfy

$$\forall a \in (1, \ldots, u) : PL_a \succ_{\mathrm{L}} PL_{a+1}, \tag{9}$$

where u is the total number of Pareto layers. Figure 3 illustrates the priority of Pareto layers based on the two conflicting objectives f_{rel} and f_{div}. The set of images in the PL_1 is also called the *Pareto optimal set* since images in the PL_1 dominate all images from the other Pareto levels.

Based on the preceding definitions, we propose a non-dominated visual diversity re-ranking approach for the image search re-ranking. In Section 3.1, the initial image retrieval produces a candidate set of N images serving as the input for NDVDR. Each image in this candidate set is evaluated in terms of a relevance score and a diversity score, which will be elaborated in Section 3.4. According to Definition 1, images are then compared based on the superiority of the two conflicting objectives. In some cases, two images may not dominate each other based on the objectives, implying that they are equally important. We define the Pareto layer as a subset of images containing such non-dominated images as specified in Definition 2.

To establish the Pareto relationship among all images, NDVDR performs non-dominated ranking in an iterative manner. Each iteration reads out one Pareto layer from the best to the worst according to Definition 3. Specifically, in the first iteration, the Pareto optimal set of images that dominate all others is identified as PL_1, which is then removed from the image list. In the subsequent iterations, the remaining images are further compared with each other to identify the remaining non-dominated images, thereby generating a new Pareto layer PL_2, and so forth. This process continues until all images have been assigned with their respective Pareto hierarchy. In this way, we obtain a set of Pareto layers $\{PL_1, PL_2, \ldots, PL_u\}$ for the images, which satisfy the Pareto optimality given in Definition 4. Afterward, the retrieval system returns the Pareto layers of images to the users one by one.

The remaining question is how to sort the specific images within each Pareto layer. Note that from the perspective of the Pareto dominance relation, images belonging to the same Pareto layer are equally important to each other. This means that for a decision-maker, any image in the same

layer can be chosen as the current optimal solution when considering all objectives. For NDVDR, within each Pareto layer, the first-ranked image is selected based on the evaluated relevance score. The images with higher relevance scores are ranked higher within the layer, ensuring that the most relevant image to the query is always given the highest priority, whereas the remaining images are directly ranked by an arbitrary sequence. Eventually, NDVDR produces a non-dominated image sequence as the final retrieval results. This image sequence represents a set of images that satisfy Pareto optimality. The main procedures of NDVDR are summarized in Algorithm 1.

3.4 Relevance and Diversity Measures

To obtain a better assessment of relevance and diversity scores for re-ranking, we utilize more advanced CNN-based visual representations instead of the binary hashing representation. For this purpose, we employ a pre-trained CNN ResNet-50[1] as the image feature extractor. We consider the results from the last activation layer, which is the layer before the fully connected layer, resulting in a 2,048-dimensional deep visual feature. The visual similarity $s(x_i, x_j)$ between two images of x_i and x_j is evaluated using the Gaussian kernel as follows:

$$s(x_i, x_j) = e^{\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right)}, \tag{10}$$

where σ is set to be the median value of all the distances between image pairs.

Relevance Measure. The relevance measure combines the deep visual feature-based similarity and the initial hash ranking information from Section 3.1. To incorporate the initial ranking information, we define an indicator $r(x_i)$ as follows:

$$r(x_i) = \frac{2 \times e^{-(\tau(x_i)-1)/z}}{1 + e^{-(\tau(x_i)-1)/z}}, \tag{11}$$

where x_i represents an image, $\tau(x_i)$ is a positional function indicating the ranking of x_i in the image list, and z is a constant parameter. Using this indicator, the top images in the initial ranking list are assigned high relevance scores, and the value decreases as the position of images goes backward. Consequently, the relevance score of an image x_i is evaluated as follows:

$$f_{\text{rel}}(x_i) = r(x_i) \times s(x_i, x_{top}), \tag{12}$$

where $top = \arg\min_j(\tau(x_j))$ represents the image ranked the first in the initial ranking list and $s(x_i, x_{top})$ represents the similarity measure between image x_i and the first images in the ranking list, and f_{rel} denotes the objective function of relevance.

The similarity matrix $S = s(x_i, x_j)$ is derived from the pre-computed deep visual similarity during offline phase. It is important to note that we directly use the distance information in S to obtain relevance scores instead of calculating the distance between the query image and the N images, which would require additional time when each new query image arrives. This mechanism assumes that the top-ranked image in the retrieval system is highly similar or identical to the query, which is crucial for an effective image retrieval system.

Diversity Measure. Most existing diversity evaluation methods only consider the similarity between the current evaluated image and the images re-ranked before it, which is not a comprehensive evaluation mechanism. For instance, it is unreasonable for the diversity evaluation of the second image in the ranking list to consider only the first image before it. For a fixed set of N images, evaluating the diversity of an image x_i should consider its similarity to all the other images

[1] https://www.vlfeat.org/matconvnet/pretrained/

in the list. Therefore, we define the diversity of image x_i as follows:

$$f_{\text{div}}(x_i) = (1-\alpha) \min_{1 \le j < \tau(x_i)} s(x_i, x_j) + \alpha \min_{\tau(x_i) < k \le N} s(x_i, x_k), \tag{13}$$

where $\alpha \in [0, 1]$ is a constant parameter and f_{div} is the objective function of diversity. The first term measures the minimal difference between x_i and the images appeared before x_i, which denotes the posterior diversity gain of x_i given all the former images. The second term calculates the minimal difference between x_i and the images appearing after x_i, which measures the prior diversity gain of x_i to the rest of the unranked images in the list. The coefficient α controls the importance of the two terms.

ALGORITHM 1: Non-Dominated Visual Diversity Re-Ranking (NDVDR)

Input: X, image dataset; q, query image; L, hash code length; z, ω, α, control parameters
Output: a re-ranking list of images $X_2 = \{PL_1, PL_2, \ldots, PL_u\}$

```
Conduct hashing retrieval;
Obtain a subset of images X_1 = {x_1, ..., x_N} by the Hamming distance, where N is defined according to Equation (4);
Evaluate each image using the objective functions: F(x_i) = (f_rel(x_i), f_div(x_i));
Set u = 1, X_res = X_1;
repeat
    set PL_u = ∅
    for id = 1 to |X_res|
        if ∄(x ∈ X_res) ≻ (x_id ∈ X_res)
            PL_u = PL_u ∪ x_id
        end if
    end for
    sort PL_u by the Hamming distance;
    remove non-dominated layer: X_res = X_res \ PL_u;
    set u = u + 1;
until all the images are divided into multiple PLs: X_res = ∅;
```

3.5 Complexity Analysis

From the preceding description, the proposed algorithm is composed of two phases: initial image retrieval and NDVDR. The computational bottleneck in the first phase is shown in Equations (2) and (3). The kernel transformation requires $O(|X|^2 N)$ and the inverse of XX^T costs $O(|X|^3)$ for Cholesky factorization. Since $N < |X|$, the computational cost of hashing can be estimated as $O(|X|^3)$. For the NDVDR of the top N images, each image is compared with the rest $(N-1)$ images in terms of the two objectives. By utilizing a fast non-dominated sorting scheme, the computational complexity of re-ranking can be $O(N^2)$.

4 EXPERIMENTS

4.1 Experimental Setting

4.1.1 Dataset. The experiments are conducted on three widely used image datasets: Cifar-100 [30], Caltech [23], and NUSWIDE [15]. For each dataset, a few categories involving too few images are not considered. Moreover, images from similar categories are merged to form a new and larger super class. For example, images from the 'helicopter' and 'airplane' classes are both considered as the 'plane' class, and images belonging to the 'tulip,' 'rose,' and 'flowers' categories are merged as the 'flower' class. The categories extracted from the datasets are summarized in Table 1.

Moreover, to simulate a more realistic web data environment, some near-duplicate or transformed images were manually added to the datasets. The image transformation included resizing,

Table 1. Details of Image Datasets

Dataset	Instances of Classes
Cifar-100	mammal, flower, food container, fruit and vegetable, people, reptile, tree, vehicle, insect, outdoor scene
Caltech	American flag, backpack, hat, truck, frog, goat, goldfish, gorilla, horse, balloon bird, ibis, motorbike, palm-tree, sunflower, swan, bike, waterfall, zebra, airplane
NUSWIDE	animal, flower, rainbow, person, waterfall, sky, buildings, ship, plane, flags, penguin, giraffe, bike, fish, elephant, aqueduct, zebra, car, tree, street, eagle, tiger, wedding, valley

Fig. 4. Illustrations of different image transformation.

cropping, flipping, and other techniques. Figure 4 displays some examples, where the leftmost image is the original image before the transformation, and the rest are its transformed versions with different cropping sizes and orientations, flipping directions, and enhancements.

Each image dataset is randomly split into three subsets: the query set, retrieval set, and labeled training set. The labeled training samples are used for the hashing training in Section 3.1. The details of each dataset are summarized as follows:

- *Cifar-100* [30]: This dataset consists of 60,000 images with 100 classes that are further grouped into 10 superclasses. The query set consists of 600 images randomly sampled from the dataset, with 50 images per category. The labeled training set contains 24,000 images, with 2,000 images per category randomly sampled from the database, and the remaining images are used as the retrieval database.
- *Caltech* [23]: We extract 20 categories with a total of 16,000 images. Approximately 30% of near-duplicate images are randomly added to the dataset, resulting in a final dataset of 20,800 images. Then, 50 images are randomly sampled from each category as queries, resulting in a query set containing 1,000 images. A total of 10,000 images with 500 images per category are further randomly sampled as the training set. The remaining 9,800 images are used as the retrieval database.
- *NUSWIDE* [15]: The extracted dataset contains 24 categories with a total of 123,578 images. After adding the duplicates, the total number of images becomes 148,294. For the experiment, 1,200 images with 50 images per category are randomly selected as the query set. A total of 12,000 images with 500 images per category are sampled from the database as the labeled training set, and the remaining images serve as the retrieval database.

4.1.2 Performance Evaluation. To examine the performance of different visual image re-ranking methods with respect to diversity and relevance, the following four metrics are used.

AP Metric. The Average Precision (AP) evaluates the relevance of images in a ranked image list to the query. It is computed as

$$AP@K(\tau) = \frac{1}{Q}\sum_{i=1}^{K} y(\tau(i)) \times \left(\frac{\sum_{j=1}^{i} y(\tau(j))}{i}\right), \tag{14}$$

where Q is the number of truly relevant images out of the total K images, $\tau(i)$ indicates the images that are located in the ith position of the ranking list, and y stores the binary label information of images respected to the query.

CR Metric. The Cluster Recall (CR) measures the diversity degree of a ranked image list by calculating the number of different subtopics among the top-ranked images. For instance, an image super class of the 'tree' may include different subtopics such as 'maple,' 'oak,' 'palm,' 'pine,' and 'willow.' A higher CR value indicates the image list covers a more diverse range of relevant topics or concepts related to the query. The CR metric is defined as

$$CR@K(\tau) = \frac{N_c}{N_{ct}}, \tag{15}$$

where N_c counts the number of topics founded in the top-ranked K result and N_{ct} counts the total number of clusters for the query.

F1 Metric. This metric computes the harmonic mean between AP and CR to evaluate the overall quality of a re-ranked image list, which is defined as

$$F1@K(\tau) = 2 \cdot \frac{AP@K \times CR@K}{AP@K + CR@K}. \tag{16}$$

ADP Metric. ADP [51] integrates both the relevance and visual diversity of an image and all the other images ranked before it in an image list. The ADP metric is computed as

$$ADP@K(\tau) = \frac{1}{Q} \sum_{i=1}^{K} y(\tau(i)) \times DIV(\tau(i)) \times \left(\frac{\sum_{j=1}^{i} y(\tau(j)) \times DIV(\tau(j))}{i} \right), \tag{17}$$

where $DIV(\tau(i)) = \min_{1 \le t < i}(1 - s(\tau(t), \tau(i)))$. Note that the ADP metric is sensitive to image representation as it requires computing the distance between any two images.

4.1.3 Comparative Methods.

- *Optimization-based methods*: DRR [51], NSGA [18], and MMR [29]. DRR is a greedy re-ranking scheme based on image similarities to boost the diversity of top-ranked images. NSGA treats the re-ranking task as an optimization problem, seeking to minimize the difference between the new ranking and the initial ranking while maximizing the diversity of images in consecutive positions. MMR tries to balance the relevance and diversity during the selection of images and maximizes diversity gain each time a new image is added to the current image list.
- *Clustering-based methods*: HC [38] and KM-ICO [20]. HC uses a hierarchical clustering algorithm with complete linkage to form image clusters. It then iteratively selects images from each cluster sorted based on their relevance. KM-ICO employs a k-Medoids algorithm to generate image clusters based on intrinsic clustering quality optimization, allowing for fine-tuning of the number of image clusters.
- *Graph-based methods*: HG [8] and CHG [9]. HG constructs a visual hypergraph to capture high-order relationships among images. Diversity is enhanced by integrating the concept of absorbing nodes into the re-ranking process. CHG incorporates collaborative representation into the construction of the hypergraph to capture the real relationships among visual images.

In the experiments, query-by-type operations are conducted based on different categories of images. The re-ranked image lists for a query generated by different methods are evaluated based on the four metrics AP@K, ADP@K, CR@K, and F1@K with $K = 20, 40, 60, 80$, and 100. Notice

Table 2. Performance of Different Image Re-Ranking Methods on the Cifar-100 Dataset

	AP					CR					F1				
	20	40	60	80	100	20	40	60	80	100	20	40	60	80	100
DRR	.499	.485	.463	.469	.453	.406	.425	.439	.444	.450	.448	.453	.451	.456	.452
NSGA	.442	.455	.427	.454	.475	.399	.408	.429	.437	.441	.419	.430	.428	.445	.457
MMR	.453	.434	.446	.449	.453	.412	.437	.433	.443	.449	.432	.436	.440	.446	.451
HC	.458	.441	.438	.437	.464	.423	.445	.452	.462	.472	.440	.443	.445	.449	.468
KM-ICO	.457	.442	.434	.438	.464	**.425**	.442	.456	.463	**.475**	.440	.442	.445	.450	.469
HG	.493	.499	.495	.509	**.519**	.402	.427	.435	.447	.456	.443	.460	.463	.476	.486
CHG	.507	.490	.493	.510	.512	.402	.429	.433	.450	.459	.449	.458	.461	.478	.484
NDVDR	**.509**	**.502**	**.504**	**.511**	.516	.421	**.447**	**.460**	**.465**	.471	**.461**	**.473**	**.481**	**.487**	**.492**

The integer portion of the results is all 0s.
The best results are in bold, and the second-best are underlined.

Table 3. Performance of Different Image Re-Ranking Methods on the Caltech Dataset

	AP					CR					F1				
	20	40	60	80	100	20	40	60	80	100	20	40	60	80	100
DRR	**.840**	.828	.825	.825	.825	.440	.455	.473	.482	.494	.578	.587	.601	.609	.618
NSGA	.805	.803	.819	.820	.822	.433	.446	.450	.467	.479	.563	.574	.581	.595	.606
MMR	.616	.619	.620	.621	.619	.446	.457	.473	.487	.503	.517	.526	.537	.546	.555
HC	.669	.674	.654	.670	.675	.443	.460	.484	.494	.501	.533	.547	.556	.569	.575
KM-ICO	.676	.673	.660	.673	.676	**.447**	**.469**	**.488**	.496	.504	.538	.553	.561	.571	.577
HG	.835	**.834**	.834	.832	**.832**	.434	.453	.464	.480	.485	.571	.587	.596	.608	.613
CHG	.834	.833	.834	**.833**	**.832**	.437	.454	.465	.479	.489	.574	.588	.597	.608	.616
NDVDR	.834	**.834**	**.835**	**.833**	**.832**	.445	.467	**.488**	**.497**	**.508**	**.580**	**.599**	**.616**	**.622**	**.631**

The integer portion of the results is all 0s.
The best results are in bold, and the second-best are underlined.

that NDVDR is a visual diversity-induced image re-ranking method, and most existing image re-ranking methods based on textual mining are not comparable. To ensure a fair comparison, all the competitors perform re-ranking based on the same initial image retrieval results, and all of them use the same visual features. The specific parameter settings are applied to NDVDR as follows. For the initial image retrieval phase, the length of hash code L is set to 64, the scaling factor ω is set to 0.25, and thus the threshold radius R ($R = \omega \times L$) is 16. All the other parameters are directly adopted from Gou et al. [21]. For the NDVDR component, the parameter z in Equation (11) is empirically set to 100 and the parameter α in Equation (13) is set to 0.5.

4.2 Overall Performance Comparison

The mean values of AP, CR, and F1 for all queries across different image categories on the three image datasets are summarized separately in Tables 2, 3, and 4.

AP Comparison. In terms of AP, it is observed that NDVDR outperforms the other competing methods on all three datasets. Among the three optimization-based methods, namely DRR, NSGA, and MMR, DRR stands out when K is small, particularly on the Caltech and NUSWIDE datasets. NSGA obtains the worst AP performance as it heavily relies on the initial rank from binary hashing. Similarly, MMR focuses solely on maximizing diversity gain based on the relevance level of the initial ranking, leading to AP degradation. For the two clustering-based methods, namely HC and KM-ICO, the quality of image clusters significantly impacts the AP performance. If too many

Table 4. Performance of Different Image Re-Ranking Methods on the NUSWIDE Dataset

	AP					CR					F1				
	20	40	60	80	100	20	40	60	80	100	20	40	60	80	100
DRR	.641	.617	.591	.600	.566	.420	.433	.452	.475	.484	.507	.509	.512	.530	.522
NSGA	.528	.534	.516	.537	.552	.410	.423	.445	.462	.476	.461	.472	.478	.496	.511
MMR	.604	.582	.576	.625	.435	.427	.435	.457	.473	**.489**	.501	.498	.509	.538	.461
HC	.595	.580	.571	.562	.582	**.428**	**.448**	.461	**.478**	.484	.498	.505	.510	.517	.528
KM-ICO	.587	.582	.571	.564	.547	.425	.447	.461	**.478**	.483	.493	.506	.510	.517	.513
HG	.619	.599	.635	.649	.657	.408	.426	.445	.468	.476	.492	.498	.523	.544	.552
CHG	.621	.615	.635	.638	.652	.410	.431	.446	.467	.478	.494	.507	.524	.540	.552
NDVDR	**.656**	**.647**	**.654**	**.652**	**.672**	.426	**.448**	**.462**	.475	.488	**.516**	**.529**	**.542**	**.550**	**.566**

The integer portion of the results is all 0s.
The best results are in bold, and the second-best are underlined.

irrelevant images are grouped into different clusters, the clustering-based methods are more likely to obtain lower AP values when iteratively selecting images from these clusters. In terms of the two graph-based methods, HG and CHG exhibit more competitive AP values compared to the other comparative methods on all three datasets.

CR Comparison. In terms of CR, all methods tend to experience growth with an increase in K since more slots are available for diverse images to be included in the re-ranked image list. The CR metric only reflects the topical diversity of all images in the re-ranked image list, regardless of their relevance to the query. Our NDVDR method demonstrates competitive performance among all methods, and the MMR, HC, and KM-ICO methods also achieve notable CR values. MMR primarily focuses on diversity when selecting new images for the re-ranked list, whereas HC and KM-ICO benefit from the formation of distinct image clusters, greatly enhancing the diversity of the top-ranked images. Therefore, the diversity enhancement of MMR, HC, and KM-ICO is more prominent than the other methods when K is small. However, as K increases, despite the higher CR values, these methods are more likely to select diverse but irrelevant images. This can be observed from their noticeably worse AP performance compared to the other re-ranking methods.

F1 Comparison. Considering the F1 metric, our NDVDR achieves the best performance on all three datasets. The HG and CHG outperform the other competing methods on the Cifar-100 and NUSWIDE datasets, whereas DRR is more competitive on the Caltech dataset. Among the three optimization-based methods, DRR and MMR generally exhibit better F1 performance compared to NSGA. The two clustering-based methods demonstrate similar performance. Among the two graph-based methods, CHG demonstrates a slight advantage on all three datasets. In summary, upon analyzing the results from Tables 2, 3, and 4, the proposed NDVDR consistently shows competitive performance across all datasets in terms of the AP, CR, and F1 metrics.

ADP Comparison. Figures 5 and 6 illustrate the average ADP@K performance for all queries per category randomly selected from the representative 20 categories of Caltech and 14 of NUSWIDE dataset accordingly. We can see that NDVDR achieves the best performance across most categories of images. The competing methods achieve similar ADP performance in several categories when K is small as there is not much room for diversity enhancement. For instance, DRR, NSGA, HG, CHG, and NDVDR obtain similar ADP@10 values for categories 'motorbike,' 'truck,' and 'waterfall,' whereas NSGA outperforms NDVDR for the category of 'American flag' on the Caltech dataset. DRR, MMR, and NDVDR generate close ADP@10 values for categories 'zebra,' 'bike,' and 'tiger' on the NUSWIDE dataset. However, the overall performance enhancement of NDVDR is more significant with an increase in K.

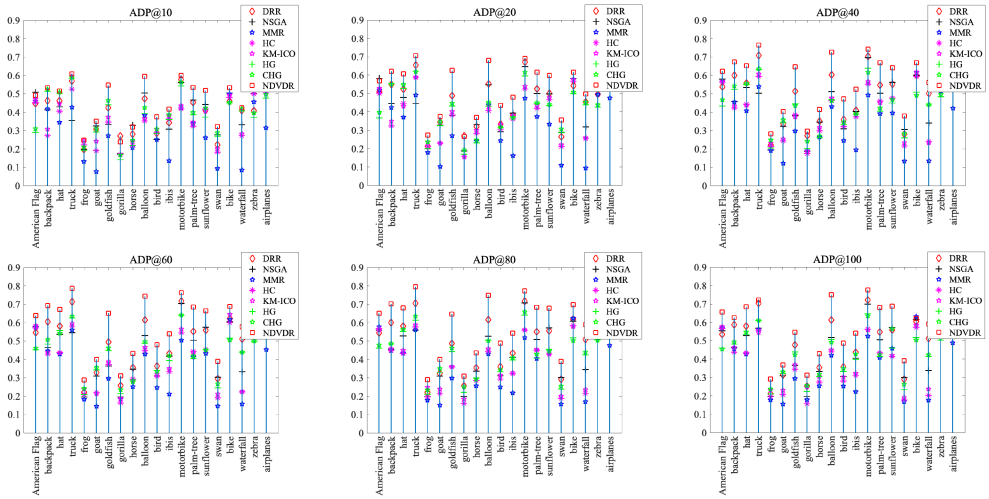

Fig. 5. Comparisons of top K images across different image categories on the Caltech dataset.

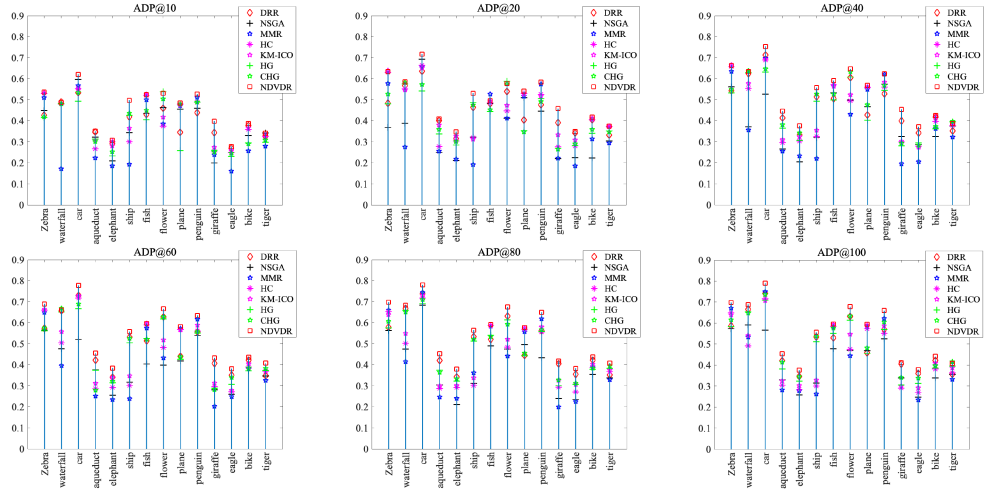

Fig. 6. Comparisons of top K images across different image categories on the NUSWIDE dataset.

From the results, we can also observe that the ADP performance of the other competing methods is unstable across different categories. In particular, for the graph-based methods, both the HG and CHG methods rely on the accuracy of graph construction where similar or duplicate images are expected to share the same edge on the graph. They try to reduce the likelihood of selecting the images that share the same edge by absorbing nodes. When the datasets contain too many similar images or duplicates, it burdens the graph-based methods by absorbing nodes with respect to diversity. For the clustering-based methods, HC and KM-ICO heavily depend on the initial image retrieval results and the quality of image clusters. If the initial image list contains many irrelevant images, the clustering process tends to produce low-quality image clusters. As a result, when iteratively selecting images from each cluster, the low-quality clusters may directly rank dissimilar images highly. Since ADP excludes all irrelevant images from evaluation even if they are possibly more visually diverse to the query image than the other relevant images, the degradation of ADP

Table 5. Time of Re-Ranking on Three Datasets (in Seconds)

	DRR	NSGA	MMR	HC	KM-ICO	HG	CHG	NDVDR
Cifar-100	1.36	547.13	1.23	<u>0.94</u>	3.35	24.63	431.83	**0.79**
Caltech	0.41	340.82	0.39	**0.21**	0.86	21.57	396.51	<u>0.35</u>
NUSWIDE	2.17	737.64	2.04	<u>1.79</u>	5.18	29.47	543.81	**1.42**

The integer portion of the results is all 0s.
The best results are in bold, and the second-best are underlined.

for HC and KM-ICO is especially noticeable as the value of K increases. For the optimization-based methods, the selection strategies are one-sided. For instance, MMR tries to maximize the diversity gain of each newly selected image and only considers images in the current re-ranked image list. Such greedy selection manner is more likely to place irrelevant images with significant visual differences ahead of other relevant images to the query. In contrast, our NDVDR integrates both relevance and diversity in an unbiased way that provides a comprehensive evaluation of all images from the initial results into the non-dominated re-ranking process. Hence, it is more stable than the others.

4.3 Discussion

4.3.1 Runtime Comparisons. Table 5 tabulates the search time required for the re-ranking process of different methods on three image datasets. Each method is repeated 20 times, and the mean value is reported. NDVDR is highly competitive in terms of efficiency. It is considerably faster than the other methods, especially on the Cifar-100 and NUSWIDE datasets. In general, the three fastest re-ranking methods are NDVDR, HC, and MMR. The HC and KM-ICO methods are influenced by the generation of image clusters while more time is consumed by KM-ICO as it performs clustering multiple times to select the suitable number of image clusters. The difference between HG and CHG depends on the graph construction process. CHG is more time consuming than HG since it employs regularized regression models to construct a hypergraph, whereas HG directly establishes connections between each image and its k-nearest neighbors. Among all methods, NSGA exhibits the highest time consumption.

4.3.2 Investigation of Image Features. The choice of image features can affect the performance of re-ranking. In this experiment, we discuss the impact of adopting different deep features in image re-ranking. Four CNN-based features, namely 2,048-dimensional ResNet-100, 4,096-dimensional AlexNet, 4,096-dimensional VGG16, and 2,048-dimensional Inception-v3, are extracted from the fully connected layer of the respective pre-trained convolution networks.[1] Additionally, recognizing the recent remarkable advancement of vision transformers [17, 19] in generating strong image features, a pre-trained vision transformer model (ViT) [17] is also employed for comparison. ViT sets the image patch size as 16×16, and the final output is 512-dimensional image features. The performance is illustrated in Figure 7. Overall, ViT obtains the best F1 results. Among the five CNN-based features, Inception-v3 further improves the results, but the maximal performance gap among these CNN-based features is only around 1.3%. We choose ResNet-50 for its moderate complexity and relatively high performance. More importantly, the flexibility of NDVDR allows for the integration of various visual features into the re-ranking process. More discriminative visual features, in particular, can capture intricate and fine-grained information about the visual content of images, thereby possibly enhancing the accuracy of relevance and diversity assessment during re-ranking. This opens up possibilities for leveraging state-of-the-art visual features to achieve even better re-ranking results and meet the evolving needs of image search and recommendation systems.

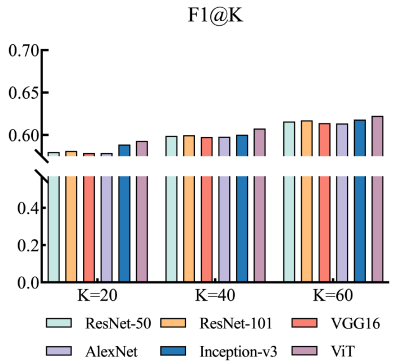

Fig. 7. Comparisons of different image features on the Caltech dataset.

Table 6. Performance of Different Initial Image Retrieval Methods on the Caltech Dataset

	F1@20	F1@40	F1@60	F1@80	F1@100
HR	.503	.517	.515	.528	.538
HR_2	.536	.518	.520	.528	.532
HR_3	.519	.526	.529	.535	.546
HR-NDVDR	.580	.599	**.616**	**.622**	**.631**
HR_2-NDVDR	.587	.594	.609	.613	.617
HR_3-NDVDR	**.592**	**.605**	.613	.617	.628

The integer portion of the results is all 0s.
The best results are in bold, and the second-best are underlined.

4.3.3 Investigation of Initial Retrieval Methods. Our NDVDR is designed to be generic and flexible, allowing for seamless integration with other image retrieval methods for the initial ranking. In this experiment, our objective revolves around showcasing the scalability and efficacy of NDVDR when conjuncted with the other two state-of-the-art image retrieval systems, namely LAH [32] and SCDH [14]. LAH is a semi-supervised hashing method designed to enhance the accuracy and efficiency of retrieval by dynamically selecting the appropriate hash code length. Similar to the HR phase in our NDVDR, LAH utilizes the GIST feature for image representation. In contrast, SCDH is a supervised hashing technique that learns binary codes for all examples in the training set. It simultaneously derives a hash function for unseen samples, leveraging deep features to represent images. As illustrated in Figure 2, the offline pre-training and initial image retrieval stages are directly replaced with the training and retrieval processes specific to LAH and SCDH separately. To differentiate between the distinct combinations, we refer to the resulting methods as HR_2-NDVDR and HR_3-NDVDR for LAH and SCDH, respectively, while keeping HR-NDVDR indicating our original method using HCSDH [21]. Table 6 summarizes the F1 performance of initial HR ranking, initial HR_2 ranking, initial HR_3 ranking, HR-NDVDR, HR_2-NDVDR, and HR_3-NDVDR on the top K images. From Table 6, we can observe variations in the performance of the initial image retrieval among the HR, HR_2, and HR_3 methods. When we conjunct the three initial retrieved results with NDVDR, HR-NDVDR, HR_2-NDVDR, and HR_3-NDVDR generally obtain similar performance. In sum, our re-ranking method is consistently effective across different initial retrieval methods. By instantiating NDVDR into the retrieval pipeline of other image retrieval systems, we can easily enhance the overall quality of the top-ranked images to provide more comprehensive and informative image results to users.

Table 7. Performance of NDVDR under Different Configurations of (L, ω) on the Caltech Dataset

	F1@20				F1@40				F1@60			
(L, ω)	16	32	64	128	16	32	64	128	16	32	64	128
0.25	.569	.575	.580	.583	.587	.593	.599	.596	.599	.608	.616	.616
0.5	.561	.570	.574	.572	.578	.581	.591	.590	.595	.606	.613	.609
0.75	.564	.572	.577	.575	.575	.582	.588	.585	.591	.597	.605	.611

The integer portion of the results is all 0s.

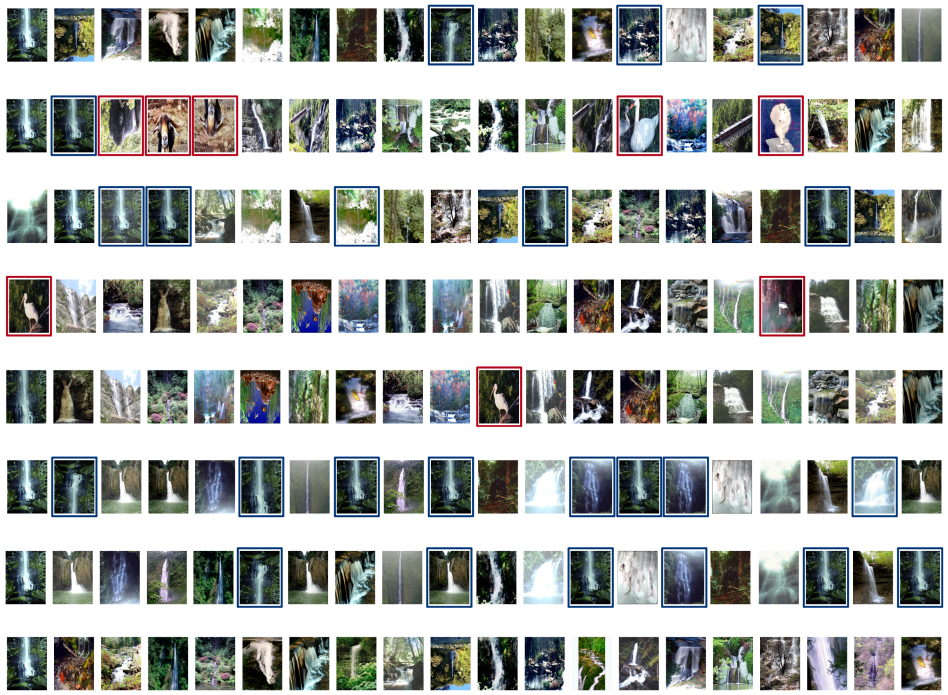

Fig. 8. Top 20 images for a query 'waterfall.' From top to bottom: DRR, MMR, NSGA, HC, KM-ICO, HG, CHG, and our NDVDR.

4.3.4 Parameter Investigation of Re-Ranking Data Size. In this section, we investigate the impact of different parameters on NDVDR. In our framework, only a subset of images from the retrieval dataset is used for re-ranking, specifically the N highest-ranked images from the initial retrieval results. The value of N is determined by a scaling factor ω and the hash code length L, as defined in Equations (4) and (5). We examine the effects of varying L and ω in Table 7. Typically, a large L can improve the precision of the initial retrieved images. Meanwhile, increasing ω expands the neighborhood radius R in Equation (4), causing more images with larger Hamming distances to be included, and potentially leading to irrelevant images being fed into the re-ranking phase. Moreover, larger L and ω settings mean more initial retrieval images are taken into account, increasing re-ranking time. Based on our experiments, we suggest using $\omega = 0.25$. Regarding the hash code length, larger L can improve similarity evaluation precision, although at the expense of higher computational cost. Therefore, we recommend using $L = 64$ for a balanced performance.

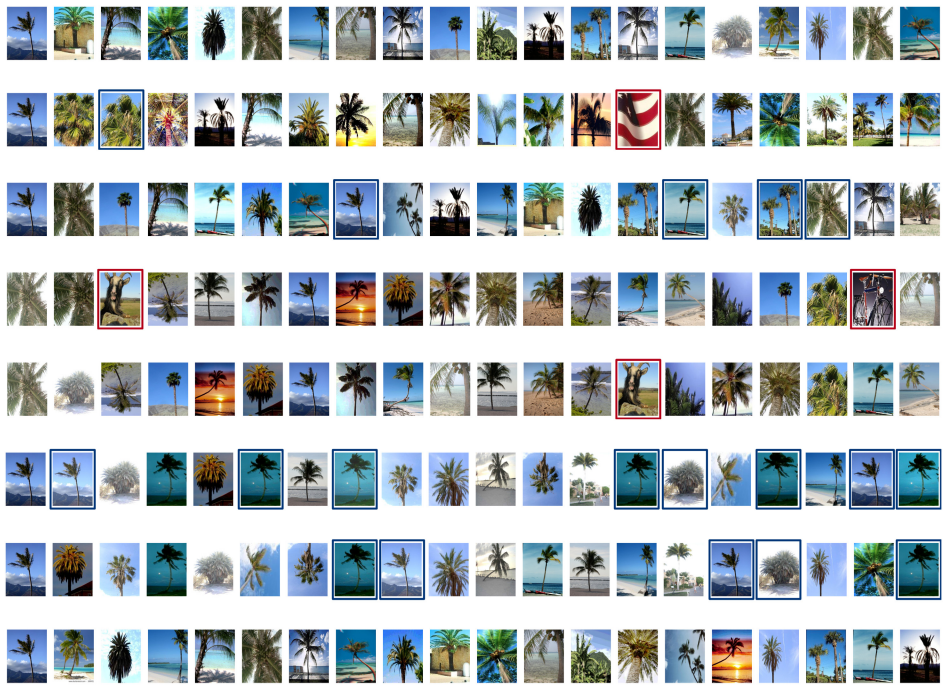

Fig. 9. Top 20 images for a query 'palm-tree.' From top to bottom: DRR, MMR, NSGA, HC, KM-ICO, HG, CHG, and our NDVDR.

4.4 Subjective Visual Comparisons

In this section, we conduct case studies on two representative query instances from the image categories of 'palm-tree' and 'waterfall' accordingly to further validate the effectiveness of our NDVDR. The top 20 images obtained from different image re-ranking methods are shown for subjective visual comparisons. In Figures 8 and 9, we highlight the irrelevant images with red squares and mark duplicates with blue squares. The quality of an image list can be subjectively evaluated from several aspects. First, it should maintain the most relevant image to the query after re-ranking. In other words, the image at the first position should remain unchanged. Second, irrelevant images should not be ranked at the top. Most importantly, the number of duplicates or similar images should be reduced after re-ranking. In Figures 8 and 9, we can see that the initial ranking contains many duplicates which definitely cannot satisfy users' expectations. The methods MMR, HC, and KM-ICO show irrelevant images. Moreover, the clustering-based methods are hard to retain the first image, which is considered to be the most relevant to the query. As for the two graph-based methods HG and CHG, more duplicates have appeared since the image datasets contain too many duplicate images. Among all the competing methods, NDVDR consistently preserves the first and most relevant image to the query. It successfully removes nearly all duplicates from the top 20 images and introduces more visually diverse images, enabling a more comprehensive response to users.

5 CONCLUSION

We proposed a non-dominated visual image re-ranking method for optimizing both relevance and diversity among top-ranked image retrieval results. In the first phase, a fast binary hashing retrieval was conducted for its efficiency to provide a coarse-grained initial retrieved image list.

Then, the second phase executed our NDVDR for re-ranking images to make the top-ranked images visually relevant and diverse to the given query image. During the re-ranking process, two objective measures of relevance and diversity were designed and computed for each image pair of the image from the initial image list and query. The Pareto optimality concept was devised to tackle the two partially conflicting objectives, and subsequently Pareto hierarchies of images were generated. The images were re-ranked according to the Pareto domination relationship among them. Experimental results proved that the proposed NDVDR outperforms other popular image search re-ranking methods in terms of the AP, CR, F1, and ADP scores. In addition, we also validated the generalization capability of the proposed framework, which opens up the possibility of integrating NDVDR into other image retrieval systems for providing more comprehensive image results to users.

REFERENCES

[1] Xaro Benavent, Angel Castellanos, Esther de Ves, Ana Garcia-Serrano, and Juan Cigarran. 2019. FCA-based knowledge representation and local generalized linear models to address relevance and diversity in diverse social images. *Future Gener. Comp. Syst.* 100 (2019), 250–265.

[2] Giulia Boato, Duc Tien Dang-Nguyen, Oleg Muratov, Naif Alajlan, and Francesco G. B. De Natale. 2016. Exploiting visual saliency for increasing diversity of image retrieval results. *Multimed. Tools Appl.* 75 (2016), 5581–5602.

[3] Boteanu Bogdan, Mihai Gabriel Constantin, and Ionescu Bogdan. 2016. LAPI retrieving diverse social images task: A pseudo-relevance feedback diversification perspective. In *Proceedings of the 2016 MediaEval Workshop*.

[4] Bogdan Boteanu, Ionuţ Mironică, and Bogdan Ionescu. 2017. Pseudo-relevance feedback diversification of social image retrieval results. *Multimed. Tools Appl.* 76 (2017), 11889–11916.

[5] Bogdan Boteanu, Ionut Mironica, Anca Livia Radu, and Bogdan Ionescu. 2014. LAPI@2014 retrieving diverse social images task: A relevance feedback diversification perspective. In *Proceedings of the 2014 MediaEval Workshop*.

[6] Mariam Bouchakwa, Yassine Ayadi, and Ikram Amous. 2020. Multi-level diversification approach of semantic-based image retrieval results. *Prog. Artif. Intell.* 9, 1 (2020), 1–30.

[7] Noura Bouhlel, Ghada Feki, and Chokri Ben Amar. 2020. Hypergraph-based image search reranking with elastic net regularized regression. *Multimed. Tools Appl.* 79, 41 (2020), 30257–30280.

[8] Noura Bouhlel, Ghada Feki, Anis Ben Ammar, and Chokri Ben Amar. 2017. A hypergraph-based reranking model for retrieving diverse social images. In *Computer Analysis of Images and Patterns*. Lecture Notes in Computer Science, Vol. 10424. Springer, 279–291.

[9] Noura Bouhlel, Ghada Feki, Anis Ben Ammar, and Chokri Ben Amar. 2020. Hypergraph learning with collaborative representation for image search reranking. *Int. J. Multimed. Inf. Ret.* 9, 3 (2020), 205–214.

[10] Rodrigo Tripodi Calumby, Marcos André Gonçalves, and Ricardo da Silva Torres. 2017. Diversity-based interactive learning meets multimodality. *Neurocomputing* 259 (2017), 159–175.

[11] Ángel Castellanos, Xaro Benavent, Ana García-Serrano, Esther de Ves, and Juan Cigarrán. 2016. UNED-UV@ retrieving diverse social images task. In *Proceedings of the 2016 MediaEval Conference*.

[12] Yair Censor. 1977. Pareto optimality in multiobjective problems. *Appl. Math. Opt.* 4, 1 (1977), 41–59.

[13] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. 2021. CrossViT: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of ICCV*. 357–366.

[14] Yong Chen, Zhibao Tian, Hui Zhang, Jun Wang, and Dell Zhang. 2020. Strongly constrained discrete hashing. *IEEE Trans. Image Process.* 29, 11 (2020), 3596–3611.

[15] Tat-Seng Chua, Jin Hui Tang, Ri Chang Hong, Hao Jie Li, Zhi Ping Luo, and Yan Tao Zheng. 2009. NUS-WIDE: A real-world web image database from national university of singapore. In *Proceedings of CIVR*. 1–9.

[16] Duc-Tien Dang-Nguyen, Luca Piras, Giorgio Giacinto, Giulia Boato, and Francesco G. B. De Natale. 2017. Multimodal retrieval with diversification and relevance feedback for tourist attraction images. *ACM Trans. Multimed. Comput.* 13, 4 (2017), 1–24.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*. 1–21.

[18] Hugo Jair Escalante and Alicia Morales-Reyes. 2013. TIA-INAOE's approach for the 2013 retrieving diverse social images task. In *Proceedings of the 2013 MediaEval Workshop*.

[19] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of ICCV*. 6824–6835.

[20] José Solenir Lima Figuerêdo and Rodrigo Tripodi Calumby. 2022. Unsupervised query-adaptive implicit subtopic discovery for diverse image retrieval based on intrinsic cluster quality. *Multimed. Tools Appl.* 81, 30 (2022), 42991–43011.

[21] Koutaki Gou, Shirai Keiichiro, and Ambai Mitsuru. 2018. Hadamard coding for supervised discrete hashing. *IEEE Trans. Image Process.* 27 (2018), 5378–5392.

[22] Burak Goynuk and Ismail Sengor Altingovde. 2020. Supervised learning methods for diversification of image search results. In *Advances in Information Retrieval*. Lecture Notes in Computer Science, Vol. 12036. Springer, 158–165.

[23] Gregory Griffin, Alex Holub, and Pietro Perona. 2007. *Caltech-256 Object Category Dataset*. California Institute of Technology.

[24] Hou Dong Hu, Yan Wang, Lin Jun Yang, Pavel Komlev, Li Huang, Xi (Stephen) Chen, Jia Pei Huang, Ye Wu, Meenaz Merchant, and Arun Sacheti. 2018. Web-scale responsive visual search at Bing. In *Proceedings of KDD*. 359–367.

[25] Bogdan Ionescu, Adrian Popescu, Anca Livia Radu, and Henning Muller. 2016. Result diversification in social image retrieval: A benchmarking framework. *Multimed. Tools Appl.* 75 (2016), 1301–1331.

[26] Bogdan Ionescu, Maia Rohm, Bogdan Boteanu, Alexandru Lucian Ginsca, and Henning Muller. 2020. Benchmarking image retrieval diversification techniques for social media. *IEEE Trans. Multimed.* 23 (2020), 677–691.

[27] Zhong Ji, Yuting Su, Yanwei Pang, and Xiaojie Qu. 2011. Diversifying the image relevance reranking with absorbing random walks. In *Proceedings of ICIG*. IEEE, Los Alamitos, CA, 981–986.

[28] Peiguang Jing, Yuting Su, Chuanzhong Xu, and Luming Zhang. 2018. HyperSSR: A hypergraph based semi-supervised ranking method for visual search reranking. *Neurocomputing* 274 (2018), 50–57.

[29] Chen Karako and Putra Manggala. 2018. Using image fairness representations in diversity-based re-ranking for recommendations. In *Proceedings of UMAP*. 23–28.

[30] Alex Krizhevskyand Geoffrey Hinton. 2009. *Learning Multiple Layers of Features from Tiny Images*. University of Toronto.

[31] Amel Ksibi, Anis Ben Ammar, and Chokri Ben Amar. 2014. Adaptive diversification for tag-based social image retrieval. *Int. J. Multimed. Inf. Ret.* 3 (2014), 29–39.

[32] Si-Chao Lei, Xing Tian, Wing W. Y. Ng, and Yue-Jiao Gong. 2023. Length adaptive hashing for semi-supervised semantic image retrieval. *Multimed. Tools Appl.* 82, 1 (2023), 1–23.

[33] Wei-Chao Lin. 2019. Aggregation of multiple pseudo relevance feedbacks for image search re-ranking. *IEEE Access* 7 (2019), 147553–147559. DOI: https://doi.org/10.1109/ACCESS.2019.2942142

[34] Wei Lu, Mengqi Luo, Zhenyu Zhang, Guobiao Zhang, Heng Ding, Haihua Chen, and Jiangping Chen. 2019. Result diversification in image retrieval based on semantic distance. *Inf. Sci.* 502 (2019), 59–75.

[35] Timo Milbich, Karsten Roth, Homanga Bharadhwaj, Samarth Sinha, Yoshua Bengio, Björn Ommer, and Joseph Paul Cohen. 2020. DiVA: Diverse visual feature aggregation for deep metric learning. In *Computer Vision—ECCV 2020*. Lecture Notes in Computer Science, Vol. 12353. Springer, 590–607.

[36] Niluthpol Chowdhury Mithun, Rameswar Panda, and Amit K. Roy-Chowdhury. 2020. Construction of diverse image datasets from web collections with limited labeling. *IEEE Trans. Circ. Syst. Vid.* 30, 4 (2020), 1147–1161. DOI: https://doi.org/10.1109/TCSVT.2019.2898899

[37] Jianbo Ouyang, Wengang Zhou, Min Wang, Qi Tian, and Houqiang Li. 2020. Collaborative image relevance learning for visual re-ranking. *IEEE Trans. Multimed.* 23 (2020), 3646–3656.

[38] Liang Peng, Yi Bin, Xi Yao Fu, Jie Zhou, Yang Yang, and Heng Tao Shen. 2017. CFM@MediaEval 2017 retrieving diverse social images task via re-ranking and hierarchical clustering. In *Proceedings of the 2017 CEUR Workshop*.

[39] Xue Ming Qian, Dan Lu, Ya Xiong Wang, Li Zhu, Yuan Yan Tang, and Meng Wang. 2017. Image re-ranking based on topic diversity. *IEEE Trans. Image Process.* 26, 8 (2017), 3734–3747.

[40] Anca Livia Radu, Bogdan Ionescu, Maria Menendez, Julian Stottinger, Fausto Giunchiglia, and Antonella De Angeli. 2014. A hybrid machine-crowd approach to photo retrieval result diversification. In *Multimedia Modeling*. Lecture Notes in Computer Science, Vol. 8325. Springer, 25–36.

[41] Vidyadhar Rao, Prateek Jain, and C. V. Jawahar. 2016. Diverse yet efficient retrieval using locality sensitive hashing. In *Proceedings of ICMR*. 189–196.

[42] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2015. Search result diversification. *Found. Trends Inf. Retr.* 9, 1 (2015), 1–90.

[43] Omar Seddati, Nada Ben Lhachemi, Stephane Dupont, and Mahmoudi Said. 2017. UMONS@MediaEval 2017: Diverse social images retrieval. In *Proceedings of the 2017 MediaEval Workshop*.

[44] Xi Shen, Yang Xiao, Shell Xu Hu, Othman Sbai, and Mathieu Aubry. 2021. Re-ranking for image retrieval and transductive few-shot classification. *Adv. Neural Info. Process. Syst.* 34 (2021), 25932–25943.

[45] Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, Alexandru Lucian Ginsca, Adrian Popescu, Yiannis Kompatsiaris, and Ioannis Vlahavas. 2015. Improving diversity in image search via supervised relevance scoring. In *Proceedings of ICMR*. 323–330.

[46] Yehui Tang, Kai Han, Chang Xu, An Xiao, Yiping Deng, Chao Xu, and Yunhe Wang. 2021. Augmented shortcuts for vision transformers. *Adv. Neural Inf. Process. Syst.* 34 (2021), 15316–15327.

[47] Sabrina Tollari. 2016. UPMC at MediaEval 2016 retrieving diverse social images task. In *Proceedings of the 2016 MediaEval Workshop*.

[48] Reinier H. Van Leuken, Lluis Garcia, Ximena Olivares, and Roelof Van Zwol. 2009. Visual diversification of image search results. In *Proceedings of WWW*. 341–350.

[49] Baptist Vandersmissen, Abhineshwar Tomar, Frederic Godin, Wesley De Neve, and Rik Van De Walle. 2014. Ghent University-iMinds at MediaEval 2014 diverse images: Adaptive clustering with deep features. In *Proceedings of the 2014 MediaEval Workshop*.

[50] Luo Wang, Xueming Qian, Xingjun Zhang, and Xingsong Hou. 2020. Sketch-based image retrieval with multi-clustering re-ranking. *IEEE Trans. Circ. Syst. Vid.* 30, 12 (2020), 4929–4943. DOI: https://doi.org/10.1109/TCSVT.2019.2959875

[51] Meng Wang, Kui Yuan Yang, Xian Sheng Hua, and Hong Jiang Zhang. 2010. Towards a relevant and diverse search of social images. *IEEE Trans. Multimed.* 12, 8 (2010), 829–842.

[52] Zhi Jing Wu, Ke Zhou, Yi Qun Liu, Min Zhang, and Shao Ping Ma. 2019. Does diversity affect user satisfaction in image search. *ACM Trans. Inform. Syst.* 37, 3 (2019), 1–30.

[53] Yan Yan, Gaowen Liu, Sen Wang, Jian Zhang, and Kai Zheng. 2017. Graph-based clustering and ranking for diversified image search. *Multimed. Syst.* 23 (2017), 41–52.

[54] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, and Miki Haseyama. 2022. Interactive re-ranking via object entropy-guided question answering for cross-modal image retrieval. *ACM Trans. Multimed. Comput.* 18, 3 (2022), 1–17.

[55] Yawen Zeng, Yiru Wang, Dongliang Liao, Gongfu Li, Weijie Huang, Jin Xu, Da Cao, and Hong Man. 2022. Keyword-based diverse image retrieval with variational multiple instance graph. *IEEE. Trans. Neural Netw. Learn. Syst.* Early access, April 28, 2022. DOI: https://doi.org/10.1109/TNNLS.2022.3168431

[56] Yanhao Zhang, Pan Pan, Yun Zheng, Kang Zhao, Yingya Zhang, Xiaofeng Ren, and Rong Jin. 2018. Visual search at Alibaba. In *Proceedings of KDD*.

Received 5 March 2023; revised 13 September 2023; accepted 15 September 2023