

# Deep ensemble architectures with heterogeneous approach for an efficient content-based image retrieval

Manimegalai A.<sup>1,2</sup>, Josephine Prem Kumar<sup>3</sup>, Nanda Ashwin<sup>4</sup>

<sup>1</sup>Visvesvaraya Technological University, Belagavi, India

<sup>2</sup>Department of Computer Science and Engineering, East Point College of Engineering and Technology, Bengaluru, India

<sup>3</sup>Department of Computer Science and Engineering, Cambridge Institute of Technology, Bengaluru, India

<sup>4</sup>Department of Internet of Things, East Point College of Engineering and Technology, Bengaluru, India

## Article Info

### Article history:

Received Jan 18, 2024

Revised Mar 3, 2024

Accepted Mar 21, 2024

### Keywords:

Content-based image retrieval

Deep learning

HybridEnsembleNet

Image retrieval processes

Text-based image retrieval

## ABSTRACT

In the field of digital image processing, content-based image retrieval (CBIR) has become essential for searching images based on visual content characteristics like color, shape, and texture, rather than relying on text-based annotations. To address the increasing demands for efficiency and precision in CBIR systems, we introduce the HybridEnsembleNet methodology. HybridEnsembleNet combines deep learning algorithms with an asymmetric retrieval framework to optimize feature extraction and comparison in extensive image databases. This novel approach, specifically custom-made for CBIR, employs a lightweight query structure skilled at handling large-scale data under resource-constrained environments. The experiments were performed on the ROxford and RParis datasets. The deep learning component of HybridEnsembleNet significantly refines the accuracy of image matching and retrieval. RParis The ROxford dataset, specifically in the medium and hard difficulty benchmarks, demonstrates an enhancement of 5.53% and 10.44%, respectively. Similarly, the RParis dataset, under medium and hard benchmarks, exhibits improvements of 3.01% and 5.83%, showcasing superior performance compared to existing models. By overcoming the traditional limitations of CBIR systems in mean average precision (mAP) metrics, HybridEnsembleNet provides a scalable, efficient, and more accurate solution for retrieving relevant images from vast digital libraries.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Manimegalai A.

Department of Computer Science and Engineering, East Point College of Engineering and Technology  
Bengaluru, India

Email: lathuramesh@gmail.com

## 1. INTRODUCTION

Every day, a significant amount of images, amounting to terabytes of data, are transmitted and stored on the internet. The inherent continuity of this process enables the formation of a substantial collection of images. The task of identifying relevant images from a vast collection poses a significant challenge, thereby generating prospects for exploring novel opportunities in multimedia research. There exist two primary approaches for retrieving images based on language and content: Text-based image retrieval (TBIR) and content-based image retrieval (CBIR) [1]. The effectiveness of TBIR relies on the textual information, known as metadata that is associated with the image. Textual data can be generated using a variety of techniques or manually inputted into a system. TBIR encounters two significant challenges within the domain of image annotation: The manual annotation process necessitates a significant investment of time and

effort. It is crucial to recognize that the interpretation of annotated data may demonstrate variability among various individuals [2].

The development of CBIR was driven by the need to address the inherent limitations of TBIR. CBIR is a methodology that utilizes visual information to enhance the process of retrieving images. CBIR is a methodology that employs various attributes, such as shape, color, and texture, to effectively classify images based on their inherent visual properties[3]. CBIR is a well-established field of study that continues to be actively researched. The phenomenon can be attributed to several factors, including the substantial increase in image datasets, the diverse range of usage scenarios, and the numerous applications associated with CBIR. In contemporary times, a multitude of search engines are employed to facilitate the storage and retrieval of extensive collections of images from the internet. These collections can reach sizes of terabytes and are accessed daily. CBIR is a specialized domain that encompasses a range of applications, which can be classified into three main categories: association search, image search, and category search [4].

The fundamental principle underlying CBIR revolves around the notion of an image's contents, specifically denoting its distinct characteristics. The process consists of three primary phases: representation, extraction, and feature selection. The primary objective of a content-based retrieval system is to efficiently distinguish and segregate the unique visual attributes that serve as defining characteristics for various forms of media, such as images, videos, and audio files [5]. The procedure of CBIR encompasses various characteristics, such as type, form, texture, and key point descriptors. The attributes of the picture dataset play a critical role in determining the feature selection process. Various color models are employed to extract color attributes. The previously mentioned models offer distinct methodologies for perceiving and representing colors, each designed to suit particular circumstances and applications. The assessment of texture within an image is of utmost importance for evaluating its material properties and overall visual representation. The process involves arranging components in various spatial positions relative to each other [6].

The concept of spatial texture organization refers to the arrangement of texture attributes within an image. The information provided presents valuable insights regarding various characteristics, including directionality, smoothness, coarseness, regularity, and uniformity. The utilization of shape attributes offers benefits in scenarios where objects possess distinct and identifiable structures, such as traffic signs, company names, and logos. Accurate extraction and representation of shape information play a crucial role in the successful implementation of CBIR applications. The effective management of images with clearly defined forms is of utmost importance. In the field of CBIR research, there has been a shift in focus among researchers towards the integration of multiple low-level features as a means to enhance system performance. In the domain of CBIR systems, it has been observed that the integration of multiple characteristics has demonstrated better efficacy compared to the sole reliance on individual features [7].

The identification of identical or similar photographs in response to a specific query image has become more challenging due to the significant growth in the number of images accessible on the internet. The utilization of manual feature extraction techniques significantly increases the complexity of this task. Deep learning algorithms are widely acknowledged as a practical and effective method for addressing this specific problem. In recent years, there has been an observed shift towards the adoption of learning-based techniques, specifically deep learning methods, instead of manual feature extraction and representation methods. These methods facilitate the automated extraction of abstract features from the data [8].

Several design options were presented to effectively accommodate the specific data type being processed. Convolutional neural networks (CNNs) are frequently employed for image data processing, while artificial neural networks (ANNs) have demonstrated their efficacy in handling one-dimensional data [9]. The application of recurrent neural networks (RNNs) [10] offers numerous benefits in the examination of time-series data. The incorporation of various advanced methodologies has facilitated the development of deep learning algorithms utilized in image retrieval. The learning paradigms discussed in this context are specifically related to network-based learning. This approach employs a diverse range of architectures, such as neural networks, convolutional networks, artificial networks, attention networks, Siamese networks, and triplet networks. Furthermore, the topic at hand encompasses various learning approaches, including supervised learning, unsupervised learning, semi-supervised learning, and self-supervised learning [11].

The performance of CBIR systems is undeniably influenced by the quality of images stored in the database. Performance degradation in CBIR systems can occur as a result of various factors, such as the presence of noise, low visibility, and insufficient texture within images. Several factors can inhibit the retrieval of relevant images that correspond to the user's query. The challenges arise from the distortion or loss of crucial visual data, which obstructs the accurate evaluation and comparison of images using CBIR techniques [12]. CBIR systems frequently encounter difficulties in achieving precise query-image matching, leading to suboptimal performance. In addition to assessing the image quality, the storage of CBIR data involves additional complexities. The implementation of effective techniques for managing image data is essential to optimize system resources and achieve rapid retrieval times. When dealing with a large number

of images, it is essential to give careful thought to storage architectures, indexing techniques, and retrieval algorithms. The choice of a data storage strategy has a direct impact on scalability, resource utilization, and retrieval speed. The effectiveness and precision of image retrieval rely significantly on the preservation, categorization, and organization of these images. Achieving an optimal balance between retrieval performance and storage efficiency is crucial for effectively managing various picture sizes, types, and feature representations [13].

To effectively address these challenges, it is vital to implement a comprehensive approach. The recommended strategy should integrate advancements in feature extraction, image enhancement, and storage technologies. At present, there is active research and practical implementation underway to enhance the performance and reliability of these systems. The main goal of their research is to develop innovative methodologies for improving image quality, mitigating visibility issues, and optimizing data storage in CBIR systems [14].

The exponential growth in the volume of images uploaded to the internet daily underscores the importance of efficient and accurate image retrieval systems. The field of CBIR is particularly crucial in this context, as it leverages the visual content of images—such as color, shape, and texture—to facilitate the retrieval process. CBIR's significance is amplified by its ability to directly analyze the visual information, avoiding the need for manual annotation. This approach is not only more aligned with how humans perceive images but also crucial for handling the sheer scale of data. With the advent of deep learning techniques, particularly CNN, CBIR systems have seen substantial improvements in identifying and classifying complex patterns within images. The motivation for continued research in CBIR is clear: to keep pace with the relentless growth of image databases and to meet the demands of diverse applications that rely on quick, accurate image retrieval—be it for digital libraries, medical diagnostics, or multimedia systems. The pursuit of more refined CBIR systems is not just an academic interest but a necessity for the infrastructure of an increasingly digital world. The contribution is mentioned here.

- Advanced asymmetric retrieval model: HybridEnsembleNet technique is designed that implement a lightweight query structure that optimizes performance under resource constraints, enhancing retrieval accuracy in CBIR systems.
- Deep learning integration: Utilizes deep neural networks for refined feature extraction, significantly improving pattern recognition and accuracy in image retrieval.
- Efficient feature embedding strategy: Employs aligned embedding spaces for query and image sets, streamlining the image comparison process for faster and more precise CBIR results.

The research organization is carried out in this paper in four sections: the first section depicts a brief overview of CBIR. The second section discusses the related work, and in the third section, the proposed methodology is designed. In the fourth section, the performance evaluation is carried out where the results are displayed in graphs and tables.

## 2. RELATED WORK

The abundance of image formats available on the internet makes it difficult to identify a particular visual item from a vast database. The retrieval of similar images based on different contents of query images is a technique that is utilized in various domains. These domains include digitally acquainted libraries, crime prevention, fingerprint identification, information systems of biodiversity, medicine, and historical place research. CBIR is a distinct approach to image retrieval that diverges from keyword-based methods by prioritizing the analysis of visual attributes within images instead of relying solely on predetermined keywords. CBIR is a technique that leverages visual elements, including color, form, and texture, to address the challenge of identifying visual entities [15].

Computer vision encompasses a wide range of applications, among them is CBIR. CBIR is a process that focuses on the retrieval of images from a database that contains a vast number of images. The objective of this study is to examine the practical application of a two-stage process for the retrieval of images based on their content. CNNs are utilized for image detection during the initial phase. The CNN demonstrates the ability to process an image efficiently within a single pass. The system can identify and classify multiple objects present in the image, where each object is assigned to a specific class. The problem of detecting multiple classes is resolved by employing a CNN. During the second phase of the process, the acquisition of relevant images takes place after the execution of object detection. The achievement of assigning priority to images within the same class is made possible through the utilization of a relevance ranking system [16].

The cloud classification system categorizes clouds into three distinct levels: high, middle, and low. The cloud categorization process employs the use of CBIR and k-means clustering algorithms. The developed approach classifies clouds into three distinct categories: low, medium, and high. The precipitation

amount is significantly influenced by the type of cloud [17]. The effect of high resolutions on search precision and result organization is not well-established and can exhibit variability. The primary aim of this study is to examine the influence of picture resolution on both search accuracy and result sorting. It is strongly recommended to resize images before adding them to the image database, especially if resizing has any effect on the images.

The process involves the identification of visual characteristics, the correlation of these features based on their effects, and the assessment of the impact of these factors on retrieval. Low-level visual features are identified to target specific perceptual components of visual data, in addition to encompassing high-level characteristics that facilitate image retrieval approaches. The primary goal of this study is to analyze the various components involved in improving the efficiency of CBIR search results [18].

A novel CBIR model that aims to efficiently retrieve images by utilizing query pictures. The proposed model employs an Adadelta-optimized residual network to improve the retrieval process. The proposed model employs a feature extractor obtained from ResNet 50 to extract a suitable set of features. Additionally, the Adadelta optimizer is utilized to effectively optimize the hyperparameters of the ResNet-50 model, resulting in enhanced retrieval performance.

The theoretical foundations and practical implementations of a CBIR system demonstrate high effectiveness. The authors provide an in-depth analysis of the concepts and discuss the real-world applications of this system. The essential components of the system include its characteristics related to colors, textures, and forms. The multilayer searching capability is achieved through the implementation of three subsequent searching processes. The proposed systems (PS) differ significantly from previous methods as they integrate all features concurrently for the single-level search of a typical CBIR system. The PS utilize a sequential approach, wherein each feature is evaluated independently. The output of one step is then used as the input for the next step, following a hierarchical pattern [19].

CBIR is a technique that distinguishes itself from keyword-based image retrieval by prioritizing the analysis of visual contents and attributes of images, including color, form, and texture. In contrast to keyword-based retrieval, which depends on explicit image descriptions, CBIR utilizes visual features to tackle the mentioned problem. The objective of this paper is to present a novel approach to picture retrieval by utilizing a hybrid feature combination technique. The technique employs the color histogram method for extracting color features and subsequently producing the color gradient. The Gabor wavelet method is utilized to extract both outer and inner edges. By implementing the aforementioned techniques, a feature vector will be generated as a result. This feature vector can then be utilized to retrieve visually similar images [20].

### 3. PROPOSED METHODOLOGY

Considering the disadvantage of traditional deep learning of heavy architecture, this research work develops HybridEnsembleNet. That combines the ensemble architecture of various deep learning models and heterogeneous retrieval approaches that make it efficient for higher effectiveness. Moreover, the ensemble network comprises the various deep learning model (deep local and global features [21], [22]).

#### 3.1. Problem definition

Assume  $\vartheta_i(.)$  and  $\vartheta_s(.)$  represents the query structure irrespectively, the visual system of the image set  $\vartheta_i(.)$  is trained and used to map the images  $\varpi$  with the feature vectors. Testing the query structure  $\vartheta_s(.)$  to process the queries  $q$  wherein the retrieval is reduced through the nearest neighbor in the embedded search space. Whereas the evaluation metric is used to enhance the performance of the retrieval system shown as  $\tau(\vartheta_s(.), \vartheta_i(.))$ . The retrieval system in symmetry, the query structure is similar to an image set as  $\vartheta_i(.) = \vartheta_s(.)$ . This deploys a powerful model to reach a high accuracy which is not satisfied through the resource-constrained scenario.

The asymmetric retrieval model, the model  $\vartheta_i(.)$  is trained and fixed. Incorporate resource-constrained scenario, this requires a compatible lightweight query set as  $\vartheta_s(.)$  that is significantly less than the  $\vartheta_i(.)$  in aspect of parameter sizes and computational complexity. The core asymmetric model for feature embeddings of the query and image set is mutually interpreted. The core asymmetric model achieves an accuracy similar in terms of the symmetric model as  $\tau(\vartheta_s(.), \vartheta_s(.)) \approx \tau(\vartheta_q(.), \vartheta_q(.))$  to ensure a balance between performance and efficiency.

#### 3.2. Deep similarity matching

The initial step involves training a feature space compressor (FSC) using the characteristics gathered by the image set. The centroids of the quantization function serve as data points for describing the structure of the space. During the training process of the query set, the image set remains in a frozen state. The query and image set are responsible for mapping each training sample into two separate embeddings. The next step

involves computing the similarities between the two embeddings and comparing them to the centroids to determine the structural similarities. Finally, this approach aims to enhance the query set by restricting the level of consistency between two structural similarities. After undergoing training, the embedding space of both the query set and the image set data points exhibit a high degree of alignment due to their shared nature. Figure 1 shows the proposed architecture.

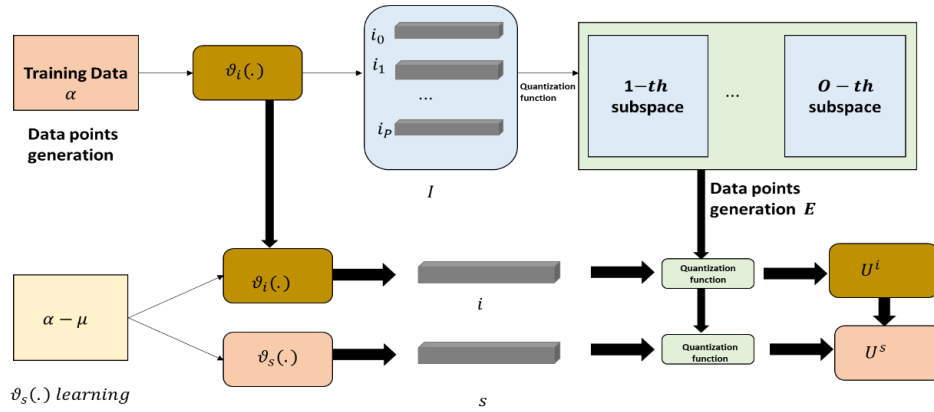


Figure 1. Proposed architecture

### 3.3. Data points generation

A comprehensive characterization of the embedding space, this approach selects the respective data points in the gallery gallery-embedded model. These data points are references in the embedded space which converts query and gallery features in similar structures. A similar approach is used for clustering to generate a series of data points. This method requires a large number of data points to specifically characterize the space structure. The clustering is adapted that requires the training samples within the computational complexity with several numbers of centroids. The large set of centroids included and the clustering cost incurred is high. The quantization function is employed here to effectively expand the data points at less cost. The training data exists here denoted by  $\alpha = \{z_1, z_2, \dots, z_p\}$  for the generation of data points. The image set is denoted as  $\vartheta_i(.)$ , first employs to extract the features as  $I = [i^1, i^2, \dots, i^P] \in T^{P \times f}$  images in  $\alpha$  as given in (1). Wherein, each feature vector  $I^k \in I$  is segmented into  $O$  separate vectors as  $w_1(I^k) \in T^{f^*}, l = 1, 2, \dots, O$  as mentioned in (2).

$$I^k = \vartheta_i(.) \in T^f, k = 1, 2, \dots, P \quad (1)$$

$$I_1^k, \dots, I_{f^*}^k, \dots, I_{f^*+1}^k, \dots, I_f^k, \quad w_1(I^k)w_0(I^k) \quad (2)$$

Here  $I_l^k$  represents the  $l$ -th feature dimension of  $I^k$ ,  $f^* = f/O$  and  $f$  is a multiple of  $O$ . When clustering is performed on each set  $[w_1(I^1); w_1(I^2); \dots; w_1(I^P)] \in T^{P \times f^*}, l = 1, 2, \dots, O$ , specifically to obtain  $E^1 \in T^{M \times f^*}$ , where  $M$  is the number of centroids. The data points in the gallery space are defined as the multiple defined as the Cartesian product. Within any centroid vector that is formed by integrating  $O$  different sub-centroid vectors. In comparison with the clustering, the quantization has distinct advantages. It is easy to generate a large number of data points  $E$ . The total number of data points directly, to store  $O * M$  sub-centroids, while training the adoption of splitting mechanism to evaluate the similarity by segments, instead of directly computing the similarities between feature vectors and the data points to reduce the overhead training.

$$E = E^1 * E^2 * \dots * E^O \in T^{M^O \times f} \quad (3)$$

#### 3.3.1. Query modelling

While learning the process of query learning, the feature query vectors and image set which is first converted into similarity structures upon evaluation against the data points. The image  $z$  is the training dataset  $\mu$ . Let  $I$  and  $s$  be the feature vectors extracted as shown as given in (4). The structure similarities are evaluated as  $U_k^i$  and  $U_k^s$  upon computation of each sub-vector as  $w_1(I)$  and  $w_1(s)$  as the adjacent centroid

vectors in the feature space compression function before training as given in (5). Table 1 shows the data points generation.

$$I = \vartheta_i(z) \in T^f, s = \vartheta_i(z) \in T^f \quad (4)$$

Split it into  $o$  – sub vectors as

$$I \rightarrow w_1(I), w_2(I), \dots, w_o(I),$$

$$S \rightarrow w_1(s), w_2(s), \dots, w_o(s)$$

$$U_k^i = [u(w_k(I), E_1^k, :), \dots, u(w_k(I), E_M^k, :)] \in T^m \quad (5)$$

$$U_k^s = [u(w_k(s), E_1^k, :), \dots, u(w_k(s), E_M^k, :)] \in T^m$$

Table 1. Data points generations algorithms

Input	Train the data $\alpha = \{z_1, z_2, \dots, z_p\}$ ; image set $\vartheta_i(\cdot)$ ; sub-vector set $O$ ; number of centroids per subvector $M$
Step 1	For each image $z_k$ in training data $\alpha$ do Utilizing the image set, extract image features by (1) Divide image features into $O$ sub-vectors $w_l(I^k) \in T^{f_s}, l = 1, 2, \dots, O$ in accordance with (2) end
Step 2	For each vector set [ $w_1(I^1); w_1(I^2); \dots; w_1(I^P)$ ] $\in T^{P \times f_s}$ do Evaluate clustering with $M$ centroid Obtaining adjacent sub-vectors as $E^l \in T^{M \times f_s}$ end
Output	Data points $E = E^1 * E^2 * \dots * E^O \in T^{M \times O \times f}$

Here  $E_l^k$  represents the  $k$  – th centroid vector within the  $l$  – th sub-space and  $u(\cdot, \cdot)$  is considered as the similarity metric. This is formulated as given in (6). By evaluating the constraints  $\gamma_e$  for the structure similarities as  $s$  and  $i$  in the embedding space of the image set. The data points are shared between the query and image set; their embedding space is aligned properly.

$$[u(w_k(I), E_l^k, :)] = \frac{E_l^k \cdot w_k(I)^V}{\|E_l^k\|_2 \|w_k(I)\|_2} \quad (6)$$

### 3.3.2. Similarity matching of query and dataset model

The asymmetric retrieval, a query set  $\vartheta_s$  this maintains feature compatibility and also ensures the structure similarity in  $i$  in the embedded space of the image set. This method focuses on ensuring the consistency of in between the two similarities as  $U_k^i$  and  $U_k^s$  for the adjacent sub-vector pair  $w_k(I)$  and  $w_k(M)$ . The Kullback-Leibler (KL) divergence is adapted to estimate the distance between  $w_k(I)$  and  $w_k(M)$ . Initially  $U_k^i$  then converted into the probability distribution as given in (7).

$$r_k^i = \left[ \frac{\exp\left(\frac{U_{k,1}^i}{\mu_i}\right)}{\sum_{n=1}^M \exp\left(\frac{U_{k,n}^i}{\mu_i}\right)}, \dots, \frac{\exp\left(\frac{U_{k,1}^i}{\mu_i}\right)}{\sum_{n=1}^M \exp\left(\frac{U_{k,n}^i}{\mu_i}\right)} \right] \quad (7)$$

$\mu_i$  is the temperature value used to control the sharpness assignment. The probability distribution corresponding to the  $k$  – th vector of the query feature  $s$  is assigned as given in (8). The similarity constraint between the two parameters for the probabilities on the same sub-centroid vectors is denoted as mentioned in (9). This consists of the cross-entropy of  $r_k^i$  and  $r_k^s$  and the loss is encountered by  $r_k^i$ . this is independent from the feature query set which does not affect the training. The final objective is defined as the sum of all the consistency in the loss adjacent to the  $O$  sub-vectors as mentioned in (10). Table 2 shows the query model training.

$$r_k^s = \left[ \frac{\exp\left(\frac{U_{k,1}^s}{\mu_s}\right)}{\sum_{n=1}^M \exp\left(\frac{U_{k,n}^s}{\mu_s}\right)}, \dots, \frac{\exp\left(\frac{U_{k,1}^s}{\mu_s}\right)}{\sum_{n=1}^M \exp\left(\frac{U_{k,n}^s}{\mu_s}\right)} \right] \quad (8)$$

$$\gamma_{kl}^k = KLL(r_k^i || r_k^s) = \sum_{n=1}^M r_{k,n}^i \log \frac{r_{k,n}^i}{r_{k,n}^s} \quad (9)$$

$$\gamma_{\text{loss}} = \sum_{k=1}^O \gamma_{\text{KLL}}^k \quad (10)$$

The centroids of the quantizing function serve as the data points in the embedded space of the image set. Upon quantizing the feature, the conversion of feature regression into an assignment task. The temperature set is  $\mu_i = 0$ , the probability  $U_k^i$  shown in (7) is one vector with only one single index at 1 index shown as  $i = \arg \max_1 (U_{k,l}^i)$ . This is further simplified as given in (11).

$$\gamma_{\text{KLL}}^k = \sum_{n=1}^M U_{k,n}^i \log \frac{U_{k,n}^i}{U_{k,n}^s} = \log \frac{1}{U_{k,l}^s} \quad (11)$$

The query set is encouraged to optimize the loss to degenerate the data point, which is the quantized feature  $i$ . The degeneration of the details feature of the image set by the query set is prevented by this. However, neglecting the discriminatory information conveyed by the associations between the feature vector and data points leads to poorer performance. To achieve the desired outcome, utilize soft assignments as the prediction target, specifically by setting  $\mu_i$  to a value greater than zero. The overall learning process is summarized in Table 2.

Table 2. Query model training algorithm

Input	The training set $\mu_s$ ; to train the image set $\theta_i(\cdot)$ ; random initializing the query set $\theta_s(\cdot)$ ; data points $E$
Step 1	For each image $z$ in training set $\mu$ do end
Step 2	Extract image features within the gallery and query set according to (4);
Step 3	Segment $i$ and $s$ into $O$ sub-vector according to (4);
Step 4	Evaluate the structure similarities as $U_k^i$ and $U_k^s$ according to (5)
Step 5	Select the consistency constraints $\gamma_e$ considering the structural similarities as $U^i$ and $U^s$ to fine-tune $\theta_s(\cdot)$ according to (11)
Step 6	end
output	Query set $\theta_s(\cdot)$ incompatibility with $\theta_i(\cdot)$

#### 4. PERFORMANCE EVALUATION

The evaluation of HybridEnsembleNet emphasizes its efficacy in enhancing image retrieval scenarios through its unique combination of ensemble architecture and heterogeneous modules. The assessment involves measuring the model's performance against relevant benchmarks, highlighting improvements in retrieval accuracy and efficiency. Additionally, a thorough analysis of HybridEnsembleNet's capabilities, such as its ability to capture both local and global features, provides insights into its effectiveness in addressing the limitations of traditional deep learning architectures for image retrieval.

##### 4.1. Dataset details

The experiments were performed on the ROxford and RParis datasets, which are widely acknowledged in the field of image retrieval [23]. The dataset comprises a total of 70 query photos, which are categorized into three distinct groups: easy, medium, and hard. The hard split is designed to specifically address challenging questions, whereas the medium split encompasses a combination of both easy and difficult questions.

##### 4.2. Results

Mean average precision (mAP) is the mean of the average precision (AP) scores for each query. In scenarios where there are multiple queries or test samples (like different objects to be detected in object detection or multiple search queries in a retrieval system), AP is calculated for each query separately, and mAP is the mean of these AP scores. mAP is a highly important metric because it considers both the precision (how many retrieved items are relevant) and the recall (how many relevant items are retrieved) across all queries. It gives a single-figure measure of quality across recall levels, making it particularly useful for evaluating systems where the retrieval of all relevant documents is important.

##### 4.2.1. Results on ROxford

Table 3 presents a comparison of various methods based on their performance on the ROxford (Medium) benchmark. The methods listed include a mix of individual approaches like deep spatial matching (DSM), image retrieval transformer (IRT), how based aggregated selective match Kernel (how+ASMK), deep orthogonal fusion of local and global features (DOLG), deep attentive local and global modeling (DALG), graph-based reasoning attention pooling with curriculum design (GRAP-CD), multiple dynamic attentions (MDA), contextual similarity distillation (CSD), tokenbased representation (TBR), deep local and

global featuresglobal (DELG), deep local and global features based  $\alpha$ -weighted query expansion (DELG global+ $\alpha$ QE), deep local and global features based geometric verification (DELG global+GV), deep local and global features based reranking transformer (DELG global+RRT), deep local and global features based graph convolution based re-ranking (DELG global+GCR). Performance scores range from 65.3 for DSM to 88.96 for PS, suggesting a progression in effectiveness or an improvement in the techniques used. DELG global and its extensions show a consistent performance in the 70s, indicating a solid baseline. Notably, the entry existing system-based graph convolution-based re-ranking (ES+GCR) scores 84.3, showing an improvisation on ES at 79.3, The increment suggests that GCR provides a substantial improvement. The highest score of 88.96, PS, stands out, possibly indicating a particularly effective method that significantly outperforms others in this benchmark. Figure 2 shows the comparison on ROxford (Medium).

Table 4 displays a set of methodological approaches evaluated on the ROxford (Hard) benchmark. The performance scores indicate how well each method copes with more challenging conditions. The scores span from a low of 31.2 for GRAP-CDto a high of 76.98 for PS, suggesting a wide range of effectiveness among the methods. Notably, traditional methods like DSM and IRT are on the lower end of the performance spectrum, while TBR scores a relatively high 66.6, indicating its robustness. The DELG global method and its variations show moderate performance, with scores generally in the 50s and low 60s, but with the GCR enhancement, it reaches 63.1. ES stands at 62.8, and its enhanced version with GCR significantly outperforms the basic version at 69.7, demonstrating the value of the GCR enhancement. The PS method outshines the others with a notable score of 76.98, which could signify a breakthrough or a particularly advanced approach to the ROxford (Hard) benchmark conditions. Figure 3 shows the comparison on ROxford (Hard).

Table 3.Comparison ofROxford (Medium)

Methods	ROxford (Medium)
DSM [24]	65.3
IRT [25]	67.2
How+ASMK[26]	79.4
DOLG [27]	81.5
DALG [28]	79.9
GRAP-CD [29]	70.8
MDA [30]	81.8
CSD [31]	77.4
TBR [32]	82.3
DELG global [21]	73.6
DELG global + $\alpha$ QE [33]	76.6
DELG global + GV [21]	79.2
DELG global + RRT [34]	78.1
DELG global + GCR [34]	82.1
ES (Existing System) [35]	79.3
ES+ GCR [35]	84.3
PS (HybridEnsembleNet)	88.96

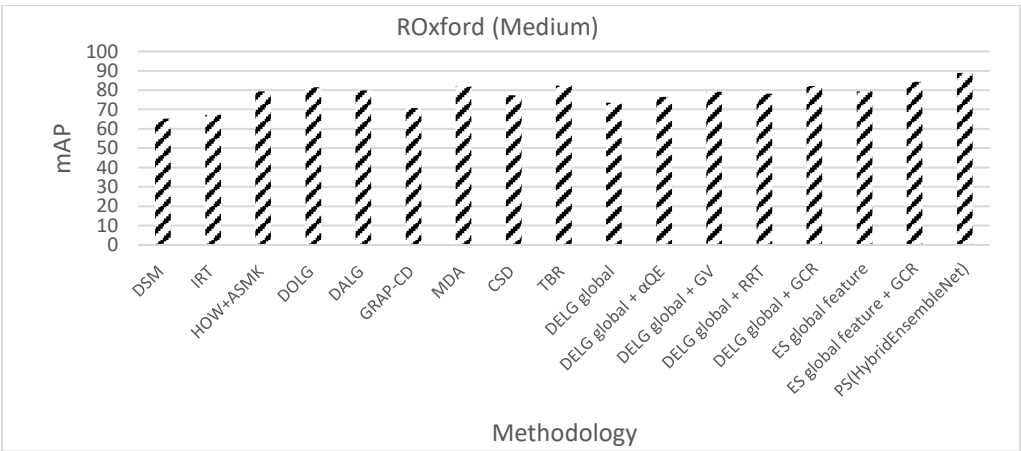


Figure 2.Comparison of ROxford (Medium)



Table 4.Comparison of ROxford (Hard)

Methods	ROxford (Hard)
DSM [24]	39.2
IRT [25]	42.8
How+ASMK[26]	56.9
DOLG [27]	61.1
DALG [28]	57.6
GRAP-CD [29]	31.2
MDA [30]	62.2
CSD [31]	59
TBR [32]	66.6
DELG global [21]	51
DELG global + $\alpha$ QE [33]	54.6
DELG global + GV [21]	57.5
DELG global + RRT [34]	60.2
DELG global + GCR [34]	63.1
ES (Existing System) [35]	62.8
ES+ GCR [35]	69.7
PS (HybridEnsembleNet)	76.98

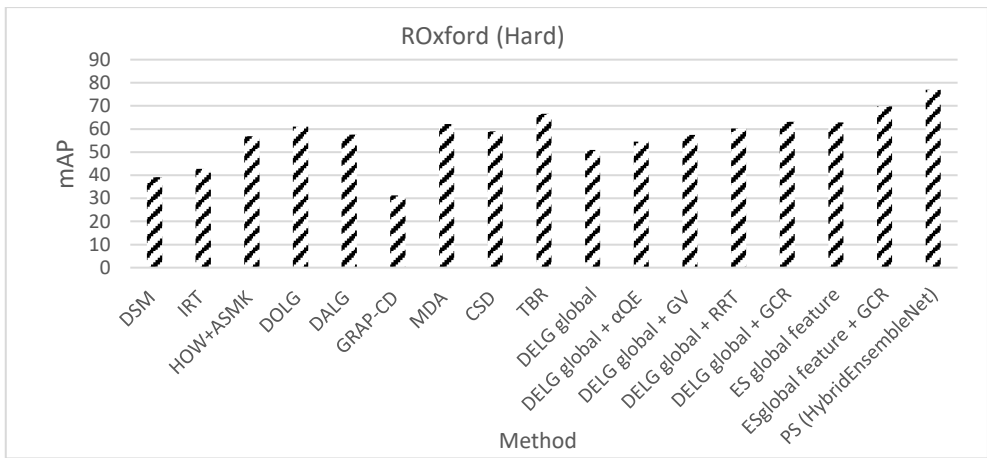


Figure 3.Comparison of ROxford (Hard)

4.2.2. RParis

Table 5 showcases various methods and their corresponding performance scores on the RParis (Medium) benchmark. The spectrum of scores is broad, with DSM at the lower end at 77.4, reflecting a base level of effectiveness, and the highest score achieved by PS at 94.67, indicating superior performance. Middle-tier scores are occupied by methods such as IRT, HOW+ASMK, and GRAP-CD, which fall between 80.1 and 81.6, suggesting moderate effectiveness. Notable high performers include DOLG and DALG, which are close competitors with scores of 91 and 90, respectively. The DELG global method, along with its variations, demonstrates strong performance, particularly with the GCR enhancement, which achieves a score of 89.2. The ES method scores 84.4, but with the addition of GCR, it significantly outperforms its unenhanced counterpart with a score of 91.9. This data underscores the impact of enhancements like GCR on method performance and highlights the PS method as potentially embodying a more advanced or efficient approach. Figure 4 shows the comparison analysis on RParis (Medium).

Table 6 and Figure 5 offer a performance evaluation of various computer vision methods on the ROxford (Hard) dataset. Scores range significantly, highlighting the varied effectiveness of these methods under challenging conditions. The lowest score, at 31.2 by GRAP-CD, suggests some methods may struggle with the dataset's complexity. In contrast, the highest score, at 76.98 by PS, indicates a notably robust approach. DELG global and its enhancements display a progressive improvement, especially with the GCR addition, which scores 63.1. Another key observation is the performance jump from ES at 62.8 to ES+ GCR at 69.7, confirming the substantial benefit of the GCR enhancement. The scores of TBR and MDA, at 66.6 and 62.2, respectively, denote methods that are more effective than the baseline but not as high as the leading scores. Overall, this table reflects the performance of the diverse methods in image retrieval tasks, with particular enhancements offering significant improvements.

Table 5.Comparison of RParis (Medium)

Methods	RParis (Medium)
DSM [24]	77.4
IRT [25]	80.1
How+ASMK[26]	81.6
DOLG [27]	91
DALG [28]	90
GRAP-CD [29]	81.2
MDA [30]	83.3
CSD [31]	87.9
TBR [32]	89.3
DELG global [21]	85.7
DELG global + $\alpha$ QE [33]	86.7
DELG global + GV [21]	85.5
DELG global + RRT [34]	86.7
DELG global + GCR [34]	89.2
ES (Existing System) [35]	84.4
ES+ GCR [35]	91.9
PS (HybridEnsembleNet)	94.67

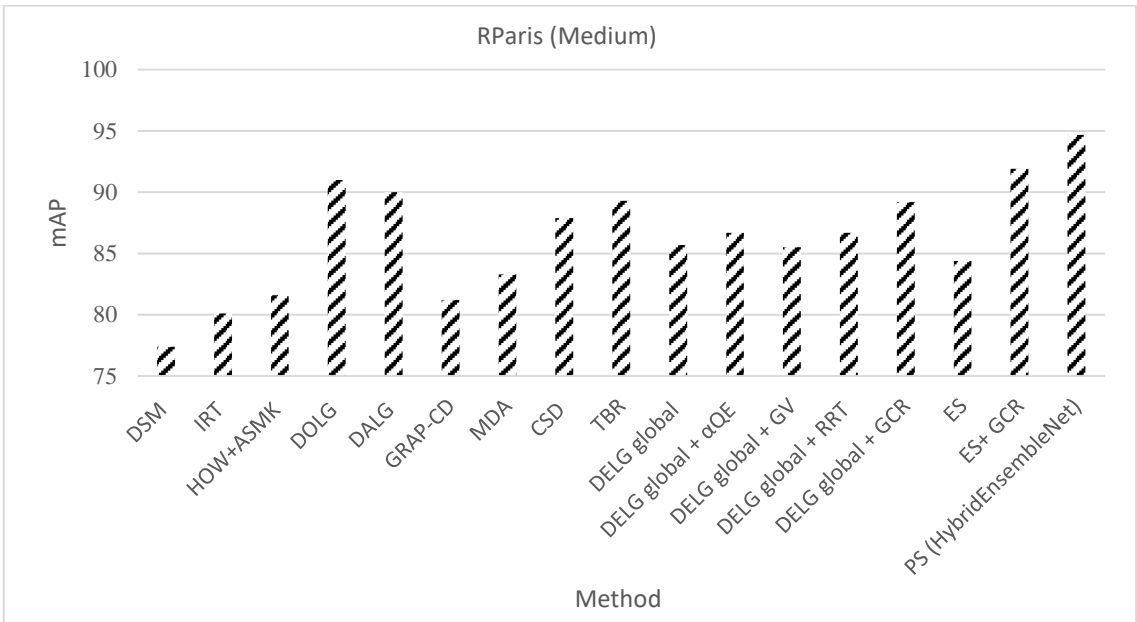


Figure 4.Comparison analysis on RParis (Medium)

Table 6.Comparison analysis on RParis (Hard)

Method	RParis (Hard)
DSM [24]	56.2
IRT [25]	60.5
How+ASMK[26]	62.4
DOLG [27]	80.3
DALG [28]	79.1
GRAP-CD [29]	62.6
MDA [30]	66.2
CSD [31]	75.7
TBR [32]	78.6
DELG global [21]	71.5
DELG global + $\alpha$ QE [33]	73.2
DELG global + GV [21]	67.2
DELG global + RRT [34]	75.1
DELG global + GCR [34]	72.4
ES (Existing System) [35]	74.4
ES+ GCR [35]	79.9
PS (HybridEnsembleNet)	84.56

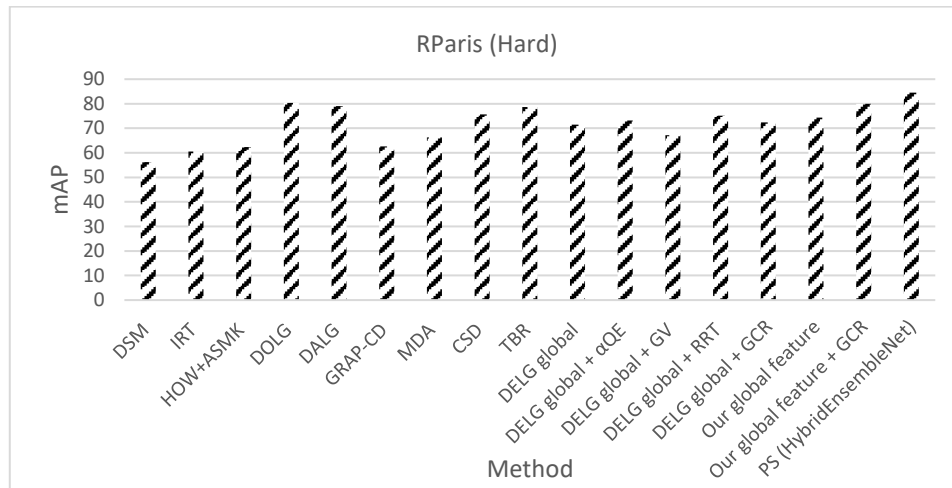


Figure 5. Comparison analysis on RParis (Hard)

#### 4.3. Comparative analysis

Table 7 compares the performance improvements of different methods on the ROxford and RParis benchmarks, with the ES serving as the baseline and the PS method representing the improved technique. Across the table, the PS method shows a positive uplift in performance. For ROxford (Medium), there's a moderate improvement of 5.53%, while a more substantial gain is observed in the ROxford (Hard) setting, where the improvement is 10.44%, indicating that the PS method significantly outperforms ES under more challenging conditions. Similarly, for RParis (Medium), the enhancement is more at 3.01%, which showcases that the PS method shows an advantage even when the baseline performance is already high. Lastly, RParis (Hard) showcases a 5.83% increase, reinforcing the trend that PS provides consistent improvisation over ES.

Table 7. Comparison analysis

Method	ES	PS (HybridEnsembleNet)	Improvisation in (%)
ROxford (Medium)	84.3	88.96	5.53
ROxford (Hard)	69.7	76.98	10.44
RParis (Medium)	91.9	94.67	3.01
RParis (Hard)	79.9	84.56	5.83

## 5. CONCLUSION





This research work presents HybridEnsembleNet methodology as a significant leap forward in the field of CBIR. By seamlessly integrating deep learning with an asymmetric retrieval model, HybridEnsembleNet addresses the critical challenges of accuracy and computational efficiency in handling large-scale image datasets. Its innovative approach to feature extraction and embedding not only enhances the precision of image retrieval but also ensures scalability and speed, crucial for modern digital applications. The successful implementation of HybridEnsembleNet underscores its potential as a transformative solution in CBIR, promising to elevate the standards for image search and retrieval technologies. This methodology not only meets the current demands of diverse CBIR applications but also lays a robust foundation for future advancements in the field, marking a pivotal moment in the evolution of image retrieval systems. As part of future work, an important focus will be on further optimizing HybridEnsembleNet to reduce image retrieval time.

## REFERENCES





- [1] D. Srivastava, B. Rajitha, and S. Agarwal, "Content-based image retrieval for categorized dataset by aggregating gradient and texture features," *Neural Computing and Applications*, vol. 33, no. 19, pp. 12247–12261, 2021, doi: 10.1007/s00521-020-05614-y.
- [2] D. Srivastava, B. Rajitha, S. Agarwal, and S. Singh, "Pattern-based image retrieval using GLCM," *Neural Computing and Applications*, vol. 32, no. 15, pp. 10819–10832, 2018, doi: 10.1007/s00521-018-3611-1.
- [3] S. Chauhan, R. Prasad, P. Saurabh, and P. Mewada, "Dominant and LBP-based content image retrieval using combination of color, shape and texture features," in *Advances in Intelligent Systems and Computing*, Springer Singapore, 2018, pp. 235–243, doi: 10.1007/978-981-10-7871-2\_23.
- [4] H. Kavitha and M. V. Sudhamani, "Content-based image retrieval using edge and gradient orientation features of an object in an image from database," *Journal of Intelligent Systems*, vol. 25, no. 3, pp. 441–454, 2016, doi: 10.1515/jisys-2014-0088.

- [5] M. Dey, B. Raman, and M. Verma, "A novel colour- and texture-based image retrieval technique using multi-resolution local extrema peak valley pattern and RGB colour histogram," *Pattern Analysis and Applications*, vol. 19, no. 4, pp. 1159–1179, Nov. 2016, doi: 10.1007/s10044-015-0522-y.
- [6] M. Verma and B. Raman, "Local neighborhood difference pattern: A new feature descriptor for natural and texture image retrieval," *Multimedia Tools and Applications*, vol. 77, no. 10, pp. 11843–11866, 2017, doi: 10.1007/s11042-017-4834-3.
- [7] M. Marinov, "Comparative analysis on different degrees of JPEG compression used in CBIR systems," *2020 XI National Conference with International Participation (ELECTRONICA)*, Sofia, Bulgaria, 2020, pp. 1–4, doi: 10.1109/ELECTRONICA50406.2020.9305154.
- [8] T. L. D. Likhitha, M. Noushika, V. S. Deepika, and V. M. Manikandan, "A detailed review on CBIR and its importance in current era," in *2021 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, Oct. 2021, pp. 124–128, doi: 10.1109/ICoDSA53588.2021.9617481.
- [9] Y. Xu, Q. Lin, J. Huang, and Y. Fang, "An improved ensemble-learning-based CBIR algorithm," *2020 Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC)*, Fuzhou, China, 2020, pp. 1–3, doi: 10.1109/CSRSWTC50769.2020.9372466.
- [10] T. Sutojo, P. S. Tirajani, D. R. I. M. Setiadi, C. A. Sari, and E. H. Rachmawanto, "CBIR for classification of cow types using GLCM and color features extraction," *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Yogyakarta, Indonesia, 2017, pp. 182–187, doi: 10.1109/ICITISEE.2017.8285491.
- [11] M. A. Aboali, I. Elmaddah, and H. E. Abdelmunim, "Neural textual features composition for CBIR," *IEEE Access*, vol. 11, pp. 28506–28521, 2023, doi: 10.1109/ACCESS.2023.3259737.
- [12] G. V. R. M. Kumar and D. Madhavi, "Stacked siamese neural network (SSiNN) on neural codes for content-based image retrieval," *IEEE Access*, vol. 11, pp. 77452–77463, 2023, doi: 10.1109/ACCESS.2023.3298216.
- [13] Z. Zhang, W. Lu, X. Feng, J. Cao, and G. Xie, "A discriminative feature learning approach with distinguishable distance metrics for remote sensing image classification and retrieval," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 889–901, 2023, doi: 10.1109/jstars.2022.3233032.
- [14] J. Pradhan, C. Bhaya, A. K. Pal, and A. Dhuriya, "Content-based image retrieval using DNA transcription and translation," *IEEE Transactions on NanoBioscience*, vol. 22, no. 1, pp. 128–142, 2023, doi: 10.1109/tnb.2022.3169701.
- [15] D. Srivastava, S. S. Singh, B. Rajitha, M. Verma, M. Kaur, and H.-N. Lee, "Content-based image retrieval: a survey on local and global features selection, extraction, representation, and evaluation parameters," *IEEE Access*, vol. 11, pp. 95410–95431, 2023, doi: 10.1109/access.2023.3308911.
- [16] G. Sumbul, J. Xiang, and B. Demir, "Towards Simultaneous image compression and indexing for scalable content-based retrieval in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022, doi: 10.1109/tgrs.2022.3204914.
- [17] Z. Xia, L. Jiang, D. Liu, L. Lu, and B. Jeon, "BOEW: A content-based image retrieval scheme using bag-of-encrypted-words in cloud computing," *IEEE Transactions on Services Computing*, vol. 15, no. 1, pp. 202–214, 2022, doi: 10.1109/tsc.2019.2927215.
- [18] J. Madake, R. Agrawal, V. Pawar, and S. Bhatlawande, "A content based image retrieval system for biodiversity system," *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*, 2023, pp. 1–6, doi: 10.1109/GCAT59970.2023.10353428.
- [19] Y. Mahajan, P. Batta, M. Sharma, and D. Saxena, "A model for content-based image retrieval using machine learning," *2023 1st International Conference on Circuits, Power and Intelligent Systems (CCPIS)*, 2023, pp. 1–6, doi: 10.1109/CCPIS59145.2023.10291361.
- [20] J. S. S. Kumar and S. M. C. Vigila, "A review on content based image retrieval techniques," in *Proceedings of the International Conference on Circuit Power and Computing Technologies, ICCPCT 2023*, Aug. 2023, pp. 1251–1256, doi: 10.1109/ICCPCT58313.2023.10245360.
- [21] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *Computer Vision—ECCV 2020: 16th European Conference*, Cham: Springer International Publishing, 2020, pp. 726–743, doi: 10.1007/978-3-030-58565-5\_43.
- [22] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," *Computer Vision – ECCV 2018*, Springer International Publishing, pp. 122–138, 2018, doi: 10.1007/978-3-030-01264-9\_8.
- [23] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Revisiting Oxford and Paris: large-scale image retrieval benchmarking," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5706–5715, 2018, doi: 10.1109/cvpr.2018.00598.
- [24] O. Simeoni, Y. Avrithis, and O. Chum, "Local features and visual words emerge in activations," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 11651–11660, 2019, doi: 10.1109/cvpr.2019.01192.
- [25] A. El-Nouby, N. Neverova, I. Laptev, and H. Jégou, "Training vision transformers for image retrieval," *arXiv-Computer Science*, pp. 1–10, 2021, doi: 10.48550/arXiv.2102.05644.
- [26] G. Tolias, T. Jenicek, and O. Chum, "Learning and aggregating deep local descriptors for instance-level recognition," *Computer Vision – ECCV 2020*, Springer International Publishing, pp. 460–477, 2020, doi: 10.1007/978-3-030-58452-8\_27.
- [27] M. Yang *et al.*, "DOLG: single-stage image retrieval with deep orthogonal fusion of local and global features," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, pp. 11772–11781, 2021, doi: 10.1109/iccv48922.2021.01156.
- [28] Y. Song, R. Zhu, M. Yang, and D. He, "DALG: deep attentive local and global modeling for image retrieval," *arXiv-Computer Science*, pp. 1–14, 2022, doi: 10.48550/arXiv.2207.00287.
- [29] X. Zhu, H. Wang, P. Liu, Z. Yang, and J. Qian, "Graph-based reasoning attention pooling with curriculum design for content-based image retrieval," *Image and Vision Computing*, vol. 115, 2021, doi: 10.1016/j.imavis.2021.104289.
- [30] H. Wu, M. Wang, W. Zhou, and H. Li, "Learning deep local features with multiple dynamic attentions for large-scale image retrieval," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11416–11425, 2021, doi: 10.1109/iccv48922.2021.01122.
- [31] H. Wu, M. Wang, W. Zhou, H. Li, and Q. Tian, "Contextual similarity distillation for asymmetric image retrieval," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9489–9498, 2022, doi: 10.1109/cvpr52688.2022.00927.
- [32] H. Wu, M. Wang, W. Zhou, Y. Hu, and H. Li, "Learning token-based representation for image retrieval," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 2703–2711, 2022, doi: 10.1609/aaai.v36i3.20173.
- [33] F. Radenovic, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, 2019, doi: 10.1109/tpami.2018.2846566.
- [34] F. Tan, J. Yuan, and V. Ordonez, "Instance-level image retrieval using reranking transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, pp. 12105–12115, 2021, doi: 10.1109/iccv48922.2021.01189.
- [35] Y. Zhang, Q. Qian, H. Wang, C. Liu, W. Chen, and F. Wang, "Graph convolution based efficient re-ranking for visual retrieval," *IEEE Transactions on Multimedia*, vol. 26, pp. 1089–1101, 2024, doi: 10.1109/tmm.2023.3276167.





**BIOGRAPHIES OF AUTHORS**

**Prof. Manimegalai A.**     holds an M.Tech. Degree from Sathyabama University and B.E. degree in Computer Science & Engineering from Golden Valley Institute of Technology. She is pursuing Ph.D. from Visveswaraya Technological University, Belagavi. She has worked at M. V. J. College of Engineering and currently working as Assistant Professor in East Point College of Engineering and Technology. Her area of specialization is image classification and retrieval, machine learning, and deep learning. She can be contacted at email: lathuramesh@gmail.com.



**Dr. Josephine Prem Kumar**     received her B.Tech. degree in Electronics and Communication Engineering and M.Tech. degree in Computer Science from Regional Engineering College (now National Institute of Technology), Warangal and Ph.D. in Computer Science and Engineering from Dr. MGR Educational and Research Institute, Dr. MGR University, Chennai. After serving ITI Limited, Bangalore for over fifteen years and Infycons Creative Software Private Ltd., Bangalore for a brief period, she has taken up the teaching profession. She has worked as Professor in MVJ College of Engineering, Bangalore and East Point Group of Institutions and is currently working as Professor-CSE in Cambridge Institute of Technology, Bangalore. She has been guiding Ph.D. students under Visveswaraya Technological University. She can be contacted at email: d\_prem\_k@yahoo.com.



**Dr. Nanda Ashwin**     working as HOD, Department CSE- (IoT & CSBT) East Point College of Engineering has about 25 years of teaching experience. She received his B.E. degree in Computer Science and Engineering M. S. in System Software and M.Tech. degree in Software Engineering with distinction from VTU, Belagavi. She has published 16 research papers in refereed international journals and 10 research papers in the proceedings of various international conferences. She has received several best paper awards for his research papers at various international conferences. Her areas of research include wireless communication, cloud computing, big data analytics, and data science. She can be contacted at email: nandaashwin@eastpoint.ac.in.