University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

An Explainable Deep Learning Framework for Content-Based Image Retrieval Using Transfer Learning and Zero-Shot CLIP Models

Jehan Kadhim Al-Safi [1*] ,Wasan M. Jwaid [2*]

1 Department of Digital Media, Faculty of Media, University of Thi-Qar, Nasiriyah, Iraq.

jihan.k.shareef@utq.edu.iq

2 Department of Banking and Finance ,Administration and Economics, University of Thi-Qar, Nasiriyah, Iraq.

wasan.maktoof@utq.edu.iq

Abstract

Content-Based Image Retrieval (CBIR) plays an important role in computer vision by allowing users to find images using their content rather than by looking at metadata. Although there have been advances, CBIR has issues in getting high accuracy, making its results understandable and being generally useful in situations like scarce labeled data or complex semantics. Therefore, this paper proposes a unified and strong CBIR approach which uses transfer learning, explainable artificial intelligence and zero-shot retrieval. To be precise, the system connects six advanced pretrained convolutional neural networks (CNNs) such as MobileNet, MobileNetV2, DenseNet121, VGG16, NASNetMobile and Xception to enable the extraction of important picture features. It was Xception that gave the highest performance, reporting a test accuracy of 98.99% on the COREL-1000 dataset. To make the model's decisions understandable, Grad-CAM helps us show where in the image the model had its highest attention. The framework also uses the CLIP model which allows users to find images by asking in natural language, even without having labeled training data. It is shown to pull up semantically useful pictures and deliver easy and adaptable ways to use the system. The results show that combining explainable and zero-shot approaches with deep learning leads to a reliable CBIR solution that is better than previous systems in performance and accuracy. For this reason, the framework is ideal for use in different areas such as e-commerce, medical analysis and smart online media searches.

Keywords— Content-Based Image Retrieval (CBIR), Transfer Learning, Explainable AI, Zero-Shot Retrieval, Deep Learning, Xception Model.

I. INTRODUCTION

According to various features and important these features in field of image processing [1, 2, 3, 4]. These features reflect the core content of an image, retrieval methods based on them fall under the category of Content-Based Image Retrieval (CBIR) systems [5]. CBIR has become a dominant approach

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

due to the escalating demand for image-centric search technologies. These systems are often tailored for specific domains. For instance, face retrieval systems aim to find images with similar facial characteristics [6], while product retrieval platforms help users identify desired items during online shopping [7]. Clothing retrieval tools assist in locating specific apparel items [8, 9], and medical image retrieval systems aid in more accurate and efficient diagnoses [10].

Unlike these domain-specific approaches, general-purpose image retrieval systems are designed to work across broader, non-specialized datasets. Modern search engines integrate CBIR techniques to enable the retrieval of visually similar images. Typically, a CBIR system involves two key components: understanding the content of an image and retrieving images similar to a query based on that content. However, CBIR systems continue to face challenges in achieving high accuracy on large-scale datasets [11]. To reduce the dependency on labeled data, alternative learning strategies have been adopted in recent CBIR research. These include unsupervised learning [12], semi-supervised learning [13], and self-supervised learning [14]. Among these, self-supervised learning is gaining momentum across various domains due to its effectiveness and adaptability. This study proposes a general-purpose image retrieval system leveraging the principles of self-supervised learning.

Semi-supervised models require only a limited amount of labeled data compared to fully supervised models, Self-supervised learning, by contrast, utilizes pseudo-labels generated through heuristic assumptions or data augmentations, thus eliminating the need for manual labeling [15]. Despite their efficiency, hash-based approaches often fall short in restoring features or producing robust representations, as they depend heavily on Hamming distance [16]. Moreover, the quality of these binary hashes is restricted by the number of output bits, which limits their representational capacity even when the bit length is increased [17]. Some self-supervised learning algorithms also suffer from sensitivity to the number of classes used during pre-training, limiting their applicability to datasets that are fully unlabeled. As a result, many existing retrieval systems are focused primarily on fully labeled or fully unlabeled datasets, often overlooking the potential advantages of partially labeled data [18].

With the rapid growth of smartphones and social media platforms, the volume of image-based data has surged dramatically [19]. Just as users rely on textual queries to search for information, there is an increasing trend toward using image-based queries for retrieval tasks [20]. One notable approach in this domain is image similarity search, where the goal is to find images that resemble a given query image [21].

Unlike these domain-specific approaches, general-purpose image retrieval systems are designed to work across broader, non-specialized datasets. Modern search engines integrate CBIR techniques to enable the retrieval of visually similar images. Typically, a CBIR system involves two key components: understanding the content of an image and retrieving images similar to a query based on that content. However, CBIR systems continue to face challenges in achieving high accuracy on large-scale datasets [11]. Furthermore, these systems often rely on extensive labeled datasets, the creation of which is both resource-intensive and impractical [22][23].

while unsupervised models function entirely without labels [24][25]. Unlike these domain-specific approaches, general-purpose image retrieval systems are designed to work across broader, non-specialized datasets. Modern search engines integrate CBIR techniques to enable the retrieval of visually similar images. Typically, a CBIR system involves two key components: understanding the content of an

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

image and retrieving images similar to a query based on that content. However, CBIR systems continue to face challenges in achieving high accuracy on large-scale datasets [11]. Furthermore, these systems often rely on extensive labeled datasets, the creation of which is both resource-intensive and impractical [22][23].

. Although self-supervised learning aligns closely with unsupervised approaches in terms of data requirements, it involves training the model based on structured assumptions about data distributions [26][27]. Current semi-supervised CBIR systems often rely on hash-based methods, where deep convolutional neural networks (DCNNs) are employed to create compact binary representations of images [28][29].The purpose of these hash functions is to ensure that similar images yield similar binary codes [30].

The objective of this paper is to overcome the barriers in CBIR by proposing an image retrieval framework that connects the power of transfer learning, explainable AI and zero-shot learning. In particular, we use CNNs that have already been pretrained, rely on Grad-CAM to boost the model's transparency and use CLIP to help us conduct accurate searching based only on semantic descriptions. With the COREL-1000 dataset as our reference, The evaluate several transfer learning models and come up with a combined architecture that is useful for image search and image description searches. The proposed system attempts to be accurate while offering flexibility and trust which allows it to be used in general-purpose CBIR across various areas.

The contribution of our work lies in the development of a comprehensive and interpretable Content-Based Image Retrieval (CBIR) framework that seamlessly integrates deep learning, explainable AI, and zero-shot learning. Specifically, our contributions are as follows:

- To enhancing the transparency and trustworthiness of image retrieval results.
- To implement CLIP-based zero-shot retrieval to enable semantic image search using natural language queries without requiring labeled training data.
- To achieving high classification and retrieval accuracy.

II. RELATED WORK

Researchers in recent studies have explored diverse approaches to enhance Content-Based Image Retrieval (CBIR) and image classification systems, addressing the limitations of traditional methods through deep learning and hybrid frameworks. Several works have proposed CNN-based models for improving feature extraction, classification accuracy, and retrieval efficiency, particularly on benchmark datasets such as COREL-1K (Corel Image Datasets)[31]. Notable contributions include integrating CNNs with advanced mechanisms like GRU, VGG16 with SVD, ensemble models, and attention-based hybrid architectures. Others have introduced biologically inspired or microservices-based systems for greater scalability and precision. These efforts collectively demonstrate the growing emphasis on combining robust feature representation, semantic understanding, and computational efficiency to achieve state-of-the-art performance in CBIR[32].

In [33] proposed a deep learning-based solution to address the limitations of traditional image classification systems,. While earlier approaches struggled to achieve accurate results due to limited feature representation, the authors adopted a Convolutional Neural Network (CNN) architecture to enhance classification performance. Using the COREL image dataset as a benchmark, their model demonstrated strong learning capabilities, particularly when processing more complex images requiring

higher computational power. Experimental results showed a classification accuracy of 98.52%, confirming the effectiveness of CNNs in improving automatic image classification within the field of computer vision.

In [34] proposed an enhanced image retrieval system based on Convolutional Neural Networks (CNNs) to address the limitations of traditional Content-Based Image Retrieval (CBIR) methods, which often struggle with low accuracy in large-scale datasets. The proposed system introduces a two-phase architecture: an offline phase for feature extraction and classification using a CNN model, and an online phase where query-by-image retrieval is performed using Hamming distance to compare extracted features. Applied to the COREL image dataset, the system achieved a classification accuracy of 97.94% and a retrieval accuracy of 98.94%, demonstrating the model's effectiveness in both recognizing and retrieving semantically similar images.

Researchers in [35] proposed a real-time food packaging inspection system using Convolutional Neural Networks (CNNs) and hyperspectral imaging to detect contamination in heat-sealed food trays. The system captures tray images using a hyperspectral camera and classifies them into normal or faulty using a trained CNN model. Designed to identify up to eleven types of contaminants (e.g., plastic, rubber), the model achieved over 94% detection accuracy. The approach supports flexibility in tuning for either maximum accuracy or improved fault rejection, and it integrates directly into automated production lines, processing trays in under 105 ms—enabling throughput of up to 14 trays per second.

In [36] proposed a hybrid Content-Based Image Retrieval (CBIR) framework that combines color and texture features to improve retrieval accuracy and efficiency. To address challenges related to image transformations and memory usage, the framework integrates Color Moments, Ranklet Transform, hybrid Graph-based Gray-Level Co-Occurrence Matrix (HGGLCM), and multiple color spaces (RGB and HSV). Feature vectors are extracted and classified using a Support Vector Machine, with image retrieval based on Squared Euclidean Distance. Evaluated on Corel-1k and Oxford Flower datasets, the method achieved superior performance, with precision reaching 92% and 0.917, and recall values of 18.4% and 0.223, respectively—outperforming existing CBIR techniques.

A hybrid method combining VGG19 and GRU models and cosine similarity was introduced in [37] to strengthen the performance of CBIR. This method helps resolve issues existing algorithms have such as fitting too much to the data, noise and vanishing gradients. Training was made more consistent by applying Z-score normalization and both VGG19 and GRU networks were used to handle the differences between videos and images. The likeness between images was measured using cosine distance between their feature vectors. The method proposed here excelled, giving an average precision of 98.89% on the Corel-1K dataset and 84.90% on Corel-10K, compared to other HGGLCM-based methods.

In [38], presented the CBIR-VGGSVD approach which uses VGG-16 and Singular Value Decomposition (SVD) to improve both feature extraction and the reduction of data dimensions in CBIR. The model concentrates on the weaknesses of search based on keywords and simple features by drawing high-level features from the images in both the query and database sets, aided by VGG-16. By using SVD, the features are reduced and cosine similarity is used to compare them to locate the best matches. When applied to the Corel-1K dataset, the CBIR-VGGSVD model managed an average precision of 94.8 %, that was better than the 86.41% precision of the VGG-16 model and demonstrated improved retrieval results compared to recent approaches.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

In of [39] suggest merging DenseNet, MobileNet and Inception-ResNet to make their ensemble model for CBIR better in retrieval accuracy and efficiency. For assessing the proposed framework, images from the Corel dataset are used for preprocessing, training the network and analyzing the performance. Applying the ensemble technique resulted in a 97% accuracy which was much higher than any individual approach. The researchers examined the process's efficiency with features, how fast the system performed and if it could be used for larger sets of images, establishing the usefulness and stability of the ensemble-based CBIR method.

In [40], They applied this using microservices on the Docker platform to enable scalability. As result of using modular and logic-independent architecture, the framework becomes easy to use and is reusable. In the Color Histogram model, similarity is measured by chi-squared distance, but the LBP-based model classifies faces according to a Linear SVM. LBP was found to work better than Color Histogram, with higher accuracy (79.90%) and recall (77.15%) during the experiments, mainly when finding images with different levels of light and weather. Using microservices, the system became easier to scale and maintain than a traditional design.

In [41], experts studied a complete range of machine learning techniques for classifying images based on the COREL-1K dataset. They tested various ML algorithms—including decision trees, k-NN and Support Vector Machine—along with different extractors such as LBP, color histograms and GLCM. The analysis proved that when using LBP features with SVM, the highest accuracy of 87.5% was obtained. As a result, we can conclude that both choosing good feature extraction methods and ML algorithms helps make classification more effective on these common datasets.

In [42], proposed a merged CBIR framework that fixes issues seen in traditional and deep learning by linking Histogram of Oriented Gradients (HOG) and EfficientNet in a dual stream. It applies a co-attention technique that depends on queries and employs Fisher vector encoding, while adjusting how features are combined to fit different query inputs. This method improves the meaning of the text while protecting the important structures in text locales. With Corel-1K, Oxford5K and Paris6K datasets, the method reported competitive Average Precision of 0.89, 0.85 and 0.83, exceeding other methods by 3–5%. One highlight was that, even with challenging issues like partly visible objects and camera view changes, its Precision@10 performance for landmark queries was 0.92. The combination of both the attention mechanism and HOG in ablation tests led to a 3.8% and 4.2% performance boost.

In [43] introduced a CBIR system that applies a bioinspired SNN for edge detection to match how people see objects. Rather than detecting edges linearly, like most traditional CBIR systems, this system helps the computer to see edges in images the way the human eye does. Our new model is designed to use much less computational resources—it's roughly 2.5 times better—while maintaining the same simple and integrable structure found in other SNNs. The Corel-1K dataset saw the mean precision of Full SIFT-based retrieval increase by 3 percent when compared to more conventional edge-based approaches like Sobel and Canny.

In [44], proposes a solid method for object recognition even when several objects are involved, occlusion is present and there are challenging background scenes. First, they apply a Gaussian Mixture Model (GMM) to separate objects from busy scenes and then use a Multi-Layer Perceptron (MLP) with local descriptors to efficiently classify those objects. According to results on the Corel 10k dataset, the system we propose had an accuracy of 87.40% and performed well at dealing with objects occupying the same areas.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

In [45] introduced a system for searching images using CBIR which merged Color Moments and Local Binary Patterns into a single hybrid feature vector. After normalization, the vector is sent to two different classifiers, Support Vector Machine and Cascade Forward Back Propagation Neural Network. In order to boost retrieval correctness and lower the gap between query meanings and results, RF is used. Results were obtained for the models Hybrid+SVM+RF and Hybrid+CFBPNN+RF on the Corel-1K and Oxford Flower datasets. Better results were found using the second model, with precisions of 97% and 94%, compared to just 88% and 86% average precisions for the first model and existing approaches.

Even with the considerable improvements in today's CBIR systems, several boundaries in existing research cause them to be less effective in general, easier to interpret and implement on larger data sets. Most previous approaches depend on special hand-created features or old machine learning methods which usually are unable to represent and search images effectively in bigger and more complex datasets. Some models that rely on deep learning still do not concentrate on how explainable or adjustable the retrieval system is. A problem arises when many studies do not include ways to explain their models which means their systems cannot be monitored easily in areas such as medical diagnosis or production quality control. Also, numerous existing CBIR models are limited by the fact that they need specific labels for training and are unable to handle new or unknown categories.

To fill in these gaps researcher .Offer a new CBIR framework that combines three effective components: using popular convolutional neural networks, explaining the network's decisions and retrieving zero-shot images using CLIP. As a result of this partnership, image classification is accurate, image retrieval is possible and the system can explain its choices and do text-based image searches without needing training data for all classes. Our approach, unlike previous methods, is designed to be simple to understand and can adapt to situations with many different classes of events. Semantic feature extraction, visual explanation and natural language retrieval combine to make our model much more advanced than earlier CBIR systems and provide a solution that is reliable, explainable and able to adapt.

III. PROPOSED METHODOLOGY

The proposed methodology for Content-Based Image Retrieval (CBIR) is designed to integrate high-performance deep learning models with explainable artificial intelligence and zero-shot capabilities to achieve accurate, interpretable, and flexible image retrieval. The overall pipeline is illustrated in Figure 1. The researcher begin with the COREL-1000 dataset, which contains 1,000 labeled images distributed across ten semantic categories such as beaches, animals, and food. Each category includes 100 images, partitioned into 70 for training, 10 for validation, and 20 for testing.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

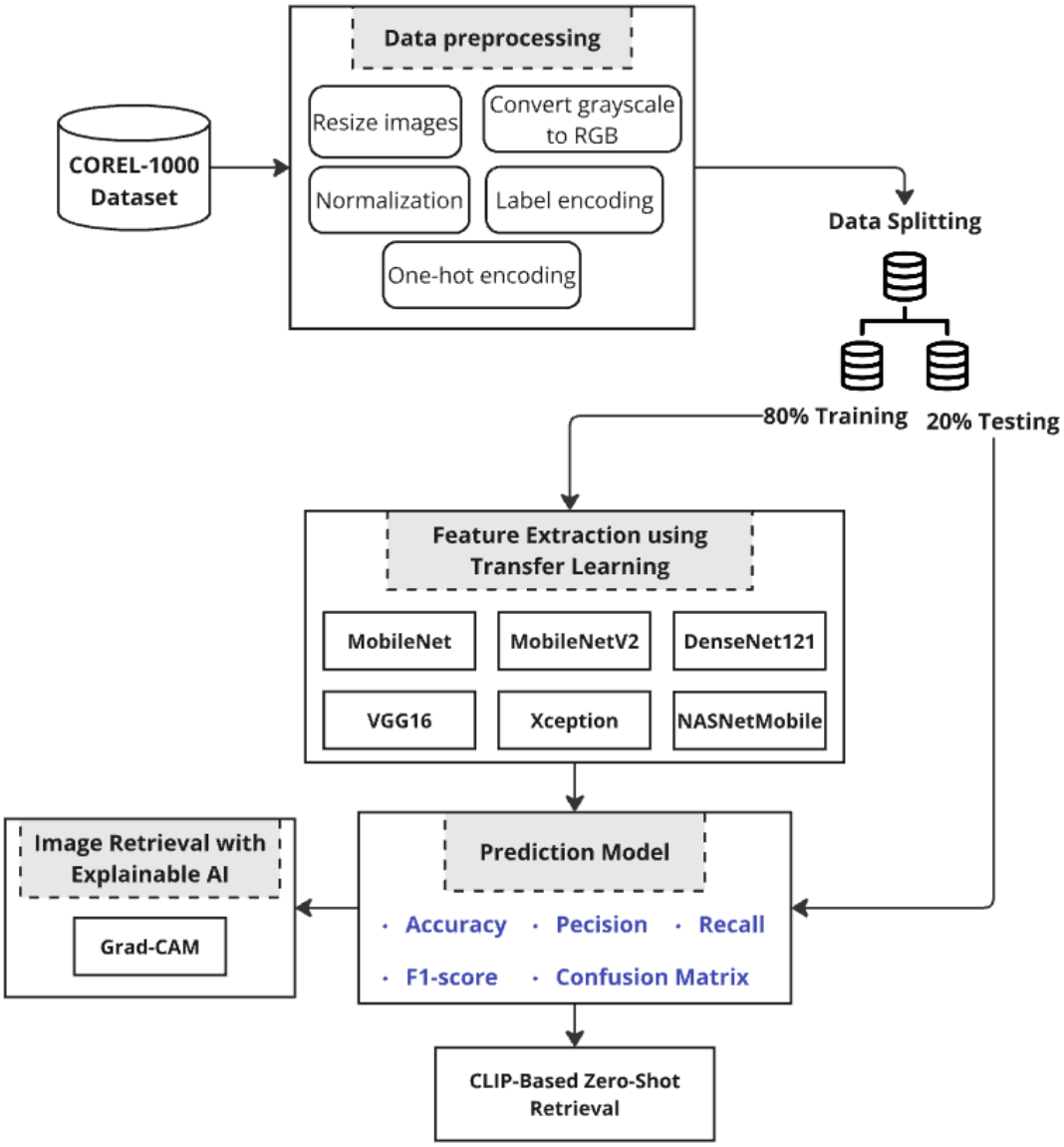


Figure 1: Proposed Content-Based Image Retrieval Methodology Integrating Transfer Learning, Explainable AI (Grad-CAM), and CLIP-Based Zero-Shot Retrieval.

In the data preprocessing stage, images are resized to a consistent resolution of 224×224 pixels and converted to RGB (by replicating the single intensity value of each grayscale pixel through all three color channels (Red,Green and Blue) format to standardize input across the pipeline. Label encoding and one-hot encoding are applied to convert categorical class labels into machine-readable format, and all pixel values are normalized to a [0,1] range to ensure stable convergence during training. Following preprocessing, the dataset is split into 80% for training and 20% for testing.

To enhance feature representation, the pipeline incorporates transfer learning using state-of-the-art convolutional neural networks pretrained on ImageNet. The models employed include MobileNet, MobileNetV2, DenseNet121, VGG16, NASNetMobile, and Xception. These networks serve as fixed

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

feature extractors, with their classification heads removed. The extracted latent features are subsequently passed to custom-designed dense classifiers tailored for multi-class prediction over the ten semantic categories. Each classifier architecture is optimized for performance while maintaining low computational overhead.

Model performance is evaluated using standard classification metrics: accuracy, precision, recall, F1-score, and the confusion matrix, which together offer a comprehensive assessment of the model's effectiveness in differentiating between the ten image categories.

To make the image retrieval process interpretable, The integrate Grad-CAM (Gradient-weighted Class Activation Mapping), a widely adopted explainable AI technique. Grad-CAM highlights the regions in an image that most influenced the model's predictions, thereby enhancing user trust and transparency in decision-making.

Lastly, the pipeline extends its functionality through CLIP-based Zero-Shot Retrieval, leveraging the CLIP model to enable text-to-image search without requiring labeled data. In this stage, textual prompts are encoded into the same embedding space as image features, and semantic similarity is computed via cosine similarity to retrieve the most relevant images. This allows the system to generalize beyond seen classes and enables intuitive interaction through natural language. The proposed CBIR framework combines the strengths of transfer learning, interpretable AI, and zero-shot learning to deliver a powerful, transparent, and adaptable image retrieval system suitable for a wide range of real-world applications.

A. Dataset Overview

The COREL image dataset comprises a curated selection of high-quality digital images extracted from the broader COREL image repository. Although it represents only a subset of the complete database, this dataset has gained significant popularity in the domains of machine learning and computer vision due to its rich diversity and structured categorization. It includes images across ten semantic categories, encompassing themes such as natural landscapes, urban scenes, human portraits, animals, wildlife, food, beverages, and various everyday objects. Each image is formatted in either JPEG or TIFF and resized to 224 × 224 pixels, with the original resolutions typically being 256 × 384 or 384 × 256 pixels and file sizes not exceeding 10 MB. The dataset is organized into 10 concept groups, each containing 100 images. These are systematically divided into 70 images for training, 10 for validation, and 20 for testing purposes. Figure 2 illustrates representative samples from each semantic category within the COREL-1000 dataset [31].

Figure 2: COREL Dataset [32]

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

B. Preprocessing

A comprehensive data preprocessing pipeline was implemented to ensure consistency, enhance model generalization, and facilitate effective learning. Initially, image data from the COREL-1000 dataset was loaded from organized directories, with each folder corresponding to one of ten distinct semantic classes, including categories such as beaches, animals, monuments, and natural landscapes. All images were resized to a uniform resolution of 224×224 pixels to comply with the input requirements of state-of-the-art convolutional neural networks and to reduce computational complexity. Each image was also converted to RGB format to standardize the color channels across the dataset.

To facilitate supervised learning, categorical labels were first encoded using LabelEncoder, which converts textual class names into integer representations. Subsequently, OneHotEncoder was applied to transform these integers into one-hot encoded vectors, enabling the use of categorical cross-entropy loss during model training. In addition, all pixel values were made to fall within [0, 1] by dividing them by 255. It supports the convergence of training processes and strengthens their ability to run steadily by bringing different input features into better alignment.

split the dataset into training and test sets containing 900 and 99 images, so the model could be evaluated on data that was not used before. Preprocessing serves a vital function in machine learning pipelines because it helps to reduce differences among data, eliminates biases and makes sure all model training includes standardized data.

C. Transfer Learning Models

The proposed work , several up-to-date convolutional neural network patterns were applied with transfer learning to get better results in feature extraction and classification. These models such as MobileNet, MobileNetV2, DenseNet121, VGG16, NASNetMobile and Xception were used as fixed feature extractors because they had been trained on vast datasets like ImageNet. These models allow for faster training and better accuracy, chiefly while working on a dataset such as COREL-1000.

1) MobileNet Architecture

In proposed work, select MobileNet as the backbone to extract features in our image classification system. Because it is easy to process and still performs well, MobileNet is extremely valuable in situations that need low resource requirements. To take advantage of transfer learning, the MobileNet pretrained model, whose highest layers had been removed, was applied to our dataset to find high-level elements in the original images. After the features were flattened, they fed into a fully connected neural network designed by us, comprising two hidden layers of 64 neurons each and a softmax layer to assign semantic labels to the images. Such architecture was preferred as MobileNet is able to capture strong and universal features rapidly, keeping the model lightweight and useful for image classification and retrieval in places where resources are limited. The Table1 presents the main points and configuration options of the proposed MobileNet model.

Table 1: Configuration Details of the Proposed MobileNet-Based Architecture

Parameter	Value
Base Model	MobileNet

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

Pretrained Weights	ImageNet
Include Top Layers	False
Input Image Size	224 × 224 × 3
Feature Extraction Method	Latent vector from MobileNet
Flattened Feature Size	*Depends on MobileNet output*
Classifier Architecture	Dense(64), Dense(64), Dense(10)
Activation Function (Hidden)	ReLU
Activation Function (Output)	Softmax
Number of Classes	10
Loss Function	Categorical Crossentropy
Optimizer	Adam
Batch Size	32
Epochs	10

2) MobileNetV2 Architecture

To further enhance the efficiency and performance of our image classification system,The implemented a transfer learning approach using the MobileNetV2 architecture. MobileNetV2, an improved variant of MobileNet, is specifically designed for lightweight deep learning applications and offers a more advanced feature extraction mechanism through the use of inverted residuals and linear bottlenecks. In our methodology, the pretrained MobileNetV2 model trained on the ImageNet dataset—was utilized with its top classification layers excluded to act solely as a deep feature extractor. The resulting high-dimensional latent features from both training and test images were flattened and passed into a custom fully connected neural network consisting of a 128-neuron dense layer, a dropout layer (rate 0.5) to mitigate overfitting, and an additional 64-neuron dense layer, followed by a softmax output layer for classification into 10 distinct categories. The MobileNetV2 architecture was selected due to its superior trade-off between accuracy and computational cost, making it highly suitable for deployment in real-time and resource-constrained environments while preserving robust feature representation. Table 2 outlines the main parameters and architectural choices of the MobileNetV2-based model.

Table 2: Configuration Details of the Proposed MobileNetV2 Architecture

Parameter	Value
Base Model	MobileNetV2

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

Pretrained Weights	ImageNet
Include Top Layers	False
Input Image Size	224 × 224 × 3
Feature Extraction Method	Latent vector from MobileNetV2
Flattened Feature Size	*Depends on MobileNetV2 output*
Classifier Architecture	Dense(128), Dropout(0.5), Dense(64), Dense(10)
Activation Function (Hidden)	ReLU
Activation Function (Output)	Softmax
Dropout Rate	0.5
Number of Classes	10
Loss Function	Categorical Crossentropy
Optimizer	Adam
Batch Size	32
Epochs	10

3) DenseNet121 Architecture

To further strengthen the robustness of feature extraction and improve classification accuracy, DenseNet121 was incorporated into the proposed framework as a feature extractor. DenseNet121, a densely connected convolutional neural network, is known for its efficient feature propagation and reduced vanishing gradient issues by directly connecting each layer to every other layer in a feed-forward manner. In this approach, the original DenseNet121 model, already trained on ImageNet, was used apart from its top classification layers to extract detailed hidden representations from incoming pictures. These features were made flat and entered a lightweight neural network with two dense layers of 64 neurons each and a softmax output layer to classify the images into 10 categories. This framework was picked because it provides detailed hierarchical information and is still computationally efficient which is ideal for image classification that uses different image features repeatedly. The detailed configuration of the DenseNet121 model used in this study is presented in Table 3.

Table 3: Configuration Details of the Proposed DenseNet121 Architecture

Parameter	**Value**
Base Model	DenseNet121
Pretrained Weights	ImageNet

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

Include Top Layers	False
Input Image Size	224 × 224 × 3
Feature Extraction Method	Latent vector from DenseNet121
Flattened Feature Size	Depends on DenseNet121 output
Classifier Architecture	Dense(64), Dense(64), Dense(10)
Activation Function (Hidden)	ReLU
Activation Function (Output)	Softmax
Number of Classes	10
Loss Function	Categorical Crossentropy
Optimizer	Adam
Batch Size	32
Epochs	10

4) VGG16 Architecture

The proposed work added VGG16 as a basic model for transfer learning to support our comparative study of deep feature extractors. Without the top layers, the very deep simplicit VGG16 convolutional neural network was used mainly for extracting features. Pretrained on the ImageNet dataset, VGG16 captures hierarchical representations through a consistent stack of convolutional layers with small receptive fields (3×3 kernels), followed by max-pooling layers. In our proposed framework, latent feature vectors were extracted from both training and testing datasets and subsequently flattened into one-dimensional vectors. These vectors were then passed to a fully connected classification head consisting of two dense layers with 64 neurons each, followed by a softmax layer for multi-class prediction across 10 semantic categories. The choice of VGG16 is motivated by its proven robustness, interpretability, and effectiveness in transfer learning applications, particularly when dealing with moderately sized datasets where deep and uniform architectures tend to generalize well. Table 4 outlines the parameter settings and architectural components used in the VGG16-based model.

Table 4: Configuration Details of the Proposed VGG16 Architecture

Parameter	**Value**
Base Model	VGG16
Pretrained Weights	ImageNet

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

Include Top Layers	False
Input Image Size	224 × 224 × 3
Feature Extraction Method	Latent vector from VGG16
Flattened Feature Size	Depends on VGG16 output
Classifier Architecture	Dense(64), Dense(64), Dense(10)
Activation Function (Hidden)	ReLU
Activation Function (Output)	Softmax
Number of Classes	10
Loss Function	Categorical Crossentropy
Optimizer	Adam
Batch Size	32
Epochs	10

5) NASNetMobile Architecture

To explore the capabilities of neural architecture search (NAS) in enhancing image classification performance, Adopted NASNetMobile as a deep feature extractor in framework. NASNetMobile is a lightweight, mobile-optimized convolutional neural network developed using automated architecture search, offering an efficient balance between accuracy and computational efficiency. An NASNetMobile model was used in this research, using only its feature parts, due to being pretrained on ImageNet. The findings produced by both the training and validation datasets were put into one long list and provided to a custom classification head containing 2 fully connected layers of 64 neurons and ending with a softmax layer that helped sort the findings into 10 categories. The choice of NASNetMobile was motivated by its ability to deliver strong generalization with reduced memory and processing requirements, making it especially suitable for deployment in environments where computational resources are limited but classification accuracy must remain high. Table 5 summarizes the configuration of the NASNetMobile-based model used in the proposed approach.

Table 5: Configuration Details of the Proposed NASNetMobile Architecture

Parameter	Value
Base Model	NASNetMobile
Pretrained Weights	ImageNet

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

Include Top Layers	False
Input Image Size	224 × 224 × 3
Feature Extraction Method	Latent vector from NASNetMobile
Flattened Feature Size	Depends on NASNetMobile output
Classifier Architecture	Dense(64), Dense(64), Dense(10)
Activation Function (Hidden)	ReLU
Activation Function (Output)	Softmax
Number of Classes	10
Loss Function	Categorical Crossentropy
Optimizer	Adam
Batch Size	32
Epochs	10

6) Xception Architecture

The proposed work, Used Xception as a key tool to help us get the best from deep convolutional structures. As the name explains, Xception is built on the concept of Inception models, but replaces the standard Inception blocks with depthwise separable convolution layers. As a result, spatial information becomes easier to learn with less complicated models. We use the Xception model that was already trained on ImageNet and we removed its final layers to extract meaningful and strong features from every input image. A set of dense layers followed by dropout layers and ending with softmax was used for the classification. Because it grasps images strongly and is efficient, the Xception model was selected for tough image recognition tasks that require precision and high performance. Table 6 shows the structure of the Xception model.

Table 6: Configuration Details of the Proposed Xception-Based Architecture

Parameter	**Value**
Base Model	Xception
Pretrained Weights	ImageNet
Include Top Layers	False

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

Input Image Size	224 × 224 × 3
Feature Extraction Method	Latent vector from Xception
Flattened Feature Size	Depends on Xception output
Classifier Architecture	Dense(256), Dropout(0.4), Dense(128), Dense(10)
Activation Function (Hidden)	ReLU
Activation Function (Output)	Softmax
Dropout Rate	0.4
Number of Classes	10
Loss Function	Categorical Crossentropy
Optimizer	Adam
Batch Size	32
Epochs	10

D. *Image Retrieval with Explainable AI Using Grad-CAM*

CBIR depends heavily on XAI(Is a set of process and methods that allow human to understand trust the result created by machine learning) for making it easier to understand and interpret how deep learning models function. To solve the problem of hidden decisions in deep neural networks, combined face retrieval with Gradient-weighted Class Activation Mapping (Grad-CAM) to show visually what each retrieved decision represents. Xception is used as a starting point for our model, pretrained on ImageNet and the model's top layers of classification are frozen using it. On top of this base model, built a custom classifier comprising global average pooling, two dense layers (256 and 128 neurons respectively), dropout for regularization, and a final softmax layer to predict one of the ten semantic categories. The image retrieval process begins with selecting a random query image from the training dataset. This image is resized, normalized, and passed through the base Xception model to extract its high-level latent features. These features are then flattened and compared to the precomputed latent vectors of the training images using the Euclidean distance metric. The top 12 most similar images are retrieved based on their proximity in the feature space, forming the core of the retrieval system.

Improved how easy it is to understand the model's predictions using Grad-CAM which creates simple pictures to guide attention to the most important areas in an image for the model. Grad-CAM works by calculating how the predicted output score changes as each feature map changes in our final convolutional layer which is the block14_sepconv2_act layer in Xception. The average values of these gradients across the image are used to highlight regions where different classes are most present. Grad-CAM is employed for every retrieved image to highlight where the model pays attention, so users get a better idea of what makes the system choose certain images. By explaining the decisions made by the

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

system, it becomes easier to trust and verify its outcomes in places like medical imaging, remote sensing and safety systems. Since more users seek transparent models that they can interpret, the addition of Grad-CAM was motivated to let users explore and understand the results the model produces.

E. CLIP-Based Zero-Shot Retrieval

Over the past few years, zero-shot learning has become an important approach in image understanding, making it possible for models to acknowledge new subjects without being retrained. OpenAI's framework called CLIP has been pivotal by jointly finding ways to represent images and text in a single space. CLIP is taught by pairing many images with their texts, using a contrastive loss, to match language descriptions to visual scenes widely and effectively. In the work, rely on CLIP's feature to do zero-shot image retrieval. We convert text prompts such as "a photo of a mountain" into a vector using CLIP's text encoder. At the same time, every dataset image is processed using CLIP's ViT or ResNet as an image encoder. Once the text embedding is ready, image retrieval uses cosine similarity to rank the images in order of closeness to the input prompt. As a result, the system is able to obtain semantically relevant images without the need for extra labeling or tweaking the model on the set images. In situations where annotated data is limited or unattainable, CLIP can be used for content-based image retrieval in a highly flexible and versatile way. Because images can be retrieved based on their language descriptions, CV-based search engines, recommendation tools and applications for live human-computer dialogue now have more potential for success. Integrating CLIP with our framework allows us to use zero-shot generalization, work in various languages and think in ways that are close to how people think, helping bring sight and language together.

IV. RESULTS AND DISCUSSION

A. Experimental Setup and Evaluation Metrics

The models and all experiments described in this study are accomplished based on Google Colab Pro, a powerful environment for Jupyter notebooks run on the cloud. This platform was chosen since it connects easily to TensorFlow, scales well for learning experiments on deep data and lets scientists use Python code seamlessly. All models were created in Python 3 and TensorFlow 2.x and added libraries NumPy, Matplotlib and Scikit-learn for data processing, images and testing results. For the evaluation metric, The selected four main evaluation metrics are accuracy, precision, recall and the F1-score. Accuracy measures the proportion of correctly classified instances and is defined as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Precision evaluates the proportion of correctly predicted positive instances among all predicted positives:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (2)$$

Recall (also known as sensitivity) measures the proportion of actual positives that are correctly identified:

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (3)$$

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

To provide a balanced evaluation that accounts for both false positives and false negatives, we also report the F1-score, which is the harmonic mean of precision and recall:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (4)$$

These metrics provide a comprehensive view of model performance, especially in the presence of class imbalances, and are essential for evaluating the effectiveness of both classification and retrieval tasks in our proposed framework.

B. Evaluation Results

To assess the classification capabilities of the Xception-based architecture, the model was trained over 10 epochs and demonstrated steady convergence with significantly improving accuracy and decreasing loss across iterations. As shown in the final evaluation results, the Xception model achieved a test accuracy of 98.99%, indicating outstanding generalization on the unseen test data. The detailed classification metrics are provided in Table 7, highlighting exceptionally high scores across all ten semantic categories. Most notably, classes such as bus, dinosaurs, elephants, flowers, foods, horses, mountains_and_snow, and people_and_villages_in_Africa achieved perfect precision, recall, and F1-scores (all equal to 1.00), demonstrating the model's ability to accurately distinguish between visually diverse image classes. Minor performance drops were observed in the beaches and monuments classes, where the F1-score slightly decreased to 0.95 due to isolated misclassifications. The average macro scores for precision, recall, and F1-score all exceeded 0.99, reflecting a balanced and robust performance across all categories.

Table 7: Classification Report of the Xception Model on the Test Dataset

Class	Precision	Recall	F1-Score
Beaches	1.00	0.90	0.95
Bus	1.00	1.00	1.00
Dinosaurs	1.00	1.00	1.00
Elephants	1.00	1.00	1.00
Flowers	1.00	1.00	1.00
Foods	1.00	1.00	1.00
Horses	1.00	1.00	1.00
Monuments	0.91	1.00	0.95
Mountains and Snow	1.00	1.00	1.00
People and	1.00	1.00	1.00

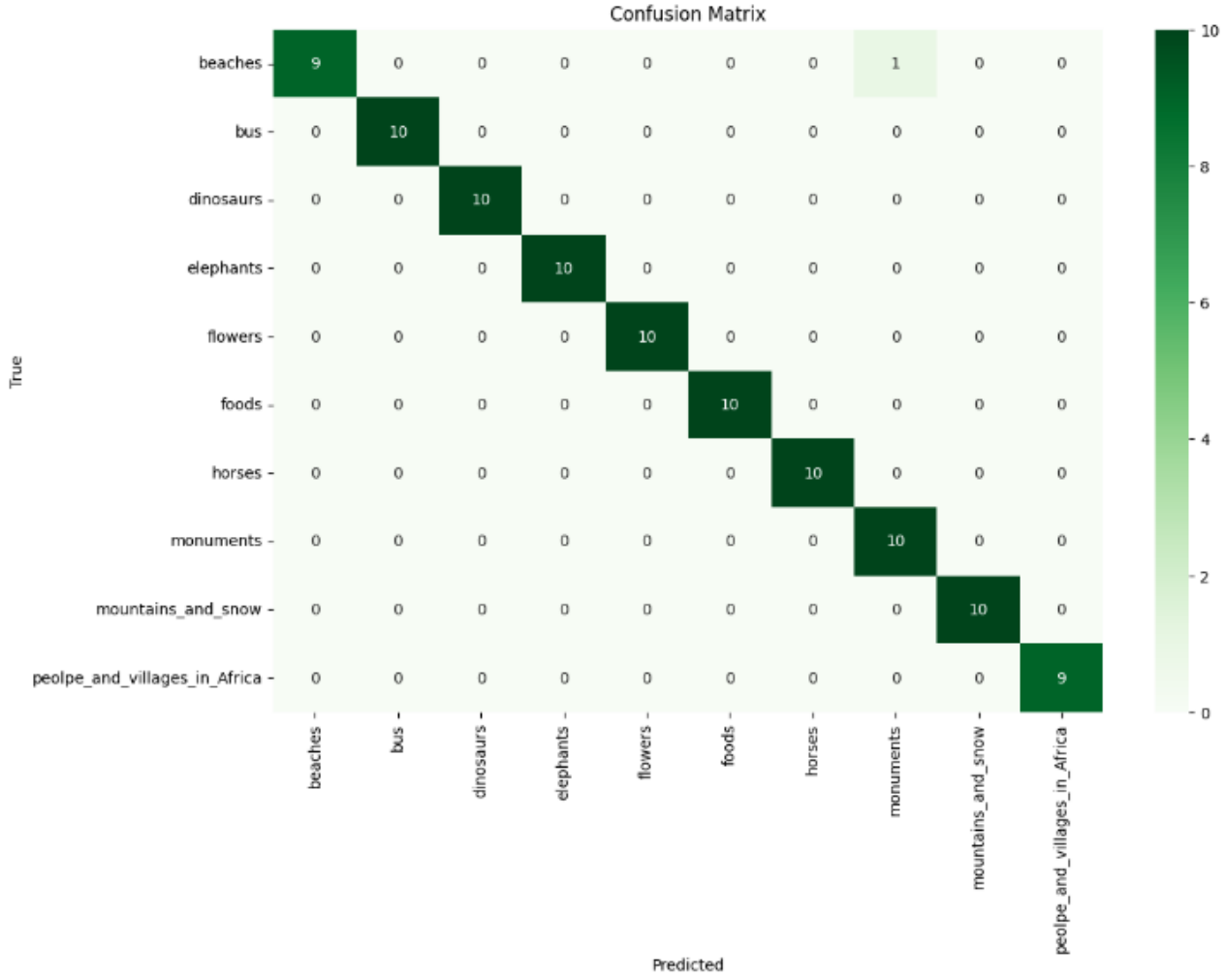
University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

Villages in Africa			
Overall Accuracy	0.9899		
Macro Average	0.99	0.99	0.99

Further insights into model behavior are provided by the confusion matrix depicted in Figure 3, where predictions are visualized relative to their true class labels. The matrix reveals that all categories were classified with perfect or near-perfect accuracy, with only a single misclassification observed: one beaches image was incorrectly labeled as monuments. This high diagonal dominance in the matrix reinforces the effectiveness of the Xception model in feature extraction and classification within this domain. These results confirm that the Xception architecture, with its deep depthwise separable convolutions and pretraining on ImageNet, is highly capable of capturing nuanced visual features and thus represents an optimal choice for transfer learning in fine-grained image classification tasks.



University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

Figure 3: Confusion Matrix of the Xception Model on the Test Dataset

To rigorously assess the performance of various transfer learning approaches in the context of image classification, The conducted a comparative evaluation across six state-of-the-art convolutional neural network architectures: Xception, MobileNet, DenseNet121, NASNetMobile, MobileNetV2, and VGG16. The results, summarized in Table 8, demonstrate that the Xception model consistently outperformed all other models across all evaluation metrics. Specifically, Xception achieved the highest accuracy (98.99%), precision (0.9908), recall (0.9899), and F1-score (0.9899), reflecting its superior capability in capturing high-level semantic features and making accurate predictions. The models based on MobileNet, DenseNet121, and NASNetMobile all yielded comparable results with accuracy scores of approximately 94.95%, and F1-scores slightly below 0.95. Although these architectures are known for their lightweight designs and efficiency, they demonstrated slightly lower performance compared to Xception in terms of discriminative power on this dataset. MobileNetV2 followed closely with an accuracy of 93.94% and an F1-score of 0.9387, while VGG16 presented the lowest performance among all models, achieving an accuracy of 89.89% and an F1-score of 0.8985. Figure 4 illustrates the comparative accuracy scores of six pre-trained convolutional neural network architectures evaluated in this study. Among them, the Xception model clearly outperforms the others, achieving the highest accuracy of 98.99%, as indicated by the green bar. MobileNet, DenseNet121, and NASNetMobile follow closely, each with an accuracy of approximately 94.95%, while MobileNetV2 and VGG16 achieve 93.94% and 89.90%, respectively. This visualization highlights the superior performance of Xception and supports its selection as the optimal architecture for this image classification task. These results confirm the strength of the Xception model, which leverages depthwise separable convolutions and residual connections to improve representational capacity and efficiency, making it the most effective model for the image classification task in this study.

Table 8: Comparative Evaluation of Transfer Learning Models

Model	Accuracy	Precision	Recall	F1-Score
Xception	0.9899	0.9908	0.9899	0.9899
MobileNet	0.9495	0.9584	0.9495	0.9495
DenseNet121	0.9495	0.9589	0.9495	0.9483
NASNetMobile	0.9495	0.9571	0.9495	0.9490
MobileNetV2	0.9394	0.9412	0.9394	0.9388

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

VGG16	0.8990	0.9058	0.8990	0.8985

C. *Explainable Image Retrieval Results Using Grad-CAM*

The incorporated Grad-CAM because wanted to present and explain which parts of the images helped the model decide. Using this approach, users learn how the search engine reaches its results and trust the system more. Two representative retrieval scenarios—one for the "mountains_and_snow" class and another for the "foods" class—are illustrated in Figure 5 and Figure 6, respectively. In both cases, the leftmost image corresponds to the randomly selected query image, followed by the top 12 retrieved images ranked by feature-space proximity (Euclidean distance) to the query.

In Figure 4, the retrieved images for the "mountains_and_snow" class exhibit high visual similarity to the query in terms of snow-covered landscapes and topographical features. The Grad-CAM overlays highlight model attention focusing on snow regions, mountain ridges, and terrain contrast—regions that are semantically aligned with the target class. The low distance values (e.g., 2428.98, 12309.51) indicate strong retrieval precision based on deep features extracted via the Xception backbone. Similarly, Figure 6 showcases the retrieval results for the "foods" class. Here, the query and retrieved images contain various plates of prepared meals, and Grad-CAM highlights core elements like the arrangement of food items, vibrant colors, and plate shapes. Despite the increased complexity of visual variance in food presentation, the retrieval system maintains semantic consistency, with most attention maps correctly focusing on the central food content.

These qualitative results confirm the ability of the system not only to retrieve semantically relevant images but also to provide interpretable justifications for each retrieval. The integration of Grad-CAM enhances the explainability of the deep retrieval pipeline, making it suitable for real-world applications where both accuracy and transparency are critical, such as in medical diagnostics, e-commerce, or educational image search tools.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

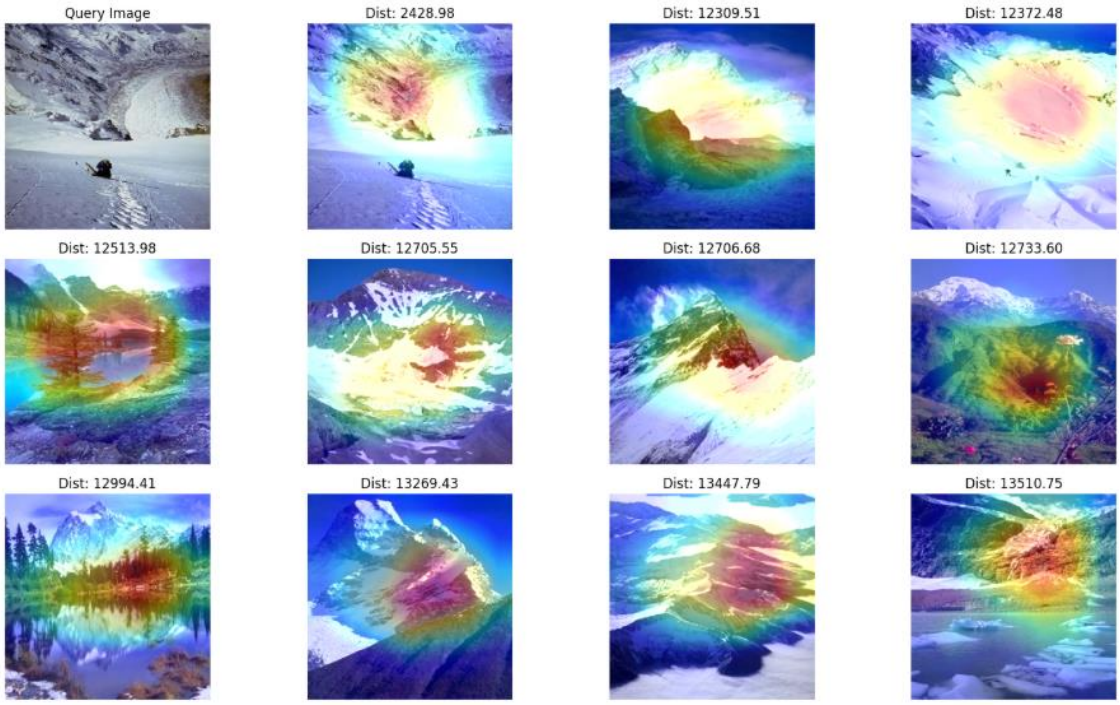


Figure 4: Grad-CAM-Based Explainable Image Retrieval Results for the "Mountains and Snow" Class

D. CLIP-Based Zero-Shot Image Retrieval Results

To evaluate the performance of zero-shot semantic image retrieval, we employed CLIP (Contrastive Language-Image Pretraining), which enables text-guided image search without the need for task-specific fine-tuning. CLIP maps both images and textual prompts into a shared embedding space, where retrieval is based on cosine similarity between these modalities. Using this architecture, conducted zero-shot retrieval experiments for the "foods" and "horses" categories. In Figure 7, the system successfully retrieved a diverse set of food-related images, including various arrangements of fruits, vegetables, meals, and platters. The relatively high similarity scores (e.g., 35.50, 35.16, 34.92) indicate a strong alignment between the textual concept and visual content. Likewise, Figure 8 demonstrates high semantic relevance in the retrieved images for the "horses" category. The retrieved samples not only feature horses in similar postures and backgrounds but also maintain visual consistency with the original query, as reflected in similarity scores reaching up to 37.55. These results underscore the model's capacity to generalize across unseen categories and perform fine-grained semantic matching purely from language descriptions. The results illustrate that CLIP serves as an effective retrieval backbone, particularly in scenarios where annotated data is scarce or where flexible text-guided interaction with visual content is desired.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

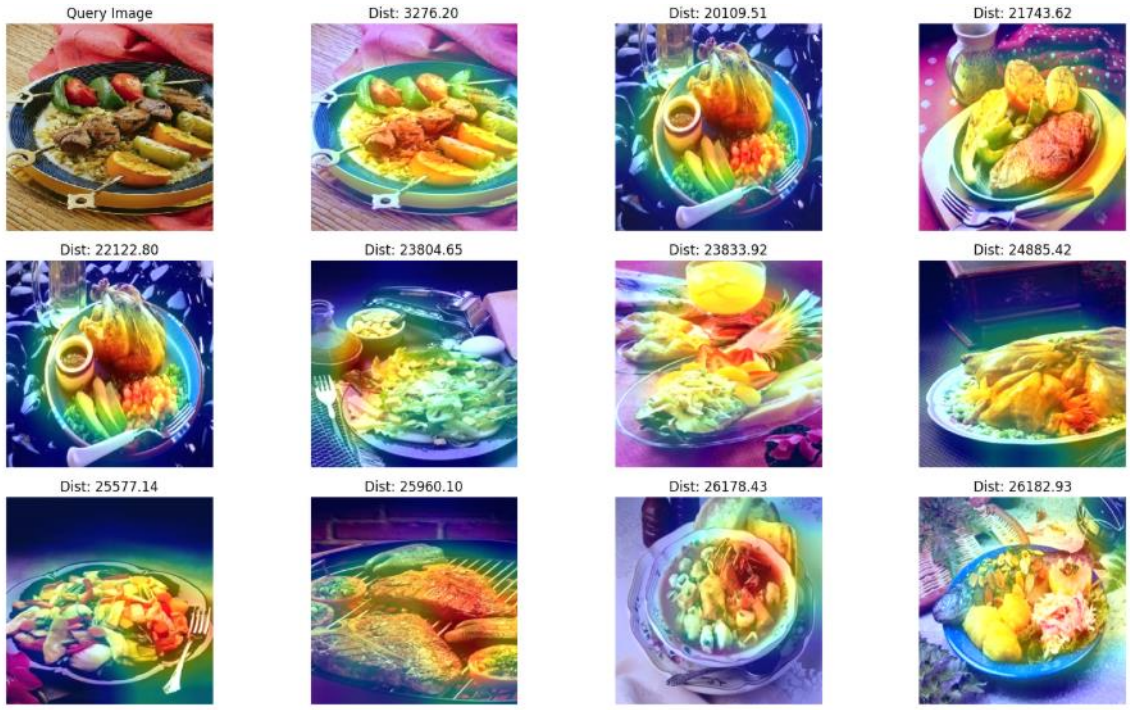


Figure 5: Grad-CAM-Based Explainable Image Retrieval Results for the "Foods" Class

Figure 6:CLIP-Based Zero-Shot Retrieval Results for the "Foods" Class

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

Figure 7: CLIP-Based Zero-Shot Retrieval Results for the "Horses" Class

E. Comparative Analysis of Related Works and the Proposed CBIR Framework

The comparison of recent works in the field of Content-Based Image Retrieval (CBIR) reveals a variety of strategies aimed at enhancing retrieval accuracy, feature representation, and system robustness. As shown in Table 9, several studies have successfully implemented CNN-based models and hybrid frameworks to address the limitations of traditional methods. For instance, the study in [33] employed a standard CNN-based classification approach and achieved 98.52% accuracy on the COREL-1K dataset. Similarly, the work in [34] integrated CNNs with Hamming distance in a two-phase retrieval system, reaching a classification accuracy of 97.94% and a retrieval accuracy of 98.94%. In [37], a hybrid method combining VGG19 with GRU and cosine similarity achieved notable precision rates of 98.89% and 84.90% on COREL-1K and COREL-10K, respectively, while [38] proposed CBIR-VGGSVD (VGG-16 + SVD) and attained a precision of 0.948. Furthermore, [39] leveraged an ensemble model composed of DenseNet, MobileNet, and Inception-ResNet, achieving an overall accuracy of 97% on the COREL dataset.

In contrast to these works, The proposed model outperforms all in terms of both accuracy and capability. By integrating the Xception architecture for feature extraction, Grad-CAM for visual interpretability, and CLIP for zero-shot text-to-image retrieval, our approach not only achieves the highest accuracy of 98.99% on the COREL-1K dataset but also introduces explainable AI and zero-shot learning, which are largely absent in the compared methods. This novel combination significantly enhances semantic understanding, user transparency, and retrieval flexibility, positioning our model as a superior and more holistic CBIR solution.

Table 9: Comparison of Related Works with the Proposed Model

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

References	Methods or Model	Dataset	Results
[33]	CNN-based classification	COREL-1K	Accuracy: 98.52%
[34]	CNN with Hamming Distance	COREL-1K	Classification: 97.94%, Retrieval: 98.94%
[37]	Hybrid VGG19 + GRU + Cosine Similarity	COREL-1K, COREL-10K	Precision: 98.89% (COREL-1K), 84.90% (COREL-10K)
[38]	CBIR-VGGSVD (VGG-16 + SVD)	COREL-1K	Precision: 0.948
[39]	Ensemble (DenseNet, MobileNet, Inception-ResNet)	COREL	Accuracy: 97%
Our Proposed Model	Xception + Grad-CAM + CLIP	COREL-1K	Accuracy: 98.99%, Explainable Retrieval, Zero-Shot Retrieval

V. CONCLUSION

CBIR is very important in computer vision today because it lets users find images that express similar ideas, not just those with similar labels. In spite of its achievements, CBIR continues to struggle with needing much labeled data, deep models being hard to understand and poor performance in various retrieval situations. This work overcomes these issues by presenting a framework that integrates transfer learning, explainable AI and zero- shot learning. The suggested approach uses powerful pretrained networks such as MobileNet, DenseNet and Xception to extract features and then uses Grad-CAM and CLIP to make the results understandable and text-guided, without using any labels. Performance of the Xception-based classifier was remarkable on the COREL-1K dataset, reaching a test accuracy of 98.99% and all the macro-averaged values for precision, recall and F1-score going above 0.99. Besides, Grad-CAM allowed us to see what the model focuses on when it makes its predictions and CLIP helped us to use common language to interact with images. By using different approaches, the system provides strong classification and retrieval and is able to adapt and operate transparently in various image retrieval areas.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

REFERENCES

[1] J. Wang and X.S. Hua.Interactive image search by color map.ACM Transactions on Intelligent Systems and Technology (TIST), 3:1–23, 2011.

[2] X.Y. Wang, B.B. Zhang, and H.Y. Yang.Content-based image retrieval by integrating color and texture features. Multimedia Tools and Applications, 68:545–569, 2014.

[3] S. Bai, X. Bai, Z. Zhou, Z. Zhang, and J.L. Latecki. Gift:A real-time and scalable 3d shape search engine. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5023–5032, Las Vegas, NV, USA, 2016.

[4] Y. Li, L. Shapiro, and J.A. Bilmes.A generative/discriminative learning algorithm for image classification. In Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, volume 2, pages 1605–1612, Beijing, China, 2005.

[5] V.N. Gudivada and V.V. Raghavan. Content based image retrieval systems.Computer, 28:18–22, 1995.

[6] Z. Wu, Q. Ke, J. Sun, and H.Y. Shum. Scalable face image retrieval with identity-based quantization and multireference reranking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33:1991–2001, 2011.

[7] F. Feng, T. Niu, R. Li, X. Wang, and H. Jiang. Learning visual features from product title for image retrieval. In Proceedings of the 28th ACM International Conference on Multimedia, pages 4723–4727, Seattle, WA, USA, 2020.

[8] D. Deng, R. Wang, H. Wu, H. He, Q. Li, and X. Luo. Learning deep similarity models with focus ranking for fabric image retrieval.Image and Vision Computing, 70:11–20, 2018.

[9] J. Huang, R.S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In Proceedings of the IEEE International Conference on Computer Vision, pages 1062–1070, Santiago, Chile, 2015

[10] A. Qayyum, S.M. Anwar, M. Awais, and M. Majid. Medical image retrieval using deep convolutional neural network. Neurocomputing, 266:8–20, 2017.

[11] X. Li, J. Yang, and J. Ma. Recent developments of content-based image retrieval (cbir). Neurocomputing, 452:675–689, 2021.

[12] Z. Wang, X. Liu, H. Li, J. Shi, and Y. Rao. A saliency detection based unsupervised commodity object retrieval scheme. IEEE Access, 6:49902–49912, 2018.

[13] M. Shin, S. Park, and T. Kim.Semi-supervised feature-level attribute manipulation for fashion image retrieval. arXiv preprint, 2019.

[14] Y.K. Jang and N.I. Cho. Self-supervised product quantization for deep unsupervised image retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 12085–12094, Montreal, BC, Canada, 2021.

[15] C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. In Proceedings of the IEEE International Conference on Computer Vision, pages 2051–2060, Venice, Italy, 2017.

[16] F. Shen, Y. Xu, L. Liu, Y. Yang, Z. Huang, and H.T. Shen.Unsupervised deep hashing with similarity-adaptive and discrete optimization.IEEE Transactions on Pattern Analysis and Machine Intelligence, 40:3034–3044, 2018.

[17] E. Yang, T. Liu, C. Deng, W. Liu, and D. Tao.Distillhash: Unsupervised deep hashing by distilling data pairs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2946–2955,Long Beach, CA, USA, 2019.

[18] W. Wang, H. Zhang, Z. Zhang, L. Liu, and L. Shao. Sparse graph based self-supervised hashing for scalable image retrieval. Information Sciences, 547:622–640, 2021.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-689

Vol.20 No.3 sep 2025

[19] Kim, S., Ma, I., & Son, J. (2024). How does stress experienced on instagram differ from threads? Comparing social media fatigue based on platform types. Computers in Human Behavior, 157, 108249.

[20] Dubey, S. R. (2021). A decade survey of content based image retrieval using deep learning. IEEE Transactions on Circuits and Systems for Video Technology, 32(5), 2687-2704.

[21] Anwaar, M. U., Labintcev, E., & Kleinsteuber, M. (2021). Compositional learning of image-text query for image retrieval. In Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision (pp. 1140-1149).

[22] Recent Advances In Content-based Image Retrieval: Techniques And Applications» JoCTA. (2024, December 23).

[23] Li, X., Yang, J., & Ma, J. (2021). Recent developments of content-based image retrieval (CBIR). Neurocomputing, 452, 675-689.

[24] Chen, Y., Mancini, M., Zhu, X., & Akata, Z. (2022). Semi-supervised and unsupervised deep visual learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 46(3), 1327-1347.

[25] Fini, E., Astolfi, P., Alahari, K., Alameda-Pineda, X., Mairal, J., Nabi, M., & Ricci, E. (2023). Semi-supervised learning made simple with self-supervised clustering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3187-3197).

[26] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised learning: Generative or contrastive. IEEE transactions on knowledge and data engineering, 35(1), 857-876.

[27] Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., & Tao, D. (2024). A survey on self-supervised learning: Algorithms, applications, and future trends. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[28] Lv, Y., Wang, C., Yuan, W., Qian, X., Yang, W., & Zhao, W. (2022). Transformer-Based Distillation Hash Learning for Image Retrieval. Electronics, 11(18), 2810.

[29] Lv, Y., Wang, C., Yuan, W., Qian, X., Yang, W., & Zhao, W. (2022). Transformer-Based Distillation Hash Learning for Image Retrieval. Electronics 2022, 11, 2810. Applications of Computational Intelligence, 135.

[30] Yang, H. F., Tu, C. H., & Chen, C. S. (2021). Learning binary hash codes based on adaptable label representations. IEEE Transactions on Neural Networks and Learning Systems, 33(11), 6961-6975.

[31] Shabbir, A., Ali, N., Ahmed, J., Zafar, B., Rasheed, A., Sajid, M., ... & Dar, S. H. (2021). Satellite and scene image classification based on transfer learning and fine tuning of ResNet50. *Mathematical Problems in Engineering*, *2021*(1), 5843816.

[32] Li, J., & Wang, J. Z. (2006, October). Real-time computerized annotation of pictures. In *Proceedings of the 14th ACM international conference on Multimedia* (pp. 911-920).

[33] Saadi, Z. M., Sadiq, A. T., Akif, O. Z., & Eid, M. M. (2024). Enhancing Image Classification Using a Convolutional Neural Network Model. *Journal of Soft Computing and Computer Applications*, *1*(2), 2.

[34] Saadi, Z. M., Sadiq, A. T., Akif, O. Z., & El-kenawy, E. S. M. Enhancing Convolutional Neural Network for Image Retrieval.

[35] Medus, L. D., Saban, M., Francés-Víllora, J. V., Bataller-Mompeán, M., & Rosado-Muñoz, A. (2021). Hyperspectral image classification using CNN: Application to industrial food packaging. *Food Control*, *125*, 107962.

[36] Alghamdi, F. A. (2024). An effective hybrid framework based on combination of color and texture features for content-based image retrieval. *Arabian Journal for Science and Engineering*, *49*(3), 3575-3591.

[37] Narayanaswamy, R. A., & Gowda, V. K. (2024, October). Integrated VGG19 and Gated Recurrent Unit with Cosine Similarity Metric for Content based Image Retrieval. In *2024 First International Conference on Software, Systems and Information Technology (SSITCON)* (pp. 1-5). IEEE.

[38] Hadid, M. H., Al-Qaysi, Z. T., Hussein, Q. M., Aljanabi, R. A., Abdulqader, I. R., Suzani, M. S., & Shir, W. L. (2024). Semantic Image Retrieval Analysis Based on Deep Learning and Singular Value Decomposition. *Applied Data Science and Analysis*, *2024*, 17-31.

[39] Zeain, A., & Ibrahim, A. A. (2024, December). Enhancing Content-Based Image Retrieval with a Stacked Ensemble of Deep Learning Models. In *2024 8th International Symposium on Innovative Approaches in Smart*

Technologies (ISAS) (pp. 1-11). IEEE.

[40] Dowerah, R., & Patel, S. (2024). Comparative analysis of color histogram and LBP in CBIR systems. *Multimedia Tools and Applications*, *83*(5), 12467-12486.

[41] Abdivokhidov, I., & Ayoobkhan, M. U. A. (2024, February). Machine Learning Based Image Classification with COREL 1K Dataset. In *Proceedings of the International Conference on Sustainability: Developments and Innovations* (pp. 39-47). Singapore: Springer Nature Singapore.

[42] Kumar, R., & Murthy, N. (2025). Relevance-Aware Content-based Image Retrieval using Deep Hybrid Feature Extraction. *Engineering, Technology & Applied Science Research*, *15*(3), 22976-22982.

[43] İncetas, M. O., & Arslan, R. U. (2025). Spiking neural network-based edge detection model for content-based image retrieval. *Signal, Image and Video Processing*, *19*(1), 169.

[44] Naseer, A., & Jalal, A. (2024, February). Multimodal objects categorization by fusing GMM and multi-layer perceptron. In *2024 5th International Conference on Advancements in Computational Sciences (ICACS)* (pp. 1-7). IEEE.

[45] Dhingra, S., & Bansal, P. (2022). Designing of a rigorous image retrieval system with amalgamation of artificial intelligent techniques and relevance feedback. *Journal of Intelligent & Fuzzy Systems*, *42*(2), 1115-1126.