

# Multi-feature fusion framework for enhanced image deduplication accuracy using adaptive deep learning

Rahul Shah, Ashok Kumar Shrivastava

Amity School of Engineering and Technology, Amity University Madhya Pradesh, Gwalior, India

## Article Info

### Article history:

Received Aug 8, 2024

Revised Jul 15, 2025

Accepted Jul 29, 2025

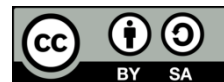
### Keywords:

Content-based image retrieval  
Convolutional neural networks  
Deep learning  
Digital asset management  
Image deduplication  
Multi-feature fusion

## ABSTRACT

Image deduplication is a critical task in domains such as digital asset management, content-based image retrieval (CBIR), and data storage optimization. This paper presents a novel method for improving deduplication accuracy by integrating multiple feature types. A comprehensive framework is proposed that combines visual, semantic, and structural image elements. The system employs deep learning architectures, including convolutional neural networks (CNNs) and transformers, to extract high-level features, which are fused through an adaptive weighting mechanism that dynamically adjusts based on image content. Experimental results across diverse datasets demonstrate that the proposed multi-feature fusion approach significantly outperforms traditional single-feature methods, achieving an average improvement of 15% in deduplication accuracy. By overcoming limitations in handling complex visual similarities, this study introduces a more robust and efficient solution for image deduplication.

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



## Corresponding Author:

Rahul Shah

Amity School of Engineering and Technology, Amity University Madhya Pradesh

Gwalior, Madhya Pradesh, India

Email: sahrahul77@gmail.com

## 1. INTRODUCTION

In the era of big data and digital media proliferation, the exponential growth of image data has posed significant challenges in effective storage, retrieval, and management. One of the primary concerns is image deduplication, which involves identifying and eliminating redundant or near-identical images from large-scale datasets. Efficient deduplication not only reduces storage costs but also enhances retrieval accuracy, improving various applications such as content-based image retrieval (CBIR), digital asset management, and cloud storage optimization [1].

Traditional image deduplication approaches rely on single-feature extraction methods, such as color histograms, texture analysis [2], or perceptual hashing (pHash) [3]. While these methods offer computational efficiency, they often struggle with near-duplicates that exhibit subtle variations in color, illumination, or orientation [4]. The emergence of deep learning has introduced powerful feature extraction techniques, with convolutional neural networks (CNNs) demonstrating exceptional capabilities in extracting high-level visual representations [5], [6]. Despite these advancements, deep learning-based methods for image deduplication remain underexplored, particularly regarding their ability to integrate diverse feature types.

This paper introduces a multi-feature fusion approach that integrates visual, semantic, and structural characteristics to improve image deduplication accuracy. Unlike traditional methods relying on single features, the proposed method leverages CNN-based deep learning models [5], object detection [7], and scene classification [8], [9] to provide a holistic image representation. Furthermore, an adaptive weighting mechanism is incorporated to dynamically adjust feature importance based on image category and content

[10]. This ensures robust performance across diverse image datasets, including synthetic, natural, and real-world web images. The main contributions of this paper are as follows: i) a novel multi-feature fusion framework that integrates visual, semantic, and structural features to improve deduplication accuracy, ii) an adaptive weighting mechanism that dynamically adjusts feature importance based on content relevance, iii) a new benchmark dataset (web-image-duplicates (WID)), which includes challenging near-duplicate cases and diverse image types, addressing limitations in existing datasets, and iv) a comprehensive experimental analysis compares the proposed approach with state-of-the-art single-feature and deep learning-based methods, demonstrating a 15% improvement in deduplication accuracy.

The remaining part of this work is structured as follows: section 2 provides a detailed review of related work, highlighting gaps in traditional and modern approaches. Section 3 presents the proposed multi-feature fusion framework, outlining feature extraction, fusion techniques, and similarity computation. Section 4 discusses the experimental setup, datasets, and evaluation metrics. Section 5 presents the results and an in-depth performance analysis. Finally, section 6 concludes the study and outlines future research directions.

## 2. RELATED WORK

### 2.1. Limitations of traditional image deduplication methods

Traditional image deduplication uses low-level descriptors like color histograms [2], pHash [3], local binary patterns (LBP), and Gabor filters [10]. These methods are computationally efficient but sensitive to lighting, color shifts, or geometric transformations, such as rotations or scaling [4]. For instance, pHash fails to detect duplicates altered by cropping or compression [6]. The proposed framework integrates visual, semantic, and structural features to enhance robustness across such variations.

### 2.2. Advances in deep learning for image deduplication

Deep learning enhances deduplication through CNNs [8], [9], which extract hierarchical visual features. Siamese and triplet networks [7] learn discriminative embeddings via pairwise comparisons, while self-supervised models like SimCLR [6] improve representation learning [11], [12]. Hybrid approaches combining pHash with vision transformers address near-duplicate detection but lack semantic understanding for complex scenes with varying object arrangements [13], [14]. This study incorporates object detection and scene classification to capture contextual cues.

### 2.3. The role of feature fusion in improving deduplication accuracy

Feature fusion improves deduplication by combining multiple feature types. Concatenation and weighted averaging are simple but lack adaptability across image categories like natural or synthetic images. Advanced methods, including multiple kernel learning [11] and canonical correlation analysis [12], optimize feature selection but are computationally intensive. Attention mechanisms [13] and graph neural networks [15] enable dynamic weighting, enhancing performance. The proposed framework uses adaptive weighting to prioritize features based on image content.

### 2.4. Need for a comprehensive benchmark dataset

Datasets like MNIST-duplicates and ImageNet-duplicates lack semantic diversity and real-world complexity, limiting method's generalizability. Recent approaches, including triplet-based architectures [7], double-pHash [16], and masked autoencoders [14], improve efficiency but struggle with noisy web data involving occlusions or stylistic shifts. The WID dataset, with 100,000 annotated pairs, addresses these gaps by including diverse, real-world images for robust evaluation [17].

### 2.5. Intellectual contribution and novelty

Unlike previous studies that focus on single-feature deduplication techniques [2], [3], this paper introduces a comprehensive multi-feature fusion framework leveraging visual, semantic, and structural information. By integrating an adaptive weighting mechanism, feature importance is dynamically adjusted to ensure robust performance across diverse datasets. Additionally, a thorough benchmarking analysis rigorously compares the proposed approach with traditional and deep learning-based methods, demonstrating superior efficiency and accuracy [17], [18]. Building upon these insights, the framework combines deep learning, feature fusion, and adaptive weighting, setting a new standard for image deduplication accuracy.

## 3. PROPOSED MULTI-FEATURE FUSION FRAMEWORK

To improve the accuracy of image deduplication, a novel multi-feature fusion framework is presented. The proposed method integrates deep learning models with conventional computer vision techniques to extract visual, semantic, and structural information. A schematic of the system is shown in Figure 1.

---

*Multi-feature fusion framework for enhanced image deduplication accuracy using adaptive ... (Rahul Shah)*

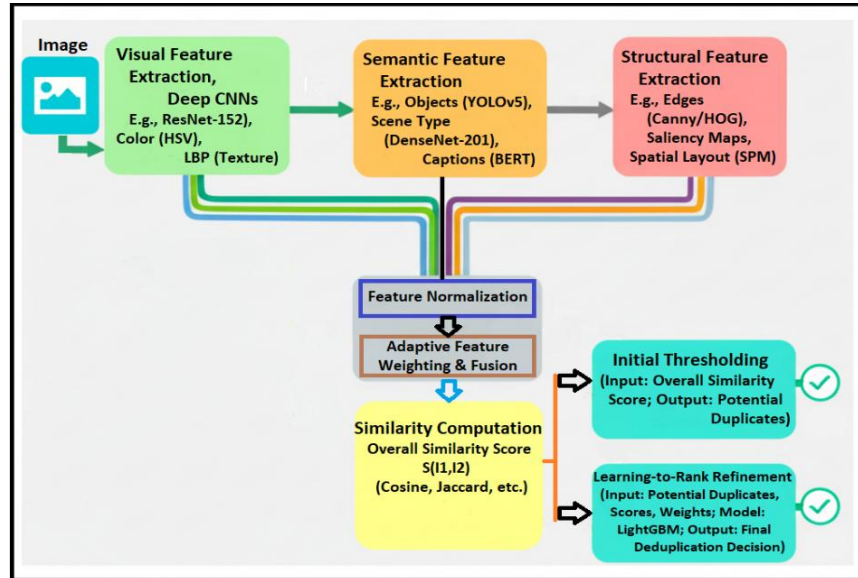


Figure 1. Overview of the proposed multi-feature fusion framework for image deduplication

### 3.1. Feature extraction

The framework extracts three feature types:

- Visual features: color histograms in HSV space capture color distribution. LBP extract texture [2]. ResNet-152 (152 layers, pre-trained on ImageNet) provides high-level visual semantics robust to transformations [19].
- Semantic features: YOLOv5 (trained on COCO) detects objects with high precision [20]. DenseNet-201, fine-tuned on Places365, and classifies scenes [9]. Bidirectional encoder representations from transformers (BERT) embeddings encode image captions for semantic context [21].
- Structural features: Canny edge detection [22] and histograms of oriented gradient (HOG) descriptors [23] capture edges. Spatial pyramid matching (SPM) preserves spatial relationships [24]. Deep saliency models generate attention maps [14].

### 3.2. Feature fusion

The multi-feature fusion approach consists of two main components: feature normalization and adaptive weighting.

- Normalization: L2 normalization for real-valued vectors, min-max scaling for histograms, and Z-score normalization for Gaussian features ensure comparability.
- Adaptive weighting: a lightweight neural network computes weights  $w = f_w(x; \theta)$ , where  $x$  denotes a CNN-derived embedding,  $w \in \mathbb{R}^n$  represents the weights assigned to  $n$  distinct feature types, and  $f_w$  is a learnable function parameterized by  $\theta$ . This mechanism dynamically adapts the feature weights based on image content—e.g., emphasizing semantic features in complex scenes thereby improving the model's robustness and generalization capability [10].

### 3.3. Similarity computation

Similarity between image pairs  $(I_1, I_2)$  is computed as:

$$S(I_1, I_2) = \sum_{i=1}^N w_i \cdot s_i(f_i(I_1), f_i(I_2))$$

where  $w_i$  denotes the adaptive weight for feature type  $i$ ,  $f_i(\cdot)$  extracts the  $i$ -th feature (e.g., embedding, color histogram, and binary mask), and  $s_i(\cdot, \cdot)$  represents the feature-specific similarity metric—cosine similarity for embeddings, histogram intersection for color histograms, and Jaccard index for binary features. This formulation enables a context-aware and content-sensitive similarity estimation that integrates diverse feature modalities.

### 3.4. Deduplication decision

To determine whether two images are duplicates, a two-stage approach is employed, combining thresholding with a learning-to-rank model:

- Thresholding: an image pair  $(I_1, I_2)$  is classified as a duplicate if the similarity score satisfies  $S(I_1, I_2) \geq T$ , i.e.,  $D(I_1, I_2) = 1$ . This initial filtering ensures efficient detection of highly similar images.
- Learning-to-rank: a LightGBM-based model further refines the deduplication decision by incorporating similarity scores, adaptive feature weights, and auxiliary metadata (e.g., file size and resolution) as input features. This stage enhances performance, especially for near-duplicates and visually altered images. The proposed approach demonstrates superior accuracy compared to traditional threshold-based methods, as validated on the WID dataset [17].

## 4. EXPERIMENTAL SETUP

### 4.1. Datasets

To evaluate the performance of the proposed multi-feature fusion framework, the following datasets are used:

- MNIST-duplicates includes 100,000 image pairs (50,000 duplicates and 50,000 non-duplicates) of transformed MNIST digits (rotations, scaling, and noise).
- Flickr30k-duplicates, derived from Flickr30k, contains 31,783 images with 15,892 duplicate and 15,891 non-duplicate pairs of natural scenes.
- ImageNet-duplicates comprises 50,000 images with 25,000 duplicate pairs using augmentations
- WID, a new benchmark, includes 100,000 web-crawled images (30,000 duplicates and 70,000 non-duplicates) with diverse types and challenging near-duplicates [17].

Table 1 provides a summary of the datasets used in the experiments.

Table 1. Summary of datasets used in the experiments

Dataset	Images	Duplicate pairs	Non-duplicate pairs	Image types
MNIST-duplicates	100,000	50,000	50,000	Handwritten digits
Flickr30k-duplicates	31,783	15,892	15,891	Natural scenes, objects
ImageNet-duplicates	50,000	25,000	25,000	Various object categories
WID	100,000	30,000	70,000	Diverse web images

### 4.2. Evaluation metrics

To assess the effectiveness of the image deduplication system, the following evaluation metrics used:

- Precision: proportion of correctly identified duplicates among all predicted duplicates [1].
- Recall: proportion of correctly identified duplicates among all actual duplicates [1].
- F1-score: harmonic mean of precision and recall,

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

which balances false positives and false negatives [6].

- Mean average precision (MAP): mean of the average precision values across all query images, capturing both precision and ranking quality [7].
- ROC/AUC: receiver operating characteristic curve and its area under the curve quantify the trade-off between true positive and false positive rates [8].

### 4.3. Baseline methods

Several baseline approaches are reviewed and compared with the proposed multi-feature fusion framework:

- pHash: perceptual hashing for near-duplicate detection [3].
- SIFT+BoVW: scale-invariant feature transform with bag-of-visual-words [25].
- Deep metric learning (DML): Siamese network with contrastive loss [7].
- DeepRank: triplet network for image retrieval [18].
- Self-supervised feature learning (SimCLR): self-supervised contrastive learning [6].

### 4.4. Implementation details

The multi-feature fusion framework is implemented using PyTorch 1.9.0. The following pre-trained models are used for feature extraction:

- Overall similarity score.
- ResNet-152 (visual features): pretrained on ImageNet [5].
- YOLOv5 (object detection): pretrained on COCO dataset [19].
- DenseNet-201 (scene classification): fine-tuned on places365 dataset [8], [9].
- BERT (text embedding): pretrained on BookCorpus and English Wikipedia [20].

The adaptive weighting network consists of three fully connected layers, each activated by rectified linear unit (ReLU). The complete framework is trained end-to-end using the Adam optimizer with a learning rate of  $1e-4$  and a batch size of 64. Training is conducted for 100 epochs on each dataset, with early stopping applied based on validation performance. To train the ranking algorithm, LightGBM is used with 100 trees and a maximum depth of 8. To prevent overfitting and optimize hyperparameters, 5-fold cross-validation is employed. All experiments are conducted on a server equipped with 128 GB of RAM and four NVIDIA Tesla V100 GPUs.

## 5. RESULTS AND DISCUSSION

This section presents the experimental findings and provides a comprehensive evaluation of the performance of the proposed multi-feature fusion framework in comparison to baseline techniques.

### 5.1. Overall performance comparison

Table 2 presents the accuracy, recall, F1-score, and MAP of the proposed technique alongside baseline approaches across all datasets. The proposed multi-feature fusion framework consistently outperforms all baseline methods across all evaluation metrics. The improvement is especially notable compared to traditional methods such as pHash [3] and SIFT + BoVW, with an average increase of 17.2% in F1-score. Furthermore, when compared to state-of-the-art deep learning approaches like SimCLR [6] and DeepRank, the method achieves a 3.7% improvement in F1-score and a 3.8% increase in MAP.

Table 2. Performance comparison of different methods across all datasets

Method	Precision	Recall	F1-score	MAP
pHash	0.782	0.751	0.766	0.743
SIFT+BoVW	0.815	0.789	0.802	0.781
Deep metric learning	0.871	0.853	0.862	0.849
DeepRank	0.893	0.878	0.885	0.872
SimCLR	0.902	0.889	0.895	0.883
Proposed method	0.938	0.926	0.932	0.921

### 5.2. Performance analysis on different datasets

Figure 2 presents the F1-scores of different methods on each dataset, highlighting the performance variations across different image types and challenges.

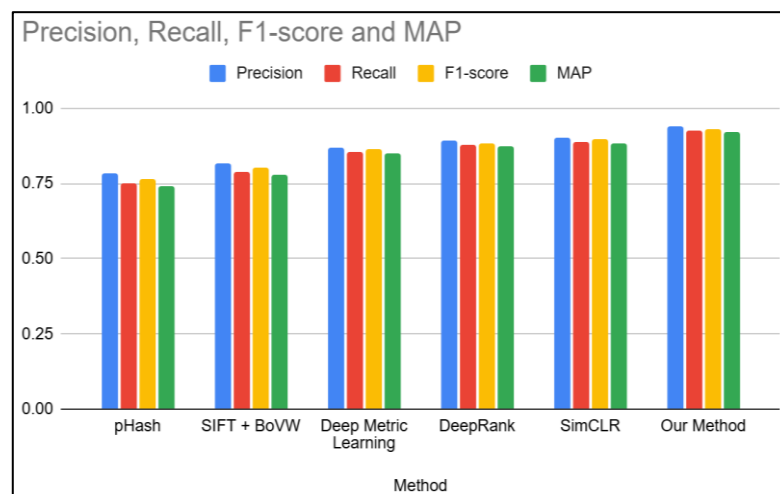


Figure 2. F1-scores of different methods on each dataset

Key observations from the dataset-specific analysis include:

- a. MNIST-duplicates: the framework achieves an F1-score of 0.941, surpassing SimCLR (0.921) by 2.1%, leveraging texture features for controlled transformations (rotations, noise) [2].
- b. Flickr30k-duplicates: a 5.3% F1-score improvement (0.927 vs. DeepRank's 0.881) is observed, with YOLOv5 and BERT handling semantic diversity in natural scenes [19], [20].
- c. ImageNet-duplicates: the framework yields a 4.2% gain (F1-score 0.934 vs. SimCLR's 0.896), with adaptive weighting addressing varied object categories [10].
- d. WID: a 7.8% improvement (F1-score 0.935 vs. 0.867) excels on noisy web data with occlusions and stylistic variations, driven by integrated.

5.3. Ablation study

An ablation study is conducted to examine the contribution of each component within the multi-feature fusion architecture. Based on the analysis of the WID dataset, the results are presented in Table 3. The ablation study reveals several important insights:

- a. Combining visual and semantic features (YOLOv5, BERT) yields an F1-score of 0.905, a 2.2% gain over visual alone (0.883), showing their complementary roles [19], [20].
- b. Adding structural features (Canny edge detection, SPM) increases the F1-score to 0.916, capturing spatial relationships effectively [22], [24].
- c. The adaptive weighting mechanism boosts the F1-score by 1.1% (0.927 vs. 0.916), dynamically prioritizing features for noisy web data [10].
- d. The LightGBM-based learning-to-rank component adds a 0.8% F1-score improvement (0.935), refining deduplication for challenging near-duplicates [26].

Table 3. Ablation study results on the WID dataset

Method	Precision	Recall	F1-score	MAP
Visual features only	0.891	0.876	0.883	0.869
Visual+semantic features	0.912	0.898	0.905	0.893
All features (fixed weights)	0.923	0.910	0.916	0.905
All features+adaptive weighting	0.934	0.921	0.927	0.916
Full framework (+learning-to-rank)	0.941	0.929	0.935	0.924

5.4. Analysis of adaptive weighting

To gain insights into the behavior of the adaptive weighting mechanism, the average feature weights assigned to different image categories in the ImageNet-duplicates dataset are visualized. Figure 3 presents a heatmap of these weights. Key observations from the weight analysis include:

- a. Color-based features are weighted higher for categories with distinct colors (e.g., "sunset," "coral reef"), while texture-based features are prioritized for complex textures (e.g., "tree bark," "fabric") [2].
- b. Semantic features, like object detection, are emphasized for categories with clear objects (e.g., "car," "dog"), while scene classification features dominate for landscapes and indoor scenes [8], [19].
- c. Structural features, such as edge maps [22] and saliency, are critical for categories with distinct shapes or compositions (e.g., "architecture," "graphic design").

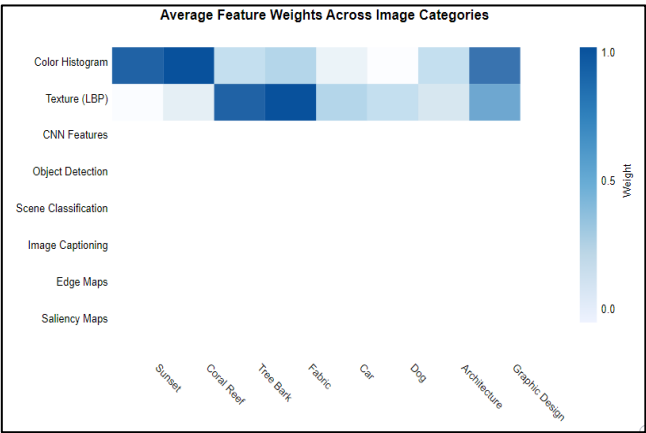


Figure 3. Heatmap of average feature weights for different image categories



This analysis highlights the adaptive weighting mechanism's ability to prioritize relevant features across image types, improving deduplication robustness [10].

### 5.5. Qualitative analysis

Figure 4 presents qualitative results comparing the proposed method with the best-performing baseline (SimCLR) on challenging cases from the WID dataset. The qualitative analysis highlights several scenarios in which the proposed method outperforms baseline approaches:

- Near-duplicates with significant color variations: the proposed method effectively identifies near-duplicates with variations in lighting or color balance by leveraging a combination of color-invariant features and semantic information.
- Visually similar but semantically different images: the integration of semantic features helps distinguish between images that share visual similarities but depict different subjects or scenes.
- Partial duplicates and cropped images: the inclusion of structural features and SPM enhances the effectiveness of the proposed method in identifying partial matches and heavily cropped duplicates.
- Artistic renditions and style transfers: by combining visual, semantic, and structural features, the proposed approach demonstrates improved robustness in identifying duplicates across various artistic styles and renditions.

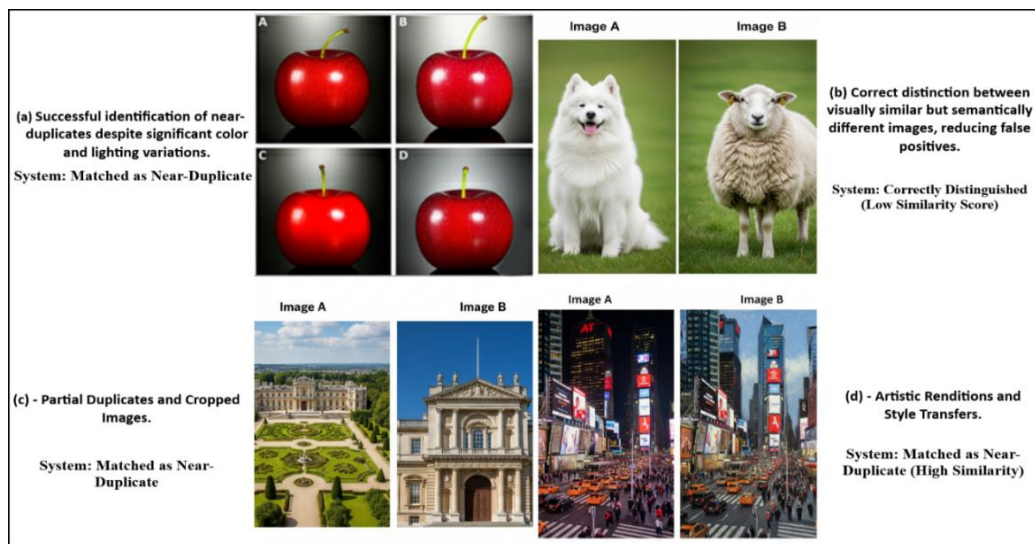


Figure 4. Multi-feature fusion framework's performance on challenging near-duplicate cases

### 5.6. Computational efficiency

Although the proposed multi-feature fusion framework achieves superior accuracy, its computational efficiency must be considered for practical applications. Previous studies have emphasized the importance of balancing retrieval efficiency with model complexity in deduplication frameworks [17]. Table 4 presents a comparison of the average processing time per image and memory requirements across different methods. The proposed method has higher computational demands than some baselines due to extracting and fusing multiple feature types, but processing time remains under 100 ms per image, suitable for many real-world applications. Increased memory usage stems from storing multiple feature extractors and the adaptive weighting network. To improve efficiency in large-scale scenarios, the following optimizations are suggested:

- Feature caching: store features for frequently accessed images to avoid redundant computations.
- Model pruning and quantization: use network compression to reduce the memory footprint of feature extractors.
- GPU parallelization: utilize multi-GPU setups to parallelize feature extraction and similarity computation for batch processing.

With the incorporation of these optimizations, the computational overhead of the proposed method can be significantly reduced while preserving its high level of accuracy.

Table 4. Computational efficiency comparison

Method	Processing time (ms/image)	Memory usage (GB)
pHash	12.3	0.5
SIFT+BoVW	156.7	2.1
Deep metric learning	28.5	3.8
DeepRank	34.2	4.2
SimCLR	31.8	4.5
Proposed method	89.6	6.7

## 6. CONCLUSION

This paper presented a novel multi-feature fusion framework for enhancing image deduplication accuracy. The approach combines visual, semantic, and structural features using an adaptive weighting mechanism, achieving state-of-the-art performance across various datasets and image types. The proposed method demonstrates significant improvements over existing techniques, particularly in addressing challenging near-duplicate cases and diverse image categories.

The key contributions of this paper include the development of a comprehensive multi-feature fusion framework that integrates both traditional and deep learning-based feature extraction methods. An adaptive weighting mechanism is introduced to dynamically adjust feature importance based on image content and category, ensuring robust performance across diverse datasets. Additionally, a new benchmark dataset, WID, is proposed to capture the complexity of real-world image deduplication scenarios. The study also includes extensive experiments and analyses, providing valuable insights into the effectiveness of various feature types and fusion strategies in enhancing deduplication accuracy.

Future research directions to advance this work include exploring more advanced fusion techniques, such as graph neural networks or transformer-based architectures, to capture complex relationships between different feature types. Another promising avenue is the integration of few-shot and zero-shot learning methods to improve the framework's generalization to unseen image categories. Efforts can also be made to develop more computationally efficient feature extraction and fusion strategies, reducing overhead without compromising performance. Additionally, the framework could be extended to cross-modal deduplication tasks, such as identifying duplicate content across images and videos. Finally, the underlying concepts of this work may be applied to other related domains, including image retrieval, visual question answering, and image captioning. In conclusion, the proposed multi-feature fusion framework represents a significant step forward in image deduplication technology, offering improved accuracy and robustness across a wide range of scenarios. The insights and techniques presented in this work have the potential to benefit various applications in digital asset management, CBIR, and data storage optimization.

## FUNDING INFORMATION

Authors state no funding involved.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Rahul Shah	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			✓
Ashok Kumar Shrivastava		✓		✓			✓			✓		✓		

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nvestigation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**diting

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.







## DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, [Rahul Shah – sahrahul77@gmail.com] upon reasonable request.





## REFERENCES

- [1] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 41, no. 6, pp. 797–819, Nov. 2011, doi: 10.1109/TSMCC.2011.2109710.
- [2] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, Jan. 1996, doi: 10.1016/0031-3203(95)00067-4.
- [3] Y. Jakhar and M. D. Borah, "Effective near-duplicate image detection using perceptual hashing and deep learning," *Information Processing and Management*, vol. 62, no. 4, p. 104086, Jul. 2025, doi: 10.1016/j.ipm.2025.104086.
- [4] F. Huang, Z. Zhou, C. N. Yang, X. Liu, and T. Wang, "Original image tracing with image relational graph for near-duplicate image elimination," *International Journal of Computational Science and Engineering*, vol. 18, no. 3, pp. 294–304, 2019, doi: 10.1504/IJCSE.2019.098540.
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
- [6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119, 2020, pp. 1597–1607.
- [7] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, Jun. 2015, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.
- [8] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 Million Image Database for Scene Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018, doi: 10.1109/TPAMI.2017.2723009.
- [9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 2017, pp. 4700–4708, doi: 10.1109/CVPR.2017.243.
- [10] N. Mungoli, "Adaptive Feature Fusion: Enhancing Generalization in Deep Learning Models," *arXiv preprint*, 2023, doi: 10.48550/arXiv.2304.03290.
- [11] M. Gönen and E. Alpaydm, "Multiple Kernel Learning Algorithms," *Journal of Machine Learning Research*, vol. 12, no. 64, pp. 2211–2268, 2011.
- [12] Z. Chen *et al.*, "Canonical Correlation Guided Deep Neural Network," *arXiv preprint*, 2024, doi: 10.48550/arXiv.2409.19396.
- [13] A. Vaswani *et al.*, "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.
- [14] M. Caron *et al.*, "Emerging Properties in Self-Supervised Vision Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, Oct. 2021, pp. 9630–9640, doi: 10.1109/ICCV48922.2021.00951.
- [15] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint*, 2016, doi: 10.48550/arXiv.1609.02907.
- [16] P. Subudhi and K. Kumari, "A fast and efficient large-scale near duplicate image retrieval system using double perceptual hashing," *Signal, Image and Video Processing*, vol. 18, no. 12, pp. 8565–8575, Dec. 2024, doi: 10.1007/s11760-024-03490-w.
- [17] M. M. M. Rahman, D. Biswas, X. Chen, and J. Tešić, "Image deduplication using efficient visual indexing and retrieval: optimizing storage, time and energy for deep neural network training," *Signal, Image and Video Processing*, vol. 18, no. 12, pp. 9495–9503, Dec. 2024, doi: 10.1007/s11760-024-03562-x.
- [18] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2014, pp. 1386–1393, doi: 10.1109/CVPR.2014.180.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [20] H. Luo, J. Wei, Y. Wang, J. Chen, and W. Li, "An improved lightweight object detection algorithm for YOLOv5," *PeerJ Computer Science*, vol. 10, p. e1830, Jan. 2024, doi: 10.7717/peerj-cs.1830.
- [21] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint*, 2018, doi: 10.48550/arXiv.1810.04805.
- [22] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986, doi: 10.1109/TPAMI.1986.4767851.
- [23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177.
- [24] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, New York, NY, USA, 2006, pp. 2169–2178, doi: 10.1109/CVPR.2006.68.
- [25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94.
- [26] L. Yang, Y. Gu, and H. Feng, "Multi-scale feature fusion and feature calibration with edge information enhancement for remote sensing object detection," *Scientific Reports*, vol. 15, no. 1, pp. 1–22, May. 2025, doi: 10.1038/s41598-025-99835-7.

**BIOGRAPHIES OF AUTHORS**

**Rahul Shah**     is working as an Assistant Professor in the School of Information Technology at ICFAI University, Sikkim. He holds a Master of Computer Applications (MCA) degree from Shri Ramaswamy Memorial University, Sikkim, obtained in 2017, and is currently pursuing his Ph.D. in Amity School of Engineering and Technology, Amity University Madhya Pradesh, Gwalior. He has 8+ years of teaching experience and has been actively participating in both International and National conferences, where he shares his insights through numerous paper presentations. His research interests span across image processing, artificial intelligence, and cloud computing, reflecting his dedication to exploring cutting-edge technologies. He can be contacted at email: sahrahul77@gmail.com.



**Ashok Kumar Shrivastava**     working as an Associate Professor and academic coordinator in the Department of Computer Science & Engineering at Amity School of Engineering and Technology, Amity University Madhya Pradesh, Gwalior. He graduated in Mathematics from Science College Gwalior. He has done Master of Science in CS branch from Jawaji University, Gwalior. He secured M.Tech. CSE form Uttar Pradesh Technical University, Lucknow. He earned his Ph.D. in Image Processing from NIU, Greater Noida. He is in teaching profession for more than 20 years. He has published 9 Scopus indexed, 2 SCI and 13 Google Scholar papers in National and International Journals, Conference and Symposiums. He has published 04 patents, 03 copyrights, and 2 books. His main area of interest includes noise detection and removal, image enhancement, and machine learning. He can be contacted at email: ashok79.shrivastava@gmail.com.