

Interpretable Region-Attribute Graph Matching Framework for Structure-Aware and Explainable Content-Based Image Retrieval

Prabhuraj Metipatil¹, Mrutyunjaya M. S.², Sanjeev Prakashrao Kaulgud³, Sunil Manoli¹

¹School of Computer Science and Engineering, REVA University, Bengaluru, Karnataka, India

²School of Engineering Technology, Svyasa Deemed to be University, Satva Global City, Campus, Bangalore, Karnataka, India

³Department of Artificial Intelligence and Machine Learning, New Horizon College of Engineering, Bengaluru, Karnataka, India

E-mails: pmetipatil@gmail.com mrutyunjayams@gmail.com sanjeev.kaulgud@gmail.com
sunil.manoli@reva.edu.in

Abstract: In recent years, Content-Based Image Retrieval (CBIR) has advanced significantly with Deep Learning (DL), yet most of the existing models rely on global descriptors and attention-based local features, which struggle to capture explicit structural relationships within images. These models often operate as black boxes and provide limited interpretability; they exhibit reduced robustness under viewpoint and geometric variations. To address these limitations, an Interpretable Region-Attribute Graph Network (IR-AGN) is proposed in this research. Specifically, each image is decomposed into semantically meaningful regions, where each node encodes visual and semantic descriptors while edges represent geometric, appearance, and contextual relations. Consequently, a joint node-edge graph matching strategy computes similarity by aligning visual descriptors and relational structures between query and retrieval images. The results of the proposed IR-AGN model exhibit a higher mean Average Precision (mAP) of 95.19% on the Revisited Paris (RParis) dataset, when compared with the existing Multi-Layer Orientation Histogram (MLOH) model.

Keywords: Content-based image retrieval, Deep Learning, Interpretable Region-Attribute Graph Network (IR-AGN), Mean Average Precision (mAP).

1. Introduction

In the digital era, Content-Based Image Retrieval (CBIR) has become a crucial field in computer vision for retrieving visually similar images from large datasets [1].

This process is carried out using the intrinsic content of images rather than metadata or textual annotations. Moreover, similarity search over high-dimensional feature vectors has become a fundamental operation in modern image retrieval systems, allowing effective and scalable retrieval of visually similar images from large-scale datasets [2]. Similarly, a high-quality image is one of the important components in this procedure for supporting the interpretation and understanding of visual data [3]. In this process, the discriminative features are extracted by measuring the inter-image similarity and ranking results based on visual relevance. However, the prior CBIR systems heavily depended on handcrafted descriptors such as Scale-Invariant Feature Transform (SIFT), Sped Up Robust Features (SURF), and color histograms, which struggle with large-scale or complex datasets. Hence, the Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have significantly improved the feature representation capability with the rapid growth of Deep Learning (DL) [4]. This process enables the retrieval systems to capture higher-level semantics, structural information, and contextual dependencies within the image content [5, 6]. However, the DL-based retrieval systems continue to suffer from interpretability and generalization challenges. Mostly, the CNN-based methods aggregate features into a single global descriptor, although compact loses fine-grained spatial relationships between regions [7, 8]. In addition, the attention-driven models heavily depend on opaque learned weights, which lack human interpretability [9]. Moreover, these attention mechanisms increase computational complexity and reduce transparency in how similar decisions are made [10, 11]. Furthermore, the Graph Convolutional Networks (GCNs) and transformer-based retrieval frameworks remain limited by their black-box nature, where the users do not explicitly identify the regions that influence retrieval performance [12, 13]. Finally, these challenges highlight the requirement for interpretable, region-aware, and computationally efficient retrieval strategies. Therefore, the aim behind the proposed Interpretable Region-Attribute Graph Network (IR-AGN) framework is to develop an explainable, attention-free CBIR system, which models the relationships between local regions for maintaining high retrieval accuracy. In this context, the approach exploits graph structures to represent the images as a set of semantically meaningful regions and explicitly encodes geometric, appearance, and conceptual relations between them [14, 15]. As a result, it captures both structural layout and semantic context without depending on learned attention maps.

This framework provides interpretability, which allows users to trace how specific regions and relationships contribute to similarity scores for higher performance and human-understandable reasoning in visual retrieval [16, 17]. The current state-of-the-art methods in CBIR focus on global and local feature

aggregation, regional attention mechanisms, and graph-based reasoning. Here, the models such as R-MAC, DELF, and NetVLAD highlight regional descriptors and pooling for efficient retrieval. In addition, the transformer-based methods integrate attention mechanisms to model inter-region dependencies for improving robustness under viewpoint or illumination changes [18]. However, these models still treat interpretability as a post-hoc process rather than an intrinsic design feature. Moreover, the graph-based approaches using GCNs or attention layers have explored relational reasoning, although the contribution of each relation is obscure. Furthermore, the explicit region-attribute graphs with measurable geometric and semantic edges for allowing retrieval decisions to be accurate and explainable. The proposed Interpretable Region-Attribute Graph Network IR-AGN model employs standard operations such as region patch extraction, cosine similarity, and semantic embeddings to capture effective features of the image. The novelty of the proposed model does not lie in individual mathematical components, as they are integrated into an explicit and interpretable Region Attributed Graph (RAG), where each node and edge carries human-readable attributes and also directly participates in analytical similarity computation. Unlike the conventional graph-based CBIR models, which heavily rely on learned graph embeddings and attention modules or GCN-propagated weights, RAG is introduced by the proposed model, in which both the nodes and edges are defined with explicit measurable descriptors rather than latent parameters. The proposed framework performs joint node-edge graph matching with analytical similarity decomposition while enabling each component of the retrieval score. Moreover, the retrieval score helps to trace visual, semantic, and geometric descriptors to specific regions and inter-region relations. The novelty of the proposed mechanism makes interpretability an intrinsic property of the retrieval mechanism instead of using post-hoc explanations, which provides the level of transparency that is overlooked by previous graph-based CBIR approaches.

The contributions of the proposed model are listed below.

- The IR-AGN model introduced a RAG model that explicitly integrates visual, semantic, and geometric information into a unified node and edge attributes. This graph-based representation preserves regional context and spatial relationships, which enables more discriminative similarity computation.
- A joint node-edge graph matching mechanism is employed to measure image similarity through interpretable relational reasoning. When compared to existing black-box attention, the matching objective decomposes similarity into visual, semantic, and structural components by enabling transparent retrieval decisions.

- The proposed IR-AGN incorporates an interpretability module that quantifies node and edge contributions, while explaining how specific regions and relationships influence retrieval results.

The organization of the research is structured as follows: Section 2 describes the literature review, Section 3 demonstrates the methodology, Section 4 specifies the experimental results, Section 5 demonstrates the research implications, and Section 6 specifies the conclusion.

2. Literature review

Shi et al. [19] presented a Canonical Correlation Analysis-based Image Retrieval Framework (CCA-IRF) to improve image similarity detection by increasing the correlation between paired image features. Initially, the suggested model extracted feature maps from both images by using a pre-trained Convolutional Neural Network (CNN) to capture deep spatial-semantic information. Subsequently, the CCA identified the projection pairs that best represent shared information between the two feature sets. Additionally, an element-wise selection provided a compact set of transformed features by using a Chernoff-information-based criterion and a binary hypothesis test to measure similarity between the images. The suggested model provided high retrieval accuracy and robustness to feature variability. However, it had computational overhead from correlation computation in a large-scale dataset.

Li, Liu and Lu [20] demonstrated a Multi-Layer Orientation Histogram (MLOH) to improve feature representation by integrating semantic and orientation cues across multiple CNN layers. Initially, the deep features were extracted from various convolutional layers to capture different semantic levels. Additionally, an iterative multilayer integration technique was employed to combine multi-scale features for improving spatial discrimination. Subsequently, Gabor filtering was used to detect and amplify the dominant orientation cues, allowing the model to encode both structural and directional information effectively. Further, the integrated representation improved global feature discriminability and retrieval precision. The suggested model had the ability to jointly capture semantic and orientation features, although the integration of multi-layer and orientation extraction led to increased computational cost.

An et al. [21] developed a Pixel Retrieval Benchmarks (PRB) to extend conventional image retrieval to the pixel level by enabling fine-grained recognition of query-related regions. Initially, the datasets were improved by professional annotators who labelled the images through multiple verification rounds to ensure

high-quality segmentation. In addition, these pixel-level annotations were structured into a benchmark dataset, which was designed to evaluate the retrieval accuracy at the pixel granularity. Subsequently, the state-of-the-art image search, matching, and segmentation models were tested to assess their performance on the new benchmarks. The utilization of a benchmark dataset provided a foundation for developing more precise and content-aware retrieval systems. However, high annotation complexity and computational cost have increased for large-scale pixel-level evaluation.

Park and Hwang [22] introduced a Hierarchical K-Means Clustering (HKMC) to accelerate CBIR by maintaining higher accuracy. Initially, the suggested model was employed to organize the global image descriptors from the dataset into structured clusters to reduce the search space. In addition, the query image was encoded as a compact descriptor and compared only with the descriptor sets, which were located in the leaf nodes of the hierarchy. Consequently, three optimized tree search algorithms were employed to balance retrieval speed and precision. The suggested model has the ability to achieve faster retrieval with higher accuracy for a large-scale dataset. However, there was an increased potential loss of retrieval precision during deep hierarchical clustering.

Zhang et al. [23] established a Deep Semantic Correlation Descriptor (DSCD) to improve the instance image retrieval accuracy by improving object localization and feature compactness. Initially, a channel semantic correlation learning mechanism was introduced to localize the targeted objects by learning inter-channel semantic relationships from an auxiliary dataset for reducing background interference. Additionally, a hierarchical attention mechanism had combined the local and global features through object and focus attention modules by generating refined feature maps with improved target relevance. Subsequently, an adaptive Principal Component Analysis (PCA)-whitening method was employed to dynamically reduce dimensionality and improve robustness of compact feature vectors. The suggested model has superior retrieval accuracy and generalization across multiple datasets, although the integration of multi-level attention led to increased computational complexity.

Liu et al. [24] demonstrated Sublimated Deep Feature Histogram (SDFH) to improve image quality by connecting all the semantic gaps among low and high-level features. Firstly, orientation-selective features were presented to match the visual cortex orientation selection mechanism and captured global shapes of the object, also reducing CNNs' texture error. Subsequently, perceptual colored features were combined to reintroduce color information rejected by classical deep features by enhancing discriminative representation. Finally, a gain whitening learning

strategy was used to refine feature representation for efficient retrievals. The model has advantages such as the ability to achieve human-aligned meaningful retrieval accuracy. However, the model faced issues as it had high computation during orientation and also color feature extraction.

Z h u et al. [25] presented a Coarse-to-Fine Compact Discriminative (CFCD) representation to achieve accurate one-step image retrieval by using image-level labels. Initially, SoftMax-based loss was designed to adjust scale and margin dynamically with each mini-batch, which improved supervision and ensured intra-class compactness. Consequently, a negative sampling strategy was used to strengthen the intra-class compactness and inter-class discrimination. Moreover, this model was capable of fusing local and global representations for accurate retrieval performance across viewpoint and occlusion differentiations. However, the model led to computational intensity during adaptive training and selection as it increased the complexity of the model.

2.1. Research gap

Although the DL has significantly improved CBIR, the existing models still rely on global descriptors and attention-based local features, which lack explicit structural reasoning and provide limited interpretability. Current graph-based approaches often depend on learned attention weights, which make their relational features challenging to justify. Moreover, the existing convolutional-based retrieval frameworks model semantic, geometric, and appearance relations together, leading to mismatched structural variations. Therefore, there remains a clear research gap that provides an interpretable and region-aware retrieval model to capture explicit region-to-region relationships while providing transparent similarity explanations.

2.2. Problem statement and objective

Despite the advancements in CBIR systems, most of the existing models depend on global descriptors and attention-based local features, which fail to capture explicit structural relationships between image regions, making their retrieval process difficult to interpret. Graph-based models also rely on learned attention weights, which complicate the true influence of regional and relational attributes. Moreover, the existing models struggle with viewpoint variations, geometric changes, and fine-grained structural dependencies while providing post-hoc explanations that lack semantic clarity. Therefore, the main objective of this research is to design an interpretable and attention-free CBIR framework that represents images through explicit graphs integrating visual, semantic, and geometric attributes.

3. Methodology

In this section, the proposed IR-AGN integrates the deep visual features with interpretable structural reasoning through the RAG model. Each image is decomposed into semantic and visual regions to represent it as a graph with nodes and edges. The nodes encode fused descriptors, and the edges capture geometric, appearance, and semantic relations. Moreover, a joint node-edge graph matching is introduced to compute similarity between the query image and structure-aware retrieval images. This analysis provides interpretability while ranking the retrieved image using the final graph similarity to return the most relevant images. The architecture of the proposed IR-AGN is presented in Fig. 1.

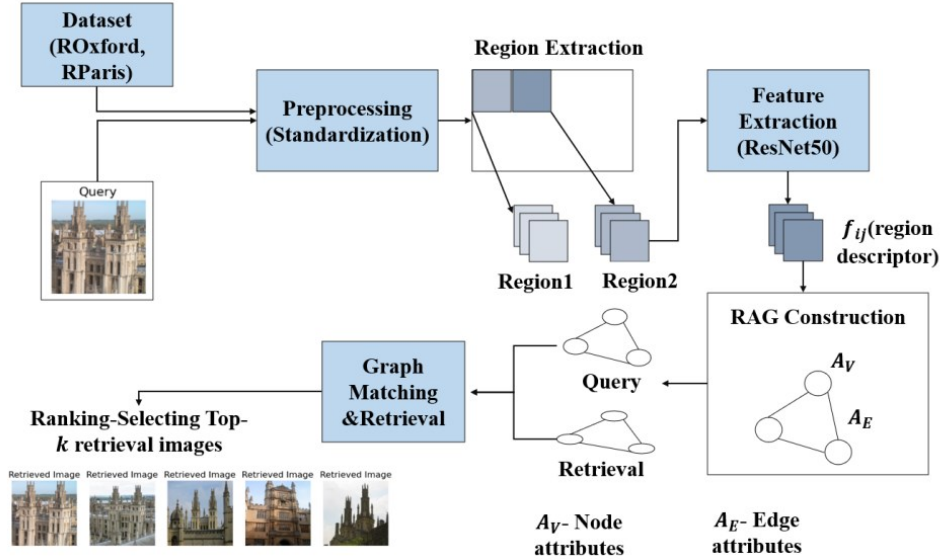


Fig. 1. The architecture of the proposed IR-AGN model for content-based image retrieval for enhanced interpretability

3.1. Data collection

The Revisited Oxford (ROxford) [26] dataset comprises 5062 images that were collected from Flickr, representing 11 landmark buildings in Oxford, United Kingdom (UK), and All Souls College (ASC). Each landmark forms a distinct class with multiple viewpoints, lighting, and scale variations. The dataset provides query images and ground truth images for evaluating instance-level and structure-aware image retrieval methods.

The Revisited Paris (RParis) dataset contains 6412 high-resolution images of 11 famous Paris landmarks such as the Eiffel Tower and Notre Dame. Each landmark corresponds to a unique class, which was captured under varying

viewpoints, weather conditions, and bounding boxes to enable accurate benchmarking of feature-based and geometry-driven retrieval models in complex urban scenes.

3.2. Data preprocessing

The input images undergo standardized preprocessing to ensure consistency and robustness before extracting features. In this context, normalization, contrast enhancement, and Gaussian filtering are considered for preprocessing. These preprocessing models simultaneously standardize the lighting conditions and reduce high-frequency noise while preserving fine geometric features over classical histogram-only techniques. Each image in the dataset $I(x, y)$ is first resized to a uniform resolution of 256×256 pixels and normalized to eliminate illumination bias. Normalization is performed channel-wise in a resized image, which is given in the equation

$$(1) \quad I'_c(x, y) = \frac{I_c(x, y) - \mu_c}{\sigma_c},$$

where μ_c and σ_c denote the mean and standard deviation of the Red, Green, and Blue (RGB) channels c in an image. Further, the image is processed with histogram equalization [27] to balance the brightness and enhance local contrast to improve the detection of architectural textures and edges. Subsequently, a Gaussian filter

$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$ is applied for smoothing the noise while preserving salient structural information, which is required for graph node generation, where $G(x, y, \sigma)$ denotes the Gaussian kernel x and y denote spatial coordinates, σ specifies the standard deviation for controlling the extent of smoothness, and π specifies a constant.

3.3. Region extraction

Each preprocessed image is divided into 64×64 overlapping patches with a stride $t = 32$ for regional-level representation, which is given in the equation

$$(2) \quad P_{ij} = I'(x + i\Delta, y + j\Delta), \quad i, j \in [1, N],$$

where P_{ij} denotes the overlapping patches, I' specifies the preprocessed image, and Δ signifies the scaling factor with 50% spatial overlapping between adjacent patches with 49 patches for a 256×256 image using a 7×7 grid size. The regional-level refinement is crucial for constructing reliable Region-Attribute Graphs (RAGs) to ensure subsequent node descriptors contain intrinsic structural and semantic content, rather than illumination and noise artifacts. In this region extraction, each region represents a localized visual entity such as a window, architecture, or tower segment. The Convolutional Neural Network (CNN) model

with Residual Network (ResNet50) is employed to extract feature maps from each region as described in the equation

$$(3) \quad F_{ij} = \phi(P_{ij}; \theta),$$

where $\phi(\cdot)$ represents the convolutional transformation and θ denotes the learned parameters. Each region descriptor is obtained with a spatial average across activation maps that is mathematically expressed in the equation

$$(4) \quad F_{ij} = \frac{1}{|F_{ij}|} \sum_{x,y} F_{ij}(x,y),$$

where f_{ij} denotes the region descriptor that contains high-level semantics in the regions of an image.

3.4. Graph construction

The region descriptor contains visual and semantic descriptors that are defined as (f_i, s_i) . This is further useful for constructing the RAG as $G = (V, E, A_V, A_E)$, where V denotes the set of nodes and E specifies the set of edges. Nodes denote the image regions, and edges connect to the spatially related regions. Furthermore, A_V refers to the node attributes and A_E defines the edge attributes. In this context, the high-level semantics s_i is derived from a pretrained visual-text encoder. In this regard, the proposed model is described as the attention-free mechanism because all relational reasoning, joint node-edge matching, and similarity computation are conducted without an attention mechanism and learned from relational weights. Although the semantic node descriptors are extracted using a pretrained visual-text encoder, these descriptors remain fixed and do not influence region importance via attention. Also, they serve solely as static semantic attributes while interpretability is achieved intrinsically from the explicit graph structure and analytically similar decomposition. The concatenation of these components of the node attributes is

$$(5) \quad v_i = [f_i; s_i] \in \mathbb{R}^{d+q},$$

where $f_i \in \mathbb{R}^d$ denote the feature vector from Resnet50 and $s_i \in \mathbb{R}^q$ specifies the semantic descriptor of the region i . The edge set $E = \{(i,j)\}$ is formed by connecting each node to its spatially nearest neighbors within the image grid while enabling geometric consistency. For each edge in the RAG, an interpretable relational attribute vector is computed and defined in the equation

$$(6) \quad a_{ij} = [d_{ij}, \tau_{ij}, \sigma_{ij}],$$

where $d_{ij} = \frac{\|c_i - c_j\|_2}{D_{\max}}$ denotes the normalized Euclidean distance between centroids c_i and c_j of the regions, where $d_{ij} \in [0,1]$, θ_{ij} specifies relative distance. Furthermore, $\tau_{ij} = \frac{\tau_{ij} - \min \tau}{\max \tau - \min \tau}$, where $\tau_{ij} \in [-1,1]$ to $[0,1]$ signifies the texture correlation between regional features and $\sigma_{ij} = \cos(s_i, s_j)$ signifies the semantic

co-occurrence between conceptual embeddings, which is transformed to $[0, 1]$ for ensuring comparability. This relational attribute signifies the edge attributes enabling a similarity function for maintaining structural consistency between the query image and retrieval images. Unlike conventional graph-based methods or attention-based CBIR models, the proposed IR-AGN model introduces a semantic-aware and explainable edge definition, which implicitly describes the visual and semantic relations between image regions rather than learning them. This mechanism enabled the proposed IR-AGN model for subsequent matching and retrieval to operate without hidden attention weights while preserving interpretability and regional-level similarity computation.

3.5. Graph matching and retrieval

Furthermore, the retrieval of images is performed after constructing the RAG by computing structural and semantic similarity between a query graph $G_q = (V_q, E_q, A_V^q, A_E^q)$ and a retrieval image graph $G_g = (V_g, E_g, A_V^g, A_E^g)$. Unlike conventional GCN or transformer-based retrieval models, which learn latent relation weights implicitly, the proposed IR-AGN model defines measurable and interpretable edge similarities based on geometric and semantic attributes. This enables the model to justify retrieval outcomes via observable relational evidence. For instance, one image is retrieved, the proposed IR-AGN model justifies the similarity by using aligned spatial layouts or consistent semantic co-occurrence instead of randomly assigning higher importance with activation functions. In addition, the main objective of image retrieval is to identify the most similar graphs in the dataset that best match the query structure. Each node in the graph corresponds to a region descriptor, which is defined as $v_i^q = [f_i^q; s_i^q]$. Similarly, each node in the retrieval image graph is specified with $v_j^g = [f_j^g; s_j^g]$. The node similarity between regions i and j is computed as a weighted sum of visual and semantic similarities, which is given in the equation

$$(7) \quad S_v(i, j) = \alpha \cdot \cos(f_i^q, f_j^g) + (1 - \alpha) \cdot \cos(s_i^q, s_j^g),$$

where $\cos(\cdot)$ denotes the cosine similarity, f_i^q and f_j^g denotes the visual feature vector of the query and retrieval image regions. In addition, s_i^q and s_j^g specifies the corresponding semantic descriptors and $\alpha \in [0, 1]$ signifies the weighting coefficient, which balances the importance of visual and semantic components. This mechanism ensures that both appearance and concept similarities contribute to transparent node matching. The structural comparison is achieved by computing an edge similarity function, as it quantifies the relationship between pairs of connected

regions in both graphs. For two edges $(i, k) \in E_q$ and $(j, l) \in E_g$, the similarity is computed as

$$(8) S_e((i, k), (j, l)) = \beta_1 e^{-|d_{ik}^q - d_{jl}^g|} + \beta_2 \cos(\theta_{ik}^q, \theta_{jl}^g) + \beta_3 \cos(\tau_{ik}^q, \tau_{jl}^g) + \beta_4 \cos(\sigma_{ik}^q, \sigma_{jl}^g),$$

where d_{ik} denotes the normalized Euclidean distance between nodes i and k , θ_{ik} specifies the relative orientation of the edges; τ_{ik} signifies the visual texture correlation, and σ_{ik} specifies the semantic co-occurrence of the connected regions. The coefficients $\beta_1, \beta_2, \beta_3, \beta_4 \geq 0$ are helpful to control the influence of geometric, orientation, appearance, and semantic relations, respectively, and k denotes the number of nearest neighbors in graph edges. Consequently, the optimal region between two graphs is determined with an assignment matrix $X \in \{0, 1\}^{|V_q| \times |V_g|}$ where $X_{ij} = 1$ indicates that node i in the query graph is matched to node j in the dataset graph. Moreover, to ensure one-to-one each query node matches at most specified as $\sum_{j=1}^{n_g} M_{ij} \leq 1$, and the M_{ij} is defined as

$$(9) M_{ij} = \begin{cases} 1 & \text{if node } i \text{ in } G \text{ is matched to node } j \text{ in } G, \\ 0 & \text{otherwise,} \end{cases}$$

similarly, each gallery node matches at most one query node, and it is defined as $\sum_{i=1}^{n_q} M_{ij} \leq 1$, by using these assignments, the model performs optimal partial matching. In addition, the binary constraint for the graph matching is defined as $M_{ij} \in [0, 1]$ and the partial permutation constraint is applied when the graphs have equal size and is given in the equation

$$(10) \sum_{j=1}^{n_g} M_{ij} = 1, \quad \sum_{i=1}^{n_g} M_{ij} = 1,$$

subsequently, the overall graph similarity is optimized using the equation

$$(11) S = \max_X \left(\sum_{i,j} S_v(i, j) X_{ij} + \sum_{(i,k) \in E_q} \sum_{(j,l) \in E_g} S_e((i, k), (j, l)) X_{ij} X_{kl} \right).$$

This similarity is employed based on one-to-one matching constraints, including $\sum_j X_{ij} \leq 1$ and $\sum_i X_{ij} \leq 1$. This objective function ensures that both node-level and edge-level similarities are jointly maximized while considering a globally consistent alignment between the two graphs. After the optimal matching, the final similarity score between the query and the retrieval graph is computed using the equation

$$(12) S(G_q, G_g) = \lambda_1 \sum_{i,j} S_v(i, j) X_{ij} n^{\lambda_2} \sum_{(i,k),(j,l)} S_e((i, k), (j, l)) X_{ij} X_{kl},$$

where λ_1 and λ_2 denotes the coefficients, which balance the contribution of node and edge similarities. The computed similarity scores for all gallery graphs are ranked in descending order, and the top-K images with the highest $S(G_q, G_g)$ values are returned as retrieval results.

3.5.1. Retrieval ranking for selecting top-K images

After computing the global similarity score $S(G_q, G_g)$ between the query graph G_q and each gallery graph G_g all gallery images are ranked in descending order in terms of similarity. The ranking function is expressed as given in the equation

$$(13) \quad R(G_q) = \text{argsort}_{G_g \in \mathcal{D}} \left(S(G_q, G_g) \right),$$

where $\mathcal{D}$ denotes gallery database, and $R(G_q)$ returns an ordered list of retrieved images. The top-K ranked images are selected as the final retrieval outcome. This ranking ensures that each retrieved image corresponds to a structurally and semantically aligned representation with the query, based on both regional-level and relational consistencies. To ensure transparency of retrieval outcomes, interpretability, and explainability are incorporated as intrinsic parts of the model. Rather than depending on post-hoc visualizations, interpretability in the proposed IR- AGN network emerges directly from graph structure and its attribute-driven similarity computation. The total graph similarity between query and gallery image is decomposed into interpretable components as given in the equation

$$(14) \quad S(G_q, G_g) = S_v^{\text{vis}} + S_v^{\text{sem}} + S_e^{\text{geo}} + S_e^{\text{app}} + S_e^{\text{sem}},$$

where $S_v^{\text{vis}} = \lambda_1 \sum_{i,j} \alpha \cos(f_i^q, f_j^g) X_{ij}$ denotes the visual feature similarity between matched regions, $S_v^{\text{sem}} = \lambda_1 \sum_{i,j} (1 - \alpha) \cos(s_i^q, s_j^g) X_{ij}$ represents the semanticsimilarity, $S_e^{\text{app}} = \lambda_2 \sum_{(i,k),(j,l)} \beta_3 \cos(\tau_{ik}^q, \tau_{jl}^g) X_{ij} X_{kl}$ captures appearance consistency, and $S_e^{\text{sem}} = \lambda_2 \sum_{(i,k),(j,l)} \beta_4 \cos(\sigma_{ik}^q, \sigma_{jl}^g) X_{ij} X_{kl}$ reflects the semantic co-occurrence similarity. Each component thus represents the distinct interpretable aspect of the retrieval process. To visually interpret the retrieval result, node-level contribution is analyzed by using the similarity difference attained by eliminating a specific node. For node v_i^q , the contribution is defined as

$$(15) \quad C(v_i^q) = S(G_q, G_g) - S(G_q \setminus v_i^q, G_g),$$

where $S(G_q \setminus v_i^q, G_g)$ is the similarity recomputed after excluding the node i . A higher $C(v_i^q)$ value indicates a greater influence of that region in determining the final match. Similarly, edge-level contribution is computed as given in the equation

$$(16) \quad C(e_{ik}^q) = S(G_q, G_g) - S(G_q \setminus e_{ik}^q, G_g),$$

where this highlights the importance of specific structural or semantic relationships. Nodes and edges with the highest contribution scores are visualized on an image to provide an intuitive explanation of “which regions” and “which relations” influence the retrieval. By using this leave-one-out marginal contribution principle, the importance of a node or edge is measured by the change in global similarity score

when that region component is removed from the graph. This subtraction-based method specifies how much each region or relation contributes to the final similarity. Quantitatively, contribution ratios are normalized to obtain a similarity decomposition as given in the equation

$$(17) \quad \rho_c = \frac{s_c}{s(G_q, G_g)} \text{ for } c \in \{ \text{vis}, \text{sem}, \text{geo}, \text{app}, \text{sem} \},$$

where ρ_c denotes the proportional contribution of each interpretive component of the final similarity. These values act as an explainable signature of each retrieval result. For instance, a pair of images may achieve 40% similarity from geometry, 35% from visual appearance, and 25% from semantic alignment, which clearly depicts why the model considered them a match. The novelty of this interpretable mechanism lies in its inherent, quantitative explanation framework. Unlike post-hoc attention heatmaps or saliency visualizations that approximate dependence, the IR-AGN produces a causal and measurable interpretability by explicitly linking each similarity component to a defined visual, geometric, or semantic factor. Furthermore, since all node and edge weights are analytically derived rather than learned, each retrieval decision is fully traceable. This attention-free interpretability allows the model to explain its reasoning in human-understandable terms, such as similar spatial layouts or consistent semantic patterns, by removing the gap between black-box deep retrieval and structured visual reasoning. The pseudocode for the proposed IR-AGN model is given in Section 3.5.2.

3.5.2. Algorithm

Algorithm 1. The Pseudocode of the Proposed IR-AGN Model for CBIR Systems

Input: Query graph $G_Q(V_Q, E_Q)$, Retrieval graph $G_I(V_I, E_I)$, Node descriptors f_{ij} , edge attributes $(d, \theta, \tau, \sigma)$, and weights $\alpha, \beta_1, \beta_2, \beta_3, \beta_4$, balance factor λ .

Output: Set of optimal matched region pairs M , final similarity score is S

For each node $g \in V_I$ **do**

 Compute visual similarity s_i

 Compute semantic similarity s_s

 Compute node similarity s_{node}

 Insert pair (q, g) into C with score s_{node}

End for

End for

foreach edge $(q_i, q_j) \in E_Q$ **do**

foreach edge $(g_i, g_j) \in E_I$ **do**

 Compute geometric distance as a relational attribute a_{ij}

 Compute texture correlation τ_{ij}

 Compute semantic co-occurrence σ_{ij}

```

                                Compute edge similarity using the Equation (8)
        End for
    End for
    while True do
        Select the highest-scoring pair  $(q, g)$  from  $C$ 
        Insert  $(q, g)$  into  $M$ 
        Mark  $(q, g)$  as matched
        Remove all pairs containing  $q$  from  $C$ 
        Remove all pairs containing  $g$  from  $C$ 
        For each matched edge  $(q_i, q_j) \in E_Q$  do
            If the corresponding edge  $(g_i, g_j)$  satisfies matching criteria, then
                Include the edge-match contribution to  $M$ 
        End for
    End while
    if  $C$  is empty then
        break # when no match is found
        Compute total node similarity  $S_{\text{node}}$ 
        Compute total edge similarity  $S_{\text{edge}}$ 
    return  $M, S(G_Q, G_I)$ 

```

4. Experimental setup

The proposed IR-AGN model was executed on a hardware and software configuration comprising Python 3.10 with PyTorch 2.2 library, and for similarity indexing. All the experiments were conducted on a workstation, which is equipped with an NVIDIA RTX 4090 GPU, an Intel Core i9 processor with 64 GB RAM, and running Ubuntu 22.04. The proposed IR-AGN model is finetuned with the Adam optimizer for 10 epochs on a dataset using contrastive triplet loss in order to adapt retrieval-specific features. The proposed IR-AGN model was trained with a learning rate of 1×10^{-4} and batch size = 16. Furthermore, all the retrieval evaluations, including graph matching and interpretability analysis, were executed for additional efficiency. The evaluation metrics include mean Average Precision (mAP), Node Contribution Score (NCS), Edge Structural Consistency (ESC), and Average Casual Drop (ACD). The NCS evaluation metric evaluates the proportion of total similarity explained by the top-5 most influential nodes. In addition, the ESC metric calculates the degree of geometric alignment between corresponding structural edges. Moreover, the ACD metric specifies the average relative drop in total similarity after removing key nodes or edges. The mathematical equation for the evaluation metrics is given in the next equations:

$$(18) \quad \text{mAP} = \frac{1}{n} \sum_{i=1}^n \text{AP}_i,$$

$$(19) \quad AP = \frac{1}{R} \sum_{k=1}^N p(k) \cdot \text{rel}(k),$$

$$(20) \quad \text{precision@}k = \frac{\text{Number of relevant items in top } k}{k},$$

$$(21) \quad \text{recall@}k = \frac{\text{Number of relevant items in top } k}{\text{Total number of relevant items in database}},$$

where AP_i specifies the Average Precision for the class i , n denotes the number of classes, and $p(k)$ denotes the precision at rank k . In addition, $\text{rel}(k) = 1$ if the item at rank k is relevant otherwise 0 and R denote the total relevant items.

4.1. Performance analysis

The performance of the proposed IR-AGN model is evaluated to quantitatively validate the interpretability and explainability when compared with two widely used baseline models for image retrieval. The baseline models include the Regional Maximum Activation of Convolutions (R-MAC) and Transformer-based Deep Local Features (Trans-DELF), which are considered and re-implemented on an identical experimental setup, such as preprocessing and hyperparameters, to ensure fair comparison. The evaluation metrics, including mAP, NCS, ESC, and ACD, are considered and illustrated in Fig. 2.

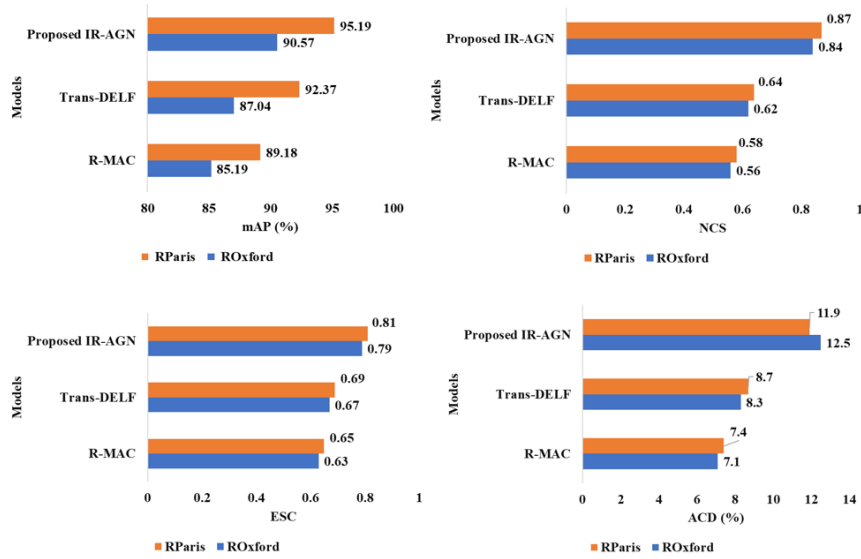


Fig. 2. Evaluating the effectiveness of interpretability and explainability of the proposed IR-AGN model over baseline models

From Fig. 2, the proposed IR-AGN model consistently demonstrates superior results across the baseline models. When compared to the R-MAC and Trans-DELF models, the proposed IR-AGN model achieved higher mAP, NCS, and ESC values on both RParis and ROxford datasets, which indicates that the retrieved results are explainable structurally and visually. Although the baseline models were

implemented on identical conditions, they struggled with limitations such as weak structural reasoning and no semantic awareness. The proposed model resolved these issues by including an explicit region relation graph to enhance robustness and interpretability. This validation significantly highlights the strong retrieval performance and provides transparent and effective explanations.

4.1.1. Generalization through Cross-Dataset validation

The generalization ability of the proposed IR-AGN model was evaluated using a cross-dataset on the Instance-Level Retrieval (INSTRE) [28] dataset. This validation is done by training the model on the ROxford and RParis dataset, then testing on unseen landmark images. The evaluation metrics, including mAP, precision, and recall, were utilized. Each model is trained and evaluated over five independent runs with different random seeds. The results were reported as a mean $\pm$ standard deviation to evaluate the variability and robustness, which is presented in Table 1.

Table 1. Cross-dataset evaluation of the proposed IR-AGN model over baseline models

Dataset	Model	mAP (%)	Precision (%)	Recall (%)
ROxford $\rightarrow$ INSTRE	R-MAC	60.32 ± 0.015	66.25 ± 0.013	62.82 ± 0.017
	Trans-DELF	63.96 ± 0.011	68.26 ± 0.012	66.18 ± 0.014
	Proposed IR-AGN	74.21 ± 0.008	80.46 ± 0.010	77.31 ± 0.012
RParis $\rightarrow$ INSTRE	R-MAC	61.75 ± 0.014	67.14 ± 0.015	64.23 ± 0.016
	Trans-DELF	64.61 ± 0.010	69.39 ± 0.011	66.83 ± 0.013
	Proposed IR-AGN	75.68 ± 0.009	81.11 ± 0.008	78.58 ± 0.010

According to Table 1, the proposed IR-AGN model achieved higher performance and consistently outperformed baseline retrieval models on the INSTRE dataset. Across all the random seeds, the proposed IR-AGN model achieved the highest mAP from 0.742 up to 0.756 with low standard deviation. The baseline models struggled to process unseen data due to a lack of explicit region relations and weak generalization. The proposed model resolved these challenges using semantic-geometric graphs to enable strong structural reasoning and improved generalizability. This validation indicates stable and transferable performance of the proposed IR-AGN model and confirms effective explicit region-relation encoding across unseen data.

4.1.2. Ablation study and statistical analysis

The ablation study mainly focuses on the individual contribution of components in the proposed IR-AGN model. The training and testing data are split into 80% and 20% where the model is evaluated on the testing set. Each variant of the model was evaluated over five independent runs with different random seeds using paired

t -test. The evaluation is reported as a mean $\pm$ standard deviation of mAP. Moreover, 95% Confidence Interval (CI) is computed with t -distribution and the significance is provided based on non-overlapping variants of CI with the entire proposed IR-AGN model CI. This CI is computed over N query images where the value of N depends on the size of the dataset, with a standard normal approximation of the sampling distribution of retrieval metrics. Results of the ablation study with statistical significance are given in Table 2.

Table 2. Ablation study and statistical analysis of the proposed IR-AGN model with different model variants

Model variant	mAP (mean $\pm$ standard deviation)	95% CI	p-value
Proposed IR-AGN model	95.19 $\pm$ 0.009	[94.88-95.91]	0.003
Without semantic node descriptors	93.87 $\pm$ 0.011	[92.85-94.89]	0.087
Without edge attributes	90.18 $\pm$ 0.014	[89.81-90.85]	0.004
Without geometric edges	88.79 $\pm$ 0.017	[87.77-89.81]	0.012
Graph matching without edge term	85.75 $\pm$ 0.019	[84.73-86.77]	0.007
Without preprocessing techniques	82.71 $\pm$ 0.023	[81.69-83.73]	0.005

From Table 2, the ablation results confirm that edge information and node-edge matching are the core variants for the model’s performance. The model’s variant removal includes removing edges, geometric relations, and disabling edge matching, which significantly demonstrates the mAP degradation with p-values < 0.005 . This result indicates that relational reasoning is a central mechanism for the image retrieval quantity. On the other hand, removing semantic node descriptors leads to a smaller and statistically non-significant reduction, where $p = 0.087$ depicts that the semantic nodes enhance robustness but are not the dominant factor for the retrieval accuracy. The role of preprocessing also influences the accuracy of image retrieval in the proposed IR-AGN model. Moreover, the training hyperparameter for all the ablation variants is fixed across runs, and only the specific component is being removed. This ensures that the performance variation is due to the removed module, not to hyperparameter changes. Altogether, the ablation and statistical analysis validate the proposed IR-AGN model’s relational graph and joint node-edge optimization strategy.

4.1.3. Computational analysis

The computational analysis of the proposed IR-AGN model is evaluated with baseline image retrieval models. This analysis evaluates runtime efficiency and resource consumption across multiple dimensions, including inference time, graph construction, and graph matching cost. The measurement is averaged over 1000 queries, which are executed on an NVIDIA RTX 4090 GPU, Intel i9 CPU, 64 GB

RAM, with identical preprocessing and batch settings. The computational results are given in Table 3.

From Table 3, the proposed IR-AGN model introduces a moderate computational overhead due to region extraction and relational graph construction. In this context, the term N/A specifies the Non-Availability of particular metrics in the published sources. The indexing time (31.7 s) is slightly higher than baseline models, but still suitable for near-real-time retrieval applications. Despite increased computation, the proposed model maintains a competitive GPU memory usage (890MB), which demonstrates that relational modeling does not impose the computational overhead. Nevertheless, the proposed model is more complex than the baseline model, including R-MAC and Trans-DELF model, but signifies higher mAP results 95.19% with interpretable image retrieval.

Table 3. Evaluating the computational efficiency of the proposed IR-AGN model over baseline models

Model	mAP (%)	Inference time (ms per 1 image)	Region extraction time (ms per 1 image)	Graph construction time (ms per 1 image)	Retrieval time (ms per 1 query)	Memory usage (MB)	Indexing time (5K images)
R-MAC	89.18	5.4	N/A	N/A	N/A	620	19.4 s
Trans-DELF	92.37	7.1	4.8	N/A	N/A	720	24.9 s
Proposed IR-AGN model	95.19	6.8	3.2	4.1	9.1	890	31.7 s

4.2. Comparative analysis

Table 4 presents the comparative evaluation of the proposed IR-AGN model across different state-of-the-art CBIR methods on the ROxford dataset under medium and hard protocols. The comparison models primarily rely on global aggregation, handcrafted orientation features, and attention-based semantic refinements. Unlike these comparison models, the proposed IR-AGN model explicitly models region attributes, semantic relations, and geometric consistency to improve robustness for variations present in the ROxford dataset.

Table 4. Comparative evaluation of the proposed IR-AGN model over existing models on the ROxford dataset

Comparative model	mAP (%)	
	Medium	Hard
CCA-IRF [19]	74.19	48.06
MLOH [20]	87.43	43.70
PRB [21]	85.70	70.30
HKMC [22]	83.40	65.30
DSCD [23]	86.60	83.90
SDFH [24]	74.23	N/A
CFCD [25]	82.42	65.06
Proposed IR-AGN	90.57	86.92

In Table 4, the proposed IR-AGN model achieves a superior performance with an mAP score of 90.57% (medium) and 86.92% (Hard) when compared to the methods across the ROxford dataset. The models, such as PRB and DSCD, perform effectively but lack relational interpretability. Furthermore, the existing CBIR models are utilized to compare with the proposed IR-AGN model on the RParis dataset, and the results are given in Table 5.

Table 5. Comparative evaluation of the proposed IR-AGN model over existing models on the RParis dataset

Comparative model	mAP (%)	
	Medium	Hard
CCA-IRF [19]	81.64	64.03
MLOH [20]	92.43	53.92
PRB [21]	92.60	72.70
HKMC [22]	90.80	76.70
DSCD [23]	91.10	86.40
SDFH [24]	84.91	N/A
CFCD [25]	91.57	81.69
Proposed IR-AGN	95.19	89.95

From Table 5, the proposed IR-AGN model achieves the highest mAP in both medium (95.19%) and hard (89.95%) protocols while outperforming all existing models. The improvement in results is achieved through the proposed IR-AGN model's explicit regional-level reasoning and interpretable relational attributes that preserve the fine-grained details of images. While the existing models, such as DSCD and PRB, provide strong localization and semantic refinement, they struggle with region-to-region relational modeling. The output sample images of the proposed model are given in Fig. 3.



Fig. 3. The sample output images of the proposed IR-AGN for image retrieval from the ROxford dataset

5. Research implications

In this research, the proposed IR-AGN model presents several meaningful research implications for advancing interpretable CBIR and computational visual understanding. Firstly, the usage of RAG graphs explicitly encodes visual, semantic, and geometric cues, where the model demonstrates the importance of shifting from global features to regional-level representations. This graph usage establishes a new direction for achieving interpretability, not only with post-hoc addition but also with an intrinsic design component. This enables the retrieval decisions to be transparently traced back to specific image region and their relationships. Consequently, the joint node-edge matching mechanism highlights the relational reasoning in achieving robust retrieval under viewpoint, scale, and illumination variations. This mechanism provides a generalizable strategy for incorporating structural constraints into visual similarity tasks, which can be extended to other domains such as scene graph retrieval and local-based visual recognition in tourism departments. Furthermore, the incorporation of explicit geometric and semantic edge attributes introduces an explainable inter-region interaction that is valuable for tourism in cultural heritage applications. Finally, the strong improvements in accuracy, generalizability, and structural consistency demonstrated across datasets validate the importance of graph-based transparent reasoning to replace dense attention mechanisms in modern retrieval systems. Collectively, these research implications positioned the proposed IR-AGN model as a foundational approach for achieving interpretability in structurally aware vision retrieval systems.

6. Conclusion

This study exploited an interpretable region-attribute-based CBIR model, which models landmark images with the proposed IR-AGN architecture. The proposed model effectively addressed major challenges in instance-level retrieval models, including weak structural reasoning and poor robustness to viewpoint and geometric variations. The interpretability and consistency gaps were resolved through explicit RAG construction, which effectively separated visual, semantic, and geometric cues while preserving complementary regional-level information without inference. Consequently, the joint node-edge graph matching further stabilizes similarity computation by capturing fine-grained spatial relationships and semantic dependencies across images. The retrieval ambiguity is resolved with a built-in interpretability module by quantifying the region and edge contributions for transparent decision-making. Finally, the integrated similarity decomposition

ensures stable reasoning and improved retrieval reliability while demonstrating explicit region-relation modeling in landmark image retrievals. From the results, the proposed IR-AGN model attained better results, including 95.19% on the RParis and 90.57% on the ROxford dataset. The future work will mainly focus on extending the proposed IR-AGN model to larger and multimodal datasets for automating semantic concept extraction using large vision-language models and optimizing graph matching for real-time scalability. Moreover, the model can be extended for real-time medical image retrieval, where interpretable regional-level reasoning helps clinicians to compare similar cases for improving reliable diagnosis while ensuring transparent decision support.

References

1. Amaithi Rajan, A., V. Vetrisevi, M. Ajitesh, R. Praveen Kumar. Secure and Fault-Tolerant Cloud-Based Framework for Medical Image Storage and Retrieval in a Distributed Environment. – Scientific Reports, Vol. **15**, 2025, No 1, 32965.
2. Emanuilov, S., A. Dimov. Billion-Scale Similarity Search Using a Hybrid Indexing Approach with Advanced Filtering. – Cybernetics and Information Technologies, Vol. **24**, 2024, No 4, pp. 45-58.
3. Rahmawati, L., S. Rustad, A. Marjuni, M. Arief, C. S. Soeleman, G. F. Shidik. Transmission Map Refinement Using Laplacian Transform on Single Image Dehazing Based on Dark Channel Prior Approach. – Cybernetics and Information Technologies, Vol. **24**, 2024, No 4, pp. 124-142.
4. Mahbod, A., N. Saeidi, S. Hatamikia, R. Woitek. Evaluating Pre-Trained Convolutional Neural Networks and Foundation Models as Feature Extractors for Content-Based Medical Image Retrieval. – Engineering Applications of Artificial Intelligence, Vol. **150**, 2025, 110571.
5. Kugunavar, S., C. J. Prabhakar. Medical Image Retrieval Using ROI Extraction and a Hybrid Bag-of-Features Model. – Journal of Electrical Systems and Information Technology, Vol. **12**, 2025, No 1, 34.
6. Khunjush, F., S. Vogler, M. Leng. Content-Based 3D Image Retrieval and a ColBERT-Inspired Re-Ranking for Tumor Flagging And Staging. – Journal of Imaging Informatics in Medicine, 2025.
7. Cep, R., M. Elangovan, J. V. N. Ramesh, M. K. Chohan, A. Verma. Convolutional Fine-Tuned Threshold Adaboost Approach for Effectual Content-Based Image Retrieval. – Scientific Reports, Vol. **15**, 2025, No 1, 9087.
8. Gautam, G., A. Khanna. Content-Based Image Retrieval System Using CNN-Based Deep Learning Models. – Procedia Computer Science, Vol. **235**, 2024, pp. 3131-3141.
9. Pacheco, S. A. B., M. Goyani, Z. G. Rehman, S. F. Rehman, T. Champaneria, S. Goyani. Enhanced Content-Based Image Retrieval Using Multivisual Features Fusion. – International Journal of Computers and Applications, Vol. **47**, 2025, No 10, pp. 835-856.

10. Salman Al-Tameemi, I. K., M. R. Feizi-Derakhshi, Z. Farhadi, A. R. Feizi-Derakhshi. Hybrid Model for Visual Sentiment Classification Using Content-Based Image Retrieval and Multi-Input Convolutional Neural Network. – International Journal of Intelligent Systems, Vol. **2025**, 2025, No 1, 5581601.
11. Denner, S., D. Zimmerer, D. Bounias, M. Bujotzek, S. Xiao, R. Stock, L. Kausch, P. Schader, T. Penzkofer, P. F. Jäger, K. Maier-Hein. Leveraging Foundation Models for Content-Based Image Retrieval in Radiology. – Computers in Biology and Medicine, Vol. **196**, 2025, 110640.
12. Rani, S., G. Kasana, S. Batra. An Efficient Content-Based Image Retrieval Framework Using Separable CNNs. – Cluster Computing, Vol. **28**, 2024, No 1, 56.
13. Moon, C., M. H. Kim, J. K. Min. Efficient Top-K Content-Based Image Retrieval Exploiting Matrix Division toward SIFT. – IEEE Access, Vol. **13**, 2025, pp. 157264-157275.
14. Taheri, F., K. Rahbar. Content-Based Image Retrieval Based on Luminance and Texture Decomposition. – The Computer Journal, 2025, pp. 1699-1710.
15. Kumar, R., N. Murthy. Relevance-Aware Content-Based Image Retrieval Using Deep Hybrid Feature Extraction. – Engineering, Technology & Applied Science Research, Vol. **15**, 2025, No 3, pp. 22976-22982.
16. Dogan, B., H. B. Mutlu, M. Yildirim, S. Yalcin, S. Aslan, N. Sampathila, O. Yildirim, E. J. Ciaccio, R. S. Tan, U. R. Acharya. Content-Based Brain Magnetic Resonance Image Retrieval and Classification with the Proposed Deep Learning and Tissue-Based System. – IEEE Access, Vol. **13**, 2025, pp. 122684-122697.
17. Narayanaswamy, R. A., V. K. Gowda, G. N. K. Murthy. Content-Based Image Retrieval Using the 2-Dimensional Convolutional Neural Network. – International Journal of Intelligent Engineering & Systems, Vol. **17**, 2024, No 4, pp. 67-76.
18. Raju, G. K., P. Padmanabham, A. Govardhan. Enhanced Content-Based Image Retrieval with Trio-Deep Feature Extractors with Multi-Similarity Function. – International Journal of Intelligent Engineering & Systems, Vol. **15**, 2022, No 6, pp. 511-525.
19. Shi, K., X. Liu, M. Alrabeiah, X. Guo, J. Lin, H. Liu, J. Chen. Image Retrieval via Canonical Correlation Analysis and Binary Hypothesis Testing. – Information, Vol. **13**, 2022, No 3, 106.
20. Li, X. P., G. H. Liu, F. Lu. Image Retrieval Using Multi-Layer Orientation Histograms. – Neural Processing Letters, Vol. **57**, 2025, No 2, pp. 33.
21. An, G., W. J. Kim, S. Yang, R. Li, Y. Huo, S. E. Yoon. Towards Content-Based Pixel Retrieval in Revisited Oxford and Paris. – In: ICCV, 2023, pp. 20507-20518.
22. Park, D., Y. Hwang. Efficient Image Retrieval Using Hierarchical k-Means Clustering. – Sensors, Vol. **24**, 2024, No 8, 2401.
23. Zhang, B. J., G. H. Liu, Z. Li, S. X. Song. Image Retrieval Using Compact Deep Semantic Correlation Descriptors. – Information Processing & Management, Vol. **61**, 2024, No 3, 103608.
24. Liu, G. H., Z. Y. Li, J. Y. Yang, D. Zhang. Exploiting Sublimated Deep Features for Image Retrieval. – Pattern Recognition, Vol. **147**, 2024, 110076.
25. Zhu, Y., X. Gao, B. Ke, R. Qiao, X. Sun. Coarse-to-Fine: Learning Compact Discriminative Representation for Single-Stage Image Retrieval. – In: ICCV, 2023, pp. 11260-11269.
26. ROxfordDataset (accessed on November 2025).
27. <https://www.kaggle.com/datasets/qiubit/roxfordparis/data>

28. W u, M., Q. Z h o n g. Image Enhancement Algorithm Combining Histogram Equalization and Bilateral Filtering. – Systems and Soft Computing, Vol. **6**, 2024, 200169.
29. W a n g, S., S. J i a n g. Instre: A New Benchmark for Instance-Level Object Retrieval and Recognition. – ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), Vol. **11**, 2015, No 3, 37.

Received: 11.03.2026; Accepted: 11.05.2026

© 2026 by the authors. Licensee IICT-BAS, Sofia, Bulgaria. This article is an open access article distributed under the terms and conditions of the **Creative Commons Attribution-NonCommercial-NoDerivativs** license.