IndexCLR: Class-Aware Contrastive Learning for Similarity Indexing

Ovais Rashid Khan
Department of Computer Science, Islamic University of Science and Technology, Kashmir, India
Khanovais.r@gmail.com (corresponding author)

Javaid Iqbal Bhat
Department of Computer Science, Islamic University of Science and Technology, Kashmir, India
Javaid.iqbal@iust.ac.in

Received: 29 September 2025 | Revised: 10 November 2025 and 19 November 2025 | Accepted: 21 November 2025

 DOI: https://doi.org/10.48084/etasr.15220

ABSTRACT

Similarity metrics and indexing are critical components of image similarity searches since they directly impact the accuracy and efficiency of retrieving similar imagery. Similarity measurements, such as Euclidean distance or cosine similarity, evaluate how closely images are compared based on simple differences in their features. More complex methods, such as contrastive loss, enhance the representation of images by emphasizing relationships in a contextual reference to other related samples. This paper introduces a framework for image similarity and indexing that utilizes class-based representations in contrastive learning to enhance retrieval. This framework uses class-based representations for context-related awareness instead of sample-level contrastive learning to refine feature representations using a contrastive learning approach. This approach reduces computational complexity and mitigates sample selection challenges while refining the similarity comparison operation of similar images using class contextual information. A bottleneck mechanism is introduced to compress high-dimensional feature spaces into compact, lower-dimensional embeddings, preserving critical semantic information while minimizing redundancy. Discrimination is further enhanced through class-based contrastive training.

Keywords-image retrieval; image similarity; contrastive learning; class-based representations; indexing; dimensionality reduction

I. INTRODUCTION

Content-based Image Retrieval (CBIR) frameworks compare images and retrieve similar ones from large databases. The similarity is based on their visual characteristics rather than relying on textual descriptions or metadata labeled by humans. These systems use both low-level and high-level features for image representation. Low-level features such as color, texture, shape, and edge information, capture basic visual properties, whereas high-level semantic features capture more complex patterns. Human perception of abstract details is captured by high-level characteristics. CBIR systems compare these features across a database to rank and retrieve the most similar images based on a query image, improving the efficiency and accuracy of the retrieval process. In recent years, a considerable number of review papers have offered comprehensive analyses of the fundamental concepts and emerging challenges in this field [1-3]. Emerging trends highlight that CBIR systems have found applications across sectors in the ever-evolving digital landscape with rapid implementation in real environments. These systems are widely used in fields such as communication, remote sensing, document assessment for watermark extraction, and clinical research for disease detection, illustrating their versatility in handling large-scale image datasets. As digital imaging technologies become more common and innovative human–computer interfaces continue to develop, the requirement for efficient and accurate retrieval systems becomes more important, highlighting the role of CBIR systems in analyzing and exploring huge image libraries.

Early CBIR systems were developed primarily using manually designed feature extraction techniques like Scale-Invariant Feature Transform (SIFT) [4], Histogram of Oriented Gradients (HOG) [5], and color histograms [6]. These methods were developed to capture specific visual features within images, allowing the computation of visual similarities based on these characteristics. Human-designed approaches struggle with large-scale datasets or widely varying visual details across images, often leading to inconsistent performance, whereas adaptation across datasets remains limited.

Machine learning methods have significantly enhanced the functionality of CBIR systems. Machine learning–based CBIR frameworks have the ability to analyze labeled datasets, determine subtle trends, and enhance their potential to determine discriminative features crucial for efficient image retrieval using techniques like Support Vector Machines (SVMs) and decision trees, rather than relying on strict rules.

Machine learning enables retrieval frameworks to adopt adaptive approaches and provides the flexibility to advance beyond static workflows by prioritizing features that improve retrieval accuracy. These systems have become significantly more suitable for determining inherent patterns and relationships in unlabeled datasets through unsupervised techniques like k-Nearest Neighbors (k-NN). CBIR systems have achieved significant improvements in retrieval accuracy and computational efficiency by implementing these machine learning techniques, which have allowed them to move from relying on manually designed features to self-directed learning and adaptation for increasingly complex visual inputs. These advancements have made it possible for CBIR systems to handle increasingly complex and diverse datasets, increasing their usefulness and efficiency in real-world situations.

II. RELATED WORK

CBIR frameworks have evolved to develop techniques that identify the discriminative characteristics of images. Fundamental characteristics include color, texture, shape, and pattern orientation, which are critical for understanding the relationships and patterns within image data, allowing systems to categorize and retrieve images based on their visual composition. Determining distinctive visual features in CBIR systems has been a primary research objective, with an emphasis on local descriptors such as SIFT, HOG [7], and Local Binary Patterns (LBP), which have been widely adopted to establish visual semantics for similarity [8]. Machine learning techniques provide significant advantages for image retrieval tasks, particularly in extracting inherent features from visual content. These models are exceptionally effective at determining fundamental characteristics and patterns that are essential for semantic description. Even in resource-constrained contexts, machine learning models are more adaptable and efficient than traditional CBIR techniques.

Studies have focused on enhancing the representational power of feature embeddings through machine learning, particularly by employing feature fusion of Gray-Level Co-occurrence Matrix (GLCM) and texture-based LBP variants to generate richer and more discriminative feature descriptors [9]. Enhancements to the traditional bag-of-words framework have been achieved by combining it with SVMs, allowing feature embeddings to be refined into more semantically meaningful forms. Such strategies illustrate that both feature fusion and machine learning–driven refinement strengthen the descriptive capacity of embeddings, reduce the semantic gap, and improve the overall effectiveness of CBIR systems [10].

Recent strategies in CBIR have focused on combining Convolutional Neural Networks (CNNs) with sparse representation techniques. Unlike traditional handcrafted feature methods, CNNs are capable of extracting deep hierarchical embeddings that effectively capture the semantic content of images, whereas sparse representation enhances the compactness and discriminative power of these embeddings. This integration not only improves retrieval accuracy but also results in more robust performance compared to conventional feature extraction approaches [11]. Furthermore, individual models often produce distinct feature embeddings and representations, each capturing different aspects of image content. By integrating multiple models, these complementary strengths can be leveraged to create more comprehensive representations, enhancing robustness and narrowing the semantic gap in CBIR. Such hybrid strategies have been shown to significantly improve retrieval effectiveness in diverse and complex datasets [12]. Another approach fine-tunes pre-trained deep neural networks using transfer learning on target datasets, producing embeddings that capture dataset-specific patterns while reducing training time and enabling effective operation with smaller or newly introduced models, making this strategy suitable for scenarios with limited computational resources or smaller datasets [13].

The length and dimensionality of feature embeddings are particularly critical in hashing-based methods. Compact embeddings, such as binary hash codes, reduce storage requirements and enable faster similarity search, whereas preserving discriminative and semantic information remains essential. The auto-encoding twin-bottleneck hashing method [14] compresses image representations into compact binary codes through a twin-bottleneck architecture, maintaining semantic and structural fidelity despite reduced dimensionality. Similarly, DistillHash [15] produces low-dimensional hash embeddings by distilling pairwise similarities in an unsupervised manner, ensuring that even short codes retain discriminative power. These studies illustrate that carefully controlling embedding length in hashing methods balances retrieval efficiency and accuracy, making it a key consideration in CBIR systems.

The integration of human expertise into CBIR systems provides an additional layer of refinement. Human input brings subjective evaluation and contextual understanding that automated algorithms may lack, improving the precision of search results [16]. The synergy between human input and automated systems enables iterative feedback processes that progressively optimize algorithmic precision [17, 18]. Such collaborative frameworks not only elevate the reliability and speed of image retrieval tools but also strengthen their adaptability to evolving real-world demands [19-21].

III. PROPOSED MODEL

A neural module is integrated into the initial phase of the framework to perform comprehensive feature extraction, enabling the recognition and encoding of patterns that span multiple levels of abstraction within the input data (Figure 1). In this stage, a CNN is employed for feature extraction. The CNN uses hierarchical layers of filters to extract semantically rich features, with each filter encapsulating high-level conceptual information essential for later-stage computational process. To bolster this feature extraction framework, a Convolutional Attention Block (CAB) is embedded within the system's design. CAB prioritizes the most influential regions while downplaying the less relevant portions, helping to extract more discriminative features. This not only enhances performance but is also computational efficient for downstream tasks. Concurrently, a spatial attention mechanism pinpoints crucial spatial regions within feature maps by producing attention maps derived through convolutional operations.

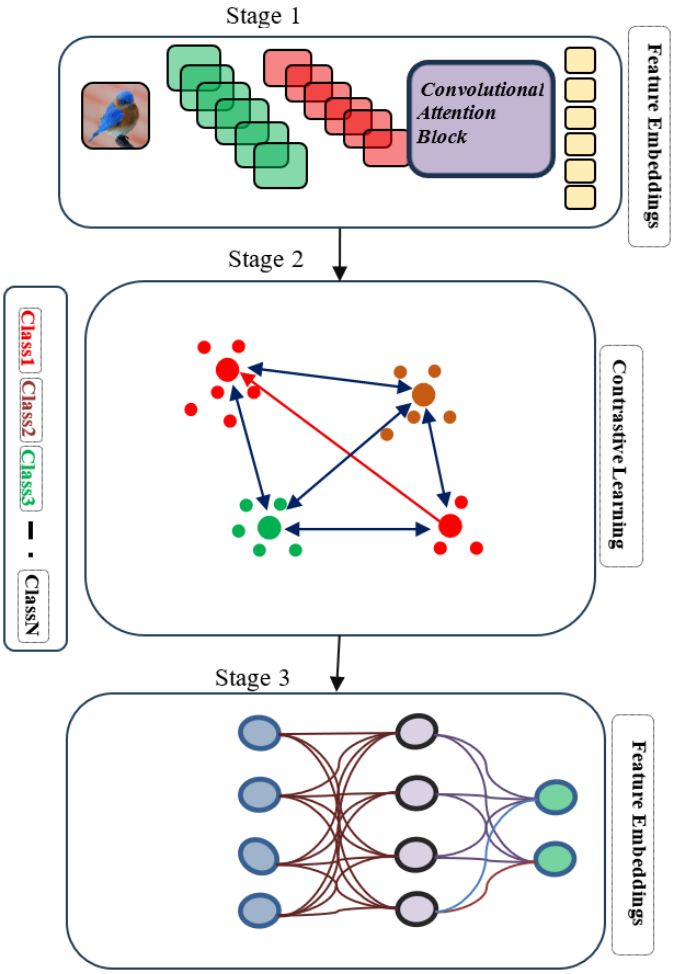

Fig. 1. Proposed model architecture.

These two attention-driven adjustments, channel reweighting and spatial focusing, are computationally fused, and the resulting adjusted feature maps are then combined with the original representations to augment the discriminative capacity of the features. The CNN's hierarchical feature abstraction and the CAB's context-aware refinement generate a robust, refined feature hierarchy that supports subsequent processing stages, ensuring operational efficacy and adaptability across heterogeneous tasks and datasets.

Architectures such as VGG19, which leverage CNNs, produce image embeddings that prove exceptionally useful for tasks such as image ranking or identifying similar visuals in retrieval systems. Through layered feature extraction and analysis, these systems achieve an in-depth grasp of visual content, closely mimicking the way humans process and interpret visual information. This hierarchical representation captures both low-level details and high-level semantic structures, making it well suited for accurately measuring image similarity and ranking images based on relevance. However, such models demonstrate high efficiency and accuracy in higher-dimensional feature spaces, whereas their performance deteriorates significantly in lower-dimensional representations. In CBIR systems, particularly in resource-constrained environments, it is crucial to achieve effective representation and retrieval using lower-dimensional feature embeddings. To enhance the discriminative power of feature embeddings in lower-dimensional spaces while maintaining efficiency, an additional phase is introduced.

In this phase, further refinement of the feature embeddings is implemented through contrastive learning to boost feature representation. Contrastive learning is a technique for learning meaningful feature representations by focusing on similarities and differences in data, and it can be effectively extended to a class-based framework. The class-based contrastive learning model operates at a higher level by leveraging class-level information, in contrast to sample-based methods that compare individual positive and negative samples. During this learning process, each class is represented by its centroid, and the model is trained to minimize the distance between image samples from the same class while maximizing the distance between samples from other classes. This approach enhances the framework's ability to generate discriminative feature representations that capture class-specific information. The class-based method offers greater computational efficiency by reducing the need to process numerous individual samples pairs, making it a scalable solution, particularly in scenarios where labeled data are available. The contrastive phase organizes the embedding space by forming distinct clusters, bringing similar instances closer together while pushing dissimilar ones farther apart. This process enhances the structure of feature representations, leading to a more well-defined and discriminative embedding space. Through comprehensive experimentation on the embedding space derived from contrastive learning, the results demonstrate some level of improvement in feature representations. However, these improvements are inadequate to significantly enhance the model's overall performance. Performance improves substantially when a neural network layer is incorporated on top of the contrastive learning architecture. The model's capacity to generalize is further improved, and the framework's overall efficacy is enhanced by the supplementary layer, which provides a deeper level of abstraction and fine-tunes the embedding space to boost its discriminative strength.

IV. RESULTS AND DISCUSSION

In this research, we performed experiments to assess the impact of intermediate embedding dimensions on model performance. The study involved evaluating the model's performance, with a particular emphasis on lower-dimensional configurations, whereas experiments were conducted to assess how reducing the dimensionality of embeddings impacts the model's ability to generalize. We evaluated the proposed framework on two datasets, including Caltech-101 [22], which has 9,149 images of 102 categories, and Corel-1000 [23], a more compact dataset with 1,000 images of 10 categories. The evaluation focused on analyzing lower-dimensional vector features in terms of retrieval precision and feature representation capabilities.

Experiments show differences in model performance when using different embedding sizes on datasets with varying class distributions. When employing 4-dimensional embeddings, models trained on datasets with limited class diversity exhibit suboptimal performance due to insufficient representational

capacity, whereas those applied to datasets with extensive class diversity also experience reduced efficacy due to inadequate discrimination among densely populated classes. These trends, summarized in Table I and shown in Figure 2, highlight distinct performance differences across the Corel-1000 and Caltech-101 datasets under varying embedding dimensions (4, 8, and 16). Increasing embeddings to 8 dimensions substantially enhances model performance, promising better discriminative representation of classes for generalization.

TABLE I. EMBEDDING-WISE AVERAGE PRECISION OF THE PROPOSED MODEL ON COREL-1000 AND CALTECH-101 DATASETS

Embedding	Corel-1000 (%)	Caltech-101 (%)
4	60	56
8	80	90
16	97	95

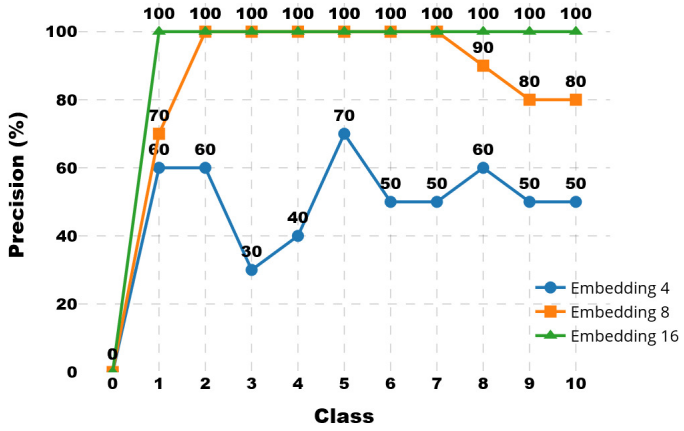

Fig. 2. Class-wise precision on Corel-1000 for embeddings 4, 8, and 16.

Furthermore, accuracy and robustness are enhanced by raising the number of embeddings to 16, demonstrating a scalable correlation between embedding dimensionality and model efficacy. Analysis of results highlights the importance of optimizing embedding dimensions to achieve both discriminative efficiency and computational efficacy, particularly when addressing datasets with varying levels of class complexity, as shown in Figure 3.

The performance of the proposed framework is validated through the experimental findings summarized in the comparison of models in Table II and illustrated in Figure 4, both of which correspond to evaluations conducted on the Corel-1000 dataset. These results demonstrate the strong retrieval capability and consistent precision of the proposed model compared to existing approaches. As shown in Figure 4, the proposed framework achieved a mean precision of 97% on the Corel-1000 dataset, surpassing previously reported methods in [24] (94.35%) and [25] (96.12%). In contrast, approaches proposed by authors in [16], [17], and [18] attained comparatively lower precision values ranging from 84.39% to 87.58%. These results highlight the improved discriminative representation and retrieval performance achieved by the proposed model in CBIR tasks.

TABLE II. COMPARISON OF AVERAGE PRECISION OF THE PROPOSED MODEL AND BENCHMARK METHODS ON COREL-1000 DATASET

Methods	Average Precision (%)
Proposed model	97.00
[24]	94.35
[25]	96.12
[16]	87.30
[11]	89.58
[18]	84.39
[17]	87.58

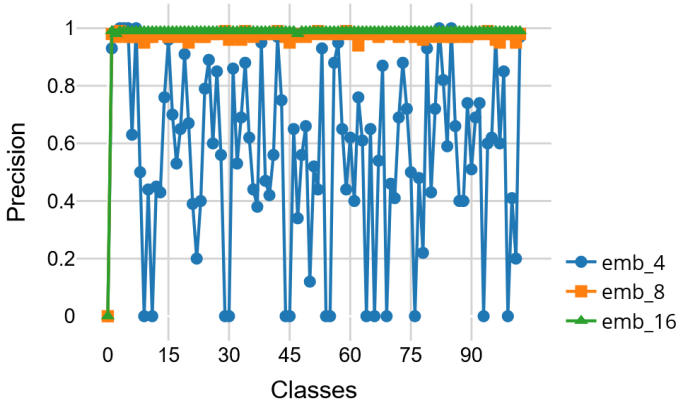

Fig. 3. Class-wise precision on Caltech-101 for embeddings 4, 8, and 16.

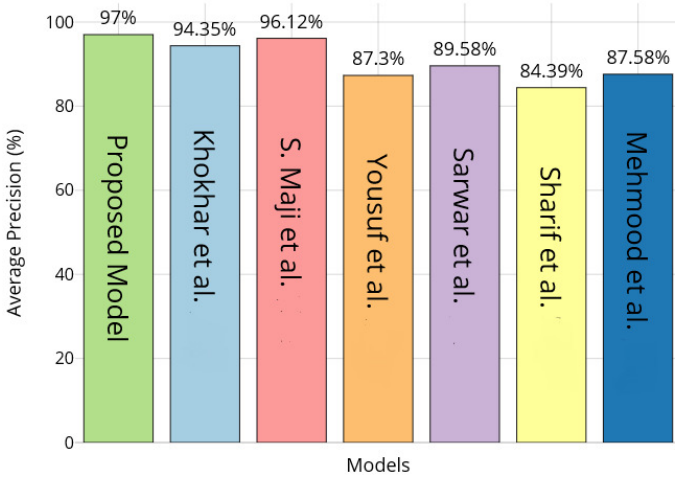

Fig. 4. Comparison of average precision on Corel-1000 for the proposed model and benchmark methods.

Similarly, Figure 5 presents the evaluation outcomes on the Caltech-101 dataset, where the proposed framework attained a mean precision of 95%. The results indicate competitive performance with a benchmark method [25], while maintaining stable retrieval accuracy across categories. The consistent performance on both datasets underscores the robustness and generalization capability of the proposed framework, confirming its potential for effective deployment in real-world CBIR applications.

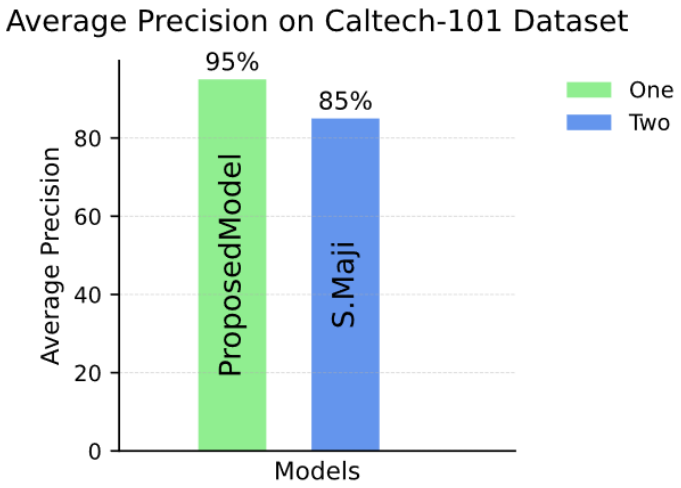

Fig. 5. Comparison of average precision on Caltech-101 for the proposed model and a benchmark method.

V. CONCLUSION

The development and evaluation of an experimental Content-based Image Retrieval (CBIR) framework employing contrastive learning to enhance similarity measurement in reduced-dimensional embedding spaces are presented in this study. The work focuses on improving the alignment between learned representations and semantic image relationships, thereby enhancing retrieval accuracy within compressed feature domains. To strengthen the discriminative capacity of feature embeddings, the framework utilizes a contrastive loss optimization strategy to train deep neural network models, enabling them to capture meaningful distinctions between data points while projecting them into lower-dimensional representations.

A key novelty of the proposed framework is the shift from traditional sample-level contrastive learning to class-level embedding contrastive learning. This design significantly reduces the complexity associated with the conventional contrastive paradigm, which often suffers from the N-pair problem due to a large number of sample pairs. By contrasting class embeddings rather than individual samples, the approach lowers training overhead while still achieving strong feature discrimination. This contributes to more efficient learning without compromising semantic separability.

The experimental assessment prioritizes critical retrieval performance measures, such as classification accuracy and average precision, using standardized datasets like Caltech-101 and Corel-1000. The evaluation explores how effectively reduced-dimensional embeddings preserve meaningful distinctions and how dimensional compression interacts with representation strength in CBIR systems. The findings demonstrate that lower-dimensional embeddings, when trained with an appropriate class-level contrastive strategy, can retain high discriminative power while significantly reducing computational cost, memory consumption, and inference time. This balance between dimensionality reduction and semantic richness is essential for developing retrieval systems that are both efficient and accurate, especially in large-scale visual search applications.

REFERENCES

[1] I. M. Hameed, S. H. Abdulhussain, and B. M. Mahmmod, "Content-based image retrieval: A review of recent trends," *Cogent Engineering*, vol. 8, no. 1, Jan. 2021, Art. no. 1927469, https://doi.org/10.1080/23311916.2021.1927469.

[2] S. R. Dubey, "A Decade Survey of Content Based Image Retrieval Using Deep Learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2687–2704, May 2022, https://doi.org/10.1109/TCSVT.2021.3080920.

[3] A. Latif *et al.*, "Content-Based Image Retrieval and Feature Extraction: A Comprehensive Review," *Mathematical Problems in Engineering*, vol. 2019, no. 1, Aug. 2019, Art. no. 9658350, https://doi.org/10.1155/2019/9658350.

[4] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004, https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp. 886–893 vol. 1, https://doi.org/10.1109/CVPR.2005.177.

[6] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, Nov. 1991, https://doi.org/10.1007/BF00130487.

[7] Y. Rashid and J. I. Bhat, "OlapGN: A multi-layered graph convolution network-based model for locating influential nodes in graph networks," *Knowledge-Based Systems*, vol. 283, Jan. 2024, Art. no. 111163, https://doi.org/10.1016/j.knosys.2023.111163.

[8] F. Baig *et al.*, "Boosting the Performance of the BoVW Model Using SURF–CoHOG-Based Sparse Features with Relevance Feedback for CBIR," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, vol. 44, no. 1, pp. 99–118, Mar. 2020, https://doi.org/10.1007/s40998-019-00237-z.

[9] Vimina E. R. and Divya M. O., "Maximal multi-channel local binary pattern with colour information for CBIR," *Multimedia Tools and Applications*, vol. 79, no. 35, pp. 25357–25377, Sept. 2020, https://doi.org/10.1007/s11042-020-09207-8.

[10] M. Garg and G. Dhiman, "A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants," *Neural Computing and Applications*, vol. 33, no. 4, pp. 1311–1328, Feb. 2021, https://doi.org/10.1007/s00521-020-05017-z.

[11] A. Sarwar, Z. Mehmood, T. Saba, K. A. Qazi, A. Adnan, and H. Jamal, "A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine," *Journal of Information Science*, vol. 45, no. 1, pp. 117–135, Feb. 2019, https://doi.org/10.1177/0165551518782825.

[12] Y. Rashid and J. I. Bhat, "Topological to deep learning era for identifying influencers in online social networks :a systematic review," *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 14671–14714, Feb. 2024, https://doi.org/10.1007/s11042-023-16002-8.

[13] A. Sezavar, H. Farsi, and S. Mohamadzadeh, "Content-based image retrieval by combining convolutional neural networks and sparse representation," *Multimedia Tools and Applications*, vol. 78, no. 15, pp. 20895–20912, Aug. 2019, https://doi.org/10.1007/s11042-019-7321-1.

[14] S. Sikandar, R. Mahum, and A. Alsalman, "A Novel Hybrid Approach for a Content-Based Image Retrieval Using Feature Fusion," *Applied Sciences*, vol. 13, no. 7, Apr. 2023, Art. no. 4581, https://doi.org/10.3390/app13074581.

[15] M. A. Mohammed, Z. A. Oraibi, and M. A. Hussain, "Content based Image Retrieval using Fine-tuned Deep Features with Transfer Learning," in *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering*, Banda Aceh, Indonesia, 2023, pp. 108–113, https://doi.org/10.1109/COSITE60233.2023.10249430.

[16] M. Yousuf *et al.*, "A Novel Technique Based on Visual Words Fusion Analysis of Sparse Features for Effective Content-Based Image Retrieval," *Mathematical Problems in Engineering*, vol. 2018, no. 1, Mar. 2018, Art. no. 2134395, https://doi.org/10.1155/2018/2134395.

[17] Z. Mehmood, T. Mahmood, and M. A. Javid, "Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine," *Applied Intelligence*, vol. 48, no. 1, pp. 166–181, Jan. 2018, https://doi.org/10.1007/s10489-017-0957-5.

[18] U. Sharif, Z. Mehmood, T. Mahmood, M. A. Javid, A. Rehman, and T. Saba, "Scene analysis and search using local features and support vector machine for effective content-based image retrieval," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 901–925, Aug. 2019, https://doi.org/10.1007/s10462-018-9636-0.

[19] Y. Shen *et al.*, "Auto-Encoding Twin-Bottleneck Hashing," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 2815–2824, https://doi.org/10.1109/CVPR42600.2020.00289.

[20] E. Yang, T. Liu, C. Deng, W. Liu, and D. Tao, "DistillHash: Unsupervised Deep Hashing by Distilling Data Pairs," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 2941–2950, https://doi.org/10.1109/CVPR.2019.00306.

[21] Y. Rao, W. Liu, B. Fan, J. Song, and Y. Yang, "A novel relevance feedback method for CBIR," *World Wide Web*, vol. 21, no. 6, pp. 1505–1522, Nov. 2018, https://doi.org/10.1007/s11280-017-0523-4.

[22] F.-F. Li, M. Andreeto, M. Ranzato, and P. Perona, "Caltech 101." CaltechDATA, Apr. 06, 2022, https://doi.org/10.22002/D1.20086.

[23] "Corel-1K, Corel-5K, Corel-10K." Kaggle. [Online]. Available: https://www.kaggle.com/datasets/amirhosseinroodaki/corel-1k-corel-5k-and-corel-10k-datasets.

[24] S. Khokhar and S. Verma, "Content Based Image Retrieval with Multi-Feature Classification by Back-propagation Neural Network," *International Journal of Computer Applications Technology and Research*, vol. 6, no. 7, pp. 278–284, July 2017, https://doi.org/10.7753/IJCATR0607.1002.

[25] S. Maji and S. Bose, "CBIR Using Features Derived by Deep Learning," *ACM/IMS Transactions on Data Science*, vol. 2, no. 3, Sept. 2021, Art. no. 26, https://doi.org/10.1145/3470568.