ISNNI: 3007-3391 HEC Category: Y

Journal of Computers and Intelligent Systems
Vol. 4, No. 3, 2026, pages 307 - 329
Journal Homepage:
https://journals.iub.edu.pk/index.php/JCIS/

Content-Based Image Retrieval Using Self-Supervised Contrastive Learning with Hybrid Backbone

Nazia Maryam[1], Saima Munawar[2], Nasir Naveed[3]
[1]Department of Computer Science, Virtual university of Pakistan

ARTICLE INFO

Article History:

Received:	July	05, 2026
Revised:	August	01, 2026
Accepted:	September	10, 2026
Available Online:	September	30, 2026

Keywords:

Content-based image retrieval (CBIR)
convolutional neural networks (CNNs)
contrastive learning
EfficientNet-B0
Facebook AI Similarity Search (FAISS)
learnable feature fusion
MongoDB
ResNet18
self-supervised learning (SSL)
SimCLR

Classification Codes:

Funding:

This research received no specific grant from any funding agency in the public or not-for-profit sector.

ABSTRACT

The exponential growth of digital image collections has driven the requirement for more efficient content-based image retrieval (CBIR) systems. Traditional methods relying on handcrafted features often fail to capture complex image semantics, hence shifting the paradigm towards utilizing CNN (convolutional neural network) based solutions along with SSL (self-supervised learning which automatically extract high-dimensional features. However, when dealing with vast and diverse datasets, retrieval performance is still not at its best.

In this study, a hybrid CBIR model that integrates pre-trained EfficientNet-B0 and ResNet18 backbones within a SimCLR-based contrastive learning framework to improve feature extraction. A learnable fusion layer is introduced to adaptively compress and emphasize the most informative dimensions across both networks. Extracted embeddings from all images are stored in a MongoDB database, allowing scalable and efficient search. Upon receiving a query image, the system uses the trained hybrid model to extract its features and performs fast similarity search using Facebook AI Similarity Search (FAISS). Retrieved results are displayed via a web interface.

The proposed model is performing well in terms of efficiency and effectiveness, scoring an average mean precision of 96% for the custom multi-class dataset and 98.03 % for the Corel-1K benchmark dataset following training through them without labels. The combination of hybrid backbone fusion, self-supervised contrastive learning, and scalable database management clearly reflects the ability of the system to provide image search efficiently.

OPEN ACCESS

© 2026 The authors published by JCIS. This is an Open Access Article under the Creative Common Attribution Non-Commercial 4.0

Corresponding Author's Email: maryammehmoodjarral@gmail.com

1 Introduction

The explosive growth of digital images in areas like e-commerce, education, healthcare, and social media has generated a persistent need for smart technologies that would be able to search and retrieve such images efficiently and effectively. This problem is addressed by Content Based Image Retrieval (CBIR) method where one can search images according to their visual attributes such as colour [Liu and Yang, 2013], texture [Singha and Hemachandran, 2012], shape [Ahmed et al., 2019], and structure [Deb and Zhang, 2004], rather than relying solely on metadata or keywords. However, the conventional CBIR methods face the problem of the semantic gap, which is the difference between the

high-level semantic interpretation of images by human users and the low-level visual features extracted by computer algorithms [Guo et al., 2008].

Recent advances in deep learning have enabled Convolutional Neural Networks (CNNs) [Mathew et al., 2020], to reduce the semantic gap by automatically learning hierarchical feature representations that capture high-level semantic information. This allows CBIR systems to retrieve images based on semantic content rather than relying solely on handcrafted low-level visual features, thereby improving retrieval accuracy. However, the performance of these models largely depends on the availability of large scale annotated data set, whose collection and annotation are both expensive and labour-intensive.

Self-supervised learning (SSL) has emerged as a promising alternative to supervised learning because it learns meaningful feature representations from unlabelled data, thereby reducing the dependence on large annotated datasets [Liu et al., 2021]. Self-supervised learning enables models to learn discriminative feature representations without requiring manual annotations. Instead, pseudo-labels are generated automatically from augmented versions of unlabelled images. Among SSL approaches that has attracted a lot of attention is contrastive learning [Wang and Qi, 2022]. In order to help models learn semantically meaningful representations, it trains them to push apart dissimilar image pairs and bring comparable image pairs closer together in the embedding space. A prominent contrastive learning approach called SimCLR [Chen et al., 2020] improves representation learning without supervision by using effective data augmentations and a projection head.

Although SimCLR has demonstrated notable performance in learning discriminative feature representations from unlabelled images, retrieval performance still depends heavily on the capability of the backbone network to extract informative visual features [Chen et al., 2020]. A single backbone may not capture all complementary visual characteristics, as different network architectures emphasize different levels of image information. Therefore, relying on a single backbone can limit the discriminative power of image embeddings, particularly for large-scale and visually diverse image retrieval tasks [Elharrouss et al., 2024].

To address this limitation, this study proposes a Content-Based Image Retrieval (CBIR) framework that integrates pre-trained EfficientNet-B0 and ResNet18 as hybrid backbones within the SimCLR framework for contrastive learning. The two backbones were selected because they learn complementary visual representations: EfficientNet-B0 effectively captures global semantic information, while ResNet18 preserves fine-grained local structural features through residual learning. By fusing these complementary representations, the proposed framework generates richer and more discriminative image embeddings, leading to improved retrieval precision and efficient retrieval on diverse image datasets. The concatenated features are passed through a fully connected fusion layer to learn a compact and discriminative joint representation. Unlike fixed linear dimensionality reduction techniques which may combine features through direct concatenation or predefined fusion strategies, the proposed framework employs a learnable fusion mechanism to transform and combine the feature representations extracted from ResNet18 and EfficientNet-B0. The fusion parameters are optimized jointly with the CNN backbones through a self-supervised contrastive learning objective. This enables the model to learn a task-specific combined representation rather than relying on a fixed feature combination. Furthermore, the proposed framework uses SimCLR-based contrastive learning, in which positive and negative samples are constructed from augmented views within the mini-batch without using category labels during training. The combination of complementary CNN representations, trainable feature fusion, and self-supervised contrastive representation learning constitutes the main methodological contribution of this work for CBIR. The effectiveness of this design choice is further validated through the ablation study presented in Experimental Evaluation Section. The image embeddings and datasets are saved on MongoDB. The FAISS algorithm is used to perform approximate nearest neighbour searches for fast retrieval.

The major contributions of this research can be summarized as:

- Hybrid Backbone Architecture: Introduced a dual-backbone feature extractor combining ResNet18 and EfficientNet-B0 within a SimCLR-based contrastive learning framework for enhanced feature representation.
- Learnable Feature Fusion Layer: Designed a fully connected fusion layer to merge heterogeneous features and reduce dimensionality, preserving the most discriminative components.
- Scalable Object-Based Storage (MongoDB): Implemented Facebook AI Similarity Search (FAISS) for fast and scalable approximate nearest neighbour search, allowing for rapid image search through the use of large data sets.

- Fast and Scalable Similarity Search Using FAISS: Integrated Facebook AI Similarity Search (FAISS) to enable end-to-end approximate nearest neighbour retrieval, ensuring rapid and scalable image search across large datasets.
- Self-Supervised Learning for CBIR: Improved the generalizability and applicability of the retrieval model by training the retrieval model without labelled data using SimCLR.
- Quantitative Evaluation and Retrieval Performance: The proposed model has obtained average mean precision of 96% for multiclass data set and 98.03% for Corel-1k data set. In addition, query latency was found to be 0.072 seconds per query for this proposed retrieval system.

2 Related Work

Content-Based Image Retrieval (CBIR) has evolved significantly since its inception in the late 1970s. Earlier CBIR systems used metadata-based approaches that depended on manually written tags and textual descriptions to retrieve images [Wilkins et al., 2005]. Nevertheless, this approach has several limitations, including inconsistencies in tagging, dependency on language, and poor scalability.

In later CBIR systems, there was a move towards image representation based on visual properties such as colour, texture, shape, and edge detection. The QBIC system developed by IBM was one of the early adopters of this idea [Flickner et al., 1995], and marked the beginning of content-based image retrieval based on low-level visual features.

The early approaches for feature engineering used handcrafted feature descriptions. One such approach included a two-level filter system consisting of Bag of Features (BoF), colour, and texture descriptors introduced by [Salih and Abdulla, 2023]. This approach uses efficient matching based on texture and colour properties after removing dissimilar images through the Bag of Features (BoF) matching method. Although texture features were obtained using Local Binary Pattern (LBP) and Discrete Wavelet Transform (DWT), and mean, entropy measures in all channels, the colour features were acquired using RGB, HSV, and YCbCr colour models. Tested on the Corel-1K database, the Top-10 precision rate is 86.6%, while Top-20 precision rate is 80.90%.

Also, [Kelishadrokhi et al., 2023] proposed a hybrid descriptor combining local texture features with human-perception-based colour features. The method was evaluated on multiple benchmark datasets including Corel-1K, Corel-5K, Corel-10K, STex, and Coloured Brodatz. On Corel-1K, the method achieved an average precision of 82.71% and recall of 53.97%, demonstrating reasonable performance but limited generalization across datasets due to handcrafted feature dependency.

Feature Selection Method for Content-Based Image Retrieval Using GA has been proposed by [Roodaki et al., 2025] This approach has made possible a remarkable decrease in the number of features by 91.42%, while improving retrieval effectiveness. The mean precision of the approach was 93.07% and mAP of 95.55% @10 and mean precision of 87.13% and mAP of 93.90% @20 using the Corel-1K dataset. Nevertheless, the inability of the approach to incorporate the end-to-end learning paradigm and use of manually designed descriptors limited its flexibility for current deep learning paradigms.

More recently,[Kumar and Murthy, 2025] proposed a relevance aware CBIR model by integrating HOG with EfficientNet through a two-stream network incorporating co-attention mechanism and Fisher vector. For Corel-1K, Oxford5K, and Paris6K datasets, the approach achieved a mAP score of 0.89, 0.85, and 0.83 respectively, and Precision @10 of 0.92. Although it performed exceptionally well, the algorithm depended on manually created features rather than end-to-end self-supervised learning approaches.

With the emergence of deep learning, CBIR shifted toward automatic feature learning.[Ranjith et al., 2023] proposed a deep learning image retrieval model (DLIM) using Inception-V3, where deep feature vectors were extracted and matched using similarity measures. Evaluated on the Corel-10K dataset, the model achieved precision of 90.37% and recall of 87.76%, showing significant improvement over handcrafted methods.

In another study, a deep feature fusion approach was proposed using image depth maps in a segmented neutrosophic domain [Taheri et al., 2024]. The image is decomposed into true, false, and indeterminate components. VGG-16 convolutional layers extract features from true and false components, while a Vision Transformer is used to process binarized depth maps and region-based segmentation. Evaluated on the Oxford5K and Paris6K datasets, the method achieved mean average precision scores of 0.824 and 0.913, respectively.

[Wangi and Makandar, 2024] developed the concept of multi-level feature fusion by employing pre-trained CNN models for enhanced retrieval precision and robustness. The experiment carried out with a Swedish leaf database proved that multi-level feature fusion worked effectively with the average retrieval precision of 92.17%.

In the study done by [Yu et al., 2024], The algorithm employs ResNet-50 as the backbone neural network for feature extraction, learns by utilizing triplet loss, and computes the cosine distance between the resulting vectors as the degree of images' similarity. The training was carried out using both triplet and cross-entropy loss functions. Based on a mAP value of 0.9366, the experiment confirmed that triplet loss outperformed cross-entropy loss in terms of retrieval efficiency.

[Mohammed et al., 2023] The authors proposed a deep learning-based Content-Based Image Retrieval (CBIR) approach using a fine-tuned VGG-19 architecture for robust feature extraction. The pre-trained VGG-19 network was adapted to the target dataset by replacing and fine-tuning its final layers to learn task-specific image representations. To improve model training, image augmentation techniques, including shifting, shearing, and flipping, were applied to the training images. Deep features were extracted from the FC7 layer, resulting in a 4096-dimensional feature representation for each image. Image retrieval was performed by calculating the Euclidean distance between the query image and database image features. The method was evaluated on the Corel-1K dataset, consisting of 1,000 images distributed across 10 categories. The reported results demonstrated average retrieval precision of 92.90.

Recently, self-supervised learning (SSL) approaches have gained attention due to their ability to learn representations without labelled data. [Diong et al., 2024] compared SimCLR and BYOL using ResNet-18 on the CheXpert dataset for medical image classification. The study revealed that the SimCLR approach performed better, as the AUC scores implied better representation learning ability.

SimCLR-Inception was developed by [Jin et al., 2023] , which is a combination of contrastive learning and Inception-V3 used for image recognition tasks. The multi weather data set is used to test the performance of the model, which gave a precision of 0.90.

Although existing studies have achieved notable improvements using handcrafted descriptors, deep learning models, hybrid feature fusion, and self-supervised learning techniques, several limitations still remain. Most approaches either rely on labelled datasets, use fixed handcrafted feature engineering, or employ single-backbone architectures that restrict representation diversity. Furthermore, hybrid methods introduce high computational complexity due to multi-stream processing and rely on non-trainable fusion mechanisms that limit adaptability.

To address these limitations, this study proposes a self-supervised CBIR framework that integrates ResNet18 and EfficientNet-B0 within a SimCLR-based contrastive learning framework using a learnable fusion layer. Additionally, a scalable retrieval pipeline is implemented using FAISS-based similarity search and MongoDB-based storage to support efficient large-scale image retrieval.

3 Used Approach

In this study, a hybrid backbone architecture for Content-Based Image Retrieval (CBIR) is proposed by integrating pre-trained EfficientNet-B0 and ResNet18 within a SimCLR-based contrastive learning framework. The concatenated feature representations taken from both backbones are combined and compressed using a learnable fusion layer that is created using fully linked layers. By dynamically balancing each network's contribution, this fusion process allows the model to concentrate on the most informative feature dimensions.

To ensure efficient large-scale retrieval, Facebook AI Similarity Search (FAISS) is utilized for fast nearest-neighbour search.

Overall, the proposed approach enhances both retrieval efficiency and precision. The complete system architecture and overall model framework are combined and illustrated in a single figure (Figure. 1) for a unified representation of the proposed system.

3.1 Data Augmentation

Data augmentation plays a fundamental role in self-supervised contrastive learning by generating multiple transformed views of the same image. Unlike supervised learning, where augmentation is mainly used to increase dataset diversity and reduce overfitting, contrastive learning utilizes augmentations to construct positive image pairs for representation learning. These transformations encourage the model to learn invariant and semantically meaningful

features under different visual variations. Figure 1 illustrates the proposed framework, in which Step 2 of the offline phase performs data augmentation.

Random augmentation operations such as vertical flipping and random rotation are applied to each input image. Let x denote the original input image:

$$x \in \mathbb{R}^{3\times128\times128}$$

Two augmented views, $x_i$and x_j, are generated as:

$$x_i = t_i(x) \quad (1)$$
$$x_j = t_j(x) \quad (2)$$

Where $t_i$and $t_j$represent stochastic augmentation transformations applied to the same image. Although the generated views appear visually different, they preserve the same semantic information. These augmented image pairs are treated as positive samples during contrastive learning and are subsequently forwarded to the hybrid backbone network for feature extraction.

3.2 Feature Extraction with Hybrid Backbone

Feature extraction is a fundamental stage in Content-Based Image Retrieval (CBIR), where raw image data is transformed into compact and discriminative feature representations. In the proposed framework, the augmented image views generated during the data augmentation stage, denoted as $x_i$and x_j, are simultaneously forwarded to two deep learning backbones: ResNet18 [Pandey and Srivastava, 2023] and EfficientNet-B0 [Hoang and Jo, 2021], as illustrated in Figure. 1, Step 3 of the offline phase.

The hybrid backbone architecture allows the model to take advantage of both networks' complementary strengths. While EfficientNet-B0 enables parameter-efficient feature extraction by compound scaling, ResNet18 offers efficient residual learning and captures hierarchical visual patterns. These architectures work together to improve the model's capacity to learn a variety of semantically rich image representations.

Let $x_i$represent an augmented input image. The feature representations extracted from EfficientNet-B0 and ResNet18 are defined as:

$$f_{eff}(x_i) \in \mathbb{R}^{B\times1280} \quad (3)$$
$$f_{res}(x_i) \in \mathbb{R}^{B\times512} \quad (4)$$

where B denotes the batch size. The extracted feature vectors are obtained after the global average pooling layer of each backbone network. The EfficientNet-B0 and ResNet18 backbones generate 1280-D and 512-D feature vectors, respectively, which are concatenated to form a 1792-D representation.

The resulting feature embeddings from both backbones are subsequently combined through the learnable fusion module to generate a unified representation for contrastive learning and image retrieval.

3.3 Learnable Feature Fusion Layer

The feature representations extracted from EfficientNet-B0 [Hoang and Jo, 2021] and ResNet18 [Pandey and Srivastava, 2023] are concatenated to form a unified hybrid feature vector:

$$\boldsymbol{z_{concat}} = [\boldsymbol{z_{eff}}; \boldsymbol{z_{res}}] \in \mathbb{R}^{\mathbf{1792}} \quad (5)$$

where $\boldsymbol{z_{eff}} \in \mathbb{R}^{\mathbf{1280}}$represents the EfficientNet-B0 feature vector and $\boldsymbol{z_{res}} \in \mathbb{R}^{\mathbf{512}}$denotes the ResNet18 feature vector which are concatenated to form a 1792-D representation. The learnable fusion strategy employed in of Figure. 1 Step 4 of the offline phase.

High-level semantic information and specific spatial features acquired from both backbones are combined in the concatenated representation. After that, this hybrid feature vector is transferred to a learnable fusion layer, which is intended to compress and refine the combined representation into a compact and discriminative embedding appropriate for image retrieval and contrastive learning tasks. The learnable fusion network transforms this representation into a 256-D fused feature, which is used as the final retrieval embedding for database indexing and image retrieval

Unlike conventional dimensionality reduction techniques that apply fixed transformations, the proposed learnable fusion mechanism adaptively learns feature importance during training. This enables the model to emphasize the most informative dimensions from each backbone while suppressing redundant information.

The fusion layer consists of:

- A fully connected layer reducing the feature dimension from 1792 to 512
- A ReLU activation function introducing non-linearity
- A second fully connected layer reducing the feature dimension from 512 to 256

The fused embedding is computed as:

$$\boldsymbol{z_{fused}} = \boldsymbol{W_2} \cdot \boldsymbol{\sigma}(\boldsymbol{W_1} \cdot \boldsymbol{z_{concat}} + \boldsymbol{b_1}) + \boldsymbol{b_2} \qquad (6)$$

where: W1, W2, b1, and b2 are learnable parameters, and $(\boldsymbol{\sigma})$denotes the ReLU activation function

$$\boldsymbol{W_1} \in \mathbb{R}^{\boldsymbol{512\times1792}}$$
$$\boldsymbol{W_2} \in \mathbb{R}^{\boldsymbol{256\times512}}$$
$$\boldsymbol{b_1} \in \mathbb{R}^{\boldsymbol{512}}$$
$$\boldsymbol{b_2} \in \mathbb{R}^{\boldsymbol{256}}$$

.

The resulting fused embedding $\boldsymbol{z_{fused}} \in \mathbb{R}^{\boldsymbol{256}}$is subsequently passed to the projection head for contrastive representation learning.

3.4 Projection Head

After feature fusion, the resulting 256-dimensional embedding is forwarded to a projection head for contrastive representation learning as shown in Figure. 1, Step 5 of the offline phase. In SimCLR-based frameworks, the projection head maps feature representations into a latent space where the contrastive loss is applied. This transformation helps improve the quality of learned embeddings by separating representation learning from the contrastive objective.

The proposed projection head consists of two fully connected layers with a ReLU activation function between them. The network transforms the fused feature vector into a compact 128-dimensional representation optimized for contrastive learning.

The projection operation is defined as:

$$a_2 = W_2 \cdot \sigma(W_1 \cdot z_2 + b_1) + b_2 \qquad (7)$$

Where: W1, W2, b1, and b2 are learnable parameter, and σ denotes the ReLU activation function.

$$W_1 \in \mathbb{R}^{512\times256}$$
$$W_2 \in \mathbb{R}^{128\times512}$$
$$b_1 \in \mathbb{R}^{512}$$
$$b_2 \in \mathbb{R}^{128}$$

The output of the projection head is a 128-dimensional latent representation that is subsequently utilized for contrastive loss computation during self-supervised training.

3.5 Contrastive Loss Function

Contrastive learning is a self-supervised learning paradigm that learns discriminative feature representations by maximizing the similarity between two independently augmented views form a positive pair while simultaneously minimizing the similarity between the other augmented views within the same mini-batch serve as negative samples in the embedding space. In the proposed framework, the Normalized Temperature-scaled Cross Entropy Loss (NT-Xent Loss) is adopted as the contrastive objective function to enforce this representation learning strategy. As illustrated in

Figure. 1, Step 6 of the offline phase, this loss function plays a central role in optimizing the feature encoder by encouraging semantically similar samples to be pulled closer while pushing dissimilar samples farther apart in the latent space.

For each input image, two augmented views are generated and treated as a positive pair, while the remaining samples within the batch serve as negative pairs. The objective of the loss function is to pull embeddings of positive pairs closer together and push embeddings of dissimilar samples farther apart in the embedding space.

The NT-Xent loss function is defined as:

$$\ell(i,j) = -\log \frac{\exp(\text{sim}(h_i, h_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(h_i, h_k)/\tau)} \qquad (8)$$

Where:

- $h_i$and $h_j$denote the latent representations of a positive pair.
- τ represents the temperature parameter.
- N denotes the batch size.
- $\mathbf{1}_{[k \neq i]}$ is an indicator function equal to 1 when $k \neq i$.

The similarity function is computed using cosine similarity:

$$\text{sim}(h_i, h_j) = \frac{h_i \cdot h_j}{\| h_i \| \| h_j \|} \qquad (9)$$

The NT-Xent loss penalizes the model when embeddings of augmented views from the same image are far apart, while encouraging representations of semantically similar images to remain close in the latent space. This enables the network to learn robust and discriminative feature embeddings suitable for image retrieval tasks.

3.6 SimCLR Training Configuration

SimCLR training employed the Lightly transformation pipeline with an input size of 128×128 pixels, a vertical-flip probability of 0.5, and a random rotation probability of 0.5. The batch size was set to 256 with 8 data-loader workers. Training used stochastic gradient descent (SGD) with an initial learning rate of 0.06, momentum of 0.9, and weight decay of 5×10^{-4}. A cosine-annealing learning-rate schedule was employed together with early stopping with a patience of 10 epochs and a maximum of 50 epochs. The NT-Xent contrastive loss used a temperature parameter of 0.5.

The hybrid backbone consisted of pretrained EfficientNet-B0 and ResNet18 networks, which were fine-tuned end-to-end during self-supervised training. The 1280-dimensional EfficientNet-B0 and 512-dimensional ResNet18 features were concatenated to form a 1792-dimensional representation. This representation was subsequently transformed into a 256-dimensional fused feature through fully connected layers. The fused feature was then passed through a SimCLR projection head. The resulting 128-dimensional projection was used for the NT-Xent contrastive loss, whereas the 256-dimensional fused feature was retained as the final retrieval representation.

3.7 Embedding Generation and Storage for Scalable Retrieval

After training, the proposed model generates dense vector embeddings for all images in the dataset. These embeddings capture the semantic and visual characteristics of images within a lower-dimensional feature space, where visually similar images are positioned closer together.

The embedding generation process is defined as:

$$emb(x) = \text{Norm}\left(\text{FusionLayer}\left(f_{eff}(x) \parallel f_{res}(x)\right)\right) \qquad (10)$$

Where:

- $f_{eff}(x)$ represents the feature vector extracted from EfficientNet-B0.
- $f_{res}(x)$ denotes the feature vector extracted from ResNet18.

- ‖ indicates feature concatenation.
- Fusion Layer$(\cdot)$ denotes the learnable fusion module.
- Norm$(\cdot)$represents feature normalization.

To maintain consistency with the training process, all images are pre-processed using the same resizing and normalization parameters based on the ImageNet mean and standard deviation values. Each pre-processed image is then passed through the trained hybrid network to generate its 256-dimensional fused embedding, which serves as the final retrieval representation.

The generated embeddings, along with their corresponding binary image data, are stored in MongoDB, which provides a scalable and flexible storage solution for large-scale image datasets. As illustrated in Figure. 1, Step 7 of the offline phase stores the trained model and embedding representations for subsequent retrieval tasks.

For efficient similarity retrieval, the generated embeddings are indexed using Facebook AI Similarity Search (FAISS), as shown in Figure. 1, Step 8 of the offline phase. During the online retrieval phase, the embedding of the query image is generated using the trained model and compared against the indexed embeddings in the FAISS database using L2 (Euclidean) distance. FAISS enables fast nearest-neighbour search, thereby facilitating efficient retrieval of visually similar images from large-scale datasets.

3.8 Online Retrieval

The online retrieval phase is responsible for retrieving and displaying visually similar images from the database in response to a user-provided query image. This phase is executed after model training and leverages the learned embedding space generated during the offline stage, as shown in the right section of Figure. 1.

3.9 Query Image Preprocessing and Embedding Generation

When a user submits a query image through the web interface, shows Figure. 1, Step 1 and 2 of the online phases. The image is first resized and normalized using the same preprocessing configuration applied during training. This ensures consistency between the training and inference distributions.

The pre-processed query image $x_q$is then passed through the trained hybrid backbone consisting of ResNet18 and EfficientNet-B0 to extract feature representations: shows Figure. 1, Step 3 of the online phase.

$$f_{res}(x_q) \in \mathbb{R}^{512}, f_{eff}(x_q) \in \mathbb{R}^{1280}$$

The extracted features are concatenated to form a unified 1792-D representation:

$$f_{combined} = \text{Concat}(f_{res}(x_q), f_{eff}(x_q)) \in \mathbb{R}^{1792} \qquad (11)$$

This combined feature vector is then passed through the learnable fusion layer to obtain a 256-D compact embedding:

$$f_{fused} = W_2 \cdot \text{ReLU}(W_1 f_{combined} + b_1) + b_2 \in \mathbb{R}^{256} \qquad (12)$$

The resulting embedding captures the most discriminative visual features of the query image in a compact latent space, making it suitable for efficient similarity comparison and retrieval.

3.10 Similarity Search

The query embedding is compared with the precomputed database embeddings using Facebook AI Similarity Search (FAISS) as illustrate in shows Figure. 1, Step 4 of the online phase. In this work, similarity is computed using the L2 (Euclidean) distance metric in the embedding space.

Let the query embedding be defined as:

$$q \in \mathbb{R}^{256}$$

and the database of indexed image embeddings be represented as:

$$\{x_i\}_{i=1}^{n}, x_i \in \mathbb{R}^{256}$$

The retrieval task aims to find the nearest embeddings to the query by solving:

$$\arg\min_{x_i \in X} \| q - x_i \|_2^2 \qquad (13)$$

The L2 (Euclidean) distance between two feature vectors $A$and $B$is defined as:

$$D(A,B) = \sqrt{\sum_{j=1}^{d} (a_j - b_j)^2} \qquad (14)$$

Where d represents the dimensionality of the embedding space.

FAISS efficiently computes these distances and ranks all database embeddings based on their similarity to the query embedding. A smaller L2 distance indicates higher similarity between images, enabling retrieval of the most visually and semantically relevant results.

3.11 Top K Similar Image Retrieval

FAISS retrieves the top-K images with the smallest L2 (Euclidean) distance to the query embedding. These retrieved indices correspond to the most similar embeddings stored in the database. The identifiers of the retrieved images are then used to fetch their corresponding binary representations from MongoDB, as shown in Figure. 1, Step 5 of the online phase. Subsequently, the retrieved images are decoded into base64 format for transmission and visualization in the web interface, along with their corresponding similarity scores, as illustrated in Figure. 1, Step 6 of the online phase. This process enables efficient and scalable retrieval of visually similar images, ensuring fast response times and accurate ranking based on embedding similarity.

.

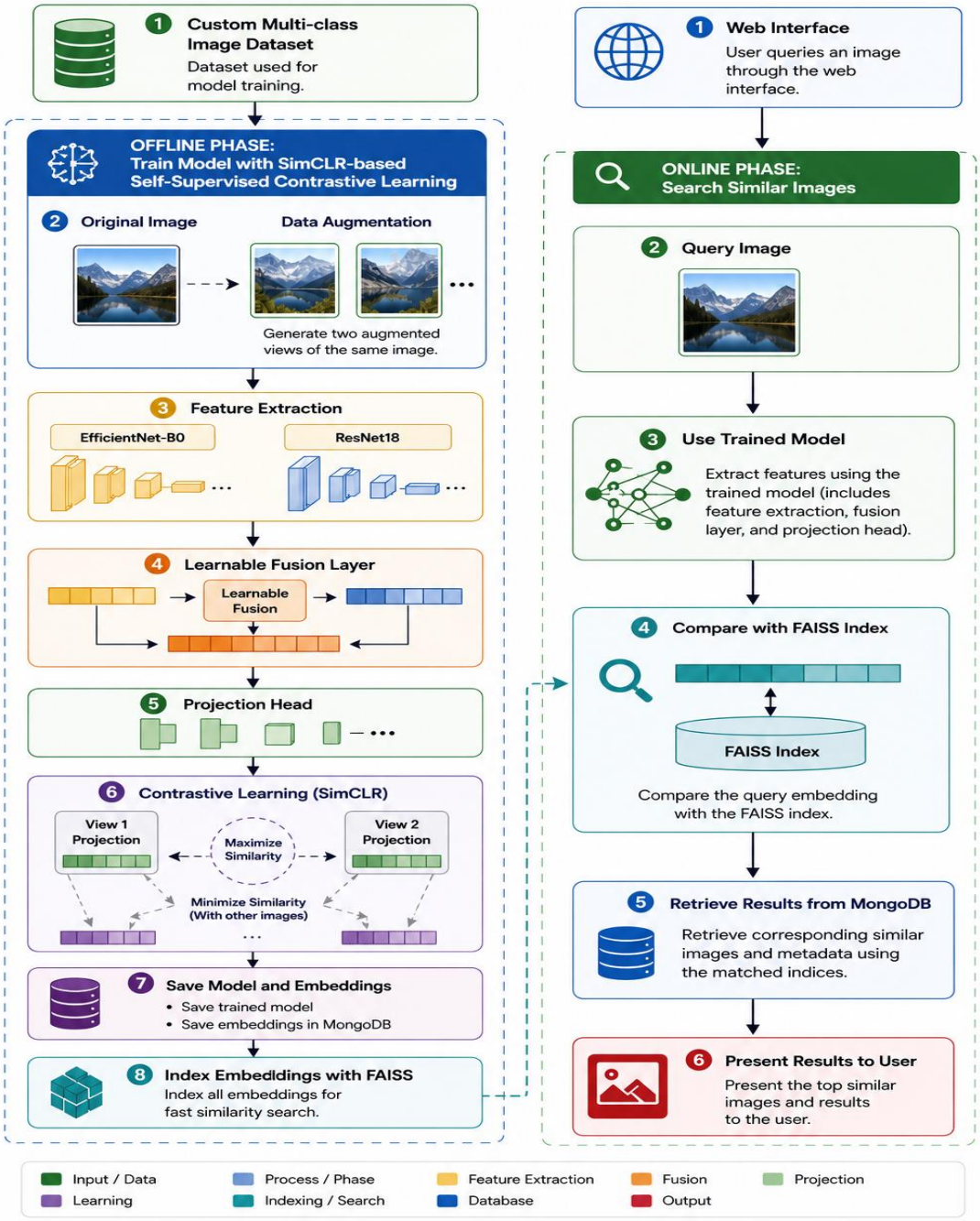


Figure 1. Overall architecture of the proposed CBIR system showing the complete pipeline from hybrid feature extraction to similarity-based image retrieval in a unified framework.

4 Experimental Evaluation

This section presents the experimental setup used to evaluate the effectiveness of the proposed hybrid backbone, learnable fusion mechanism, and self-supervised contrastive learning framework. A series of experiments were conducted to evaluate the retrieval performance of the proposed framework. In addition, an ablation study was performed to investigate the impact of different backbone architectures and feature fusion strategies on retrieval performance. The proposed method was also compared with existing state-of-the-art CBIR approaches.

4.1 Dataset Description

In this research, the proposed Content-Based Image Retrieval (CBIR) framework is evaluated using two datasets: a custom multi-class dataset and the Corel-1K benchmark dataset. The use of both datasets allows the system to be evaluated in terms of real-world generalization as well as standard benchmark comparison.

Custom Multi-Class Dataset: A custom multi-class image dataset was developed by collecting images from multiple publicly available datasets on Kaggle. The dataset comprises eight visually diverse categories: aeroplanes [Commercial Aircraft Dataset, n.d.], apples [‘Fruits Classification’, n.d.], bananas [‘Fruits Classification’, n.d.], buses [‘Bus Dataset’, n.d.], cycles [‘Bike Ads (images, prices, specifications)’, n.d.], lavender flowers [‘Flower Classification’, n.d.], roses [‘Flower Classification’, n.d.], and sunflowers [‘Flower Classification’, n.d.]. Each category contains 1,500 images, resulting in 12,000 images in the database set.

For retrieval evaluation, 30 images were randomly selected from each class to form an independent query set of 240 images, while the remaining images were assigned to the database set shown in Table1. The query images were excluded from the indexed database. The resulting database and query partitions were kept fixed throughout the experiments. Sample images from the eight classes are shown in Figure 2.

The dataset was intentionally constructed from different sources to introduce high variability in visual appearance, including differences in background, illumination, object orientation, scale, and viewpoint. This ensures both intra-class variation, where images within the same category exhibit different visual conditions, and inter-class separability, where different categories remain visually distinct.

The dataset's moderate scale and diversity make it ideal for self-supervised contrastive learning, in which the model learns discriminative feature representations without the need for human annotations. The model can learn reliable and significant embeddings for retrieval tasks in this scenario, where augmented views of the same image form positive pairs and images from different categories operate as negative pairs.

Corel-1K Benchmark Dataset: The Corel-1K [‘Corel-1K’, n.d.] is a popular benchmark dataset for assessing systems that use content-based image retrieval. The database set contains 1,000 images from 10 classes, with 100 images per class, while the independent query set contains 100 images, with 10 images per class shown in Table1. As shown in Figure 3, the dataset includes diverse categories such as animals, flowers, buildings, buses, beaches, and indoor scenes. The database and query images were taken from their respective predefined folders, and no additional random splitting was performed. The query images were excluded from the indexed database during retrieval evaluation.

Because of its balanced class distribution and distinct category structure, Corel-1K is frequently utilised in the CBIR literature and is appropriate for assessing retrieval performance under controlled circumstances. However, it is relatively small in scale compared to modern datasets, which limits its ability to represent real-world diversity.

This study compares the suggested strategy with current state-of-the-art CBIR techniques documented in the literature using Corel-1K as a benchmark dataset. This makes it possible to fairly assess retrieval performance in a controlled experimental environment. The testing/query images were excluded from the indexed database. During retrieval evaluation, each query was compared only against the indexed database images, ensuring that the query image itself was not present in the retrieval database.

Table 1. Dataset splitting and query configuration used for retrieval evaluation.

Dataset	Classes	Database Images	Images/Class	Query Images	Query/Class	Query Excluded from Database
Corel-1K	10	1,000	100	100	10	yes
Custom Multi-Class Dataset	8	12,000	1,500	240	30	yes



Figure 2. This figure illustrations sample images from different categories from various publicly available datasets on Kaggle highlighting the diversity of image types used in the evaluation.



Figure 3. Samples of different images from Corel 1K dataset

4.2 Experimental Setup

Experiments were conducted using an Intel Core i7 CPU, 16 GB RAM, Windows 11, and an NVIDIA Tesla T4 GPU through Google Colab. All images were resized to 128×128 pixels and normalized using the ImageNet mean and standard deviation. The Corel-1K dataset was used for benchmarking, while the custom dataset was used to evaluate retrieval performance on a more diverse image collection. A fixed random seed of 1 was used for reproducibility during model training.

4.3 Evaluation Metric

The precision rate, which is the ratio of relevant images recovered to all images retrieved, is one of the most well-known confusion matrices used in CBIR to assess performance. Precision, also known as Positive Predictive Value, quantifies the percentage of accurately detected positive samples among all samples projected to be positive, as shown in the following equation.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (15)$$

- True Positive (TP): A sample that is positive and correctly identified as positive by the model.
- False Positive (FP): A sample that is negative but incorrectly identified as positive by the model.

Precision @K:

is an evaluation metric used in information retrieval to measure the proportion of relevant items among the top-K items retrieved

$$\text{Precision@K} = \frac{\text{No of relavant items in the top K results}}{\text{K}} \quad (16)$$

In classification term:

$$\text{Precision@K} = \frac{\text{TP@K}}{\text{TP@K} + \text{FP@K}} \quad (17)$$

- True Positives@K: Relevant items found within the top K list.
- False Positives@K: Irrelevant items found within the top K list.
- K: The cutoff rank limit chosen for the evaluation.

Mean Precision

Mean Precision (MP) represents the average Precision@K across all query images belonging to a particular class. It can be calculated as:

$$MP_c = \frac{1}{N_c} \sum_{q=1}^{N_c} Precision@K_q \quad (18)$$

Where

- MP_c= Mean Precision for class c
- N_c= Number of queries in class c
- Precision@K$_q$= Precision@K obtained for query q.

Average of Mean Precision

Average Mean Precision is calculated by averaging the Mean Precision values obtained for all classes. It provides an overall measure of retrieval precision across the complete dataset and is defined as:

$$\text{Average Mean Precision} = \frac{1}{C} \sum_{c=1}^{C} MP_c \quad (19)$$

Where

- MP_cdenotes the Mean Precision of class c
- Crepresents the total number of classes.

4.4 Evaluation Using Corel-1K Benchmark Dataset

The proposed framework is evaluated using the Corel-1K benchmark dataset under a standardized experimental setting to ensure fair comparison with existing state-of-the-art Content-Based Image Retrieval (CBIR) approaches. The evaluation assesses retrieval effectiveness using mean precision achieved by different deep learning architectures and feature fusion strategies for content-based image retrieval (CBIR)

For the evaluation presented in Table 2 and Figure 4, the top 20 retrieved images for each query image were considered. It presents a comparative analysis of individual backbone architectures, namely ResNet-18, ResNet-50, and EfficientNet-B0, together with hybrid representations formed by combining ResNet-18 and EfficientNet-B0 using simple feature concatenation, Kernel Principal Component Analysis (Kernel PCA), and the proposed learnable fusion mechanism. The numerical results for each method were independently obtained in the present study under the same experimental conditions. For a fair comparison, all configurations reported in this analysis use the same SimCLR-based

self-supervised contrastive learning framework, database/query partition, image preprocessing, feature extraction protocol, retrieval procedure, and evaluation criteria.

An image is considered relevant to a query when it belongs to the same semantic class as the query image, whereas images from different classes are considered non-relevant. The Mean Precision for each class was calculated based on the relevant images retrieved within the top 20 results, and the final Average Mean Precision was obtained by averaging the class-wise Mean Precision values across all classes.

The experimental results reveal a progressive improvement in retrieval performance across the evaluated architectures. The standalone backbone networks achieve competitive performance, with average mean precision values of 83% and 84% for ResNet18 and ResNet50, respectively. Even though the residual learning helps in improving the feature extraction ability, the increase in the depth of the model does not significantly improve the retrieval performance

EfficientNet-B0 exhibits relatively higher average mean precision of 86.14%, proving the efficiency of compound scaling to balance depth, width, and image size. It produces compact but rich feature representations, making it more suitable for retrieval-based feature extraction than architectures relying solely on increased depth.

In order to study more about the effects of multi-fusion feature integration, a hybrid architecture using simple feature concatenation further increases the average mean precision to 90%, indicating that combining complementary representations from ResNet-18 and EfficientNet-B0 can improve retrieval performance. A Kernel PCA-based feature fusion strategy achieves 94%, it is found that there is a great improvement. This enhancement indicates that integrating complementary feature spaces significantly improves discriminative capability by combining information extracted from multiple backbone networks. However, Kernel PCA performs a fixed and non-learnable transformation that operates independently of the retrieval objective. Consequently, it cannot dynamically adapt feature importance during training or optimize feature representations according to dataset-specific characteristics.

However, the proposed model utilizes a learnable fusion layer that optimizes both networks' feature contributions to form a hybrid architecture. The advantage of this learnable fusion approach is that it enables the model to prioritize informative feature dimensions while filtering out irrelevant and noise features.

Therefore, under the specified evaluation protocol, the proposed model Hybrid ResNet-18 + EfficientNet-B0 with Learnable Fusion and SimCLR achieves the highest average mean precision of 98.03%. This improvement is observed across most of the evaluated classes, including Beaches (97.8%), Buildings (96.8%), Mountains (98.9%), and Food (91.9%). The proposed approach also achieves 100% mean precision for Buses, Dinosaurs, Elephants, Roses, and Horses. The gain is not just incremental, but structural, it shows the effectiveness of learnable fusion compared with statistical transformations in retrieval-based representation learning tasks. Unlike Kernel PCA, the proposed fusion mechanism is explicitly optimized with respect to the retrieval embedding space, resulting in stronger alignment between feature learning and retrieval objectives.

In summary, the findings in Tables 2 and Figure 4 strongly indicate that the suggested approach consistently outperforms all comparative configurations on the Corel-1K benchmark dataset, thereby validating its effectiveness, discriminative capability, and reliability for CBIR applications.

Table 2. Comparative analysis of class-wise mean precision (%) for individual backbones and hybrid ResNet-18–EfficientNet-B0 feature fusion strategies with SimCLR using Corel 1K dataset.

Class Name	ResNet-18 +SimCLR Mean precision	ResNet-50 +SimCLR	EfficientNet-B0 +SimCLR	Hybrid (ResNet-18+EfficientNet-B0) Simple Concatenation +SimCLR	Hybrid (ResNet-18+EfficientNet-B0) with Kernel PCA +SimCLR	Hybrid (ResNet-18+EfficientNet-B0) with Learnable Fusion

						+SimCLR (Proposed)
Africans	90	75.6	84.5	90	86	94.9
Beaches	76	90	89.3	91	95	97.8
Buildings	67.6	68.9	65	70	84.1	96.8
Buses	85	85.2	100	100	100	100
Dinosaur	100	100	100	100	100	100
Elephant	70.4	80	82.2	92.4	100	100
Roses	100	90.2	97.3	100	100	100
Horses	95	100	100	100	100	100
Mountains	80	79.5	83	84.8	90	98.9
Food	66	70.6	60.1	71.8	84.9	91.9
Average	83	84	86.14	90	94	98.03

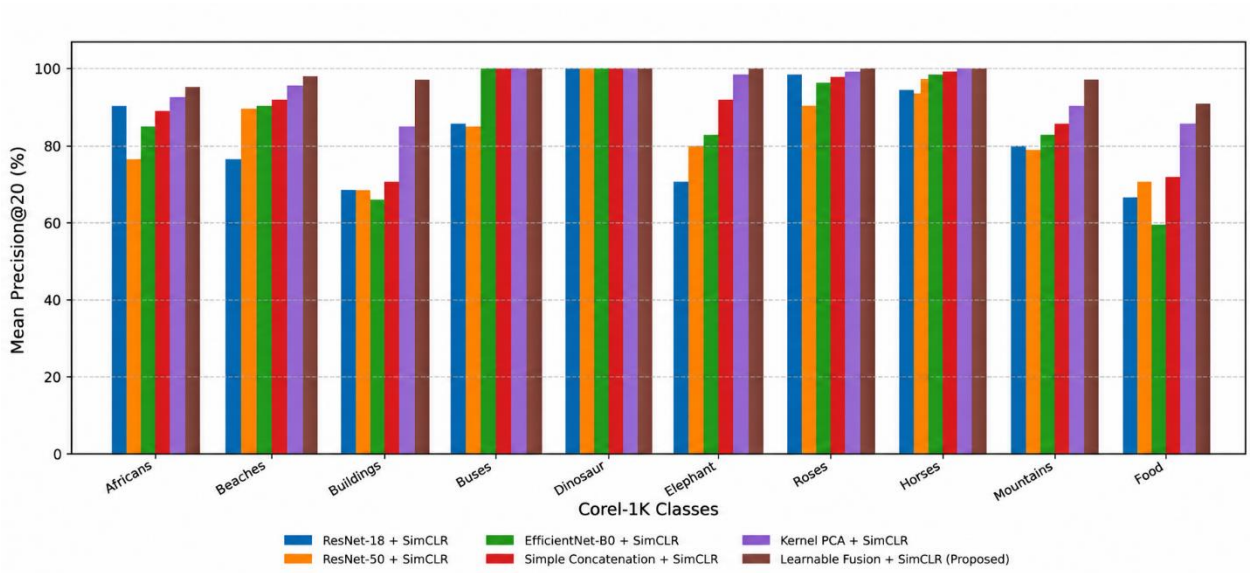


Figure 4. Class-wise Mean Precision@20 comparison of different backbone and feature-fusion configurations on the Corel-1K dataset, considering the top 20 retrieved images for each query.

4.5 Evaluation Using Custom Multi-Class Dataset

To further evaluate the practical applicability of the proposed framework, experiments are conducted on a custom multi-class dataset comprising diverse image categories with higher visual variability and inter-class similarity. Due to differences in the way the objects look, their background environment, illumination conditions, and compositions, the specially designed dataset presents retrieval challenges not present in benchmarking datasets.

The Mean Precision for each class was calculated based on the relevant images retrieved within the top 20 results, and the final Average Mean Precision was obtained by averaging the class-wise Mean Precision values across all classes. According to the results obtained from the experiment as presented in Figure 5 and Table 3, it can be seen that the retrieval average mean precision is consistently good across all scenarios evaluated. For standalone backbone networks, EfficientNet-B0 yields higher retrieval average mean precision compared to the ResNet model architecture due to its ability to learn discriminative features effectively.

The hybrid fusion models incorporate complimentary feature data from several backbone networks to further enhance retrieval performance. In particular, the proposed learnable fusion framework achieves the best overall results, obtaining average mean precision of 96%.

Because the custom dataset is more complicated and variable than the Corel-1K benchmark test, there is a minor decrease in retrieval average mean precision. However, the suggested framework shows significant durability and generalisation capabilities for practical CBIR applications while maintaining very competitive performance.

Evaluation is carried out utilising unseen question images that were not used during training in order to guarantee impartial testing. The findings verify that the suggested approach maintains consistent retrieval performance in more difficult retrieval contexts and successfully generalises beyond benchmark datasets.

Table 3. Comparative analysis of class-wise mean precision (%) for individual backbones and hybrid ResNet-18–EfficientNet-B0 feature fusion strategies with SimCLR using Multiclass dataset.

Class Name	ResNet-18 +SimCLR	ResNet-50 +SimCLR	EfficientNet-B0 +SimCLR	Hybrid (ResNet-18+EfficientNet-B0) Simple Concatenation +SimCLR	Hybrid (ResNet-18+EfficientNet-B0) with Kernel PCA +SimCLR	Hybrid (ResNet-18+EfficientNet-B0) with Learnable Fusion +SimCLR (Proposed)
Aeroplan	82.8	83.1	90	92	93	97.5
cycle	84.7	88	89	91	93.4	100
Buses	85.6	89.2	90.2	94	95	100
Lavander	70	71	70.5	75.5	80	92.8
Sunflower	74.4	75	77.7	84.6	86.2	93.6
Roses	85.3	86	91	92	95	97.2
Apples	83.9	84.1	85	86	88.4	95
Banana	74.1	80	78.7	81	81	92
Average	80.01	82.05	84	87	89	96

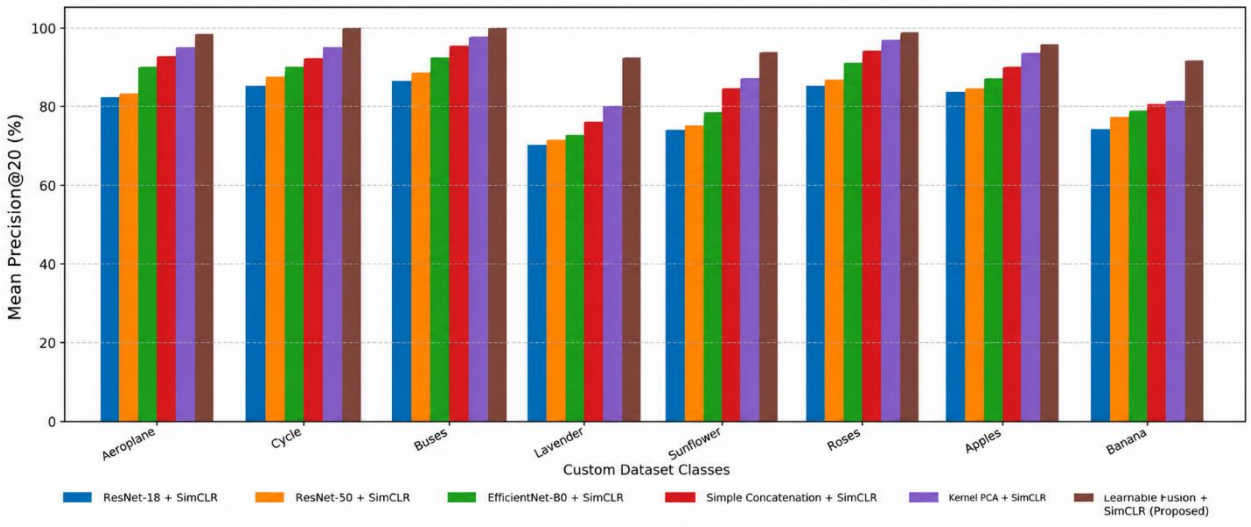


Figure 5. Class-wise Mean Precision@20 comparison of different backbone and feature-fusion configurations on the custom dataset, considering the top 20 retrieved images for each query.

4.6 Comparative Analysis with Existing Techniques Using Corel-1K Benchmark Dataset

Comparative evaluation is carried out on the Corel-1K benchmark dataset, a common dataset used for evaluating image retrieval systems, in order to thoroughly examine the performance of the suggested content-based image retrieval (CBIR) architecture. The proposed methodology is evaluated using e Mean Precision for each class was calculated based on the relevant images retrieved within the top 20 results, and the final Average Mean Precision was obtained by averaging the class-wise Mean Precision values across all classes. Average mean precision that has been considered to be a holistic measure of evaluation considering its ability to measure correctness and relevance of ranking.

Table 4 presents a class-wise comparison of the retrieval performance of the proposed method with the methods reported by Roodaki et al. (2025), Mohammed et al. (2023), and Salih and Abdulla (2023) across ten categories of the Corel-1K dataset. Comparison with previously published CBIR methods, only literature results reported using the top 20 retrieved images, are considered directly comparable. The proposed method demonstrates consistently high performance across the evaluated categories, achieving mean precision values between 91.9% and 100%. In particular, it achieves 100% for the Buses, Dinosaur, Elephant, horses and Roses categories, while 97.8% is obtained for the Beaches and 98.9 for Mountain categories. The value of 96.8% ,94.9% is observed for the Building and Africans categories. The lowest value of 91.9% is observed for the Food category. Overall, the proposed method achieves an average mean precision of 98.03%, compared with 87.13%, 92.90%, and 80.90% for the compared methods, respectively. These class-wise results demonstrate the consistency of the proposed retrieval representation across different semantic categories.

Table 5 provides an overall comparison of the proposed method with existing CBIR approaches on the Corel-1K dataset using Average mean Precision@20 whereas the graph comparing the average mean precision of the compared approaches has been shown in Figure 6.The proposed method achieves an average mean precision of 98.03%, whereas the literature-reported methods achieve 87.13% for [Roodaki et al., 2025], 80.90% for [Salih and Abdulla, 2023] and 92.90% for [Mohammed et al., 2023].

The proposed framework combines SimCLR-based self-supervised representation learning with ResNet-18 and EfficientNet-B0 through a learnable fusion strategy, provides an opportunity to extract semantically rich and highly discriminative feature vectors without the need for labelled data.

The results indicate that the proposed representation provides high retrieval precision on the Corel-1K dataset. Since the existing results are taken from the respective literature and the proposed result was obtained in the present study, the comparison should be interpreted as a reference-based comparison rather than a controlled reimplementation of all competing methods. The compared methods employ different feature extraction and learning strategies; therefore,

the comparison focuses on retrieval performance under the common evaluation protocol rather than assuming identical methodologies.

Table 4. Class-wise average precision (%) comparison of the proposed method with existing CBIR methods on the Corel-1K dataset.

Class Name	[Roodaki et al., 2025]	[Mohammed et al., 2023]	[Salih and Abdulla, 2023]	proposed
Africans	86.65	92	81.3	94.9
Beaches	67.7	71	65.4	97.8
Buildings	79.9	100	66.7	96.8
Buses	96.4	100	95.3	100
Dinosaur	100	98	99.8	100
Elephant	83.35	100	73.6	100
Roses	98.45	100	91.7	100
Horses	96.6	100	95.8	100
Mountains	73.4	100	67.3	98.9
Food	88.85	68	72.5	91.9
Average	87.13	92.90	80.90	98.03

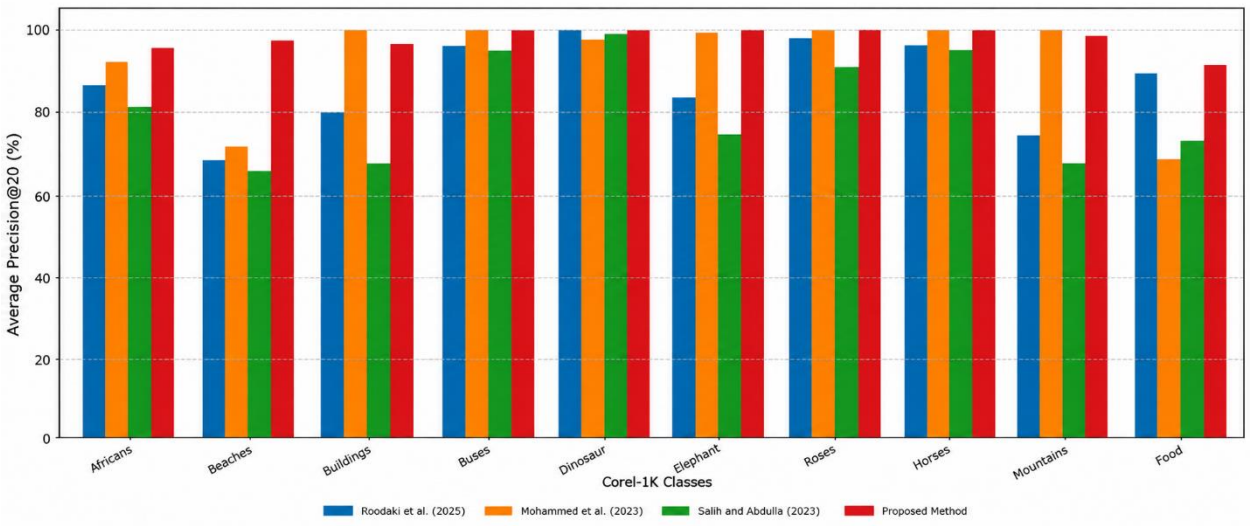


Figure 6. Class-wise average precision (%) comparison of the proposed method with existing CBIR methods on the Corel-1K dataset.

Table 5. Class-wise Average Precision@20 comparison between literature-reported CBIR methods and the independently obtained results of the proposed method on the Corel-1K dataset.

Reference	Corel-1K Dataset			
	Proposed Model	Feature Strategy	Result Source	Average Precision @20 (%)
[Roodaki et al., 2025]	Genetic Algorithm-based CBIR	Handcrafted feature selection	Literature-reported	87.13
[Salih and Abdulla, 2023]	Two-layer hybrid CBIR	BoF + LBP + DWT + Color features	Literature-reported	80.90
[Mohammed et al., 2023]	Fine-tuned VGG-19	Deep CNN features (FC7, 4096-D)	Literature-reported	92.90
Proposed Method (Ours)	SimCLR + ResNet18 + EfficientNet-B0 with learnable fusion	Deep self-supervised representation learning	This study	98.03

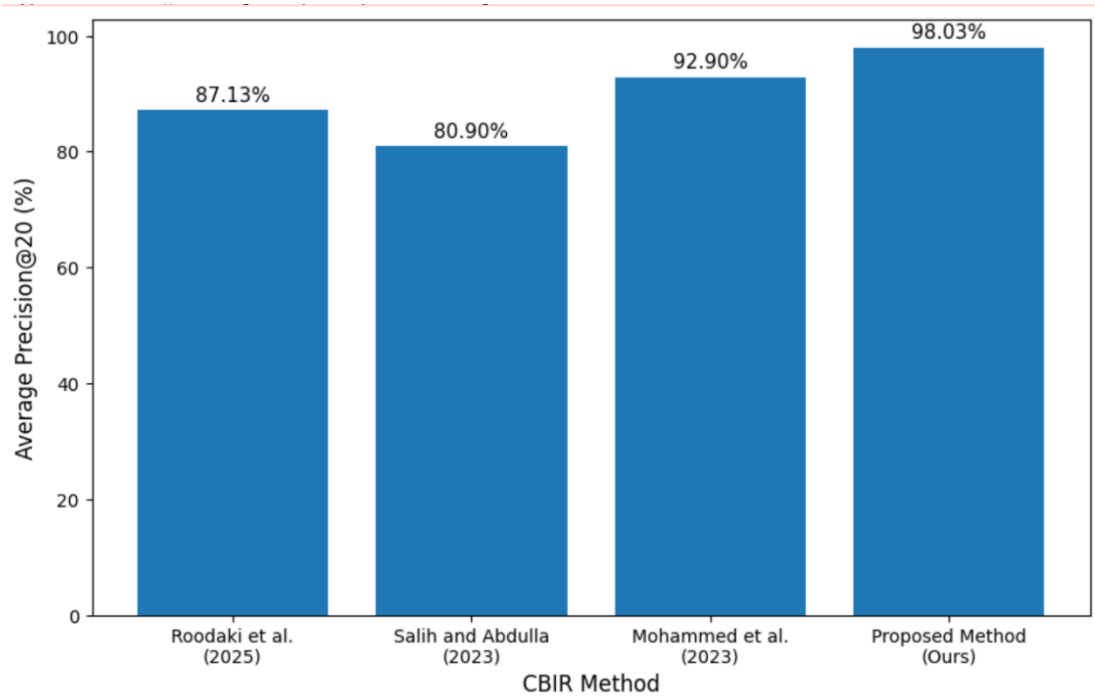


Figure 7. Comparison of Average Precision@20 (%) between the proposed method and literature-reported CBIR methods on the Corel-1K dataset.

4.7 Retrieval Time Analysis

In order to assess the computational efficiency of the proposed content-based image retrieval framework, the overall retrieval latency of the complete pipeline is analysed. The retrieval efficiency of the proposed CBIR system was evaluated using a database containing 1,000 indexed images. The execution time is computed from the point when a query image is sent to the system until the retrieval of the top 50 matching images. For the total retrieval time, we use the time

measurement function provided by the Python language, where we start counting right after receiving the query image input and stop counting once we obtain the similarity score and retrieve the top 50 images. The reported time represents the complete retrieval operation, including query processing, feature extraction, FAISS similarity search, retrieval of the corresponding database images, and display of the top 50 results. Model loading and FAISS index construction were excluded as one-time operations. The timing experiment measured the complete end-to-end retrieval operation; therefore, separate independently measured times for model inference, FAISS search, and database retrieval are not reported.

As can be seen in Fig. 8, our proposed framework takes around 0.072 seconds to perform all steps for each query image input.

Total retrieval time is calculated using Python's time module, with the timer being triggered right away after obtaining the query image and stopping after getting the similarity score as well as the top 50 images. From Fig. 8 below, the proposed system takes about 0.072 seconds to carry out the whole CBIR process for each query image.

Experimental observations suggest that the process of image preprocessing and embedding creation using SimCLR is the most computationally intensive stages of the pipeline, while the FAISS similarity search and MongoDB embedding retrieval add minimal latency to the process. The results demonstrate that the proposed framework achieves low retrieval time suitable for end-to-end retrieval and web-based visual search applications, confirming its effectiveness for scalable CBIR deployment.

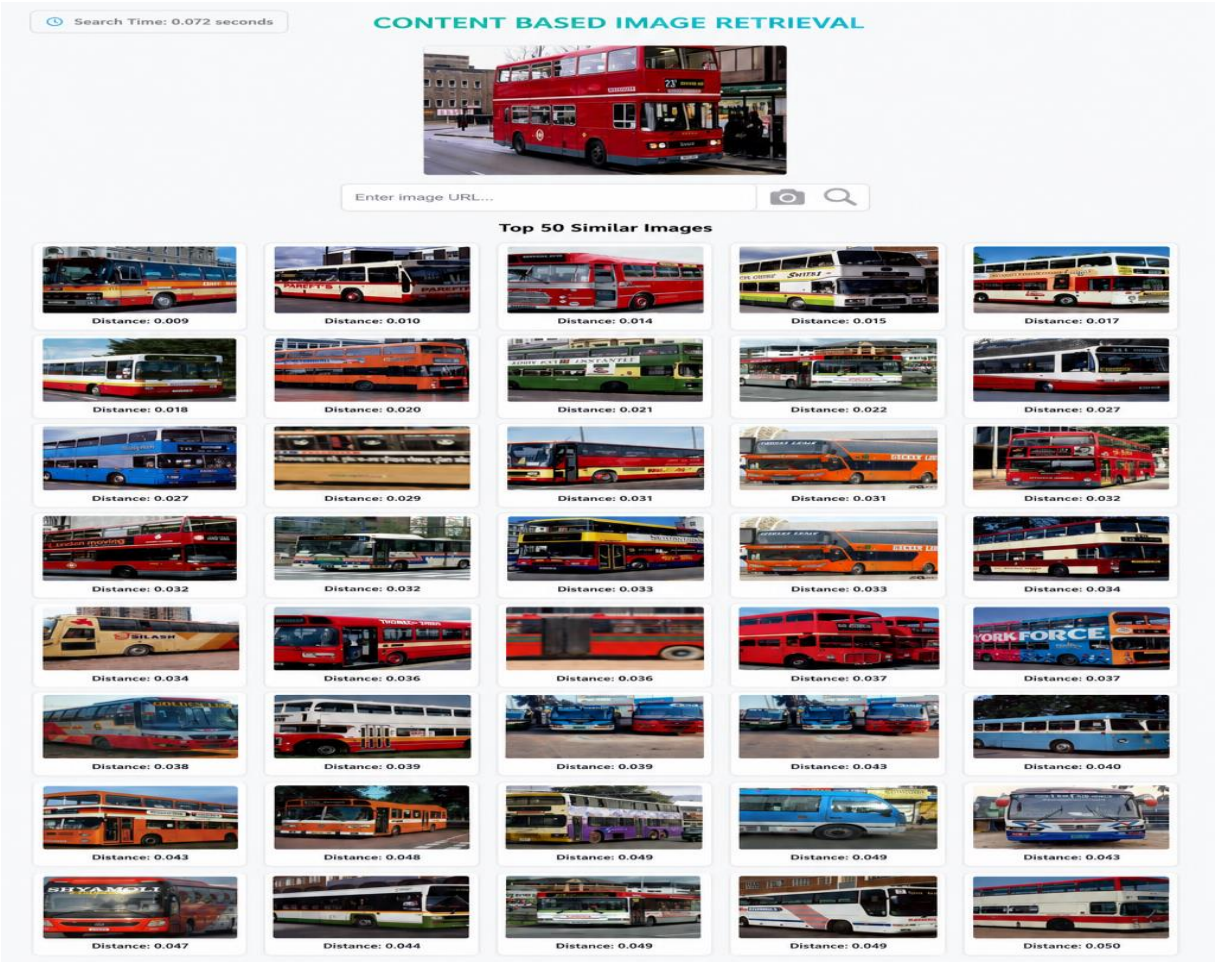


Figure 8. Retrieval time analysis of the proposed CBIR framework for retrieving the top 50 similar images.

5 Conclusion

This research presents a CBIR framework based on self-supervised contrastive learning with a hybrid backbone and a learnable feature fusion mechanism. By exploiting the complementary feature extraction capabilities of ResNet18 and EfficientNet-B0, the proposed framework generates compact and discriminative image embeddings, leading to improved retrieval end-to-end and efficient image retrieval.

In contrast to the conventional CBIR techniques that use hand-crafted features or pre-determined methods of fusion, the proposed technique uses an adaptive fusion layer that adapts the weights of relevance of the learned features from both backbones. This makes for better learning representation and retrieval effectiveness.

The success of the suggested model is demonstrated by performing thorough experiments on two different datasets: one is the custom large multi-class dataset, while the other is the standard benchmark dataset called Corel-1K. The experiment results show that our approach consistently outperforms all three individual backbone architectures, namely ResNet-18, ResNet-50, and EfficientNet-B0, together with hybrid representations formed by combining ResNet-18 and EfficientNet-B0 using simple feature concatenation, Kernel Principal Component Analysis (Kernel PCA), and the proposed learnable fusion mechanism. In both experiments, our approach outperforms others based on Mean Precision for each class and Average Mean Precision metrics.

In addition, the results obtained after a comparative study using the recently developed benchmark test set of Corel-1K prove that the proposed method continues to perform competitively by delivering higher Average Mean Precision, validating the effectiveness of the learnt hybrid representation model. Also, the results of retrieval latency demonstrated that the complete process of content-based image retrieval is effective, delivering a latency rate of around 0.072 seconds per query.

Overall, the proposed framework effectively addresses the challenge of learning discriminative image representations without relying on large manually labelled datasets. By integrating a hybrid backbone with self-supervised contrastive learning, it achieves a balance between robust feature representation and efficient retrieval performance. Transformer-based architectures could be considered in future research, as well as memory-efficient indexing.

6 Acknowledgements

In this regard, the author wishes to extend their heartfelt thanks to Dr. Saima Munawar for the invaluable help, guidance, and encouragement she provided throughout this process, which greatly assisted in the completion of this assignment. The dedication and keen eye towards academic progress are highly appreciated.

It is also necessary to thank the Virtual University of Pakistan for providing all the required facilities that enabled the successful completion of this assignment.

7 References

1. [Ahmed et al., 2019] Ahmed, K. T., Ummesafi, S., Iqbal, A.: ‘Content based image retrieval using image features information fusion’; Information Fusion, 51 (2019), 76–99.
2. [‘Bike Ads (images, prices, specifications)’, n.d.] (n.d.). Retrieved from https://www.kaggle.com/datasets/tysonpo/bike-ads-images-prices-specifications
3. [‘Bus Dataset’, n.d.] (n.d.). Retrieved from https://www.kaggle.com/datasets/samuelayman/bus-dataset
4. [Chen et al., 2020] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: ‘A simple framework for contrastive learning of visual representations’; In International conference on machine learning. PmLR (2020), 1597–1607.
5. [Commercial Aircraft Dataset, n.d.] Commercial Aircraft Dataset: ‘Commercial Aircraft Dataset’; (n.d.). Retrieved from https://www.kaggle.com/datasets/nelyg8002000/commercial-aircraft-dataset
6. [‘Corel-1K’, n.d.] (n.d.). Retrieved from https://www.kaggle.com/datasets/amirhosseinroodaki/corel-1k-corel-5k-and-corel-10k-datasets
7. [Deb and Zhang, 2004] Deb, S., Zhang, Y.: ‘An overview of content-based image retrieval techniques’; In 18th International Conference on Advanced Information Networking and Applications, 2004. AINA 2004. (Vol. 1). IEEE (2004), 59–64.

8. [Diong et al., 2024] Diong, Z. Y., Lim, W. Y., Goh, C. P.: ‘Self-supervised learning in medical diagnostics: An examination of simclr and byol in image classification’; In 2024 3rd International Conference on Digital Transformation and Applications (ICDXA). IEEE (2024), 210–214.
9. [Elharrouss et al., 2024] Elharrouss, O., Akbari, Y., Almadeed, N., Al-Maadeed, S.: ‘Backbones-review: Feature extractor networks for deep learning and deep reinforcement learning approaches in computer vision’; Computer Science Review, 53 (2024), 100645.
10. [Flickner et al., 1995] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., et al.: ‘Query by image and video content: The QBIC system’; Computer, 28, 9 (1995), 23–32.
11. [‘Flower Classification’, n.d.] (n.d.). Retrieved from https://www.kaggle.com/datasets/utkarshsaxenadn/flower-classification-5-classes-roselilyetc
12. [‘Fruits Classification’, n.d.] (n.d.). Retrieved from https://www.kaggle.com/datasets/utkarshsaxenadn/fruits-classification
13. [Guo et al., 2008] Guo, C., Ma, Q., Zhang, L.: ‘Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform’; In 2008 IEEE conference on computer vision and pattern recognition. IEEE (2008), 1–8.
14. [Hassan et al., 1855] Hassan, R. Q., Sultani, Z. N., Dhannoon, B. N.: ‘Content-based image retrieval based on corel dataset using deep learning’; Int J Artif Intell ISSN, 2252 (1855), 8938.
15. [Hoang and Jo, 2021] Hoang, V.-T., Jo, K.-H.: ‘Practical analysis on architecture of EfficientNet’; In 2021 14th International Conference on Human System Interaction (HSI). IEEE (2021), 1–4.
16. [Jin et al., 2023] Jin, M., Zhang, Y., Cheng, X., Ma, L., Hu, F.: ‘SimCLR-inception: an image representation learning and recognition model for robot vision’; In Asian conference on pattern recognition. Springer (2023), 137–147.
17. [Kelishadrokhi et al., 2023] Kelishadrokhi, M. K., Ghattaei, M., Fekri-Ershad, S.: ‘Innovative local texture descriptor in joint of human-based color features for content-based image retrieval’; Signal, Image and Video Processing, 17, 8 (2023), 4009–4017.
18. [Koonce, 2021a] Koonce, B.: ‘EfficientNet’; In Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization. Springer (2021a), 109–123.
19. [Koonce, 2021b] Koonce, B.: ‘ResNet 50’; In Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization. Springer (2021b), 63–72.
20. [Kumar and Murthy, 2025] Kumar, R., Murthy, N.: ‘Relevance-Aware Content-based Image Retrieval using Deep Hybrid Feature Extraction’; Engineering, Technology & Applied Science Research, 15, 3 (2025), 22976–22982.
21. [Liu and Yang, 2013] Liu, G.-H., Yang, J.-Y.: ‘Content-based image retrieval using color difference histogram’; Pattern Recognition, 46, 1 (2013), 188–198.
22. [Liu et al., 2021] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: ‘Self-supervised learning: Generative or contrastive’; IEEE Transactions on Knowledge and Data Engineering, 35, 1 (2021), 857–876.
23. [Mathew et al., 2020] Mathew, A., Amudha, P., Sivakumari, S.: ‘Deep learning techniques: an overview’; In International conference on advanced machine learning technologies and applications. Springer (2020), 599–608.
24. [Mohammed et al., 2023] Mohammed, M. A., Oraibi, Z. A., Hussain, M. A.: ‘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 (COSITE). IEEE (2023), 108–113.
25. [Pandey and Srivastava, 2023] Pandey, G. K., Srivastava, S.: ‘ResNet-18 comparative analysis of various activation functions for image classification’; In 2023 International Conference on Inventive Computation Technologies (ICICT). IEEE (2023), 595–601.
26. [Ranjith et al., 2023] Ranjith, E., Parthiban, L., Latchoumi, T. P., Kumar, S. A., Perera, D. G., Ramaswamy, S.: ‘An effective content based image retrieval system using deep learning based inception model’; Wireless Personal Communications, 133, 2 (2023), 811–829.
27. [Roodaki et al., 2025] Roodaki, A., Sotoodeh, M., Moosavi, M. R.: ‘Genetic Algorithm-Based Feature Selection From High-Dimensional Descriptors for Improved Content-Based Image Retrieval’; IEEE Access, 13 (2025), 198034–198053.
28. [Salih and Abdulla, 2023] Salih, F. A. A., Abdulla, A. A.: ‘Two-layer content-based image retrieval technique for improving effectiveness’; Multimedia Tools and Applications, 82, 20 (2023), 31423–31444.
29. [Singha and Hemachandran, 2012] Singha, M., Hemachandran, K.: ‘Content based image retrieval using color and texture’; Signal & Image Processing, 3, 1 (2012), 39.
30. [Taheri et al., 2024] Taheri, F., Rahbar, K., Beheshtifard, Z.: ‘Content-based image retrieval through fusion of deep features extracted from segmented neutrosophic using depth map’; The Visual Computer, 40, 10 (2024), 6867–6881.

31. [Wang and Qi, 2022] Wang, X., Qi, G.-J.: ‘Contrastive learning with stronger augmentations’; IEEE Transactions on Pattern Analysis and Machine Intelligence, 45, 5 (2022), 5549–5560.
32. [Wangi and Makandar, 2024] Wangi, K., Makandar, A.: ‘Cnn pre-trained model using the fusion of features for cbir framework’; In 2024 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI). IEEE (2024), 1–5.
33. [Wilkins et al., 2005] Wilkins, P., Ferguson, P., Smeaton, A. F., Gurrin, C.: ‘Text based approaches for content-based image retrieval on large image collections’; In 2nd European workshop on the integration of knowledge, semantics and digital media technology (EWIMT 2005). IET (2005), 281–288.
34. [Yu et al., 2024] Yu, S., Hu, T., Gu, M., Li, T., Zhang, D.: ‘Content-Based Galaxy Image Retrieval Using Convolutional Neural Networks’; In 2024 4th International Conference on Consumer Electronics and Computer Engineering (ICCECE). IEEE (2024), 1–4.