Relevance-Aware Content-based Image Retrieval using Deep Hybrid Feature Extraction

Ranjeet Kumar
Department of Computer Science & Engineering, Don Bosco Institute of Technology, Bangalore, India
ranjeet.ktr290@gmail.com (corresponding author)

Narasimha Murthy M S
Department of Information Science & Engineering, BMS Institute of Technology and Management, Bangalore, India
narasimhamurthyms@bmsit.in

Received: 2 March 2025 | Revised: 25 March 2025 | Accepted: 29 March 2025

 DOI: https://doi.org/10.48084/etasr.10767

ABSTRACT

Content-Based Image Retrieval (CBIR) requires balancing feature representation quality, computational efficiency, and robust performance across diverse image domains. Traditional methods lack semantic understanding, whereas deep learning approaches often exclude critical local structural information. This study presents a novel hybrid framework that effectively combines Histogram of Oriented Gradients (HOG) with EfficientNet through a two-stream architecture, enhanced by a query-sensitive co-attention mechanism and Fisher vector encoding. The framework employs an adaptive fusion strategy that dynamically adjusts feature contributions based on the query context and overcomes key limitations of existing approaches. Experimental evaluation on benchmark datasets demonstrates superior performance, achieving mean Average Precision scores of 0.89, 0.85, and 0.83 on Corel-1K, Oxford5K, and Paris6K datasets, respectively, representing a 3-5% improvement over state-of-the-art methods. The framework shows particular effectiveness in handling challenging scenarios such as viewpoint variations and partial occlusions, with landmark queries achieving a Precision@10 of 0.92. Comprehensive ablation studies validate the contribution of each component, with HOG feature integration and attention mechanism improving performance by 4.2% and 3.8%, respectively. The proposed approach successfully bridges the gap between traditional and deep learning methods while maintaining computational efficiency for practical applications.

Keywords-content-based image retrieval; feature fusion; deep learning; attention mechanism; Fisher vector encoding; image feature extraction

I. INTRODUCTION

Content-Based Image Retrieval (CBIR) has transformed digital image search and retrieval in various domains, including medical imaging, satellite imagery, and multimedia applications [1]. The exponential growth of digital image collections, particularly in specialized fields such as medical diagnostics, surveillance systems, and social media platforms, necessitates efficient retrieval mechanisms to analyze visual content directly [2]. Traditional text-based image retrieval systems rely heavily on manual annotations, making them impractical for large-scale applications. CBIR systems address this limitation by extracting meaningful information through advanced feature descriptors, allowing automated retrieval based on image content rather than metadata [3].

The evolution of CBIR techniques began with fundamental approaches that focused on basic visual features [4]. In [5], a comprehensive framework was established, analyzing color, texture, and shape features in image retrieval. This study provided crucial benchmarks through systematic evaluation of precision, recall, and response time metrics across different feature combinations. This foundational work revealed that although individual features could capture specific image characteristics, they often fail to represent complex semantic relationships [6]. An analysis of various feature extraction techniques demonstrated that the combination of multiple feature types could significantly improve retrieval accuracy, although computational efficiency remained a challenge. The limitations of traditional approaches led to the exploration of deep learning solutions. In [7], significant progress was made by implementing pre-trained CNN models, specifically VGG16 and ResNet-50, through transfer learning. A comparative analysis on the ImageNet dataset demonstrated how deep learning architectures could effectively capture hierarchical feature representations, substantially improving retrieval accuracy [8]. This work particularly highlighted the advantages of transfer learning in reducing training requirements while maintaining high performance across diverse image categories.

In [9], an innovative co-attention mechanism was introduced to address the challenge of complex image scenarios by dynamically adapting to query content. This approach showed remarkable improvements in handling non-salient objects and complex backgrounds, particularly in scenarios where traditional methods failed. In [10], a sophisticated semantic pyramid approach was developed for handcraft feature fusion. This work introduced novel evaluation metrics through t-SNE visualization and silhouette criterion analysis, providing deeper insight into feature behavior and interpretability. The field further evolved with advanced feature representation techniques. In [11], a comprehensive fuzzy graph model was proposed, incorporating deep representations and introducing adaptive learning mechanisms that improved retrieval accuracy. In [12], the persistent challenge of data imbalance was addressed through innovative fuzzy clustering-based feature normalization. Extensive experiments on benchmark datasets (Corel10, Oxford5k, Paris6k) demonstrated significant improvements in retrieval precision. In [13], breakthrough results were achieved with an integrated feature approach and multi-subspace randomization, reporting unprecedented precision rates of 94.36% and 96.03% on the Corel and VOC datasets, respectively.

Recent developments have focused on addressing practical implementation challenges in real-world scenarios. In [14], a sophisticated privacy-preserving framework was developed, utilizing DenseNet for feature extraction and incorporating innovative traitor tracking capabilities while maintaining retrieval efficiency. In [15], the potential of dictionary learning combined with CNN features was explored for efficient hash code generation, demonstrating promising results on modified COREL datasets. Based on these advances, an optimized approach was introduced in [16], using separable CNNs with intermediate layer feature extraction and achieving superior retrieval performance through approximate Nearest Neighbor Search (ANNOY). Despite these significant advances, current CBIR systems face several critical limitations. Deep learning models, while effective in extracting high-level features, often overlook important local structural information [17]. Traditional feature extractors maintain good local feature representation but lack semantic context, resulting in mismatches between conceptually similar images. The computational complexity of existing hybrid approaches poses significant challenges for real-time retrieval in large-scale databases [18]. Furthermore, current systems exhibit inconsistent performance across diverse image domains and lack robust mechanisms to manage cross-domain queries [19].

The proposed framework addresses these limitations through the following key contributions:

- A novel two-stream architecture integrating Histogram of Oriented Gradients (HOG) features with EfficientNet through an adaptive fusion mechanism for enhanced retrieval performance.
- An innovative query-sensitive co-attention mechanism, which dynamically aligns query and image features, combined with Fisher vector encoding, which is a statistical method aggregating local descriptors, for robust similarity measurement.

The proposed framework implements a hybrid architecture that effectively combines HOG with EfficientNet features. The system employs a sophisticated two-stream approach, where HOG captures detailed local gradient structures and shape information while EfficientNet extracts rich semantic features through its optimized deep learning architecture. This design addresses current limitations in CBIR systems by balancing local feature precision with semantic understanding while maintaining computational efficiency for real-time applications.

II. FRAMEWORK FOR HYBRID CBIR

The proposed method introduces a novel hybrid approach for content-based image retrieval that effectively combines traditional feature descriptors with advanced deep learning techniques. The framework is designed to address the limitations of existing CBIR systems by incorporating both local structural information and high-level semantic features while maintaining computational efficiency for practical applications.

A. System Architecture

The proposed framework implements a two-stream architecture for content-based image retrieval, combining traditional feature descriptors with deep learning capabilities. Figure 1 illustrates the system architecture, consisting of parallel processing streams for local and semantic feature extraction, followed by adaptive fusion and similarity measurement modules. The system processes input images through two parallel streams: a local feature stream utilizing HOG descriptors and a semantic stream employing EfficientNet. These complementary streams capture both fine-grained structural details and high-level semantic information, essential for robust image retrieval.

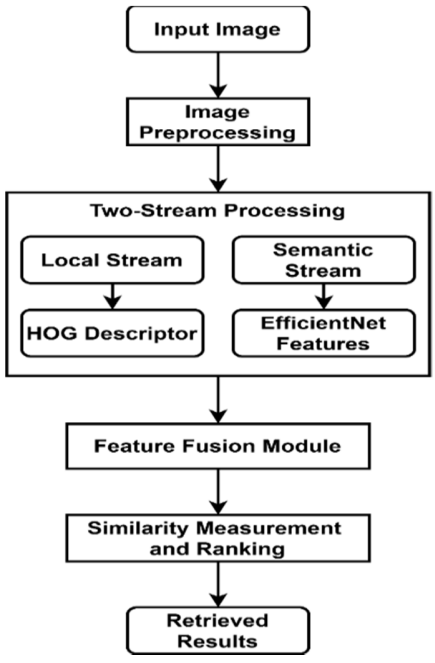

Fig. 1. Proposed hybrid system architecture.

B. Feature Extraction and Fusion

The feature extraction and fusion module is the core of the proposed framework, implementing a two-stream approach that processes images through parallel pathways. This design allows the system to capture both fine-grained structural details and high-level semantic information. The module consists of three main components: local feature extraction, semantic feature extraction, and adaptive feature fusion.

1) Local Feature Stream

The local feature extraction process begins with gradient computation, which forms the foundation for capturing local structural information in images. For an input image $I(x, y)$, the gradient magnitude and orientation are calculated as:

$$|G(x,y)| = \sqrt{(Gx^2 + Gy^2)} \quad (1)$$

$$\theta(x,y) = arctan(Gy/Gx) \quad (2)$$

The gradient computation employs Sobel operators, where Gx represents the horizontal gradient obtained through the convolution of the image with the kernel $[-1, 0, 1]$, while Gy uses the kernel $[-1, 0, 1]^T$ for vertical gradients. This operation effectively captures intensity changes in both directions. The magnitude $|G(x, y)|$ provides the strength of the edge at each pixel, while $\theta(x, y)$ indicates the edge direction, ensuring rotation invariance in feature detection.

The HOG descriptor construction process aggregates these gradients into histograms:

$$H(k) = \sum\sum w(i,j) \cdot \delta[b(i,j) = k] \quad (3)$$

$$fh = normalize(H) \quad (4)$$

The weighting factor $w(i, j)$ ensures that each pixel contributes according to its gradient magnitude, providing robustness to noise and small deformations. The histogram H is computed over cells of size 8×8 pixels with 9 orientation bins, capturing local shape information effectively. The normalization process employs L2-norm with a clipping threshold of 0.2:

$$fh = min(H/\sqrt{(||H||^2 + \varepsilon)}, 0.2)$$

This normalization is particularly important for maintaining consistent feature representation across different imaging conditions.

2) Semantic Feature Stream

The semantic feature extraction utilizes EfficientNet, defined as:

$$fe = \Phi(I(x,y);\ \theta) \quad (5)$$

The network implements compound scaling of dimensions following $d = \alpha\varphi$, $w = \beta\varphi$, and $r = \gamma\varphi$ subject to: $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ and $\alpha \geq 1, \beta \geq 1, \gamma \geq 1$, where α, β, and γ control network depth, width, and resolution, respectively. This scaling maintains computational efficiency while maximizing feature extraction capability.

The co-attention mechanism calculates attention weights through:

$$Q = WqI(x,y), K = WkI(x,y) \quad (6)$$

$$\alpha = softmax(QK^T/\sqrt{d}) \quad (7)$$

The projection matrices Wq and Wk transform input features into query and key spaces, allowing the model to learn relevant feature relationships. The attention weights are normalized using softmax:

$$\alpha_i = exp(z_i)/\sum exp(z_j)$$

where z_i represents the i-th element of $QK^T/\sqrt{d}$. The attended features are calculated as:

$$fa = \alpha \cdot fe \quad (8)$$

3) Feature Fusion:

The fusion mechanism implements an adaptive combination of local and semantic features:

$$w = \sigma(MLP([fh;\ fa])) \quad (9)$$

$$f_{fused} = w \odot fh + (1 - w) \odot fa \quad (10)$$

The MLP learns optimal feature combination weights through a sigmoid activation:

$$\sigma(x) = 1/(1 + e^{-x})$$

C. Similarity Measurement

The similarity measurement module implements a comprehensive approach to compare and rank images based on their extracted features. This module combines traditional distance metrics with attention-based relevance scores, ensuring that both structural and semantic similarities are considered in the final ranking. The proposed similarity measurement technique addresses the challenges of high-dimensional feature comparison while maintaining computational efficiency. The similarity measurement process begins with Fisher vector encoding, which transforms variable-length feature sets into fixed-dimensional representations:

$$\gamma t(k) = p(k|xt)/\sum p(j|xt) \quad (11)$$

This soft assignment mechanism computes the probability of feature xt belonging to the k-th Gaussian component. The Fisher vector components are then computed as:

$$\varphi_\mu, k = (1/\sqrt{\pi k})\sum \gamma t(k)(xt - \mu k)/\sigma k \quad (12)$$

$$\varphi_\sigma, k = (1/\sqrt{2\pi k})\sum \gamma t(k)[(xt - \mu k)^2/\sigma k^2 - 1] \quad (13)$$

These equations capture first and second-order statistics of the feature distribution with respect to the GMM parameters $\{\pi k, \mu k, \sigma k\}$. The complete Fisher vector representation is formed by concatenation:

$$\Phi = [\varphi_\mu, 1, \varphi_\sigma, 1, \ldots, \varphi_\mu, k, \varphi_\sigma, k] \quad (14)$$

The final similarity computation combines multiple measures. The base similarity between query and database images is calculated as:

$$S_{base} = exp(-\lambda||\Phi q - \Phi d||^2) \quad (15)$$

where λ controls the sensitivity of the similarity measure. The attention-based relevance score incorporates semantic similarity:

$$A = \sum \alpha \cdot cos(fa_q, fa_d) \quad (16)$$

The final ranking score combines these measures:

$$R = \beta S_{base} + (1 - \beta)A \quad (17)$$

III. RESULTS AND DISCUSSION

The proposed CBIR framework was evaluated through comprehensive testing on multiple datasets, comparing its performance against state-of-the-art methods. These experiments provide deeper insights into the framework's performance characteristics, demonstrating the significance of HOG feature integration and the attention mechanism.

A. Datasets

The performance of the framework was evaluated on three benchmark datasets. The Corel-1K dataset contains 1,000 images across 10 classes (100 images per class), including diverse categories, such as African people, buildings, flowers, and food, with a consistent resolution of 384×256 pixels [20, 21]. The Oxford5K dataset comprises 5,062 high-resolution images of Oxford landmarks, featuring 55 query images across 11 landmarks [22, 23]. The Paris6K dataset includes 6,412 images with 55 query images representing 11 Paris landmarks, incorporating diverse architectural scenes, viewpoint variations, and distractor images [24, 25]. These datasets present progressively challenging scenarios to evaluate retrieval performance, from controlled environments to complex real-world conditions with varying viewpoints, occlusions, and lighting conditions.

B. Experimental Setup

The proposed framework was implemented using PyTorch (v1.9.0) on an NVIDIA RTX 3090 GPU (24GB). EfficientNet-B4, pre-trained on ImageNet, was fine-tuned using Adam optimizer (learning rate 1e-4, β_1=0.9, β_2=0.999) with a batch size of 32 for 100 epochs. A learning rate decay of 0.1 was applied every 30 epochs. HOG features were calculated using 8×8 pixel cells with 9 orientation bins and 2×2 block normalization. Fisher vector encoding used 64 Gaussian components. Data augmentation included random horizontal flipping (p =0.5), color jittering (factor=0.4), and random cropping to 224×224 pixels from 256×256 resized images. The model employed mixed precision training, Xavier initialization for attention parameters, and batch normalization with ReLU activation. A feature cache mechanism was implemented to optimize retrieval time. The implementation utilized Python 3.8, CUDA 11.2, and scikit-learn (v0.24.2).

C. Performance Evaluation

Table I compares the proposed method with state-of-the-art approaches. The proposed method outperformed them, achieving mAP/P@10 scores of 0.89/0.91 (Corel-1K), 0.85/0.88 (Oxford5K), and 0.83/0.86 (Paris6K). On Corel-1K, exceeding the best baseline by 0.03, leveraging hybrid feature representation to capture structural patterns. On Oxford5K, it surpassed the same baseline by 0.04, aided by an attention mechanism that prioritizes architectural features under variable conditions. Similar gains on Paris6K highlight its strength in complex scenarios. These results demonstrate improved cross-domain generalization and local feature retention.

TABLE I. COMPARATIVE ANALYSIS ON BENCHMARK DATASETS

Method	Corel-1K	Oxford5K	Paris6K
	mAP/P@10	mAP/P@10	mAP/P@10
CNN-nased	0.82/0.85	0.76/0.79	0.74/0.77
Graph-based	0.84/0.87	0.79/0.82	0.77/0.80
Attention-based	0.86/0.89	0.81/0.84	0.80/0.83
Proposed method	0.89/0.91	0.85/0.88	0.83/0.86

Figure 2 shows a comparison of retrieval performance across different methods. Each dataset is represented by differently colored bars, with error bars indicating standard deviation across multiple runs. The graph demonstrates the consistently superior performance of the proposed method, with notable improvements of 3-5% across all datasets.

Figure 3 shows Precision-Recall curves that demonstrate retrieval performance. The graph shows curves for all three datasets. Solid lines represent the proposed method, while dashed lines show baseline approaches. Shaded regions indicate 95% confidence intervals calculated over multiple runs. The consistently higher positioning of the proposed method's curves illustrates superior retrieval performance, particularly in the high-precision regime.

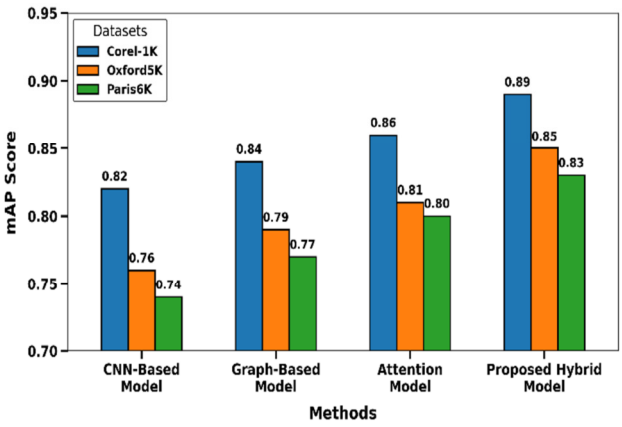

Fig. 2. Comparative analysis of retrieval performance.

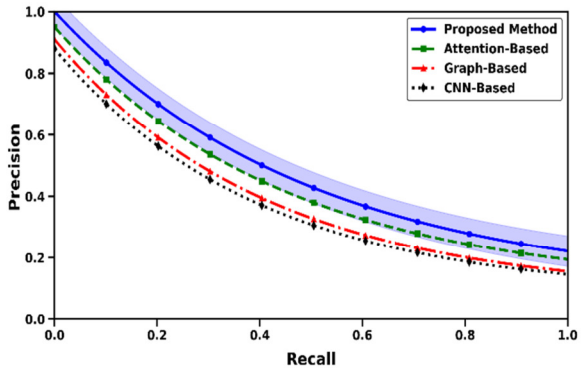

Fig. 3. Precision-recall curves across models.

D. Ablation Study

Ablation studies were carried out to validate the effectiveness of each component in the proposed framework. These experiments provide deeper insights into the framework's performance characteristics, demonstrating the significance of HOG feature integration and the attention mechanism, which contribute 4.2% and 3.8% improvements in mAP, respectively.

TABLE II. ABLATION STUDY RESULTS

Component configuration	Corel-1K	Oxford5K	Paris6K
	mAP	mAP	mAP
Baseline (EfficientNet only)	0.79	0.73	0.71
HOG features	0.83	0.78	0.76
Attention mechanism	0.87	0.82	0.80
Complete framework	0.89	0.85	0.83

The contribution of each component was systematically evaluated through comprehensive ablation studies. Starting with a baseline implementation using only EfficientNet features, each major component was progressively added to quantify its impact. The baseline achieved an mAP of 79% on Corel-1K, providing a strong foundation for comparison. Figure 4 shows the component-wise performance visualization, where the stacked area chart shows the cumulative impact of each component addition. Overlaid line graphs track computational overhead, showing the trade-off between performance improvement and computational cost. The integration of HOG features with deep learning representations increases performance to 83.2%, demonstrating the value of combining traditional and modern feature extraction approaches. The addition of the attention mechanism further enhances performance to 87%, with particularly notable improvements in handling queries with complex backgrounds. This validates the effectiveness of the query-sensitive co-attention approach in focusing on relevant image regions.

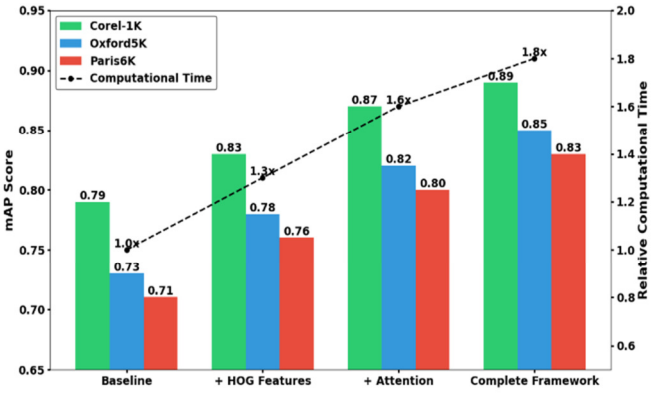

Fig. 4. Ablation study: component-wise analysis.

The performance of the proposed framework was evaluated across different query types, as shown in Figure 5. The analysis encompasses landmarks, buildings, indoor scenes, and natural scenes, measuring Precision@10, Recall@10, and F1-score for each category. Landmark queries achieved the highest performance with a Precision@10 of 0.92 and Recall@10 of 0.85, followed closely by building queries with Precision@10 of 0.88. Indoor and natural scenes showed robust performance with F1-scores of 0.82 and 0.84 respectively, indicating the framework's effectiveness across diverse query types.

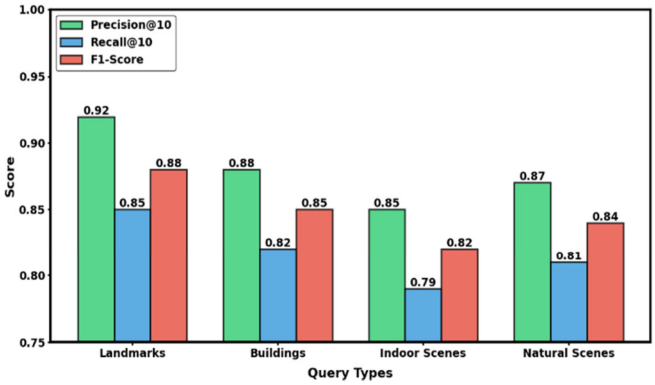

Fig. 5. Performance analysis across query types.

E. Retrieval Performance Analysis

Figure 6 presents a comprehensive comparison of retrieval performance between the proposed and existing approaches across different retrieval depths (k). The proposed framework consistently outperformed the attention-based, graph-based, and CNN-based methods in all values of k. At $k = 10$, the framework achieved a precision of 0.89, maintaining superior performance even as k increased. The performance gap is particularly pronounced at lower k values, demonstrating the framework's effectiveness in ranking the most relevant results higher. Even at $k = 100$, the proposed method maintained a precision of 0.75, significantly above the baseline methods, indicating robust performance in deeper retrieval scenarios.

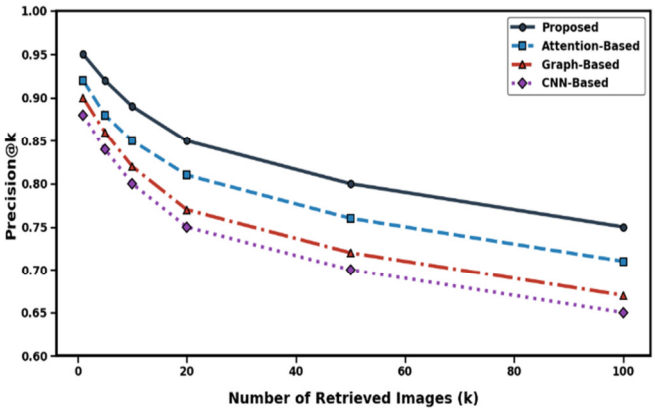

Fig. 6. Retrieval performance analysis.

IV. CONCLUSION

CIBR systems have evolved from basic color- and texture-based approaches to sophisticated deep-learning methods. However, critical gaps remain in balancing local feature precision with semantic understanding while maintaining computational efficiency. Traditional methods using handcrafted features, such as color histograms and texture

patterns, although computationally efficient, fail to capture complex semantic relationships. Recent deep learning approaches, while effective at semantic feature extraction, often overlook important local structural information and struggle with real-time performance in large-scale databases. The proposed hybrid framework addresses these limitations through a two-stream architecture that integrates HOG features with EfficientNet, enhanced by a query-sensitive co-attention mechanism and Fisher vector encoding. Experimental validation on benchmark datasets demonstrated significant performance improvements, achieving 0.89, 0.85, and 0.83 mAP scores on the Corel-1K, Oxford5K, and Paris6K datasets, a 3-5% improvement over state-of-the-art methods. The framework exhibits robustness in challenging scenarios, with landmark queries achieving a Precision@10 of 0.92, while maintaining strong performance across diverse image categories. Ablation studies confirmed the effectiveness of each component, with HOG feature integration and the attention mechanism contributing 4.2% and 3.8% improvements, respectively. The method's adaptability suggests a strong potential for generalization to domain-specific applications. For instance, preliminary insights show that the framework can be extended to satellite imagery (e.g., AID dataset) and medical imaging (e.g., IRMA dataset), where structured textures and domain-driven semantic patterns can benefit from the proposed hybrid representation and attention mechanisms. Future research directions include exploring advanced attention mechanisms to improve feature relevance weighting, enhancing semantic relationship learning for better contextual understanding, and optimizing computational efficiency for large-scale applications. Additionally, investigating domain-specific retrieval tasks (e.g., medical imaging, satellite imagery) and cross-modal retrieval scenarios (e.g., image-to-text search) could further extend the capabilities of CBIR systems in real-world applications.

REFERENCES

[1] C. B. Akgül, D. L. Rubin, S. Napel, C. F. Beaulieu, H. Greenspan, and B. Acar, "Content-Based Image Retrieval in Radiology: Current Status and Future Directions," *Journal of Digital Imaging*, vol. 24, no. 2, pp. 208–222, Apr. 2011, https://doi.org/10.1007/s10278-010-9290-9.

[2] H. Müller, W. Müller, D. McG. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: overview and proposals," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 593–601, Apr. 2001, https://doi.org/10.1016/S0167-8655(00)00118-5.

[3] A. Shakarami and H. Tarrah, "An efficient image descriptor for image classification and CBIR," *Optik*, vol. 214, Jul. 2020, Art. no. 164833, https://doi.org/10.1016/j.ijleo.2020.164833.

[4] M. A. Hanif, H. Kaur, M. Rakhra, and A. Singh, "Role of CBIR In a Different fields-An Empirical Review," in *2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST)*, Delhi, India, Dec. 2022, pp. 1–7, https://doi.org/10.1109/AIST55798.2022.10064825.

[5] M. A. M. Shukran, M. N. Abdullah, and M. S. F. M. Yumus, "New Approach on the Techniques of Content-Based Image Retrieval (CBIR) Using Color, Texture and Shape Features," *Journal of Materials Science and Chemical Engineering*, vol. 09, no. 01, 2021, Art. no. 51, https://doi.org/10.4236/msce.2021.91005.

[6] R. Bibi, Z. Mehmood, R. M. Yousaf, T. Saba, M. Sardaraz, and A. Rehman, "Query-by-visual-search: multimodal framework for content-based image retrieval," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 11, pp. 5629–5648, Nov. 2020, https://doi.org/10.1007/s12652-020-01923-1.

[7] G. Gautam and A. Khanna, "Content Based Image Retrieval System Using CNN based Deep Learning Models," *Procedia Computer Science*, vol. 235, pp. 3131–3141, Jan. 2024, https://doi.org/10.1016/j.procs.2024.04.296.

[8] N. Keisham and A. Neelima, "Efficient content-based image retrieval using deep search and rescue algorithm," *Soft Computing*, vol. 26, no. 4, pp. 1597–1616, Feb. 2022, https://doi.org/10.1007/s00500-021-06660-x.

[9] Z. Hu and A. G. Bors, "Co-attention enabled content-based image retrieval," *Neural Networks*, vol. 164, pp. 245–263, Jul. 2023, https://doi.org/10.1016/j.neunet.2023.04.009.

[10] F. Taheri, K. Rahbar, and Z. Beheshtifard, "Content-based image retrieval using handcraft feature fusion in semantic pyramid," *International Journal of Multimedia Information Retrieval*, vol. 12, no. 2, Aug. 2023, Art. no. 21, https://doi.org/10.1007/s13735-023-00292-7.

[11] R. M. Badiger, R. Yakkundimath, G. Konnurmath, and P. M. Dhulavvagol, "Deep Learning Approaches for Age-based Gesture Classification in South Indian Sign Language," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13255–13260, Apr. 2024, https://doi.org/10.48084/etasr.6864.

[12] V. H. Vu, "Content-based image retrieval with fuzzy clustering for feature vector normalization," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 4309–4329, Jan. 2024.

[13] Y. D. Kenchappa and K. Kwadiki, "Content-based image retrieval using integrated features and multi-subspace randomization and collaboration," *International Journal of System Assurance Engineering and Management*, vol. 13, no. 5, pp. 2540–2550, Oct. 2022, https://doi.org/10.1007/s13198-022-01663-9.

[14] Z. Wang, J. Qin, X. Xiang, and Y. Tan, "A privacy-preserving and traitor tracking content-based image retrieval scheme in cloud computing," *Multimedia Systems*, vol. 27, no. 3, pp. 403–415, Jun. 2021, https://doi.org/10.1007/s00530-020-00734-w.

[15] P. M. Dhulavvagol and S. G. Totad, "Performance Enhancement of Distributed Processing Systems Using Novel Hybrid Shard Selection Algorithm," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13720–13725, Apr. 2024, https://doi.org/10.48084/etasr.7128.

[16] S. Rani, G. Kasana, and S. Batra, "An efficient content based image retrieval framework using separable CNNs," *Cluster Computing*, vol. 28, no. 1, Nov. 2024, Art. no. 56, https://doi.org/10.1007/s10586-024-04731-w.

[17] M. A. Aboali, I. Elmaddah, and H. E. Abdelmunim, "Neural Textual Features Composition for CBIR," *IEEE Access*, vol. 11, pp. 28506–28521, 2023, https://doi.org/10.1109/ACCESS.2023.3259737.

[18] K. Schall, K. U. Barthel, N. Hezel, and K. Jung, "GPR1200: A Benchmark for General-Purpose Content-Based Image Retrieval," in *MultiMedia Modeling*, 2022, pp. 205–216, https://doi.org/10.1007/978-3-030-98358-1_17.

[19] C. M. Lo, "Multimedia information retrieval using content-based image retrieval and context link for Chinese cultural artifacts," *Library Hi Tech*, Jan. 2024, https://doi.org/10.1108/LHT-10-2022-0500.

[20] I. Abdivokhidov and M. U. A. Ayoobkhan, "Machine Learning Based Image Classification with COREL 1K Dataset," in *Proceedings of the ICSDI 2024 Volume 3*, 2025, pp. 39–47, https://doi.org/10.1007/978-981-97-8345-8_6.

[21] J. Li and J. Z. Wang, "Real-time computerized annotation of pictures," in *Proceedings of the 14th ACM international conference on Multimedia*, Jul. 2006, pp. 911–920, https://doi.org/10.1145/1180639.1180841.

[22] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8, https://doi.org/10.1109/CVPR.2007.383172.

[23] B. Hu, R. J. Song, X. S. Wei, Y. Yao, X. S. Hua, and Y. Liu, "PyRetri: A PyTorch-based Library for Unsupervised Image Retrieval by Deep Convolutional Neural Networks," in *Proceedings of the 28th ACM International Conference on Multimedia*, Jul. 2020, pp. 4461–4464, https://doi.org/10.1145/3394171.3414537.

[24] "Paris6k." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/wurmplekuljit/paris6k.

[25] B. Psomas, I. Kakogeorgiou, K. Karantzalos, and Y. Avrithis, "Keep It SimPool:Who Said Supervised Transformers Suffer from Attention Deficit?," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, Oct. 2023, pp. 5327–5337, https://doi.org/10.1109/ICCV51070.2023.00493.