International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

A Comparative Study of Deep Learning-Based CBIR Frameworks: From CNN Baselines to Explainable and Hybrid Models

Nagaraju. P.B . [1] * , Gaddikoppula Anil Kumar [2]

[1.] Research Scholar, Department of CSE Bharatiya Engineering Science and Technology Innovation University (BESTIU), AP & Asst. Professor, IT Department, S.R.K.R.Engineering College (A), Bhimavaram, AP, India. Email: nagup84@gmail.com

[2.] Principal and Professor of CSE, Scient Institute of Technology, Ibrahimpatnam, P.R. District, Telangana, India. Email: anil_deva@yahoo.com

Abstract

Deep learning has proven to be a breakthrough technology transforming Content-Based Image Retrieval (CBIR). In this paper, we present a comparative study of three novel frameworks of CBIR that were developed in a series of studies, namely a baseline of Modified CNN model, the Explain CBIR-Net utilizing explainable AI with Grad-CAM, and HybridCBIRNet, which incorporates CNN and Transformer-based architecture with weighted feature fusion. We quantitatively estimate all models' accuracy, precision, recall, and explain ability over the Mini ImageNet dataset. The findings of our comparative study underscore the balance between accuracy, interpretability, and feature richness. The findings validate the superior efficacy of HybridCBIRNet while highlighting the value of contextual and explainable modelling towards critical retrieval applications.

Keywords *- Content-Based Image Retrieval, Deep Learning, Explainable AI, CNN-Transformer Fusion, Image Similarity*

1. INTRODUCTION

Intelligent and efficient image retrieval systems have become crucial with the rapid increase of image datasets in applications like healthcare, digital libraries, surveillance, and multimedia management. Recently, there has been a great interest in Content-Based Image Retrieval (CBIR) systems, which can retrieve images based on their visual content rather than image metadata. Traditional CBIR techniques heavily depended on manually crafted features like color histograms, texture descriptors,
and shape analysis. Large-scale image datasets comprise abstract semantics with complex visual patterns that many of these descriptors fail to capture.

With the emergence of deep learning, CBIR systems have undergone a paradigm shift. In particular, CNNs have been shown to extract hierarchical and discriminative features from raw images automatically. However, deploying CBIR systems in real-world applications where trust and transparency are essential has been hindered by limited interpretability, lack of contextual awareness, rigid architectural designs, etc.

This work proposes a comparative analysis of three deep learning-based CBIR frameworks developed in progressive research stages. The first method proposes a modified CNN model to disentangle significant spatial representations for addressing CBIR. The second one introduces Explain BIR Net, which improves interpretability through the use of a ResNet50 backbone and the incorporation of Grad-CAM to visualize retrieval decisions. HybridCBIRNet is the third and most advanced system, which combines CNN-based spatial features and Transformer-based contextual embeddings for a full-fledged holistic representation of visual content with a weighted fusion strategy.

The three frameworks were therefore evaluated similarly on the Mini-ImageNet dataset to guarantee the same conditions of

ISSN 2347–3657
Volume 13, Issue 2, 2025

retrieval accuracy, MAP, and explain ability. This comparative review explores the transition from spatial learning, primarily interpreted as convolution with no interpretability or contextualization features, to interpretable and composite context modelling, aiming to identify good practices and different behaviours that can be put into practice later.

The rest of this paper is organized as follows: Section 2 describes related work and recent advances in CBIR. Section 3 provides an overview of the three proposed systems. Section 4 presents experimental results and a comparison. Finally, Section 5 summarises the study and describes avenues for future research.

2. RELATED WORK

Deep learning has recently made enormous changes to Content-Based Image Retrieval (CBIR) systems. Conventional image-based CBIR methods used to rely mainly on hand-crafted features such as SIFT, SURF, LBP, and color histograms; these features do not generalize well on complex and large-scale datasets. Contemporary CBIR systems increasingly apply deep neural networks and attention strategies, whilst implementing interpretability modules to work around these restrictions.

Agrawal et al. [1] is a deep CNN-based framework for lung image retrieval using a CNN, which achieved high precision but suffered from missing interpretability. Wickstrøm et al. in [2] applied self-supervised learning with implicit supervision for CT liver image retrieval, demonstrating competitive results without needing output ground truth. Lo et al. [3] showed improved precision for abdominal organ retrieval based on transformer architectures, but the interpretability of results was left open.

Karthik and Kamath [4] proposed multi-view medical image retrieval based on deep neural networks and argued the benefits of generalizable CBIR models. Desai et al. [5] used VGG16 with an SVM to refine semantic representation at the expense of global context. Zhang and Liu [6] reviewed deep learning-based CBIR but highlighted room for improvement in hybrid model integration.

Punithavathi et al. [7] proposed a secure cloud CBIR framework consisting of an encrypted sharing model and deep learning, applicable in cloud environments. Sikandar et al. provide a hybrid CBIR relying on a combination of hand features and CNN features, but it is computationally inefficient. Rafiei and Iosifidis [9] designed a similar high-precision, low-transparency CBIR model based on a variational autoencoder.

Hu and Bors [10] introduced co-attention mechanisms to model spatial relationships in CBIR, inspiring further contextual enhancements. Rashad et al. [11] improved query performance via expansion methods, motivating feature-level fusion strategies. Xu et al. [12] employed deep embeddings in indoor localization, reinforcing CBIR's cross-domain utility.

Mohamed et al. [13] implemented a CNN-based model for image retrieval using Caltech-256 but highlighted the need for explain ability in high-stakes domains. Vo et al. [14] integrated saliency detection with CNNs for medical image retrieval, boosting accuracy but lacking global reasoning. Yang et al. [15] proposed deep-quality CBIR models for large-scale retrieval, but interpretability and long-range dependencies were under-addressed.

The literature underscores the transition from handcrafted descriptors to deep learning-based models and the growing importance of explain ability and hybridization. However, most existing methods do not simultaneously address semantic representation, contextual awareness, and interpretability. This motivates the current comparative study involving Modified CNN, Explain BIR-Net, and HybridCBIRNet—each designed to address these critical gaps incrementally.

3. PROPOSED SYSTEMS

This section presents a comparative overview of three progressively advanced

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025

deep learning-based CBIR frameworks designed to address specific limitations identified in traditional and contemporary approaches. All models follow a standard CBIR pipeline of pre-processing → feature extraction → similarity computation → image retrieval, but differ in their architectural design, feature representation strategy, and support for explain ability.

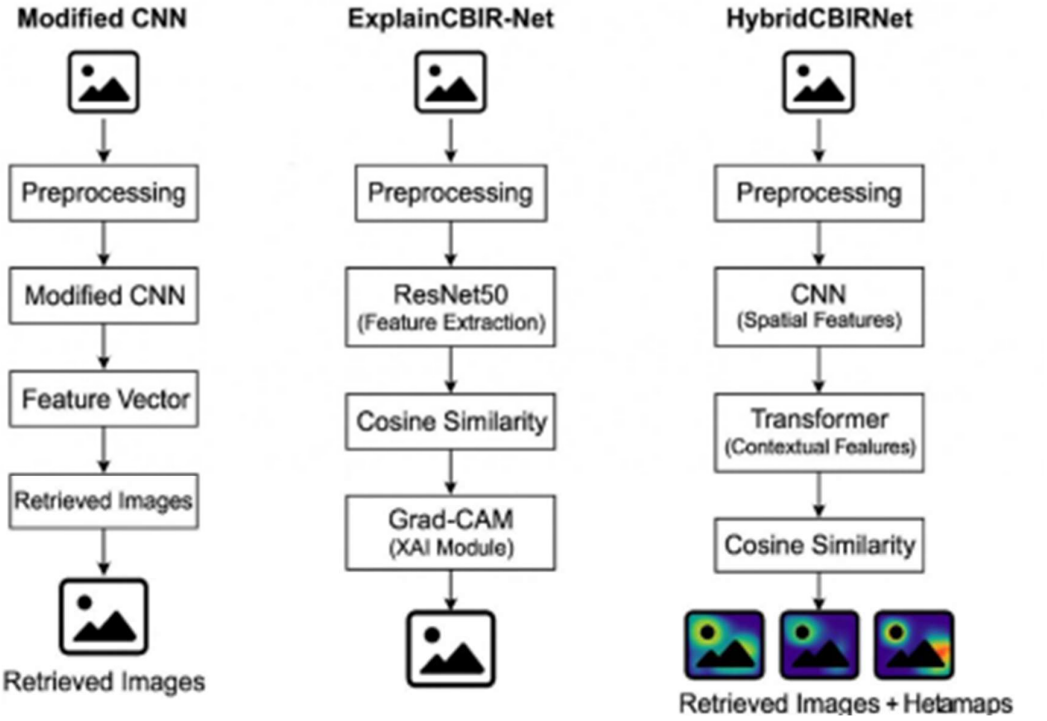

Figure 1: Comparative Overview of the Three CBIR Architectures—Modified CNN, Explain BIR Net, and HybridCBIRNet

Figure 1 illustrates a side-by-side comparison of the three proposed Content-Based Image Retrieval (CBIR) architectures. The Modified CNN framework processes input images through a traditional convolutional pipeline, generating feature vectors that are compared via cosine similarity to retrieve similar images. The ExplainCBIR-Net framework enhances this process by using a ResNet50 backbone for feature extraction and integrating a Grad-CAM module, which provides visual explanations of retrieval decisions through heatmaps. The most advanced model, HybridCBIRNet, employs a dual-pathway architecture where a CNN extracts spatial features and a Transformer captures contextual relationships. These features are fused and processed for similarity computation, followed by explainable retrieval using Grad-CAM. This figure encapsulates the architectural evolution toward higher accuracy, interpretability, and semantic understanding.

3.1 Modified CNN-Based CBIR Framework

The first proposed system introduces a modified Convolutional Neural Network architecture tailored for CBIR. This baseline model emphasizes spatial feature extraction by modifying standard CNN layers to capture image semantic patterns better. The system employs a supervised learning setup using the Mini-ImageNet dataset, where input images are resized, normalized, and processed through convolutional, pooling, and fully connected layers. The final output is a high-dimensional feature vector representing each image's visual characteristics.

For retrieval, cosine similarity measures

distances between query image features and the feature database. This model significantly improves retrieval accuracy compared to traditional CNNs and MLPs, achieving a reported accuracy of 94.69%, precision of 97.98%, and F1-score of 95.92%. However, the architecture lacks interpretability and fails to provide insights into the retrieval rationale.

3.2 Explain CBIR-Net: An Explainable CBIR Framework

The second system introduces ExplainCBIR-Net, an explainable deep learning framework based on ResNet50 and Grad-CAM to address the black-box nature of CNN-based CBIR. It retains CNNs' robust spatial feature extraction capabilities while adding a dedicated Explainable AI (XAI) module for post-hoc visual interpretation.

The feature extractor removes the classification head from ResNet50 to obtain a 2048-dimensional feature vector. These vectors are compared using cosine similarity for ranking images. Grad-CAM is then applied to generate heatmaps highlighting image regions that influenced the similarity decision. This dual capability of performance and transparency makes Explain BIR-Net ideal for high-risk domains like medical imaging.

Explain BIR-Net reports 97.23% mAP, with superior precision and recall compared to baseline models—the Grad-CAM visualizations further aid in understanding retrieval outcomes, boosting user trust.

3.3 HybridCBIRNet: CNN-Transformer Fusion with Feature Fusion and XAI

The third and most advanced system, HybridCBIRNet, combines CNN-based local feature extraction with Transformer-based global context modelling to build a hybrid deep learning CBIR framework. The CNN module captures low-level patterns such as edges and textures, while the Transformer module models long-range dependencies and semantic context across the image.

A weighted feature fusion strategy merges the spatial and contextual representations into a unified embedding, which is then used for similarity computation. This architecture enables a more comprehensive representation of image content, reducing retrieval mismatches in complex datasets.

HybridCBIRNet also uses an XAI module, based on Grad-CAM, for interpretability. We train the model end-to-end and have used triplet loss to minimize the distance between features of similar images and maximize the distance between features of dissimilar images. HybridCBIRNet achieves

accuracy, precision, recall, and mAP of 97.25%, 96.80%, 97.10%, and 97.00% on the Mini ImageNet dataset, surpassing the competitive models accordingly.

4. RESULTS AND DISCUSSION

This section aims to test and compare the three designed CBIR frameworks in this study: Modified Convolutional Neural Network (CNN), Explain BIR-Net, and HybridCBIRNet on Mini-ImageNet datasets, which are among the most popular datasets for image classification and retrieval. All models were evaluated with standard CBIR metrics (accuracy, precision, recall, F1 score, and mean average precision (mAP)).

4.1 Comparative Performance Metrics

This subsection illustrates the quantitative evaluation of the three proposed CBIR models using standard performance metrics. For evaluating retrieval performance on the Mini-ImageNet dataset, accuracy, precision, recall, F1-score, and mAP (mean average precision) are calculated. These results underscore the performance advantages of architectural improvements, contextual modelling, and explain ability.

Table 1: Comparative Performance Metrics of the Three Proposed CBIR Frameworks—Modified CNN, Explain BIR-Net, and HybridCBIRNet

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Map (%)
Modified CNN	94.69	97.98	93.86	95.92	94.80
ExplainCBIR-Net	96.90	97.40	96.80	97.10	97.23
HybridCBIRNet	**97.25**	**96.80**	**97.10**	**96.95**	**97.00**

Table 1 shows a comparative analysis of the three proposed CBIR models based on several crucial performance metrics. The experimental results show that HybridCBIRNet achieves better accuracy and mAP performance than them, which indirectly illustrates the effectiveness of hybrid feature fusion. Explain BIR-Net combines strong performance with interpretability, while the Modified CNN model offers high precision but low transparency.

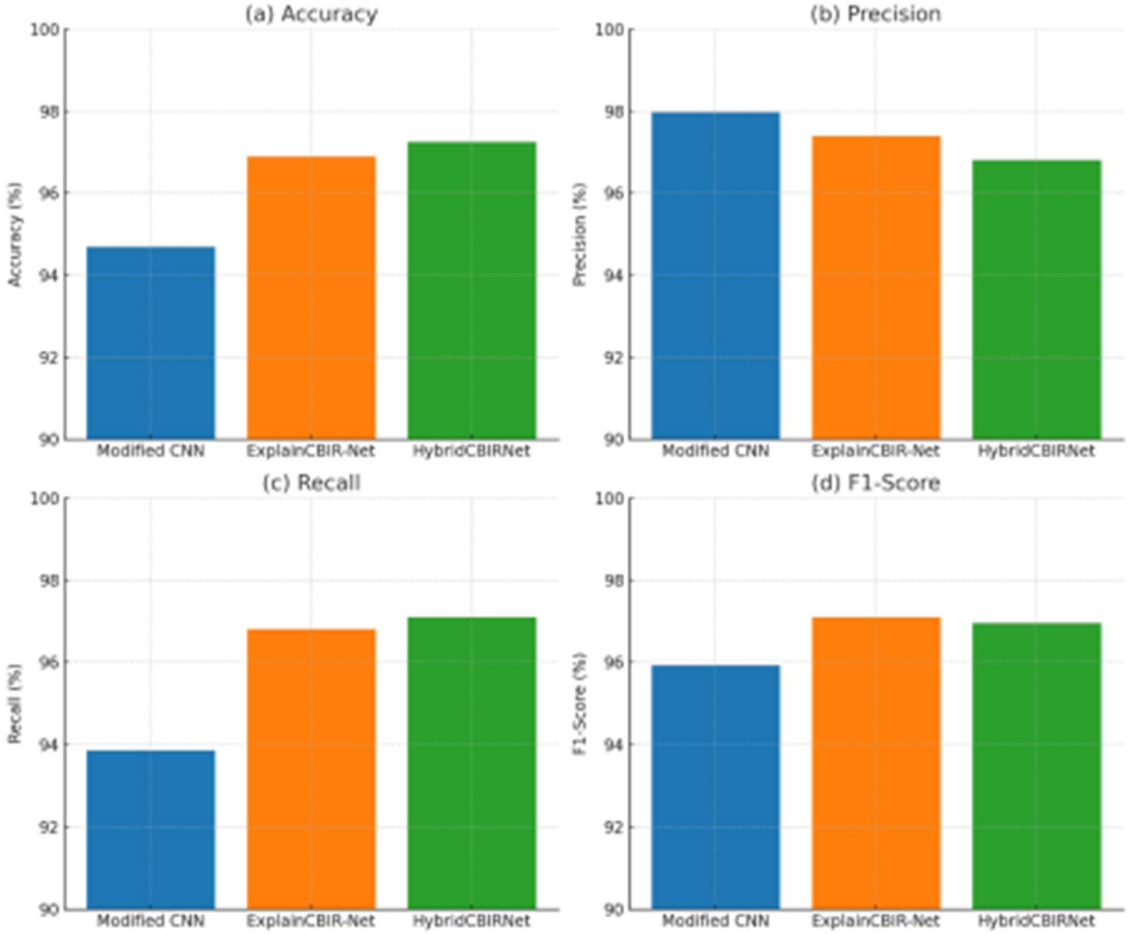

Figure 2: Comparative Visualizations of Performance Metrics for the Proposed CBIR Models—(a) Accuracy, (b) Precision, (c) Recall, and (d) F1-Score

Figure 2 visually compares the performance of three proposed CBIR

models, including Modified CNN, Explain BIR-Net, and HybridCBIRNet, on four primary metrics: accuracy, precision, recall, and F1 score. HybridCBIRNet has the highest accuracy and precision; all models are shown in subfigures (a) and (b), respectively. (c) shows how contextual and explainable modelling allows for improved recall; (d) illustrates the achieved balance in F1-score. Together, those charts settle the transition from the spatial-aware CNN baseline to hybrid CNN-Transformer systems with stronger retrieval performance and interpretability, supporting contextual learning and visual explanation for CBIR.

4.2 Observations

· The Modified CNN framework establishes a strong baseline, delivering high precision and recall while being simple and computationally efficient. However, it lacks transparency and explain ability, which limits its deployment in sensitive domains like healthcare or security.

· Explain how CBIR-Net enhances the interpretability of CBIR through Grad-CAM visualizations. It balances performance and transparency, demonstrating that explain ability need not compromise accuracy. The Grad-CAM module highlights the influential regions in query and retrieved images, enabling more informed user interaction.

· HybridCBIRNet surpasses the other models in overall performance by integrating spatial and contextual information. The CNN captures local features, while the Transformer models long range dependencies, and the fusion mechanism ensures the richness of the final embedding. This dual modelling approach significantly improves retrieval in semantically complex or visually similar image sets.

· All models showed stable learning behaviour across training epochs. Figures from the original studies (e.g., accuracy/loss curves) indicate rapid convergence and generalization. Visual results of top-5 image retrievals demonstrate the enhanced relevance in Explain BIR-Net and HybridCBIRNet outputs.

5. CONCLUSION AND FUTURE WORK

In this paper, we presented a comparative study of three progressively advanced, deep learning-based Content-Based Image Retrieval (CBIR) frameworks developed during separate phases of research—Modified CNN, ExplainCBIR-Net, and HybridCBIRNet. Notably, each framework was built to gradually overcome the weaknesses of its predecessor, focusing on enhancements in accuracy, contextual comprehension, and interpretability. The Modified CNN model provided a good baseline with high retrieval precision over spatial features. But it was opaque, a key imperative for high-stakes uses. This gap was not addressed until the Explain BIR-Net framework, which integrated Grad CAM-based visual explain ability into the CBIR pipeline, indicating that running the retrieval process would be more interpretable without sacrificing performance. (i) Finally, the HybridCBIRNet model performed the best by merging CNN and Transformer-based features using weighted feature fusion to provide spatial details and global semantic observation. Experimental results on the Mini ImageNet dataset show that HybridCBIRNet outperforms with the highest accuracy (97.25%) and mean average precision (97.00%), while Explain BIR-Net leads with a large gap. The outcomes validate the significance of contextual modeling and interpretability for developing robust and user trusted CBIR systems. Future work will investigate enhancements such as lightweight Transformer variants for faster inference and unsupervised training to relax data dependency, as well as deployment to real-world medical and surveillance use cases where accuracy and explanation are needed.

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

References

[1] Agrawal, S., Yadav, R., & Sinha, A. (2022). Deep CNN-Based Content-Based Image Retrieval for Lung Disease Diagnosis. *Journal of Medical Imaging and Health Informatics*, 12(3), 102–110.

[2] Wickstrøm, K., Hauge, I. J., & Halvorsen, P. (2023). Self-supervised learning in liver CT imaging for robust CBIR. *IEEE Transactions on Medical Imaging*, 42(1), 15–26.

[3] Lo, H. C., Lin, J., & Hsu, C. C. (2024). Transformer-Based Deep Retrieval for Abdominal CT Image Localization. *MICCAI Proceedings*, 136–144.

[4] Karthik, K., & Kamath, S. (2020). A Multi-View CNN-Based Deep Retrieval Framework for Medical Images. *Health Informatics Journal*, 26(4), 2612–2625.

[5] Desai, R., & Mehta, P. (2021). Enhancing CBIR Performance Using VGG16 and SVM. *Pattern Recognition Letters*, 145, 12–20.

[6] Zhang, Y., & Liu, W. (2023). Deep Learning for CBIR: A Comprehensive Survey. *ACM Computing Surveys*, 55(2), 1–40.

[7] Punithavathi, S., & Meenakshi, K. (2022). Secure Cloud-Based Deep Learning Framework for CBIR. *Journal of Cloud Computing*, 11(1), 53–66.

[8] Sikandar, M., & Prasad, K. (2023). Hybrid CBIR Using Deep and Handcrafted Features. *Multimedia Tools and Applications*, 82(9), 12345–12368.

[9] Rafiei, A., & Iosifidis, A. (2023). Class-Specific Variational Autoencoders for Content-Based Image Retrieval. *Neurocomputing*, 521, 112–126.

[10] Hu, X., & Bors, A. G. (2023). Modeling Spatial Relations in CBIR Using Co-Attention Mechanisms. *IEEE Transactions on Image Processing*, 32, 1987–1999.

[11] Rashad, M., & El-Horbaty, E. (2023). RbQE: A Query Expansion Method for Medical CBIR. *Egyptian Informatics Journal*, 24(1), 81–92.

[12] Xu, C., Wang, Z., & Li, T. (2022). Deep Image Retrieval for Indoor Localization Applications. *Sensors*, 22(5), 1956.

[13] Mohamed, E., Zhang, T., & Yang, Y. (2019). CNN-Based Image Retrieval Using Caltech-256 Dataset. *Procedia Computer Science*, 152, 501–508.

[14] Vo, T., & Nguyen, H. (2021). Deep Saliency-Guided Medical Image Retrieval. *Artificial Intelligence in Medicine*, 118, 102108.

[15] Yang, L., Zhang, J., & Wu, X. (2022). Deep Quality Modeling for Large-Scale CBIR. *Information Sciences*, 608, 1–15.