Content Based Deep Learning Image Retrieval: A Survey

Chi Zhang
University of Chinese Academy of Sciences
Institute of Automation, Chinese Academy of Sciences
Beijing, China
zhangchi2021@ia.ac.cn

Jie Liu*
Institute of Automation, Chinese Academy of Sciences
Beijing, China
jie.liu@ia.ac.cn

ABSTRACT

With the development of digital technology, various fields generate and share a large amount of visual content. Image retrieval is a hot research direction in the field of computer vision. Efficient and accurate retrieval of query content from massive data is the ultimate form pursued by image retrieval technology. In recent years, the rise of deep learning technology has promoted the rapid development of the field of computer vision. Due to the powerful expressive ability of deep features on image content, image retrieval based on deep learning has become the most cutting-edge research direction in CBIR technology. This paper summarizes the relevant research on the classic deep learning image retrieval technology in recent years, first introduces the form of the CBIR problem, and then lists the classic datasets in this field. Afterwards, content-based deep image retrieval methods are reviewed from the perspectives of network models, deep feature extraction, and retrieval types. Finally, summarize the problems to be solved urgently in the current research, and look forward to the future research direction.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**; **Matching**.

KEYWORDS

Content Based Image Retrieval, Deep Learning, Convolution Neural Network

ACM Reference Format:
Chi Zhang and Jie Liu. 2023. Content Based Deep Learning Image Retrieval: A Survey. In *2023 the 9th International Conference on Communication and Information Processing (ICCIP) (ICCIP 2023), December 14–16, 2023, Lingshui, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3638884.3638908

1 INTRODUCTION

1.1 Background

With the rapid growth of digital content in the modern Internet, retrieving image content in the wide-area Internet has become increasingly difficult and complex. Relying on the powerful expressive ability of deep features to visual information, the performance of content-based deep image retrieval technology has been proved to be able to greatly surpass the traditional content-based image retrieval (CBIR) methods. Besides, it has made achievements in copyright protection, search engine and many other fields. This paper reviews the classic methods of deep image retrieval technology, summarizes the challenges and problems that need to be solved in current image retrieval tasks, and analyzes the future research direction of this technology.

1.2 Image Retrieval Task

The purpose of image retrieval is to search an image database for images that are similar or homologous to an input image. Image retrieval can be divided into text-based image retrieval (TBIR) and content-based image retrieval (CBIR) according to the way of describing image content [65].

TBIR uses manual annotation or semi-automatic annotation of image recognition technology to describe the image content, and forms keywords to describe the image content for each image. In the retrieval phase, the user retrieves the annotated images from the image library through keywords. In addition, this method is easy to implement. Due to the existence of manual or image recognition technology annotations, the accuracy of the algorithm is relatively high, and it has a good application prospect in the face of small and medium-scale image search problems.

Due to the time-consuming and labor-intensive manual annotation of TBIR, the process is easily affected by factors such as the knowledge level of the annotator, language use, and subjective judgments, resulting in problems such as differences in text descriptions and pictures. In order to solve the semantic gap between the high-level semantics and low-level visual features of retrieved images, both academia and industry have made efforts to develop CBIR. With the continuous improvement of deep learning theory, CBIR has made great progress. In large-scale image retrieval, the CBIR task is to search for the most relevant content to a given query data in a large image collection, which mainly includes two stages of feature extraction and similarity measurement. Compared with TBIR, which uses unstructured data, namely text, as the annotation method, the use of deep features enables CBIR to overcome the shortcomings of TBIR and improve retrieval efficiency.

2 BENCHMARKS AND METRICS

2.1 Classic Benchmarks

The learning process of deep learning algorithms is basically data-driven, and the datasets can not only provide training samples for the algorithm to learn, but also a benchmark for fair comparison of

*Corresponding author.

ICCIP 2023, December 14–16, 2023, Lingshui, China

ACM ISBN 979-8-4007-0890-9/23/12.
https://doi.org/10.1145/3638884.3638908

Table 1: Common benchmarks.

Dataset	classes	scale	Scenario
NUS-WIDE[10]	21	2.7×10^5	instance-level
MS-COCO[32]	80	1.2×10^5	multimodal retrieval
Flickr30k[41]	-	3×10^4	multimodal retrieval
GLD v2[61]	2×10^5	5×10^6	instance-level
XMarket[6]	5471	1.8×10^5	category-level
CUB200-2011[55]	200	1.2×10^4	category-level
Aircraft[34]	102	1×10^4	fine-grained
Paris-6k[40]	12	6,000	instance-level
Oxford5k[39]	11	5,000	instance-level
UKBench[37]	2550	1×10^4	instance-level
Holidays[20]	500	1,500	instance-level
Sketchy[46]	125	8.8×10^4	sketch retrieval
Fashion-IQ[18]	3	7.8×10^4	interactive retrieval

various algorithms. Commonly used datasets are listed in Table 1. The Google Landmarks Dataset v2 [61] contains more than 5×10^6 images and 2×10^5 different instance labels, including more than 4×10^6 images in the training set, 7×10^5 images in the reference set, and 1×10^5 images in the test set. GLDv2 is the largest landmark dataset, containing annotated images of man-made and natural landmarks. NUS-WIDE [10] is a multi-label definition dataset about image text matching, which contains 2.7×10^5 pictures, and each picture contains an average of 2~5 labels. The MS-COCO [32] dataset contains 1.2×10^5 images, and each image contains at least 5 sentence annotations. Flickr30k [41] contains more than 30,000 pictures, and each picture contains 5 sentence annotations. Oxford-5k [39] consists of more than 5,000 images of 11 Oxford buildings. Sketchy [46] contains 125 sketch image pairs of different categories, each category contains 100 images.

2.2 Evaluation Methods

Choosing an appropriate evaluation formula in image retrieval tasks depends on two factors: the algorithm itself and the problem domain. At present, the commonly used evaluation metrics of CBIR include Recall, Precision, F-score. The recall rate refers to the percentage of images correctly retrieved by the retrieval system to the total number of relevant images in the dataset, and the calculation formula is shown in Equation 1:

$$R = \frac{T}{T + M} \tag{1}$$

where T represents the number of correctly retrieved samples, and M represents the number of samples not returned in the dataset related to the query image. Precision refers to the percentage of images correctly retrieved by the retrieval system to the total number of retrieved images, and the calculation formula is shown in Equation 2:

$$P = \frac{T}{T + F} \tag{2}$$

where F represents the number of samples retrieved that are not related to the query sample. In general, R and P are contradictory, and the recall rate and precision rate can be judged according to the requirements for image retrieval tasks in specific fields. The F-score refers to the weighted harmonic mean of the recall rate and the precision rate, and the calculation formula is shown in Equation 3:

$$F = \frac{(1 + \beta^2)PR}{\beta^2(P + R)} \tag{3}$$

where β is a parameter to adjust the weight of recall rate and precision rate. If a higher precision rate is required, β will be reduced, and if a higher recall rate is required, β will be increased. When $\beta = 1$, R and P are equally important, that is, F1-score. The higher the F1 value, the better the retrieval performance of the system. In addition to the F1, mAP (mean Average Precision) is also one of the important indicators to evaluate the overall performance of the retrieval system.

3 DEEP CBIR

The deep image retrieval technology is generally based on the image features extracted by the deep neural network for vector retrieval, because the features contain the semantic content of the image, so the deep image retrieval belongs to the content-based image retrieval [65].

3.1 Deep Image Retrieval

3.1.1 Category-level Retrieval. The main task of category-level image retrieval is to retrieve any image of the same category as the query image. Sharma et al. [47] proposed a supervised discriminative distance learning method that outperforms baselines in category-based image retrieval tasks. Meng et al. [35] performed feature extraction and matching at the class level, and proposed a new image retrieval method based on merged regions. [63] proposed a cross-domain representation learning framework, which achieved strong performance in category-level image retrieval.

3.1.2 Instance-level Retrieval. The goal of instance-level image retrieval is to find images containing specific instances in the query image, which may be captured under different background conditions. To achieve accurate and efficient retrieval in large-scale image databases, the core task of instance-level image retrieval is to obtain compact and discriminative feature representations of images. [44] developed a deep CNN-based baseline for instance retrieval using local feature extraction based on CNN representations.

Other approaches to image instance retrieval include local convolutional feature packs [36], instance-aware image representation methods [25], and hashing models for deep multi-instance ranking [9], etc. Amato et al. [2] introduced a deep feature representation method based on scalar quantization, and proved the effectiveness of the method on instance-level retrieval benchmarks. Krishna et al. found that models trained using contrastive methods outperformed pretrained baselines trained on ImageNet in retrieval tasks. Bai et al. [4] proposed an unsupervised framework that focuses on instance objects in images, called adversarial instance-level image retrieval. It is the first time that adversarial training is used in the retrieval process of instance-level image retrieval tasks, which can significantly improve retrieval accuracy without increasing time cost.

3.1.3 Fine-grained Retrieval. Xie et al. [62] proposed the concept of fine-grained image search. Driven by deep learning technology,

more and more fine-grained image retrieval methods based on deep learning have been proposed [31, 75, 76]. [56] proposed a deep ranking model that learns a fine-grained image similarity model directly from images. Ahmad et al. [1] proposed an object-oriented feature selection mechanism for pre-training CNN's deep convolutional features. The model uses a locality-sensitive hashing method to enable fine-grained retrieval in large-scale surveillance datasets.

3.1.4 Cross-modal Retrieval. With the application of deep neural networks in the field of image retrieval research, cross-modal retrieval has received extensive attention. The two modalities of image and text are very common in the field of retrieval. When the data of one modality is given, the cross-modal retrieval task needs to find several corresponding or closest data to the given modality in the space of another modality.

Multimodal retrieval methods include deep visual semantic hashing [7], self-supervised adversarial hashing [27], deep cascaded cross-modal ranking model [59], deep mutual information maximization algorithm [16]. Dey et al. [11] proposed a cross-modal deep network structure that allows text and sketches to be used as query input, and uses an attention model to retrieve multiple objects in the query. Lee et al. [26] studied the image-text matching problem and proposed a stacked cross-attention mechanism that uses image regions and words in sentences as context to discover complete potential alignments and infer image-text similarities. Wang et al. [60] proposed a cross-modal adaptive information transfer model consisting of cross-modal information aggregation and cross-modal gating fusion to adaptively explore the interaction between images and sentences in text-image matching. Chaudhuri et al. [8] proposed a remote sensing cross-modal retrieval framework based on deep neural networks. Sumbul et al. [50] proposed a new self-supervised cross-modal image retrieval method, which does not require any labeled training images, can still effectively maintain the similarity between modalities and between modalities, and eliminate the differences between modalities.

3.1.5 Sketch-based Retrieval. Sketch based image retrieval (SBIR) is essentially cross-modal information retrieval. Researchers have established effective SBIR algorithms from three aspects: deep multimodal feature generation, cross-modal correlation modeling, and similarity function optimization. Eitz et al. [13] benchmarked SBIR. Qi et al. [42] proposed SBIR based on Siamese CNN architecture. Song et al. [49] constructed a new fine-grained SBIR (FG-SBIR) model by introducing attention modules, shortcut connection fusion blocks and high-order learnable energy functions. Pang et al. [38] first discovered and solved the cross-category FG-SBIR generalization problem, defined FG-SBIR cross-category generalization as a domain generalization problem, and proposed an unsupervised learning method to model a general visual sketch feature flow shapes, automatically adapting to new categories. [67] proposed a zero-shot SBIR (ZS-SBIR) benchmark for retrieval of classes that were not trained. Dey et al. contributed a large-scale ZS-SBIR dataset QuickDrawerExtended [12] to the community.

Other approaches to SBIR include a cross-domain representation learning framework [63], a CNN-based semantic reranking system [57], and semantically aligned pairwise recurrent consensus generative networks [169]. Bhunia et al. [5] designed a cross-modal retrieval framework FG-SBIR based on reinforcement learning to solve the problem of taking a long time to draw sketches. Torres et al. [54] utilized the uniform manifold approximation and projection (UMAP) for dimensionality reduction, proposing the use of compact feature representations in the SBIR environment. Sain et al. [45] proposed a SBIR model that can adapt to the agnostic drawing style in view of the diversity of styles of different users when drawing sketches. Yu et al. [68] first defined and solved the problem of fine-grained instance-level image retrieval using freehand sketches, and provided a large-scale fine-grained sketch dataset.

3.1.6 Conversational Image Retrieval. Conversational image retrieval can gradually clarify the user's retrieval intention according to the interactive user response, and achieve more accurate retrieval. Liao et al. [30] proposed a knowledge-aware multimodal dialogue model that considers the semantic and domain knowledge contained in visual content. Guo et al. [17] introduce an interactive image search method based on deep learning, which enables users to provide feedback through natural language. On this basis, Zhang et al. [71] proposed a constraint-enhanced reinforcement learning framework to effectively incorporate users' preferences over time. Zhang et al. [72] proposed a reward-constrained recommendation framework for text-based interactive recommendation. Yuan et al. [69] proposed a multi-turn natural language feedback text framework that can effectively handle conversational fashion image retrieval. Kaushik et al. [23] introduced a multi-view conversational image search system, developed a reinforcement learning model based on the initial running state, incentives, and sessions, and predicted the images provided to the user through a customized search algorithm.

3.2 DNNs For CBIR

The most representative models for the feature extraction in image retrieval include VGG [48], GoogLeNet [51], ResNet [19] and EfficientNet [52].

3.2.1 VGG. VGG [48] has more convolutional layers than AlexNet [24], and VGG-16 and VGG-19 are the most widely used versions, consisting of 13 and 16 convolutional layers, respectively. The strategy of VGG is to deepen the number of layers of the convolutional neural network. The experimental results show that within a certain range, deepening the network can effectively improve the performance of the model.

3.2.2 GoogLeNet. GoogLeNet [51] designs an inception module, which can construct a sparser CNN structure. By using different sizes of convolution kernels to capture different sizes of receptive fields, the last layer uses a global mean pooling layer to replace the fully connected layer, reducing model parameters. Compared with AlxeNet and VGGNet, the GoogLeNet model is deeper and wider, with fewer model parameters and higher learning efficiency. Deeper architectures are beneficial to learn higher-level abstract features, thereby reducing the semantic gap.

3.2.3 ResNet. ResNet [19] converts a normal CNN network into a residual network using skip connections, and ResNets have fewer convolution filters than VGGNets. ResNet uses skip connections

or just skips some layers to avoid the problem of gradient disappearance. The skip connections act as gradient highways, allowing gradients to flow undisturbed.

3.2.4 EfficientNet. Compared with the traditional model random scaling, EfficientNet [52] uses the composite coefficient technology to balance the ratio of the three dimensions of width, depth and image resolution. In addition, 7 versions of different scales have been developed, and experiments have shown that its performance exceeds most convolutional neural networks and is more efficient.

3.3 Deep Feature

The feature extraction based on deep learning is mainly carried out by the fully connected layer or convolutional layer. The model can extract the global features from the fully connected layer, or local features from the convolutional layer, or combine the two methods. Specifically, the way of feature fusion includes layer level and model level [65].

3.3.1 Deep Feature Selection. The convolution extracts local features, and the fully connection reassembles the previous local features into complete features through the weight matrix, thus representing the global features of the image. After the features extracted by the fully connected layer are reduced and standardized by PCA, the similarity between images can be measured. However, using fully connected layer features alone may limit image retrieval accuracy. Song et al. pointed out that establishing a direct connection between the first fully connected layer and the last one can achieve a coarse-to-fine improvement [49]. Furthermore, since the fully connected layers represent image-level features, they lack local geometric invariance. To this end, Song et al. also extract local features on a finer scale to solve the background clutter problem. Because the lack of geometric invariance will affect the robustness of features to image transformation, such as image cropping, occlusion and so on. To this end, researchers proposed to use intermediate convolutional layers to solve this problem [3, 44, 70].

Features are usually aggregated using pooling operations, where sum/average pooling and max pooling are the two simplest pooling methods. Pooling the features extracted by the convolutional layer can effectively reduce the number of parameters and enhance the robustness of feature representation. In addition, pooling methods such as R-MAC [53], CroW [22], SPoC [3] and GeM pooling [43] can also effectively improve the retrieval performance of image features.

3.3.2 Deep Feature Fusion. Feature fusion is to combine the strengths of different features to achieve complementary advantages. [33] merge multiple deep global features from different fully connected layers. Li et al. [29] applied the R-MAC coding scheme to the 5 convolutional layers of VGG-16 and concatenated them into multi-scale feature vectors. Wang et al. [58] selected all convolutional layers of VGG-16 to extract image feature representations to achieve multi-feature fusion, and this method is more robust than using only single-layer features.

In fine-grained image retrieval, in order to emphasize the decisive role of local features, Yu et al. used low-level features to refine the ranking results of high-level features instead of directly connecting multi-layer features. Through the mapping function, low-level features are used to measure the fine-grained similarity between the nearest neighbor images that have the same semantics as the query and the image. Gong et al. [15] proposed a multi-scale orderless pooling CNN, which extracts and encodes CNN features from different layers, and then connects the aggregated features of different layers to measure images. Li et al. [73] proposed a multi-layer orderless fusion (MOF) algorithm on the basis of multi-scale orderless pooling, and the experiments on the Holiday and UKBench datasets proved that the performance is better. Zhang et al. [28] fused the index matrix generated by two features extracted from the same CNN, which has low computational complexity. Yang et al. [66] gave up the two-stage retrieval and proposed a deep orthogonal local and global (DOLG) feature fusion framework for end-to-end image retrieval. The image retrieval performance of this method was verified on the Oxford and Paris datasets.

Fusing the features of different models requires the complementarity between the models. Simonyan et al. [48] introduced a fusion strategy within the convolution model, fusing VGG-16 and VGG-19 to improve the feature learning ability of VGG. Yang et al. [64] introduced dual-stream attention in CNN to achieve image retrieval. This method can calculate image similarity by retaining salient content and suppressing irrelevant regions like humans, and achieved strong image retrieval performance. Zheng et al. [74] believed that fusion between models can bridge the gap between intermediate and high-level features, so combined VGG-19 and AlexNet to learn combined features. Ge et al. [14] proposed a multi-level feature fusion method to improve the feature representation of high-resolution remote sensing image retrieval. Jiang et al. [21] proposed an image retrieval method based on image feature fusion and discrete cosine transform. They compared methods based on shallow feature fusion and deep feature fusion, and the experiments on Oxford dataset show that both methods can improve the performance of the retrieval system. According to the order of fusion and prediction, feature fusion can be divided into early fusion and late fusion. Among them, early fusion first fuses features, and then performs image retrieval on the fused unique feature representation [33, 64, 66]. Late fusion improves retrieval performance by combining retrieval results with different features [28].

4 CONCLUSION

This paper reviews the research progress of CBIR based on deep learning, expounds the connection between each method and summarizes the representative methods. CBIR based on deep learning has become a hot research direction at this stage. Researchers have produced a lot of innovative work and made great progress in retrieval accuracy and retrieval efficiency, but many new problems have also emerged. First of all, feature selection and extraction are the basis of CBIR. How to select appropriate features to reflect the semantics contained in images has always been the first problem in the past, present and future. In addition, in the face of the increase in the dimensionality of feature vectors brought about by feature fusion, dimension reduction technology is worthy of further study, because only low-dimensional and good discriminative features can guarantee retrieval performance and efficiency. How to use low to medium feature vector dimensions to express images

is still a big problem. Secondly, data-driven is one of the characteristics of deep learning. Specific retrieval tasks require specific datasets as benchmarks, and the introduction of various types of datasets has become an urgent need for researchers. At this stage, the CBIR method focuses on static datasets and is difficult to apply to incremental scenarios. With the increase of new data, how to make the trained system perform incremental learning is a problem worth considering. Finally, the ultimate goal of image retrieval is people-oriented, and how to use feedback technology to achieve user satisfaction with minimal iteration still needs further research [65].

ACKNOWLEDGMENTS

This work has been supported by the National Key R&D Program of China under Grant NO.2022YFC3302400 and the Key Laboratory of Digital Rights Services, which is one of the National Science and Standardization Key Labs for Publication Industry.

REFERENCES

[1] Jamil Ahmad, Khan Muhammad, Sambit Bakshi, and Sung Wook Baik. 2018. Object-oriented convolutional features for fine-grained image retrieval in large surveillance datasets. *Future Generation Computer Systems* 81 (2018), 314–330.

[2] Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, and Lucia Vadicamo. 2020. Large-scale instance-level image retrieval. *Information Processing & Management* 57, 6 (2020), 102100.

[3] Artem Babenko and Victor Lempitsky. 2015. Aggregating local deep features for image retrieval. In *Proceedings of the IEEE international conference on computer vision*. 1269–1277.

[4] Cong Bai, Hongkai Li, Jinglin Zhang, Ling Huang, and Lu Zhang. 2021. Unsupervised adversarial instance-level image retrieval. *IEEE Transactions on Multimedia* 23 (2021), 2199–2207.

[5] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. 2020. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9779–9788.

[6] Hamed Bonab, Mohammad Aliannejadi, Ali Vardasbi, Evangelos Kanoulas, and James Allan. 2021. Cross-market product recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 110–119.

[7] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S Yu. 2016. Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1445–1454.

[8] Ushasi Chaudhuri, Biplab Banerjee, Avik Bhattacharya, and Mihai Datcu. 2020. CMIR-NET: A deep learning based model for cross-modal retrieval in remote sensing. *Pattern recognition letters* 131 (2020), 456–462.

[9] Gang Chen, Xiang Cheng, Sen Su, and Chongmo Tang. 2020. Multiple-instance ranking based deep hashing for multi-label image retrieval. *Neurocomputing* 402 (2020), 89–99.

[10] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*. 1–9.

[11] Sounak Dey, Anjan Dutta, Suman K Ghosh, Ernest Valveny, Josep Lladós, and Umapada Pal. 2018. Learning cross-modal deep embeddings for multi-object image retrieval using text and sketch. In *2018 24th international conference on pattern recognition (ICPR)*. IEEE, 916–921.

[12] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. 2019. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2179–2188.

[13] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. 2010. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE transactions on visualization and computer graphics* 17, 11 (2010), 1624–1636.

[14] Yun Ge, Zihong Yang, Zihan Huang, and Famao Ye. 2021. A multi-level feature fusion method based on pooling and similarity for HRRS image retrieval. *Remote Sensing Letters* 12, 11 (2021), 1090–1099.

[15] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. 2014. Multi-scale orderless pooling of deep convolutional activation features. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*. Springer, 392–407.

[16] Chunbin Gu, Jiajun Bu, Xixi Zhou, Chengwei Yao, Dongfang Ma, Zhi Yu, and Xifeng Yan. 2022. Cross-modal image retrieval with deep mutual information maximization. *Neurocomputing* 496 (2022), 166–177.

[17] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. *Advances in neural information processing systems* 31 (2018).

[18] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris. 2019. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *arXiv preprint arXiv:1905.12794* 1, 2 (2019), 7.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[20] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2008. Hamming embedding and weak geometric consistency for large scale image search. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*. Springer, 304–317.

[21] DaYou Jiang and Jongweon Kim. 2021. Image retrieval method based on image feature fusion and discrete cosine transform. *Applied Sciences* 11, 12 (2021), 5701.

[22] Yannis Kalantidis, Clayton Mellina, and Simon Osindero. 2016. Cross-dimensional weighting for aggregated deep convolutional features. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I 14*. Springer, 685–701.

[23] Abhishek Kaushik, Nicolas Loir, and Gareth JF Jones. 2021. Multi-view conversational search interface using a dialogue-based agent. In *European Conference on Information Retrieval*. Springer, 520–524.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).

[25] Hanjiang Lai, Pan Yan, Xiangbo Shu, Yunchao Wei, and Shuicheng Yan. 2016. Instance-aware hashing for multi-label image retrieval. *IEEE Transactions on Image Processing* 25, 6 (2016), 2469–2479.

[26] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*. 201–216.

[27] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. 2018. Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4242–4251.

[28] Ying Li, Xiangwei Kong, Liang Zheng, and Qi Tian. 2016. Exploiting hierarchical activations of neural network for image retrieval. In *Proceedings of the 24th ACM international conference on Multimedia*. 132–136.

[29] Yang Li, Yulong Xu, Jiabao Wang, Zhuang Miao, and Yafei Zhang. 2017. Ms-rmac: Multiscale regional maximum activation of convolutions for image retrieval. *IEEE Signal Processing Letters* 24, 5 (2017), 609–613.

[30] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*. 801–809.

[31] Kevin Lin, Fan Yang, Qiaosong Wang, and Robinson Piramuthu. 2019. Adversarial learning for fine-grained image search. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 490–495.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[33] Yu Liu, Yanming Guo, Song Wu, and Michael S Lew. 2015. Deepindex for accurate and efficient image retrieval. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. 43–50.

[34] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151* (2013).

[35] Fanjie Meng, Dalong Shan, Ruixia Shi, Yang Song, Baolong Guo, and Weidong Cai. 2018. Merged region based image retrieval. *Journal of Visual Communication and Image Representation* 55 (2018), 572–585.

[36] Eva Mohedano, Kevin McGuinness, Noel E O'Connor, Amaia Salvador, Ferran Marques, and Xavier Giró-i Nieto. 2016. Bags of local convolutional features for scalable instance search. In *Proceedings of the 2016 ACM on international conference on multimedia retrieval*. 327–331.

[37] David Nister and Henrik Stewenius. 2006. Scalable recognition with a vocabulary tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. Ieee, 2161–2168.

[38] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. 2019. Generalising fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 677–686.

[39] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 1–8.

[40] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 1–8.

[41] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.

[42] Yonggang Qi, Yi-Zhe Song, Honggang Zhang, and Jun Liu. 2016. Sketch-based image retrieval via siamese convolutional neural network. In *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2460–2464.

[43] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2018. Fine-tuning CNN image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence* 41, 7 (2018), 1655–1668.

[44] Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. 2016. Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications* 4, 3 (2016), 251–258.

[45] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. 2021. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8504–8513.

[46] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–12.

[47] Gaurav Sharma and Bernt Schiele. 2015. Scalable nonlinear embeddings for semantic category-based image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. 1296–1304.

[48] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[49] Jingkuan Song, Tao He, Lianli Gao, Xing Xu, and Heng Tao Shen. 2017. Deep region hashing for efficient large-scale instance search from images. *arXiv preprint arXiv:1701.07901* (2017).

[50] Gencer Sumbul, Markus Müller, and Begüm Demir. 2022. A Novel Self-Supervised Cross-Modal Image Retrieval Method in Remote Sensing. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2426–2430.

[51] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[52] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[53] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2015. Particular object retrieval with integral max-pooling of CNN activations. *arXiv preprint arXiv:1511.05879* (2015).

[54] Pablo Torres and Jose M Saavedra. 2021. Compact and effective representations for sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2115–2123.

[55] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

[56] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1386–1393.

[57] Luo Wang, Xueming Qian, Yuting Zhang, Jialie Shen, and Xiaochun Cao. 2019. Enhancing sketch-based image retrieval by cnn semantic re-ranking. *IEEE transactions on cybernetics* 50, 7 (2019), 3330–3342.

[58] Qi Wang, Jinxiang Lai, Zhenguo Yang, Kai Xu, Peipei Kan, Wenyin Liu, and Liang Lei. 2019. Improving cross-dimensional weighting pooling with multi-scale feature fusion for image retrieval. *Neurocomputing* 363 (2019), 17–26.

[59] Yanfei Wang, Fei Huang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. 2020. Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval. *Pattern recognition* 100 (2020), 107148.

[60] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5764–5773.

[61] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2575–2584.

[62] Lingxi Xie, Jingdong Wang, Bo Zhang, and Qi Tian. 2015. Fine-grained image search. *IEEE Transactions on multimedia* 17, 5 (2015), 636–647.

[63] Dan Xu, Xavier Alameda-Pineda, Jingkuan Song, Elisa Ricci, and Nicu Sebe. 2018. Cross-paced representation learning with partial curricula for sketch-based image retrieval. *IEEE Transactions on Image Processing* 27, 9 (2018), 4410–4421.

[64] Fei Yang, Jia Li, Shikui Wei, Qinjie Zheng, Ting Liu, and Yao Zhao. 2017. Two-stream attentive CNNs for image retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*. 1513–1521.

[65] Hui Yang and Shuicai Shi. 2023. Survey of Research on Content-Based Image Retrieval Technology. In *Software Guide*.

[66] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. 2021. Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *Proceedings of the IEEE/CVF International conference on Computer Vision*. 11772–11781.

[67] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. 2018. A zero-shot framework for sketch based image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 300–317.

[68] Qian Yu, Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. 2021. Fine-grained instance-level sketch-based image retrieval. *International Journal of Computer Vision* 129 (2021), 484–500.

[69] Yifei Yuan and Wai Lam. 2021. Conversational fashion image retrieval via multiturn natural language feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 839–848.

[70] Joe Yue-Hei Ng, Fan Yang, and Larry S Davis. 2015. Exploiting local features from deep networks for image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 53–61.

[71] Ruiyi Zhang, Tong Yu, Yilin Shen, Hongxia Jin, and Changyou Chen. 2019. Text-based interactive recommendation via constraint-augmented reinforcement learning. *Advances in neural information processing systems* 32 (2019).

[72] Ruiyi Zhang, Tong Yu, Yilin Shen, Hongxia Jin, Changyou Chen, and Lawrence Carin. 2020. Reward constrained interactive recommendation with natural language feedback. *arXiv preprint arXiv:2005.01618* (2020).

[73] Zhizhong Zhang, Yuan Xie, Wensheng Zhang, and Qi Tian. 2019. Effective image retrieval via multilinear multi-index fusion. *IEEE Transactions on Multimedia* 21, 11 (2019), 2878–2890.

[74] Liang Zheng, Yali Zhao, Shengjin Wang, Jingdong Wang, and Qi Tian. 2016. Good practice in CNN feature transfer. *arXiv preprint arXiv:1604.00133* (2016).

[75] Xiawu Zheng, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, Feiyue Huang, and Yanhua Yang. 2018. Centralized Ranking Loss with Weakly Supervised Localization for Fine-Grained Object Retrieval.. In *IJCAI*. 1226–1233.

[76] Xiawu Zheng, Rongrong Ji, Xiaoshuai Sun, Baochang Zhang, Yongjian Wu, and Feiyue Huang. 2019. Towards optimal fine grained retrieval via decorrelated centralized loss with normalize-scale layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9291–9298.