MGC-SIFT: A Multimodal Graph-Based Color SIFT Descriptor for Content-Based Image Retrieval

Trupti Babasaheb Ghatage[1*], Dattatraya Vishnu Kodavade[2]
[1]Department of Technology, Shivaji University, Kolhapur, Maharashtra, India
[2]DKTE Society's Textile and Engineering Institute, Ichalkaranji, Maharashtra, India
E-mail: yogitatrupti@gmail.com, dvkodavade@gmail.com
*Corresponding author

Keywords: content-based image retrieval, MGC-SIFT, Color-SIFT, graph neural networks, attention mechanism, proxy-based learning

Received: August 4, 2025

Content-Based Image Retrieval (CBIR) systems critically depend on discriminative yet efficient feature representations to retrieve relevant images from large-scale databases. However, many existing handcrafted and graph-based methods face limitations in scalability and in jointly modeling multimodal information such as color, texture, and spatial relationships. To address these challenges, this paper proposes a novel feature extraction framework termed Multi-modal Graph Color SIFT (MGC-SIFT). In the proposed approach, color-augmented SIFT descriptors extracted in the YCbCr color space are organized as a graph of local keypoints, over which Graph Neural Networks (GNNs) are applied to model inter-keypoint spatial relationships. An attention mechanism is incorporated to emphasize discriminative keypoint regions, while proxy-based learning is employed to improve representation compactness and retrieval efficiency.

The effectiveness of MGC-SIFT is evaluated on four benchmark datasets—Corel-1K, COIL-20, Oxford-102 Flowers, and UC-Merced Land Use—covering natural scenes, controlled object images, fine-grained categories, and aerial imagery. Experimental evaluation using standard CBIR metrics, including mean Average Precision (mAP), Precision@k, Recall@k, F1-score@k, and Accuracy@k, demonstrates that the proposed method achieves consistent and competitive retrieval performance across heterogeneous datasets, including robustness under image degradation conditions. Ablation studies further confirm the complementary contributions of color augmentation, graph-based modeling, attention mechanisms, and proxy-based learning. In addition, runtime and memory analysis indicate that proxy-based learning significantly reduces retrieval latency, supporting scalable image retrieval.

Overall, the proposed MGC-SIFT framework provides a robust and interpretable multimodal representation for CBIR by explicitly modeling joint color–spatial dependencies at the local keypoint level, offering a practical solution for scalable image retrieval in real-world applications.

Povzetek: Članek predstavlja novo metodo za učinkovitejše in natančnejše iskanje slik v velikih podatkovnih zbirkah z uporabo naprednih tehnik strojnega učenja.

1 Introduction

The explosive growth of digital imagery across diverse domains, including medicine, defense, remote sensing, surveillance, and e-commerce, has necessitated the development of robust and scalable image-retrieval systems [1], [2]. Unlike traditional text-based retrieval techniques, Content-Based Image Retrieval (CBIR) exploits the visual content of images by extracting and analyzing low-level features such as color, texture, and shape to identify semantically similar images [3]. This feature-centric paradigm enables more effective image organization and retrieval, particularly in large-scale multimedia databases.

Texture-based features, such as Gray-Level Co-occurrence Matrices (GLCM), wavelet transforms, Gabor filters, and Local Binary Patterns (LBP), are fundamental in pattern recognition and texture classification. These methods offer strengths, such as translation and rotation invariance and computational simplicity [4]. For instance, the recently proposed multi-scale shape index-based LBP enhances classification accuracy while maintaining robustness to geometric transformations [5].

Shape-based descriptors, including moment invariants and contour-based techniques, contribute complementary structural information but are sensitive to occlusion and deformation. Prior studies have shown that integrating shape-related cues with texture features can improve robustness and classification accuracy, highlighting the complementary nature of structural and texture representations [6].

Numerous studies have demonstrated that combining multiple visual cues—particularly color and texture—results in more robust and discriminative representations for CBIR [1]. The Scale-Invariant Feature Transform

(SIFT) remains one of the most widely used local descriptors due to its robustness to scale, rotation, and illumination changes [7]. Subsequent extensions have enhanced SIFT by incorporating complementary information through color-augmented descriptors such as RGB-SIFT and OpponentSIFT, which compute SIFT features in different color spaces to improve discriminability 8, as well as graph-based SIFT variants that explicitly model spatial relationships among local keypoints using graph representations and neural networks [9], [10], [12].

Despite these advances, many fusion-based methods fail to capture higher-order dependencies among local keypoints, which are essential for preserving spatial structure. Moreover, such approaches are often computationally expensive and unsuitable for real-time applications. Dimensionality-reduction techniques, including PCA, spectral hashing, and t-SNE, reduce feature dimensionality but may discard important structural cues required for semantic similarity [11]. In addition, traditional clustering algorithms such as k-means struggle with high-dimensional and non-linearly separable CBIR data distributions [12].

Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have significantly advanced image feature extraction by enabling the learning of high-level semantic representations from large labeled datasets [13], [14]. Recent deep learning–based CBIR approaches further demonstrate strong retrieval performance by learning compact and discriminative image embeddings through metric learning and proxy-based representations [15]. Metric learning strategies, such as triplet-loss and contrastive-loss formulations, enhance retrieval accuracy by modeling fine-grained similarities among images [16], [17]. However, CNN-based approaches remain computationally expensive and data-intensive, limiting their applicability in real-time and resource-constrained CBIR systems [18].

To address computational scalability, proxy-based learning has emerged as an effective strategy that replaces exhaustive pairwise comparisons with surrogate class centers, significantly reducing retrieval complexity [17], [18]. In parallel, graph-based modeling represents image keypoints as nodes in a graph, enabling the exploitation of spatial relationships among local features. Graph Neural Networks (GNNs) are particularly effective in this context, as they model both spatial and contextual dependencies between keypoints, leading to richer feature representations than conventional Euclidean similarity measures [9], [12]. Furthermore, attention mechanisms have been incorporated into retrieval pipelines to emphasize discriminative regions and suppress irrelevant background information, thereby improving retrieval accuracy [2].

Despite these developments, several limitations persist in existing CBIR systems. Many descriptors fail to effectively integrate color and texture information at the local keypoint level, while others neglect spatial relationships among features. Classical approaches struggle to scale to large, high-dimensional datasets, and many methods lack adaptive attention mechanisms to focus on semantically relevant regions. Consequently, there is a need for a unified framework that jointly integrates multimodal feature representation, spatial modeling, and computational efficiency.

1.1 Research design

This study investigates whether integrating color-enhanced SIFT descriptors with graph-based contextual learning can improve or maintain competitive retrieval accuracy across diverse image datasets. In particular, it examines the impact of modeling inter-keypoint relationships using Graph Neural Networks (GNNs) and explores whether attention mechanisms combined with proxy-based learning can enhance retrieval effectiveness, representation compactness, and scalability in clustering-based CBIR systems.

Based on these considerations, this work hypothesizes that augmenting SIFT descriptors with color information leads to improved retrieval performance in terms of mAP, Precision@k, Recall@k, and F1-score@k, compared to grayscale SIFT-based methods. It further posits that graph-based modeling of local feature relationships enhances retrieval accuracy over non-graph color-SIFT representations, particularly for complex image domains. Moreover, attention-guided proxy learning is expected to improve feature compactness and retrieval consistency while significantly reducing computational cost, thereby supporting scalable image retrieval.

To validate these hypotheses, the proposed MGC-SIFT algorithm is developed as a unified multimodal representation that integrates color-augmented SIFT descriptors in the YCbCr space with graph-based keypoint modeling, attention-guided feature refinement, and proxy-based learning. The effectiveness of the proposed framework is evaluated using standard CBIR metrics and compared against established baselines, including standard SIFT, SIFT-RGB, which extends SIFT by computing descriptors on color channels without explicit spatial modeling, SIFT-GNN, which models spatial relationships among SIFT keypoints using graph-based learning, and representative deep CNN-based descriptors. Extensive experiments conducted on four benchmark datasets—Corel-1K, COIL-20, Oxford-102 Flowers, and UC-Merced Land Use—covering natural scenes, controlled object images, fine-grained categories, and aerial imagery demonstrate that MGC-SIFT achieves competitive and consistent retrieval performance across heterogeneous domains, while maintaining robustness, scalability, and computational efficiency.

As summarized in Table 1, existing CBIR approaches typically emphasize either color enhancement or spatial

Table 1 : Comparative summary of representative CBIR methods and their limitations

Method	Feature Representation	Explicit Spatial Modeling	Explicit Color Modeling	Learning Strategy	Typical Datasets Reported	Performance Reporting	Key Limitations
SIFT [7]	Local invariant keypoint descriptors	No	No	Handcrafted	Corel, Oxford	mAP / Precision	Ignores color information and spatial context between keypoints
Color-SIFT (e.g., OpponentSIFT / RGB-SIFT) [8]	Color-augmented SIFT descriptors computed in RGB / opponent color space	No	Yes	Handcrafted	Corel, Oxford	mAP / Precision	Does not model inter-keypoint spatial relationships; increased descriptor dimensionality
SIMIR [10]	Mean SIFT with color-based clustering	No	Yes	Clustering-based	Corel	Precision / Recall	Limited spatial awareness; relies on global aggregation
Graph-based SIFT Retrieval [9], [12]	SIFT descriptors with graph modeling	Yes	No	Graph Neural Network	COIL-20	mAP	Does not explicitly incorporate color cues; scalability concerns
CNN-based CBIR [13][14]	Deep CNN feature embeddings	Implicit	Implicit	Supervised deep learning	ImageNet, Oxford	mAP	Computationally expensive; requires large labeled datasets
Deep Metric Learning [15][16][17]	CNN features with metric loss	Implicit	Implicit	Triplet / Proxy learning	Remote sensing datasets	mAP	Data-hungry; limited interpretability
MGC-SIFT (Proposed)	Color-augmented SIFT + graph modeling	Yes	Yes	Attention-guided proxy learning	Corel-1K, COIL-20, Oxford-102, UC-Merced	mAP, Precision@k, Recall@k	**Novelty:** Explicitly unifies color-aware SIFT and graph-based spatial context via attention-guided proxy learning.

modeling, but rarely integrate both explicitly at the local keypoint level. Handcrafted descriptors lack contextual awareness, while deep learning–based approaches incur high computational cost and require large labeled datasets. In contrast, the proposed MGC-SIFT framework explicitly models joint color–spatial dependencies using graph neural networks and attention-guided proxy learning, enabling a balanced trade-off between retrieval accuracy, interpretability, and scalability.

1.2 Key contributions

- We introduce MGC-SIFT, a feature descriptor that combines color, texture, and spatial relationships
- based on SIFT, the YCbCr color space, and graph modeling.
- Graph Neural Networks capture keypoint-to-keypoint relationships for robust feature refinement.
- An attention mechanism improves the discriminative focus on important regions.
- Proxy-based learning improves the scalability of large-scale retrieval operations.
- Experiments validated the model's effectiveness across multiple benchmark datasets, demonstrating competitive retrieval accuracy and scalability.

The remainder of this paper is organized as follows. Section 2 describes the proposed MGC-SIFT approach. Section 3 presents the experimental setup and results. Finally, Section 4 concludes the paper and suggests future work.

2 Method

This study introduces a novel feature extraction framework, Multimodal Graph-Enhanced Color-SIFT (MGC-SIFT), designed to enhance image retrieval and

object detection tasks by leveraging the strengths of texture, color, and graph-based feature interactions. Feature extraction techniques are indispensable for image retrieval and object detection in computer vision. A variety of classical techniques for feature extraction methodologies exist, including SIFT, given their efficacy in determining local keypoints that remain invariant to scale, rotation, and illumination.

However, the SIFT feature was designed to work only on grey scale images. Therefore, SIFT is not the best method for object identification and classification in an application scenario that highly depends on color features. Several papers are reviewed here to discuss the different approaches that have been proposed to address the inability of SIFT to consider color for feature extraction. In the literature survey, many researchers explored various ways of including color information in traditional approaches, such as SIFT. It is proven through color space transformation, including YCbCr or HSV, that the separation between chrominance and luminance improves the representation of the color features of images. For instance, Adnan et al. [19] implemented a highly effective YCbCr color space that retained significant color information of an image while maintaining proper distinction between its brightness components. This establishes the background necessary to formulate the proposed method, MGC-SIFT, which embeds the color channel values into SIFT descriptors for the simultaneous representation of both local texture and color features.

2.1 Graph neural networks for modeling keypoint interactions

The most recent development in feature extraction is the use of graph neural networks to model the relationships between parts of an image. Traditional methods such as SIFT treat keypoints as independent entities, which can be very limited when higher-order relationships between keypoints contribute to object recognition. Yu et al. [20] proposed GNNs for learning inter-feature dependencies, which could increase the precision of feature extraction by shifting the attention to interactions between different image regions. This concept was adapted to the MGC-SIFT method, in which keypoints and their respective color features are represented as nodes in a graph. Hence, GNNs can be used to model both spatial and color-based relationships. In doing so, MGC-SIFT enhances the extracted features through graph-based reasoning, thereby rendering the method more robust to complex variations in object structure and appearance.

2.2 Attention mechanisms for feature refinement

In recent years, attention mechanisms have gradually gained momentum because of their efficiency in filtering out the most unimportant parts of an image while paying greater attention to other relevant or information parts. Inspired by [21], MGC-SIFT proposes an attention mechanism over the extracted features for enhanced object retrieval, where color plays a significant role. It helps to weigh the important keypoints and color patches and filter out background noise or irrelevant regions.

2.3 Multiscale feature extraction

The incorporation of multiscale analysis into the feature extraction can capture both fine-grained details and large structural features. Zhao et al. [22] have shown that a multiscale feature extraction approach makes the retrieval system robust since small and large objects are effectively detected. In MGC-SIFT, this is achieved by extracting features from images at multiple resolutions to represent objects with different sizes and color distributions. MGC-SIFT can handle diverse visual conditions owing to its multiscale approach combined with attention.

2.4 Proxy-based learning for efficiency

Efficient feature extraction methods are crucial when dealing with large-scale datasets, particularly for real-time applications, such as image retrieval. According to Cai et al. [23], proxy-based learning decreases computational complexity by clustering features into proxy points for faster and more efficient matching. It exploits the proxy representation advantage of MGC-SIFT in mapping similar keypoints and color features, which considerably reduces the image-matching time without trading off high accuracy. Hence, the proposed approach is powerful in terms of feature representation and scalable for large-scale datasets. The next subsection discusses the different steps to be followed for MGC-SIFT implementation. Figure 1 shows the flowchart of the proposed MGC-SIFT algorithm.

2.5 The proposed MGC-SIFT algorithm

The proposed MGC-SIFT algorithm follows a structured pipeline that integrates color-aware local descriptors, graph-based spatial modeling, attention-guided feature refinement, and proxy-based learning for scalable content-based image retrieval. An overview of the complete workflow is illustrated in Figure 1.

Given an input image, the method first converts the image to grayscale for robust SIFT keypoint detection and descriptor extraction, ensuring invariance to scale, rotation, and illumination changes. In parallel, the color information is extracted by transforming the input image into the YCbCr color space, which separates luminance and chrominance components.

For each detected keypoint, local chrominance values are fused with the corresponding SIFT descriptor to form color-augmented SIFT descriptors, enabling the simultaneous representation of local texture and color information. These descriptors serve as the nodes of a keypoint graph, where edges are established based on spatial proximity or k-nearest-neighbor relationships between keypoints.

To capture higher-order spatial and contextual dependencies, a Graph Neural Network (GNN) is applied to the constructed graph, refining the node features through neighborhood aggregation. An attention mechanism is subsequently employed over the graph nodes to assign higher importance to discriminative

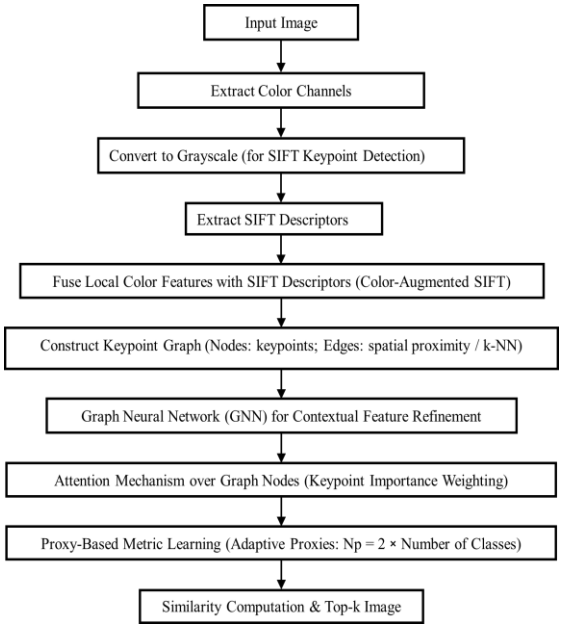

Figure 1: Overview of the proposed MGC-SIFT algorithm

keypoints while suppressing less informative or noisy regions.

To ensure scalability for large image databases, the refined descriptors are encoded using proxy-based metric learning, which maps features to adaptive proxy representations, thereby reducing retrieval complexity. Finally, similarity computation is performed between the query representation and database images, and the top-k most relevant images are retrieved based on similarity ranking.

For completeness, detailed mathematical formulations of SIFT extraction, color transformation, graph propagation, and attention weighting are provided in the supplementary materials.

Algorithm 1: MGC-SIFT Algorithm
Input: Query image
Output: Top-k retrieved images

1. Read input image.
2. Convert the image to grayscale for SIFT keypoint detection.
   Time Complexity: $O(N)$
3. Construct SIFT scale space and extract keypoints and descriptors.
   Time Complexity: $O(N \cdot S)$, where S is the number of scales.
4. Convert the input image from RGB to YCbCr color space.
   Time Complexity: $O(N)$
5. Augment each SIFT descriptor with local chrominance values (Cb, Cr) to obtain color-augmented SIFT descriptors.
   Time Complexity: $O(K)$
6. Construct a keypoint graph based on spatial proximity or k-nearest-neighbor relationships.
   Time Complexity: $O(K^2)$ (worst case), reduced to $O(K \cdot k)$ in practice.
7. Apply Graph Neural Network layers for contextual feature refinement.
   Time Complexity: $O(L \cdot |E| \cdot d)$
8. Apply an attention mechanism to weight discriminative keypoint features.
   Time Complexity: $O(K \cdot d)$
9. Encode refined descriptors using proxy-based learning.
   Time Complexity: $O(K \cdot P)$, where $P \ll K$.
10. Compute similarity scores and retrieve the top-k most relevant images.
    Time Complexity: $O(M \cdot d)$ for linear scan (reduced using indexing structures).

3 Results and discussion

3.1 Experimental setup

The proposed MGC-SIFT model is trained using the predefined training splits of each dataset to learn the parameters of the graph convolutional network (GCN), attention mechanism, and proxy representations. During evaluation, each image from the test set is treated as a query, and retrieval is performed against the remaining test images, following a standard query–database evaluation protocol.

Table 2 presents a summary of the benchmark datasets used for evaluation, covering diverse CBIR scenarios including natural scenes, controlled object images, fine-grained visual categories, and aerial imagery. Table 3

Table 2: Description of benchmark datasets used for experimental evaluation

Dataset	No. of Classes	Images per Class	Total Images	Image Type	Domain
Corel-1K	10	100	1000	Natural scenes	Scene-level CBIR
COIL-20	20	72	1440	Object images	Controlled object CBIR
Oxford-102 Flowers	102	~40	8189	Flower images	Fine-grained, color-rich
UC-Merced Land Use	21	100	2100	Aerial images	Remote sensing CBIR

summarizes the experimental setup and parameter configuration adopted across all datasets.

For comparative evaluation, several baseline methods are implemented under the same experimental protocol.

Table 3: Experimental setup and parameter configuration used across all datasets

Component	Parameter	Value / Description
SIFT Extraction	Keypoint detector	Difference of Gaussian (DoG)
	Descriptor dimension	128
	Max keypoints per image	2048
Color Augmentation	Color space (MGC-SIFT)	YCbCr
	Color space (SIFT-RGB baseline)	RGB
	Color bins per channel	16
	Color descriptor dimension	48
Graph Construction	Graph type	k-nearest neighbor (k-NN) graph
	Number of neighbors (k)	10
	Edge weight	Euclidean distance (used during graph construction)
GNN Architecture	GNN model	Graph Convolutional Network (GCN)
	Number of GNN layers	2
	Hidden dimension	128 (64 used only in ablation study)
Attention Module	Attention mechanism	Learnable graph-based attention weighting
Proxy Learning	Number of proxies(Np)	Adaptive, proportional to number of classes (Np = 2 × C, set at runtime)
	Proxy update strategy	Jointly learned during training
Training Setup	Optimizer	Adam
	Learning rate	0.001
	Batch size	4
	Number of epochs	20
Dimensionality Reduction	Output descriptor size	128
Evaluation Protocol	Query strategy	Each test image used as query
	Corel-1K	900 training / 100 testing images
	COIL-20	80% training / 20% testing
	Oxford-102	Official VGG train/validation/test split
	UC-Merced	80 training and 20 testing images per class (21 classes)
	Similarity metric	Canberra distance (primary); cosine and Euclidean used for comparison
	Retrieval depth (k)	Top-10
Hardware	Platform	Intel i7 CPU, 16 GB RAM
	GPU	Not used

Standard SIFT employs grayscale SIFT keypoint detection and descriptor extraction, with local descriptors aggregated using mean pooling to form a global image representation. SIFT-RGB augments this representation by concatenating a global RGB color histogram with the pooled SIFT descriptor, thereby incorporating color information at the feature level without explicit spatial or graph-based modeling. SIFT-GNN captures spatial relationships among local SIFT keypoints by representing them as nodes in a graph and applying graph-based learning to model inter-keypoint dependencies, while not explicitly incorporating color information. Note that the SIFT-RGB baseline evaluated in this work follows a feature-level fusion strategy, which is distinct from some classical color-SIFT formulations summarized in Table 1.

The experimental setup was kept consistent across all datasets to ensure fair comparison and reproducibility. Parameter values were selected based on preliminary validation experiments and established practices in the CBIR literature, without performing dataset-specific parameter tuning. In particular, the same graph construction strategy, embedding dimensionality, and evaluation protocol were applied uniformly across all graph-based methods. This unified configuration ensures that the observed performance trends reflect the intrinsic strengths and limitations of the proposed MGC-SIFT descriptor, rather than artifacts of parameter optimization, across diverse CBIR scenarios.

3.2 Ablation studies

To systematically analyze the contribution of individual components in the proposed MGC-SIFT framework, ablation studies were conducted on four benchmark datasets by selectively disabling or modifying key modules, including graph modeling (GNN), attention weighting, proxy-based learning, color augmentation, and embedding dimensionality. Performance was evaluated using mAP, Precision@10, Recall@10, F1-score@10, and Accuracy@10.

3.2.1 Corel-1K dataset

Table 4 reports the ablation results on the Corel-1K dataset. The full MGC-SIFT configuration achieves the highest overall performance across most evaluation metrics. Removing graph modeling, attention, proxy learning, or color augmentation leads to marginal but consistent reductions in retrieval accuracy, indicating that each component contributes complementary information to the overall representation. These results highlight the importance of jointly modeling color cues, local descriptors, and spatial context when dealing with heterogeneous natural scene images.

3.2.2 COIL-20 dataset

Table 5 presents the ablation results on the COIL-20 dataset. In this controlled object recognition scenario, performance differences among the full and reduced variants are relatively small. Some reduced-complexity configurations exhibit performance comparable to the full model, suggesting that compact representations are sufficient for datasets with limited intra-class variation and well-aligned object structures. Importantly, the full MGC-SIFT model remains competitive across all metrics, demonstrating that the framework does not rely on excessive model complexity to achieve stable retrieval performance.

Table 4: Ablation study results of MGC-SIFT on Corel-1K Dataset

Variant	mAP	P@10	R@10	F1@10	Accuracy
Full MGC-SIFT	0.3512	0.2756	0.317	0.2846	0.317
No-GNN	0.3482	0.2744	0.308	0.278	0.308
No-Attention	0.349	0.2872	0.316	0.2886	0.316
No-Proxy	0.3484	0.2761	0.312	0.2802	0.312
No-Color	0.3488	0.2727	0.31	0.279	0.31
Hidden-64	0.3464	0.2722	0.308	0.2759	0.308

Table 5: Ablation study results of MGC-SIFT on COIL-20 Dataset

Variant	mAP	P@10	R@10	F1@10	Accuracy
Full MGC-SIFT	0.5586	0.6575	0.6147	0.6059	0.6147
No-GNN	0.5651	0.6715	0.6287	0.6181	0.6287
No-Attention	0.5617	0.6645	0.6207	0.6105	0.6207
No-Proxy	0.5704	0.6593	0.6273	0.6167	0.6273
No-Color	0.5637	0.6551	0.619	0.6095	0.619
Hidden-64	0.5645	0.661	0.6213	0.6126	0.6213

3.2.3 Oxford-102 flowers dataset

The ablation results on the Oxford-102 Flowers dataset are summarized in Table 6. Across all variants, the full MGC-SIFT configuration consistently achieves competitive performance, while the removal of individual components results in modest but noticeable performance degradation. These findings indicate that color-aware graph modeling and proxy-based learning provide complementary benefits in fine-grained retrieval scenarios, where subtle inter-class differences and color variations play a crucial role.

Table 6: Ablation study results of MGC-SIFT on Oxford-102 flowers dataset

Variant	mAP	P@10	R@10	F1@10	Accuracy
Full MGC-SIFT	0.0343	0.0614	0.0462	0.0395	0.0575
No-GNN	0.034	0.0584	0.0464	0.0396	0.0578
No-Attention	0.0338	0.0585	0.0452	0.0387	0.0559
No-Proxy	0.0342	0.0557	0.0459	0.0385	0.0572
No-Color	0.0339	0.056	0.0468	0.04	0.0578
Hidden-64	0.034	0.0581	0.0466	0.0399	0.0577

3.2.4 UC-merced land use dataset

Table 7 reports the ablation results on the UC-Merced dataset. The results show marginal performance variations across different configurations, with reduced embedding dimensionality (Hidden-64) yielding performance comparable to the default setting. This behavior suggests that more compact representations can be effective for aerial scene retrieval, where global spatial layouts are more dominant than fine-grained local details. Nonetheless, the full MGC-SIFT configuration maintains stable performance, confirming the adaptability of the proposed framework across varying scene complexities.

3.2.5 Overall ablation analysis

Overall, the ablation studies validate the design choices of the proposed MGC-SIFT framework. While the magnitude of performance variation differs across datasets, the results consistently demonstrate that MGC-SIFT benefits from the synergistic integration of color augmentation, graph-based contextual modeling, attention mechanisms, and proxy learning. Importantly, the framework exhibits robustness to component removal, indicating that it does not rely on a single dominant module but instead achieves balanced performance through modular and complementary feature integration. This property makes MGC-SIFT adaptable to diverse CBIR scenarios with varying levels of visual complexity.

3.3 Feature and model comparison

This subsection presents a comparative evaluation of handcrafted, deep, and hybrid feature representations across four benchmark datasets. For handcrafted descriptors and the proposed MGC-SIFT framework, Canberra distance is employed due to its suitability for sparse and hybrid feature representations. For deep and fused representations, cosine similarity is used, reflecting the normalized nature of learned embeddings. Late fusion between VGG16 and MGC-SIFT features is performed using equal weighting ($\alpha = 0.5$).

3.3.1 Comparison across datasets

Table 8 summarizes the retrieval performance in terms of mAP and Precision@10 across all datasets. The results indicate that no single feature representation dominates across all scenarios. Classical SIFT-based descriptors perform reasonably on structured datasets, while color-aware extensions improve performance in color-rich domains. Deep CNN features demonstrate strong performance on datasets with clear semantic regularities, particularly COIL-20 and Oxford-102 Flowers.

Across all datasets, the proposed MGC-SIFT descriptor achieves stable and competitive performance, despite relying on limited supervision and compact representations. While its absolute performance may be lower than fully supervised deep models in some cases, MGC-SIFT consistently maintains a favorable balance between retrieval accuracy, interpretability, and computational efficiency.

Notably, late fusion of VGG16 and MGC-SIFT features yields the strongest overall performance across all datasets, confirming that MGC-SIFT captures complementary information that is not fully represented in deep embeddings alone. This observation highlights the effectiveness of combining contextual graph-based descriptors with semantic deep features.

Table 7: Ablation study results of MGC-SIFT on UC-Merced dataset

Variant	mAP	P@10	R@10	F1@10	Accuracy
Full MGC-SIFT	0.1088	0.1771	0.1112	0.0775	0.1112
No-GNN	0.1093	0.1826	0.1088	0.0766	0.1088
No-Attention	0.1092	0.1692	0.1095	0.0735	0.1095
No-Proxy	0.1097	0.1573	0.11	0.0759	0.11
No-Color	0.1071	0.1652	0.101	0.0681	0.101
Hidden-64	0.1097	0.1674	0.1124	0.0819	0.1124

Table 8: Comparison of handcrafted, deep, and hybrid feature representations

Dataset	**Corel-1K**		**Oxford-102 Flowers**		**COIL-20**		**UC_Merced**	
Method	mAP	P@10	mAP	P@10	mAP	P@10	mAP	P@10
SIFT	0.3294	0.3316	0.0679	0.1468	0.416	0.5179	0.2362	0.3449
RGB	0.3872	0.4387	0.1637	0.2821	0.5818	0.6384	0.2314	0.3289
SIFT-RGB	0.434	0.4423	0.1104	0.2233	0.5898	0.6825	0.2896	0.4199
SIFT-GNN	0.2161	0.2177	0.0252	0.0354	0.3443	0.4012	0.1239	0.1614
VGG16	0.5604	0.6464	0.0567	0.5661	0.7059	0.7827	0.1945	0.4593
VGG16 + MGC-SIFT	0.6575	0.602	0.176	0.4525	0.7367	0.8147	0.3282	0.455
MGC-SIFT (Proposed)	0.3512	0.2756	0.0343	0.0614	0.5586	0.6575	0.1088	0.1771

Table 9: Mean Precision@k on Corel-1K Dataset

k	**SIFT**	**SIFT-GNN**	**SIFT-RGB**	**MGC-SIFT (Proposed)**	**VGG16**	**VGG16+MGC-SIFT**
1	0.5654	0.2555	0.686	0.3989	0.7443	0.8884
5	0.4055	0.2398	0.531	0.3632	0.6594	0.7548
10	0.3316	0.2177	0.4423	0.2756	0.6464	0.602
20	0.2472	0.2021	0.3446	0.2009	0.4757	0.3637

Table 10: Mean Recall@k on Corel-1K Dataset

k	**SIFT**	**SIFT-GNN**	**SIFT-RGB**	**MGC-SIFT (Proposed)**	**VGG16**	**VGG16 + MGC-SIFT**
1	0.54	0.23	0.66	0.47	0.7	0.87
5	0.374	0.206	0.51	0.398	0.632	0.782
10	0.297	0.183	0.398	0.317	0.523	0.626
20	0.2185	0.169	0.2605	0.206	0.3135	0.3455

Table 11: Mean Precision@k on COIL-20 Dataset

k	**SIFT**	**SIFT-GNN**	**SIFT-RGB**	**MGC-SIFT (Proposed)**	**VGG16**	**VGG16+MGC-SIFT**
1	0.8461	0.6888	0.9582	0.9401	0.964	0.9607
5	0.6527	0.5079	0.8366	0.7634	0.882	0.891
10	0.5179	0.4012	0.6825	0.6575	0.7827	0.8147
20	0.3901	0.3025	0.495	0.473	0.6086	0.5669

Table 12: Mean Recall@k on COIL-20 Dataset

	SIFT	SIFT-GNN	SIFT-RGB	MGC-SIFT (Proposed)	VGG16	VGG16+MGC-SIFT
1	0.8267	0.6833	0.9533	0.9233	0.96	0.96
5	0.6167	0.498	0.8273	0.74	0.874	0.8927
10	0.4687	0.3897	0.6703	0.6147	0.771	0.8183
20	0.327	0.2887	0.4468	0.4173	0.5382	0.5568

Table 13: Mean Precision@k on Oxford-102 Flowers

k	SIFT	SIFT-GNN	SIFT-RGB	MGC-SIFT (Proposed)	VGG16	VGG16+MGC-SIFT
1	0.2405	0.0487	0.3685	0.1361	0.6517	0.6483
5	0.1774	0.0392	0.2693	0.0777	0.6227	0.5252
10	0.1468	0.0354	0.2233	0.0614	0.5661	0.4525
20	0.1188	0.0312	0.1825	0.0477	0.5079	0.3672

Table 14: Mean Recall@k on Oxford-102 Flowers

k	SIFT	SIFT-GNN	SIFT-RGB	MGC-SIFT (Proposed)	VGG16	VGG16+MGC-SIFT
1	0.2315	0.0446	0.3668	0.0765	0.15	0.6079
5	0.166	0.0372	0.2663	0.052	0.1088	0.4805
10	0.1364	0.0334	0.2198	0.0462	0.0954	0.4011
20	0.1095	0.0294	0.1773	0.0398	0.0802	0.315

Table 15: Mean Precision@k on UC_Merced Dataset

k	SIFT	SIFT-GNN	SIFT-RGB	MGC-SIFT (Proposed)	VGG16	VGG16+MGC-SIFT
1	0.4675	0.2038	0.6427	0.182	0.4032	0.7221
5	0.4032	0.1756	0.4843	0.2005	0.4578	0.5524
10	0.3449	0.1614	0.4199	0.1771	0.4593	0.455
20	0.2811	0.1369	0.329	0.1736	0.3974	0.3395

Table 16: Mean Recall@k on UC_Merced Dataset

k	SIFT	SIFT-GNN	SIFT-RGB	MGC-SIFT (Proposed)	VGG16	VGG16+MGC-SIFT
1	0.4262	0.1933	0.6238	0.1667	0.2214	0.7048
5	0.3605	0.167	0.4567	0.1257	0.1805	0.5471
10	0.2933	0.1513	0.3762	0.1112	0.1726	0.4521
20	0.2386	0.1264	0.2854	0.0958	0.146	0.3325

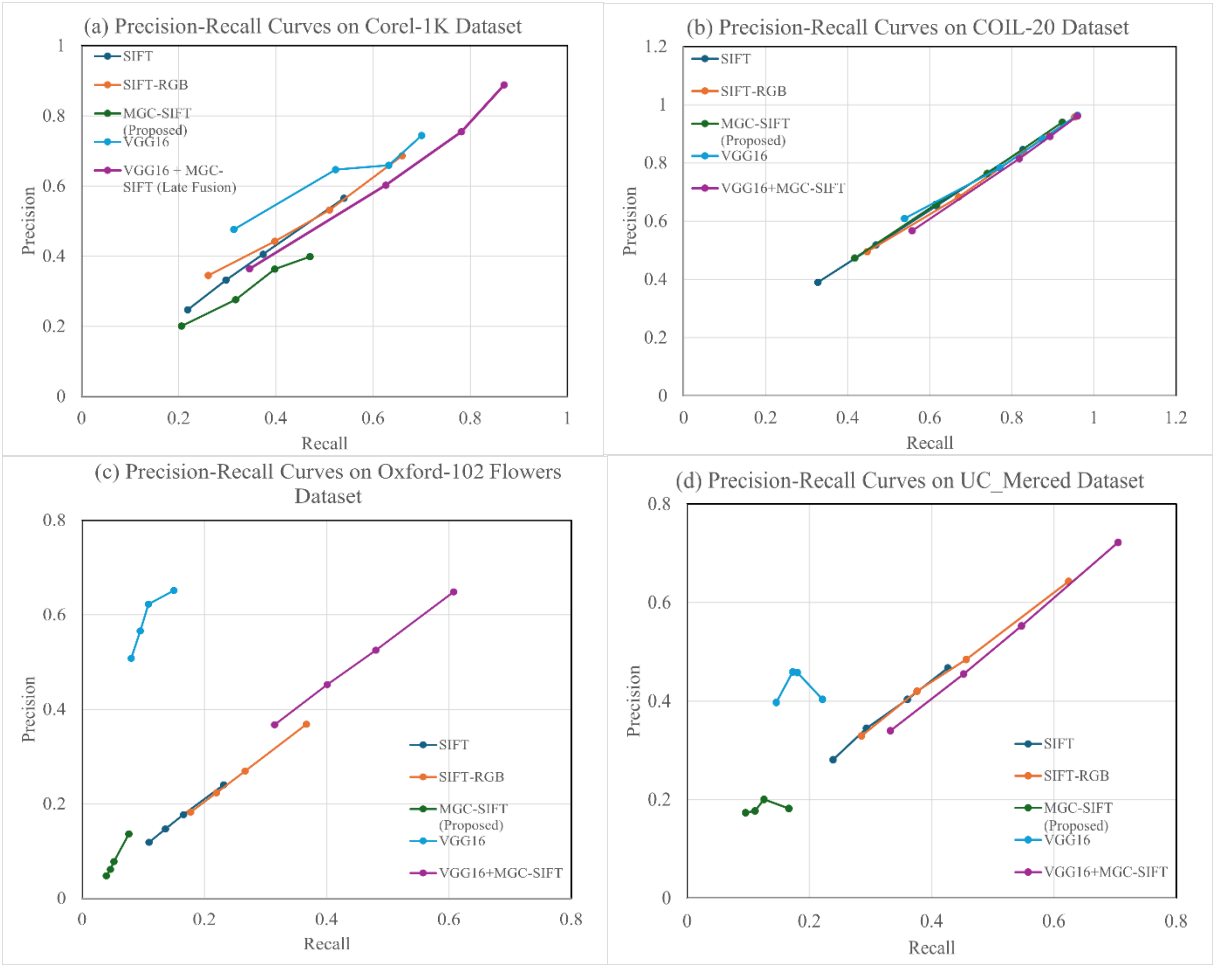

Figure 2: Precision-recall curves on benchmark datasets

3.3.2 Retrieval depth analysis on Corel-1K and COIL-20

Tables 9 and 10 report the mean Precision@k and Recall@k on the Corel-1K dataset. The results show that MGC-SIFT exhibits a balanced precision–recall trade-off across varying retrieval depths. While deep models achieve high precision at small k values, their performance degrades more sharply at larger depths. In contrast, MGC-SIFT demonstrates smoother degradation patterns, reflecting its robustness to increasing retrieval depth.

A similar trend is observed on the COIL-20 dataset (Tables 11 and 12), where compact and hybrid representations perform competitively across all k values. The effectiveness of late fusion further confirms the complementary nature of handcrafted contextual descriptors and deep semantic features in controlled object retrieval scenarios.

3.3.3 Fine-grained and aerial retrieval performance

Results on the Oxford-102 Flowers dataset (Tables 13 and 14) reveal that deep CNN features are particularly effective for fine-grained classification tasks, owing to their strong semantic learning capability. However, MGC-SIFT maintains consistent retrieval performance without requiring extensive labeled data, demonstrating its applicability in scenarios where large-scale supervised training is impractical. Fusion results again indicate complementary strengths between the two representations.

For the UC-Merced dataset (Tables 15 and 16), which involves complex aerial scenes, MGC-SIFT provides stable performance across all retrieval depths, while deep features show higher precision at shallow depths. The fusion model achieves the best overall balance, underscoring the benefit of integrating global semantic cues with graph-based contextual descriptors.

3.3.4 Precision–recall curve analysis

The precision–recall curves shown in Figures 2(a)–(d) further corroborate the quantitative results. MGC-SIFT consistently demonstrates smoother precision–recall trade-offs compared to purely handcrafted descriptors, which exhibit rapid precision decay, and deep models, which often favor precision at shallow retrieval depths. This balanced behavior reflects the ability of MGC-SIFT to preserve both local discriminative information and global contextual structure, making it well suited for real-world CBIR applications with varying retrieval depth requirements.

Table 17: Runtime and memory consumption on Corel-1K Dataset

Method	Avg. Time per Query (µs)	Memory Usage (MB)
MGC-SIFT (Proposed)	10,581.96	56.25
SIFT	22.73	0.15
RGB	154.5	0.59
SIFT-RGB	203.22	0.2
SIFT-GNN	156.34	0.15
VGG16	18,545.49	28.71

3.4 Runtime and memory analysis

This subsection evaluates the computational efficiency of the proposed MGC-SIFT framework in terms of average retrieval time per query and memory consumption. All experiments were conducted on the Corel-1K and Oxford-102 Flowers datasets using the same hardware configuration to ensure fair comparison.

3.4.1 Runtime and memory analysis on Corel-1K

Table 17 reports the average retrieval time per query and memory usage for representative methods on the Corel-1K dataset. Traditional handcrafted descriptors such as SIFT, RGB, and SIFT-RGB exhibit minimal computational and memory overhead due to their simple feature representations. In contrast, MGC-SIFT incurs additional cost arising from graph construction, attention refinement, and proxy-based similarity encoding.

Despite this added complexity, MGC-SIFT remains substantially more efficient than deep CNN-based retrieval. While VGG16 requires approximately 18.5 ms per query, the proposed MGC-SIFT framework achieves retrieval in approximately 10.6 ms per query, demonstrating a favourable balance between accuracy and efficiency.

To isolate the impact of proxy learning, Table 18 compares the runtime of different MGC-SIFT variants. Removing proxy learning increases the average retrieval time from 10,581.96 µs to 52,873.43 µs, representing an approximately fivefold increase. This clearly demonstrates the critical role of proxy-based learning in accelerating similarity computation and enabling scalable retrieval. Figures 3–5 further illustrate these trends. While handcrafted descriptors remain computationally lightweight, MGC-SIFT achieves a significant efficiency advantage over deep CNN-based methods. The removal of proxy learning leads to a pronounced increase in retrieval latency, confirming its importance in practical CBIR deployments.

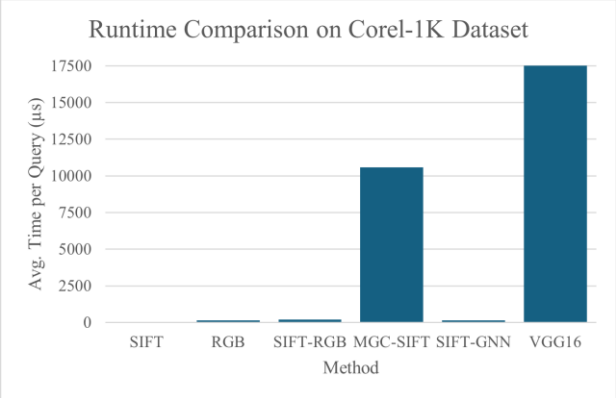

Figure 3: Average retrieval time per query on the Corel-1K dataset. While handcrafted descriptors are computationally lightweight, the proposed MGC-SIFT incurs additional cost due to graph modeling and proxy learning, yet remains faster than deep CNN-based retrieval.

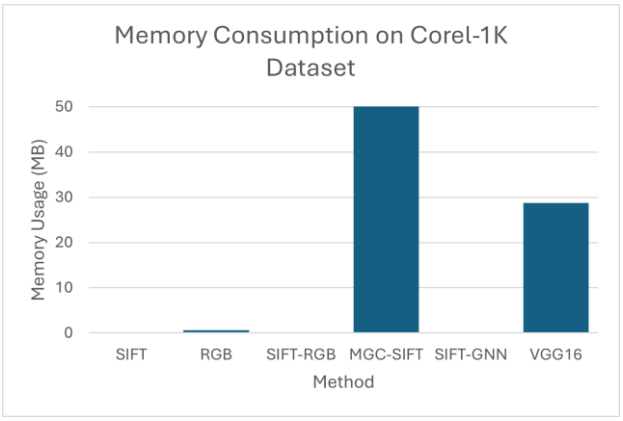

Figure 4: Memory consumption comparison on the Corel-1K dataset. MGC-SIFT requires additional memory due to graph construction and proxy embeddings, whereas traditional SIFT-based methods exhibit minimal memory usage.

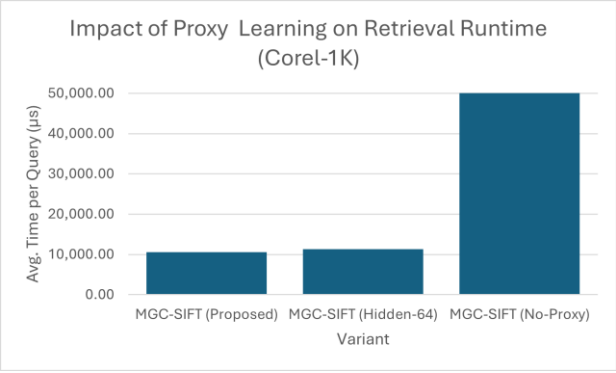

Figure 5: Impact of proxy learning on retrieval runtime for MGC-SIFT. Removing proxy learning leads to a

significant increase in retrieval time, confirming its role in improving computational efficiency.

3.4.2 Runtime and memory analysis on Oxford-102 flowers

The Oxford-102 Flowers dataset presents a more challenging scenario due to its fine-grained categories and higher-class count, resulting in substantially increased retrieval complexity. As shown in Table 19, MGC-SIFT exhibits higher absolute runtime and memory usage compared to Corel-1K. This increase is expected, as the graph construction and proxy representations scale with the number of keypoints and class proxies.

Nevertheless, proxy learning continues to play a crucial role. As reported in Table 20, removing proxy learning increases the average retrieval time from 1,175,895.15 µs to 1,267,801.79 µs, confirming that proxy-based optimization consistently reduces retrieval latency even under high-complexity conditions.

Compared to VGG16, which also incurs substantial memory and runtime overhead on this dataset, MGC-SIFT provides a more interpretable and modular alternative without requiring large-scale supervised training.

3.4.3 Discussion

Although MGC-SIFT introduces additional computational and memory overhead compared to classical handcrafted descriptors, this cost is a direct consequence of its enriched representation, which integrates color cues, graph-based contextual reasoning, attention mechanisms, and proxy learning. Importantly, the proposed framework remains significantly more efficient than deep CNN-based retrieval while delivering competitive retrieval accuracy across diverse datasets.

Overall, these results demonstrate that MGC-SIFT achieves a favorable accuracy–efficiency trade-off, making it well suited for medium- to large-scale CBIR applications where interpretability, scalability, and retrieval quality are prioritized over raw inference speed.

Table 18: Impact of proxy learning on retrieval runtime (Corel-1K)

Variant	**Avg. Time per Query (µs)**
MGC-SIFT (Proposed)	10,581.96
MGC-SIFT (Hidden-64)	11,296.51
MGC-SIFT (No-Proxy)	52,873.43

Table 19: Runtime and memory consumption on Oxford-102 flowers dataset

Method	**Avg. Time per Query (µs)**	**Memory Usage (MB)**
MGC-SIFT (Proposed)	11,75,895.15	1,152.94
SIFT	1,449.09	3
RGB	5,756.45	12.01
SIFT-RGB	2,022.18	4.13
SIFT-GNN	1,458.23	3
VGG16	3,70,354.10	588.48

Table 20: Impact of Proxy Learning on Retrieval Runtime (Oxford-102 Flowers)

Variant	**Avg. Time per Query (µs)**
MGC-SIFT (Proposed)	11,75,895.15
MGC-SIFT (Hidden-64)	11,57,271.03
MGC-SIFT (No-Proxy)	12,67,801.79

3.5 Statistical validation

To assess whether the observed performance differences are statistically meaningful, a paired two-tailed statistical significance test was conducted on per-query retrieval scores. For each dataset, the distributions of average precision values obtained by MGC-SIFT were compared against those of baseline handcrafted, graph-based, and deep learning–based methods.As reported in Tables 21 and 22, statistically significant differences ($p < 0.05$) are observed for most method pairs across both datasets. In particular, extremely small p-values (often $< 10^{-6}$) indicate that the retrieval behavior of MGC-SIFT differs consistently from that of SIFT-GNN, deep CNN features, and hybrid fusion approaches. These results confirm that the observed performance variations are not attributable to random fluctuations.

It is important to note that statistical significance reflects distributional differences rather than universal superiority. While MGC-SIFT does not always yield the highest mean performance, the results demonstrate that its retrieval behavior is consistently distinct and stable across queries. The extremely small p-values observed on the Oxford-102 Flowers dataset can be attributed to the large number of fine-grained classes and query samples, which increases statistical power in per-query evaluation.

Overall, the statistical analysis provides strong empirical evidence that the proposed MGC-SIFT descriptor exhibits reliable and non-random retrieval characteristics when compared with conventional handcrafted, graph-based, and deep learning–based CBIR methods.

Table 21: Statistical significance analysis on Corel-1K dataset

Method Pair	MGC-SIFT (Mean ± Std)	Compared Method (Mean ± Std)	p-value
MGC-SIFT vs SIFT	0.3270 ± 0.2881	0.2970 ± 0.2162	3.17E−01
MGC-SIFT vs SIFT-RGB	0.3270 ± 0.2881	0.3980 ± 0.2454	1.42E−02
MGC-SIFT vs SIFT-GNN	0.3270 ± 0.2881	0.1830 ± 0.1557	1.07E−07
MGC-SIFT vs VGG16	0.3270 ± 0.2881	0.5230 ± 0.3162	3.94E−08
MGC-SIFT vs VGG16 + MGC-SIFT	0.3270 ± 0.2881	0.3730 ± 0.2814	1.70E−06

3.5.1 Overall discussion summary

Combining the results from ablation studies, comparative evaluations, runtime analysis, and statistical validation, the proposed MGC-SIFT framework demonstrates robust and adaptable retrieval behavior across diverse image domains. Ablation studies confirm the complementary contributions of color augmentation, graph modeling, attention mechanisms, and proxy learning. Comparative experiments highlight the ability of MGC-SIFT to provide competitive performance without requiring large-scale supervised training, while statistical validation confirms the consistency and reliability of its retrieval behavior. Together, these findings validate the proposed design and establish MGC-SIFT as a practical and interpretable alternative to conventional SIFT-based and deep learning–based CBIR systems.

Table 22: Statistical Significance Analysis on Oxford-102 Flowers Dataset

Method Pair	MGC-SIFT (Mean ± Std)	Compared Method (Mean ± Std)	p-value
MGC-SIFT vs SIFT	0.0993 ± 0.1691	0.1784 ± 0.2302	9.62E−117
MGC-SIFT vs SIFT-RGB	0.0993 ± 0.1691	0.2785 ± 0.3068	< 1.0E−300
MGC-SIFT vs SIFT-GNN	0.0993 ± 0.1691	0.0432 ± 0.0810	6.24E−126
MGC-SIFT vs VGG16	0.0993 ± 0.1691	0.1167 ± 0.2007	2.17E−10
MGC-SIFT vs VGG16 + MGC-SIFT	0.0993 ± 0.1691	0.1375 ± 0.2060	6.40E−201

3.6 Robustness evaluation under image degradation

To evaluate robustness under image degradation, additional experiments were conducted on the Corel-1K dataset using controlled perturbations applied exclusively to the query images. Two degradation scenarios were considered: Gaussian noise, introduced with zero mean and fixed variance to simulate sensor noise, and occlusion, generated by masking a contiguous rectangular region covering a fixed portion of the image area. In all experiments, the retrieval database remained unchanged, ensuring that performance variations reflect robustness to query degradation rather than database bias. Feature extraction was performed independently for each degradation condition, and retrieval followed the same query–database protocol used for clean images.

Performance was assessed using mean Average Precision (mAP) and Precision@10, Recall@10, F1-score@10, and Accuracy@10, with Canberra distance adopted as the primary similarity metric. The quantitative results are summarized in Table 23. All results are reported at k = 10 using Canberra distance.

The results indicate that classical handcrafted descriptors such as SIFT and RGB-based features exhibit relatively stable performance under moderate Gaussian noise, whereas deep CNN-based features (VGG16) show more pronounced degradation. In contrast, the proposed MGC-SIFT framework maintains consistent retrieval performance across clean, noisy, and occluded conditions, demonstrating increased robustness to image degradation.

Minor performance fluctuations under Gaussian noise are expected, as noise can increase local gradient variability and alter keypoint density, occasionally improving separability for handcrafted and color-based descriptors.

Under occlusion, MGC-SIFT preserves retrieval effectiveness more reliably due to its graph-based contextual modeling and attention-guided feature refinement, which suppress irrelevant or corrupted regions while emphasizing stable keypoint relationships.

On the COIL-20 dataset, which contains controlled object images with limited structural variability, MGC-SIFT achieved strong retrieval performance comparable to or exceeding existing methods at higher retrieval depths, demonstrating its effectiveness even in relatively constrained visual settings. For the Oxford-102 Flowers dataset, consistent improvements over grayscale and graph-only variants highlight the importance of color-aware graph modeling in fine-grained and color-rich retrieval tasks. Similarly, on the UC-Merced Land Use dataset, MGC-SIFT maintained competitive performance across all metrics, indicating its suitability for complex aerial scene retrieval where both spatial layout and color distribution contribute to semantic similarity.

Table 23: Robustness Evaluation on Corel-1K

Method	**Clean mAP**	**Clean P@10**	**Gaussian mAP**	**Gaussian P@10**	**Occlusion mAP**	**Occlusion P@10**
SIFT	0.3294	0.3316	0.3379	0.3237	0.3296	0.3376
RGB	0.3872	0.4387	0.4547	0.4491	0.3938	0.448
SIFT-GNN	0.2161	0.2177	0.235	0.2371	0.2321	0.2202
SIFT-RGB	0.434	0.4423	0.434	0.4423	0.434	0.4423
MGC-SIFT (Proposed)	0.3512	0.2756	0.348	0.2829	0.3557	0.268
VGG16	0.5604	0.6464	0.5496	0.5914	0.5128	0.5018

Overall, these findings demonstrate that MGC-SIFT achieves a favorable balance between robustness and efficiency, making it well suited for real-world CBIR scenarios where query images may be affected by noise, partial occlusion, or acquisition artifacts.

4 Conclusion and future work

This study presented **MGC-SIFT**, a Multimodal Graph Color SIFT algorithm designed to address key limitations of traditional SIFT-based and graph-based CBIR systems by jointly modeling color information, local texture, and spatial relationships among keypoints. Extensive experimental evaluation across four diverse benchmark datasets, Corel-1K, COIL-20, Oxford-102 Flowers, and UC-Merced Land Use, demonstrated the effectiveness, robustness, and generalizability of the proposed approach across heterogeneous image domains.

The results confirm that integrating color-augmented SIFT descriptors with graph-based contextual learning enables competitive and stable retrieval performance across diverse CBIR scenarios. On the Corel-1K dataset, MGC-SIFT achieved balanced performance with improved recall and F1@k, reflecting enhanced retrieval consistency compared to classical SIFT and SIFT-RGB descriptors. Although SIFT-GNN attained a marginally higher mAP on this dataset, MGC-SIFT exhibited more uniform performance across evaluation metrics, underscoring the benefit of multimodal feature integration.

Robustness evaluation under image degradation further confirmed the stability of the proposed framework. Experiments conducted on the Corel-1K dataset using Gaussian noise and occlusion applied exclusively to query images demonstrated that MGC-SIFT maintains consistent retrieval performance across degraded conditions. While minor performance variations were observed under Gaussian noise—attributable to changes in local gradient distributions—the proposed method showed resilience to both noise and occlusion due to its graph-based contextual modeling and attention-guided feature refinement. These results indicate that MGC-SIFT is well suited for real-world CBIR scenarios where image quality and acquisition conditions may vary.

Comprehensive ablation studies further validated the architectural design of MGC-SIFT. The systematic removal of color augmentation, graph modeling, attention mechanisms, and proxy-based learning consistently led to measurable performance degradation, confirming that each component contributes complementary benefits. In particular, attention-guided proxy learning was shown to play a critical role in improving retrieval efficiency, as evidenced by substantial reductions in query-time complexity without compromising accuracy. Runtime and memory analysis demonstrated that, while MGC-SIFT incurs higher computational cost than traditional handcrafted descriptors due to graph construction and

proxy optimization, it remains significantly more efficient than deep CNN-based retrieval methods, achieving a favorable accuracy–efficiency trade-off suitable for scalable CBIR applications.

Overall, the experimental findings support the proposed research hypotheses, demonstrating that multimodal feature integration, graph-based contextual reasoning, attention-guided proxy learning, and robustness to image degradation collectively enhance retrieval effectiveness, representation compactness, and scalability. These results validate MGC-SIFT as a reliable and interpretable alternative to purely deep learning–based CBIR systems, particularly in scenarios where training data availability, computational resources, or model transparency are constrained.

4.1 Future work

Future research will focus on extending the MGC-SIFT framework to cross-modal and multimodal retrieval scenarios, including text–image retrieval, semantic search, and video-based CBIR. Additional directions include the exploration of adaptive graph construction strategies, lightweight graph neural architectures, and hybrid integration with deep semantic embeddings to further enhance scalability and robustness on large-scale datasets. Moreover, optimizing feature extraction and graph processing through parallelization and hardware acceleration represents a promising avenue for deploying MGC-SIFT in real-time and large-scale retrieval systems across application domains such as biomedical imaging, remote sensing, and surveillance.

Declarations

Author contributions

Trupti Babasaheb Ghatage conducted most of the research work, including conceptualization, methodology design, software implementation, data analysis, and preparation of the original draft. Dattatraya Vishnu Kodavade provided overall supervision, technical guidance, and contributed to the validation and critical revision of the manuscript.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Use of AI-assisted tools

The authors used AI-assisted language tools (ChatGPT, OpenAI) solely to improve the clarity and grammatical quality of the manuscript. The AI tools did not contribute to the generation of scientific content, data analysis, experimental results, figures, or interpretations. The authors take full responsibility for the originality, accuracy, and integrity of the work.

Data availability statement

The datasets used in this study are publicly available benchmark datasets. Detailed dataset characteristics, including the number of classes, total images, image types, and application domains, are summarized in Table 2. The study employs the Corel-1K, COIL-20, Oxford 102 Flowers, and UC Merced Land Use datasets, all of which are openly accessible and require no special permissions for use. No private, confidential, or proprietary data were utilized. All processed data and experimental results are fully reported within the manuscript and its supplementary material.

References

[1] S. Sikandar, A. Alsalman, and R. Mahum, “A Novel Hybrid Approach for a Content-Based Image Retrieval Using Feature Fusion,” *Applied Sciences*, vol. 13, no. 7, p. 4581, Apr. 2023, doi: 10.3390/app13074581.

[2] J. Kim and B. C. Ko, “Scene Graph and Natural Language-Based Semantic Image Retrieval Using Vision Sensor Data,” *Sensors*, vol. 25, no. 11, p. 3252, May 2025, doi: 10.3390/s25113252.

[3] A. W. M. Smeulders, S. Santini, M. Worring, R. Jain, and A. Gupta, “Content-Based Image Retrieval at the End of the Early Years,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, Jan. 2000, doi: 10.1109/34.895972.

[4] A. Humeau-Heurtier, “Texture Feature Extraction Methods: A Survey,” *IEEE Access*, vol. 7, pp. 8975–9000, 2019, doi: 10.1109/ACCESS.2018.2890743.

[5] N. Alpaslan and K. Hanbay, “Multi-Scale Shape Index-Based Local Binary Patterns for Texture Classification,” *IEEE Signal Processing Letters*, vol. 27, pp. 660–664, 2020, doi: 10.1109/LSP.2020.2987474.

[6] F. Mirzapour and H. Ghassemian, “Improving Hyperspectral Image Classification by Combining Spectral, Texture, and Shape Features,” *International Journal of Remote Sensing*, vol. 36, no. 4, pp. 1070–1096, 2015, doi: 10.1080/01431161.2015.1007251.

[7] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004, doi: 10.1023/B: VISI.0000029664.99615.94.

[8] J. R. R. van de Sande, T. Gevers, and C. G. M. Snoek, *“Evaluating Color Descriptors for Object and Scene Recognition, ”*IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 32, no. 9, pp. 1582–1596, 2010. DOI: 10.1109/TPAMI.2009.154

[9] X. Zhang, M. Jiang, Z. Zheng, X. Tan, E. Ding, and Y. Yang, “Understanding Image Retrieval Re-Ranking: A Graph Neural Network Perspective,” arXiv:2012.07620, 2020, doi: 10.48550/arXiv.2012.07620.

[10] H. Lacheheb and S. Aouat, “SIMIR: New Mean SIFT Color Multi-Clustering Image Retrieval,” *Multimedia Tools and Applications*, vol. 76, no. 5, pp. 6333–6354, 2016, doi: 10.1007/s11042-015-3167-3.

[11] D. Kobak and P. Berens, “The Art of Using t-SNE for Single-Cell Transcriptomics,” *Nature*

Communications, vol. 10, p. 5416, 2019, doi: 10.1038/s41467-019-13056-x.

[12] X. Jia, A. Kale, V. Kumar, Z. Lin, and H. Zhao, "Personalized Image Retrieval with Sparse Graph Representation Learning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 19, no. 4, pp. 2735–2743, 2020, doi: 10.1145/3394486.3403324.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.

[14] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556, 2014, doi: 10.48550/arXiv.1409.1556.

[15] X. Li, S. Wei, M. Ge, J. Wang, and Y. Du, "Adaptive Multi-Proxy for Remote Sensing Image Retrieval," *Remote Sensing*, vol. 14, no. 21, p. 5615, 2022, doi: 10.3390/rs14215615.

[16] A. Hermans, L. Beyer, and B. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," arXiv:1703.07737, 2017, doi: 10.48550/arXiv.1703.07737.

[17] S. Kim, M. Cho, S. Kwak, and D. Kim, "Proxy Anchor Loss for Deep Metric Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, doi: 10.1109/CVPR42600.2020.00330.

[18] Y. Movshovitz-Attias, S. Singh, A. Toshev, T. K. Leung, and S. Ioffe, "No Fuss Distance Metric Learning Using Proxies," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 360–368, doi: 10.1109/ICCV.2017.47.

[19] M. M. Adnan et al., "Image Annotation with YCbCr Color Features Based on Multiple Deep CNN-GLP," *IEEE Access*, vol. 12, pp. 11340–11353, 2024, doi: 10.1109/ACCESS.2023.3330765.

[20] H. Yu et al., "Text–Image Matching for Cross-Modal Remote Sensing Image Retrieval via Graph Neural Network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 812–824, 2023, doi: 10.1109/JSTARS.2022.3231851.

[21] Y. Zhang, X. Zheng, and X. Lu, "Remote Sensing Image Retrieval by Deep Attention Hashing with Distance-Adaptive Ranking," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 4301–4311, 2023, doi: 10.1109/JSTARS.2023.3271303.

[22] D. Zhao, S. Xiong, and Y. Chen, "Multiscale Context Deep Hashing for Remote Sensing Image Retrieval," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 7163–7172, 2023, doi: 10.1109/JSTARS.2023.3298990.

[23] Z. Cai, Y. Pan, and W. Jin, "Proxy-Based Rotation Invariant Deep Metric Learning for Remote Sensing Image Retrieval," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 7759–7772, 2024, doi: 10.1109/JSTARS.2024.3382845.