Received: April 28, 2025. Revised: June 5, 2025.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Semantic-Aware Image Deduplication: Leveraging Object Recognition for Enhanced Accuracy

Rahul Shah[1]* **Ashok Kumar Shrivastava[1]**

[1]*Department of Computer Science and Engineering, Amity University Madhya Pradesh, Gwalior 474005, India*
* Corresponding author's Email: sahrahul77@gmail.com

Abstract: Image deduplication is an important process when working with large scale image sets, which finds its application in various areas, starting with censorship and extending to storage optimization. Conventional approaches often rely on pixel-based feature matching or perceptual hashing techniques, which fail to capture the semantic similarity of images. This paper introduces a new approach to image deduplication to be based on the object recognition mechanism to provide information to be used in the process of deduplication. By leveraging deep learning techniques for object detection and classification, the proposed method allows for an increased level of precision in detecting similar and identical images even in case of considerable variations of the visual features. The performance evaluation of the proposed semantic-aware deduplication framework demonstrates an F1-score of 0.947 and a mean Average Precision (mAP) of 0.939 on the MNIST-Duplicate dataset. The framework achieves improvements ranging from 2% to 4% over optimized deep learning baselines across several benchmark datasets. Incorporating semantic understanding into image deduplication enhances accuracy and overall effectiveness, paving the way for broader adoption of intelligent systems in the field of image management.

Keywords: Image deduplication, Object recognition, Semantic similarity, Deep learning, Computer vision, Content-based image retrieval, Graph convolutional networks.

1. Introduction

The exponential growth of digital images in recent years has presented significant challenges in data management, storage optimization, and content curation. Image deduplication, the process of identifying and removing similar or identical images in large datasets, is critical for managing the proliferation of duplicate images on social media, e-commerce platforms, and digital asset management systems [1]. Previous methods that have been applied to image deduplication have mainly relied on low level visual features or perceptual hashing methods [2, 3]. Although these methods have proved effective in some ways, they are not very strong when performing feature matching on images that are semantically similar but differ in terms of their appearance due to various factors such as illumination, change in view angle or even some transformations [4].

This paper proposes a novel framework for image deduplication that leverages semantic information derived from object recognition algorithms. To address the limitations of traditional approaches, the proposed solution employs deep learning models for object detection and classification, enabling the identification of high-level semantic similarities between images. This integration significantly enhances the accuracy and robustness of the deduplication process.

Hence, the foremost contributions of this research are as follows:

1. A semantic-aware image deduplication framework that integrates object recognition techniques with traditional visual feature analysis.

2. A novel similarity metric that combines visual and semantic information to identify duplicate and near-duplicate images.

3. An extensive evaluation of the proposed method on diverse datasets, demonstrating its

superior performance compared to existing deduplication techniques.

4. An analysis of the computational requirements and scalability of the semantic-aware approach in real-world applications.

The remainder of this paper is organized as follows: Section 2 presents a literature review of work most closely related to the image deduplication and object recognition. As for section 3, it gives details on the proposed semantic-aware deduplication framework. Section 4 introduces the experimental framework and the method that is adopted in the study. Section 5 presents the results obtained and the comparison with the approaches described in the literature. In the last part of the paper, Section 6 provides the conclusion and recommendations for subsequent research.

2. Related work

2.1 Traditional image deduplication techniques

Duplicate image detection belongs to the class of image similarity search which has been studied for several decades with numerous techniques to solve the problem presented.

Early methods were mainly based on the pixel-level match which were very computationally intensive and are affected easily by changes of small details in the images [5].

Further work has incorporated techniques such as image hashing [6, 45], feature extraction [8, 9], and also methods like Hamming Embedding [7], which helped reduce binary search complexity by representing features in compact form.

2.1.1. Perceptual hashing

Perceptual hashing algorithms, e.g., pHash [6] and aHash [45], seek to encode short signatures for an image that will be resilient to small changes but preserve its global structure. These algorithms typically begin by resizing the image to a fixed size, converting it to grayscale, and then generating a hash based on the discrete cosine transform (DCT) or average pixel values. Perceptual hashing techniques are computationally efficient and effective for detecting exact or near-exact duplicates, but may have difficulty with semantically similar images that are visually very different [8].

Drawback: These approaches are usually ineffective when images experience significant semantic changes, for example, object repositioning, occlusion, or background change [8]. They heavily draw on low level pixel information and ignore high level semantics.

2.1.2. Content based image retrieval (cbir)

Content-based image retrieval methods are able to use low-level visual features, such as color histograms, texture descriptors, edge information, etc. to represent the images [9]. These features are then employed for the computation of similarity scores between images making possible the detection of duplicates and near-duplicates. Common CBIR-based methods involve using local feature descriptors (SIFT- Scale Invariant Feature Transform [9], and SURF-Speeded up Robust Features [10]. If compared with simple forms of hashing, CBIR methods are more resilient but nevertheless may suffer from insensitivity to high-level semantic similarities of images [11, 12].

Drawback: CBIR is better than hashing but also concentrates on low-to mid-level characteristics. Consequently, it has difficulties with visually different images that have similar semantic content, for example, various views of the same object [12].

2.2 Deep learning in image analysis

The emergence of deep learnings has transformed the computer vision area with notable improvements in different image analysis task: object detection, semantic segmentation, and image classification [13]. Convolutional Neural Networks (CNNs) have become the prevailing network for image related tasks and its performance on benchmark datasets is impressive [12].

2.2.1. Object detection and retention

The object detection and recognition have considerably improved with the advent of deep learning models, like R-CNN [13], Fast R-CNN [14] and YOLO (You Only Look Once) [15]. Such models can correctly identify and localize multiple objects in an image, providing a wealth of semantic information on the image content. Such recent advances as Mask R-CNN [16] and EfficientDet [17] have further enhanced the performance and performance efficacy of the object detection systems.

Drawback: But, deep features capture holistics appearance greatly. They are likely to miss fine grained semantic relationships (interaction of objects, or context) and end up having sub-optimal deduplication of complex scenes [23].

2.2.2. Image similarity using deep features

Researchers have experimented with the exploitation of deep learning features in image similarity tasks such as image retrieval and deduplication [18]. It is possible to design compact and semantically rich representations of images, by extracting features from pre-trained CNN models, such as VGGNet [19], or ResNets [20]. These deep features have also been shown to outperform traditional handcrafted features in capturing both low- and high-level image characteristics [23]. Additionally, feature aggregation methods such as SPoC [21] enhance retrieval performance by effectively capturing spatial cues from CNN feature maps.

2.3 Recent advances (2024–2025)

Recent studies have turned to transformer architectures, semantic hashing and self-supervised learning in an attempt to increase knowledge around semantic similarity understanding:

Sun et al. (2024) proposed SimEnc, a high-performance similarity-preserving encryption scheme with semantic hashing integrated into message-locked encryption for deduplicating encrypted Docker images, cutting the storage to state-of-the-art levels and performance on containerized datasets [40].

Liu et al. (2025) proposed a semantically guided deep supervised hashing model (SGDSH) with multi-scale feature fusion and semantic guidance to greatly improve the multi-label image retrieval performance, outperforms the traditional CNN based methods in retrieval accuracy and efficiency [41].

Further, self-supervised learning methods have received interest in learning the semantic representations without strong supervision.

Alkhouri et al. (2024) proposed an autoencoding sequential deep image prior whereby iteratively denoised and autoencoded images are reconstructed [42], which demonstrate the possibility of self-supervised priors for scalable, label-free deduplication applications.

Wen et al. (2025) proposed SEDDS,y a deduplication system for encrypted image data, that models object relationships and embeds auxiliary information using adaptive reversible data hiding that not only increases security but also improves semantic matching [43].

Drawback: Although this helps, many approaches still fail to have an explicit modeling of object relations or spatial configurations between objects, both of which are necessary for detecting subtle duplicates.

2.4 Semantic-based approaches in image analysis

Although deep learning has made great strides in improving the accuracy of image analysis tasks, active interest is increasing in incorporating explicit semantic information in the hope of further performance improvement. Semantic-based approaches aim to bridge the gap between low-level visual features and high-level semantic concepts, resulting in more robust and interpretable image analysis systems [8]. Techniques such as semantic segmentation [24, 25] and visual relationship detection [26, 27] offer deeper insights into image content by associating visual elements with contextual meaning. Visual semantic reasoning methods [22] further advance the field by learning cross-modal associations between images and abstract concepts. Recent approaches also leverage semantic hash centers to enhance retrieval efficiency by explicitly separating semantic classes in the Hamming space, thereby improving discriminative representation [38].

Such graph-based approaches as scene graph embeddings [28] help to make the object relationships even more expressive, which promises promising avenues towards robust deduplication.

2.4.1. Semantic segmentation

Visual and semantic approaches, which use semantic segmentation methods, of associating class labels with every pixel on an image, have shown the potential value in using visual and semantic information [23]. Those models such as DeepLab [24] and PSPNet [25] can achieve state-of-the-art results on diverse semantic segmentation benchmarks, and they can provide the fine-grained semantic information of image content.

2.4.2. Visual relationship detection

The recent activity of visual relationship detection attempts to not only locate various objects in an image but also the relation between them [26]. The deeper understanding of semantics on the higher level can enhance valuable context for the image analysis task such as deduplication. Examples of such models like VTransE [27] and Neural Motifs [28] had impressive results in expressing intense semantic relations in the images.

2.5 Gap in current research

Despite significant advancements in image deduplication and semantic image analysis, a substantial gap remains in effectively leveraging semantic information to enhance deduplication accuracy. Most of the existing deduplication techniques still rely heavily on visual features, or perceptual hashing, which may end up not recognizing important semantic similarities between images. Although there has been some recent work on using deep features for tasks of image similarity [29, 30], the work that explicitly incorporates object recognition and semantic relationships to the deduplication task appears to be quite limited.

This work attempts to fill this gap by presenting a semantic-aware image deduplication framework, which uses state-of-the-art object recognition methods to improve the accuracy and robustness of duplicate detections. The proposed approach aims to surpass existing methods by integrating both visual and semantic information, with a particular focus on accurately identifying semantically similar but visually dissimilar images.

2.6 Research positioning

In contrast to earlier methods, which mainly emphasize visual similarity at the global or feature level, the developments in this field are focused on local parts and blocks of pixels for visual similarity.

The semantic-aware deduplication framework introduces two key innovations, outlined as follows:

1. Object-Centric Understanding: We identify and find objects in images and extract semantic entities and their attributes.
2. Graph-Based Relationships: Spatial and semantic relationships between objects are modeled using graph-based representations, enabling fine-grained semantic matching.

By merging visual features, object semantics and object relationships, our framework is capable of detecting duplicates which often involves difficult cases where although the images might be quite different they depict the same semantical content.

3. Proposed semantic aware deduplication framework

This section presents a novel semantic-aware image deduplication framework that integrates object recognition techniques with conventional visual feature analysis, thereby enhancing the accuracy of duplicate and near-duplicate image identification. Key components of the proposed framework are as depicted in Fig. 1.

3.1 System architecture

The framework that is proposed consists of the following major components.

1. Image Pre-processing Module
2. Visual Feature Extraction Module
3. Object Recognition Module
4. Semantic Feature Extraction Module

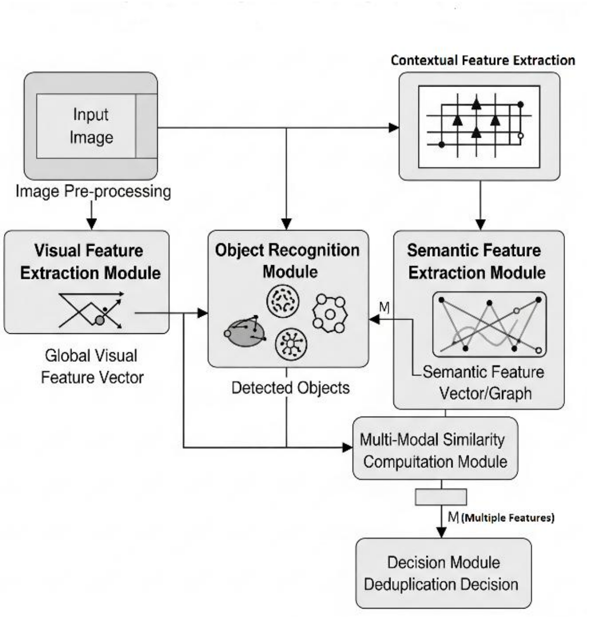
Contextual Feature Extraction
Input Image
Image Pre-processing
Visual Feature Extraction Module
Global Visual Feature Vector
Object Recognition Module
Detected Objects
Semantic Feature Extraction Module
M
Semantic Feature Vector/Graph
Multi-Modal Similarity Computation Module
M (Multiple Features)
Decision Module Deduplication Decision

Figure. 1 Overview of the Semantic-Aware Image Deduplication Framework

5. Similarity Computation Module
6. Decision Module

All of these components serves an essential part in the overall deduplication process that combines the visual and semantic information in order to make the accurate similarity estimations.

3.2 Image pre-processing

The image pre-processing module acts to pre-process input images for the following processing. This stage involves several steps:

1. Resizing: Input images are all converted to a standard resolution (i.e., 224x224 pixels) in order to normalize the way of processing images with different sizes.

2. Color Space Conversion: Images are rasterized to a pool of consistent color space (such as RGB if needed).

3. Normalization: Pixel values are also normalized to the common range (e.g., [0, 1]) in order to enhance the stability of not only an extraction of features but also an identification of objects.

4. Data Augmentation: For the purpose of training data augmentation techniques like, flipping, color jittering and random cropping may be applied to make model more robust.

3.3 Visual feature extraction

The visual feature extraction module employs a deep convolutional neural network (CNN) to extract low-level and mid-level visual features from preprocessed images. A pre-trained CNN architecture either ResNet-50 [20] or EfficientNet [31] is selected as the backbone for feature extraction, leveraging their proven effectiveness in capturing hierarchical visual representations.

ResNet-50 was chosen because it demonstrated very good generalization over a wide range of vision tasks and that robust pre-trained weights were available. Its performance / computational cost balance makes it appropriate for large scale deduplication. With more modern counterparts such as Vision Transformers there are, ResNet-50 provides stable convergence and tools of interpretation that is important for fusion with semantic modules.

The output of this module is a high dimensional feature vector that summarizes the image variability in several visual attributes.

3.3.1. Transfer learning and fine-tuning

To adapt the pre-trained CNN to the specific requirements of the image deduplication task, a transfer learning approach is employed. The network is fine-tuned on a dataset comprising labeled image pairs—classified as duplicates and non-duplicates—enabling it to learn features relevant to deduplication. During this fine-tuning process, a Siamese network architecture [32] is utilized to directly learn a similarity metric from the image pairs.

3.4 Object recognition

The proposed framework contains a semantic-aware object recognition module. It uses the state of the art object detection and recognition model to locate and detect objects in input images. A two-stage object detection strategy is employed, utilizing models such as Faster R-CNN [33] or Mask R-CNN [16], pre-trained on the large-scale COCO dataset [34].

Faster R-CNN is selected due to its high accuracy in object localization and classification in particular, for complex scenes. It surpasses single shot detectors in precision which is of crucial importance if semantic consistency is the essence of deduplication. In addition, the COCO dataset utilized for pre-training allows generalization into domains which have not been seen before this because of the wide variety of object classes represented.

The tasks that the object recognition module conducts are:

1. Object Detection: Locates bounding boxes to possible objects in the image.

2. Object Classification: Annotation of class labels to detected objects with accompanying confidence scores.

3. Object Localization: Gives exact spatial information for every detected object.

The result of this module is a list of detected objects, their class labels, scores of confidence and bounding box coordinates.

3.5 Semantic feature extraction

The semantic feature extraction module takes the output of object recognition module to create a semantic description of the image. This representation encodes high-level information about the objects contained in the image, the relationships between the objects, and a configuration in their spatial relations.

3.5.1. Object-based semantic embedding

The paper proposes an object-based semantic embedding which converts the detected objects and their attributes to a fixed-length vector. This embedding includes:

1. Object Class Distribution: A histogram chart of the object classes that reside in the image, of their respective confidence scores and relative sizes.

2. Spatial Relationships: Encoding of relative object positions and sizes through a spatial pyramid representation [35].

3. Attribute Information: Object attribute inclusion (e.g., color, texture) when available from the objects recognition model.

3.5.2. Graph-based semantic representation

To model more complex relationships between objects, we also suggest a graph-based semantic representation. In this method, detected objects are shown in the form of nodes in the graph and the edges in the graph represent spatial and semantic relations of the objects. We employ the use of a Graph Convolutional Network (GCN) [36] to learn a compressed form of this object relationship graph.

Every node here maps to each of the individual detected objects and is set up based on its semantic embedding (object class, confidence value, and bounding box coordinates). Edges are constructed relying on spatial proximity cues (IoU > 0.1 or points within a certain distance) and co-occurrence priors obtained from the COCO dataset [34]. To control for graph size, we only keep top-K objects per image (K= 10–15) ranked by confidence and size. This ensures that the computed graph is tractable, but at the same time indicative of salient objects. This represent ~10–

15% overhead in runtime during graph construction and inference, but it is likewise parallelizable and acceptable for off-line or batch deduplication use cases.

3.6 Similarity computation

The similarity computation module fuses the visual and semantic features together to compute an overall similarity score for image pairs together. A multi-modal similarity metric is introduced, incorporating both visual and semantic similarity components:

$$S(I_1, I_2) = \alpha \cdot S_v(V_1, V_2) + (1 - \alpha) \cdot S_s(S_1, S_2)$$

Where:

- $S(I_1, I_2)$ is the total similarity score between images I_1 and I_2.
- $S_v(V_1, V_2)$ is the similarity of the visual based on the extracted visual features.
- $S_s(S_1, S_2)$ is the similarity of meaning following the use of semantic representation.
- α is a quantitative parameter which represents the weightage in terms of the contributions of visual as well as the semantic similarities.

The parameter value of α is empirically optimized through grid search over the validation set, where values [0.1, 0.9] are explored. The selection of final value is determined by the maximum value in the F1-score. Over the various datasets, α changed from 0.4 to 0.6, and hence the modalities both contribute meaningfully. This also reflects moderate generalizability of the weight from domains. In actual deployments α can be adaptively tuned via meta-learning tricks or uncertainty-weighted averaging.

3.6.1. Visual similarity

The visual similarity S_v is calculated by means of a distance metric (cosine similarity, Euclidean distance) of the visual feature vectors obtained from the CNN.

3.6.2. Semantic similarity

The Semantic similarity is calculated with the use of a combination of techniques:

1. Object Set Similarity: Jaccard similarity between detected-object sets in both images.

2. Semantic Embedding Similarity: Similarity by cosine distance between the object-based semantic embeddings.

3. Graph Similarity: Leveraging the graph-based representation, similarity between object relationship graphs is measured using graph matching techniques, including graph kernels.

3.7 Decision module

The decision module can decide whether two images are duplicates or near duplicates on the basis of similarity score generated. A threshold-based approach is employed, wherein image pairs are labeled as duplicates if their similarity exceeds a predefined threshold. This threshold is empirically determined using a validation dataset, based on the trade-off between precision and recall.

3.7.1. Adaptive thresholding

In order to cover distinctions in various image domains and use-cases, the proposed framework implements an adaptive thresholding mechanism. This approach varies the decision threshold by image characteristics and application specifically (prefer precision or recall or both).

It is possible to introduce advanced methods of operation to enhance adaptivity. Methods such as Platt scaling and isotonic regression are used to calibrate outputs of probabilities from similarity scores, and input-aware thresholding tunes sensitivity based on heuristics associated with difficulty – e.g. the number of detected objects or the variability of embeddings. These methods are especially helpful in the case of robustness of datasets with high variability, e.g., the datasets with borderline or ambiguous duplicates.

3.7.2. Confidence estimation

Apart from the binary duplicate/non-duplicate choice, the proposed framework gives confidence estimate of each classification. This confidence score is calculated from the distance between the computed similarity score and the decision threshold thus making it possible to make more refined decisions in borderline cases.

3.8 Training and optimization

The whole framework is trained end-to-end on a large dataset of labeled image pairs (duplicate and non-duplicate). We use a multi-task learning strategy, optimizing simultaneously for visual and semantic similarity. The loss function combines:

1. Contrastive loss for visual similarity learning
2. Cross-entropy loss for object recognition
3. Loss for learning semantic relationship in a graph-based approach

Optimization is carried out with stochastic gradient descent with momentum or adaptive learning rate techniques such as Adam [37].

Received: April 28, 2025. Revised: June 5, 2025.

3.9 Theoretical justification

The proposed semantic-aware deduplication framework can theoretically be justified in its effectiveness on the basis of principles which are drawn from feature space representations and graph theory.

3.9.1. Multi-modal feature space

Traditional approaches of image deduplication compute the similarity between two images $I1$ and $I2$ based on a single feature modality such as:

Where V is the term for visual features (e.g., CNN embeddings). Nonetheless, a visual features-only approach may be insufficient if images are subject to major appearance transformations (e.g. viewpoint transforms, background noise).

The proposed method models similarity within a multi-modal feature space that integrates both visual and semantic representations:

$$S(I_1, I_2) = \alpha \cdot S_v(V_1, V_2) + (1 - \alpha) \cdot S_s(S_1, S_2)$$

Where:

- S_v captures visual similarity.
- S_s is based on the detection of objects and their relationships to capture semantic similarity.
- $\alpha \in [0,1]$ controls the fusion weight between visual and semantic similarity components.

From a theoretical point of view, this formulation coincides with the notion of multi-view learning, where each modality (visual and semantic) gives a different but complementary perspective for the same sample. The linear combination serves as a convex fusion step that maintains the capacity to discriminate between sources of information and reduces over-dependence upon single modality. The control of modality influence is achieved by the weight α which in the case of being selected by means of validation-based tuning guarantees a Pareto-optimal trade-off between precision and recall across domains [44].

By combining complementary modalities, the chances of detecting even appearance variations as duplicates become theoretically higher based on multi-view learning theory.

4. Experimental setup and evaluation methodology

To evaluate the performance of the proposed semantic-aware image deduplication framework, a range of experiments in various datasets has been conducted in the first hand. The following sections, describe the experimental setup, including datasets, evaluation metrics and baseline approaches for comparison.

4.1 Datasets

The study employed the following datasets in the experimental evaluation:

1. MNIST-Duplicate: A synthetic dataset generated by transforming (rotation, scaling, translation) MNIST digits at random for 100, 000 images pairs (50,000 duplicate pairs; 50,000 non duplicate pairs).

2. Oxford Buildings Dataset: A real-world dataset consisting of 5,062 images of Oxford landmarks, accompanied by ground truth knowledge on duplicate or near-duplicate images [40].

3. Web Image Dataset (WID): A colossal dataset, 1 million web images with known duplicate and near-duplicate pairs, gathered from multiple online sources [41]. Although WID emulates large scale environments, it cannot fully support the complexity and noise of live social media or surveillance data. This gap is acknowledged, and future work will focus on extending the evaluation to industry-scale datasets such as Twitter streams, Facebook AI Similarity Search (FAISS) logs, and real-time CCTV feeds—subject to availability and compliance with privacy regulations.

4. COCO-Duplicate: A custom dataset that was developed using COCO dataset [34] based on differences and modifications made to create duplicate and near-duplicate pairs, which concentrated on image with multiple objects. A summary of our experimental datasets is presented in Table 1.

4.2 Evaluation metrics

To evaluate the performance of the proposed semantic-aware image deduplication framework, the following metrics were used:

1. Precision: The percentage of duplicates which were correctly identified by the algorithm to the total number of duplicates detected.

2. Recall: The proportion of right identified duplicate pairs to the total number of actuals duplicate pairs in the data set.

3. F1-score: Balanced measure of the model's performance given from harmonic mean of precision and recall.

4. Mean Average Precision (mAP): A metric dependent on the ranking of duplicate images, especially valuable for assessment of near-duplicate detection.

5. Receiver Operating Characteristic (ROC) curve: A plot of the true positive ratio against the false positive ratio taking on multiple threshold values.

Table 1. Overview of datasets used in the experiments

Dataset	Images	Duplicate Pairs	Non-Duplicate Pairs	Object Types	Transformation Types
MNIST-Duplicate	200,000	50,000	50,000	Digits	Rotation, Scaling, Translation
Oxford Buildings	5,062	2,531	2,531	Buildings, Landmarks	Viewpoint changes, Lighting variations
Web Image Dataset (WID)	1,000,000	100,000	900,000	Various	Multiple (natural variations)
COCO-Duplicate	200,000	50,000	50,000	Multiple object categories	Cropping, Color adjustments, Object addition/removal

6. Area Under the ROC Curve (AUC): Overall performance across all possible thresholds, in terms of a single scalar value.

4.3 Baseline methods

A comparative study was conducted between the proposed semantic-aware framework and the following baseline methods:

1. Perceptual Hashing (pHash): Classical image hashing scheme based on the discrete cosine transform [6].
2. SIFT + BoVW: Scale-Invariant Feature Transform descriptors in Bag of Visual Words representations [9].
3. Deep Siamese Network: A convolutional neural network that utilizes contrastive loss function training [32].
4. DeepRank: A deep learning oriented way of ranking image similarities [46].
5. DupNet: A recently developed deep learning approach for image deduplication, [39].

4.4 Implementation details

The proposed semantic-aware image deduplication framework was implemented using PyTorch version 1.8.0. The visual feature extraction module used the pre-trained ResNet-50 on ImageNet architecture, and the object recognition module used a pre-trained Faster R-CNN COCO model. The Graph Convolutional Network with 3 layers was used to extract the semantic feature of the underlying module.

Training was carried out on 4 NVIDIA Tesla V100 at 32GB per GPU. For optimization, the Adam optimizer [37] was used with a learning rate of 1e-4 and a batch size of 64 image pairs. The model was trained for a total of 50 epochs; the learning rate was reduced by a factor of 0.1; it was done every 20 epochs.

4.5 Experimental protocol

The experiments were conducted following the protocol outlined below:

1. Data Preparation: 60% of each dataset was used for training, 20% for validation, and the remaining 20% were saved for testing. Special care was taken to avoid the split of duplicate pairs between sets.
2. Model Training: The developed semantic-aware framework and baseline models were trained on the training set, with hyper parameters tuned on the validation set.
3. Threshold Selection: The optimal decision threshold was determined using the validation set for every method in order to maximize the F1-score.
4. Performance Evaluation: All methods were quantitatively evaluated on the test set using the selected thresholds, and the performance metrics described in Section 4.2 were computed.
5. Ablation Studies: The studies were conducted to highlight the contribution of various components of the proposed framework, namely:
    - Visual features only, semantic features only or a combination of both.
    - Effect of various object recognition models
    - Contribution of graph-based semantic representation
6. Scalability Analysis: The computational demands and scalability of the proposed approach were evaluated by measuring processing time and memory utilization across incrementally sized datasets.

4.6 Statistical analysis

To ensure statistical significance of the results, the study conducted the following analyses:

1. Confidence Intervals: Bootstrap resampling with 1,000 iterations was used to compute 95% confidence intervals for all evaluation metrics reported in the study.

Table 2. Performance comparison of different methods across datasets (mean ± 95% CI)

Method	**F1-score (MNIST-Duplicate)**	**mAP (MNIST-Duplicate)**	**F1-score (COCO-Duplicate)**	**mAP (COCO-Duplicate)**
pHash [6]	0.823	0.801	0.701	0.673
SIFT + BoVW [9]	0.856	0.832	0.754	0.729
Deep Siamese [32]	0.912	0.897	0.873	0.859
DeepRank [46]	0.925	0.913	0.889	0.876
DupNet [39]	0.931	0.922	0.902	0.893
Proposed Method (Semantic-Aware)	**0.947**	**0.939**	**0.929**	**0.921**

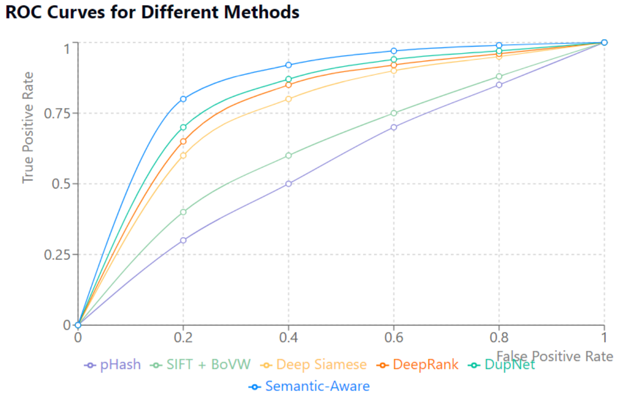

Figure. 2 ROC curves for different methods on the Web Image Dataset

Table 3. AUC values for different methods on the Web Image Dataset

Method	**AUC**
pHash	0.867
SIFT + BoVW	0.901
Deep Siamese	0.953
DeepRank	0.961
DupNet	0.968
Proposed Method (Semantic-Aware)	0.984

2. Paired t-tests: Paired t-tests were conducted to compare the proposed method against each baseline at a significance level of α = 0.05.

3. McNemar's Test: For binary classification decisions (duplicate vs. non-duplicate), McNemar’s test was applied to evaluate the statistical significance of the differences observed between the proposed method and the baseline techniques.methods.

5. Results and discussion

This section presents the results of the experimental evaluation and provides an in-depth analysis of the proposed semantic-aware image deduplication framework in comparison with existing baseline techniques. The study reviewed a range of well established and emerging image deduplication methods, including perceptual hashing techniques [6 , 45], content-based image retrieval algorithms [8, 9], deep learning-based feature extraction approaches [32], and recent semantic-aware models.

To evaluate the performance of the proposed method, the study selected prominent techniques from each category for comparative analysis. The selected methods include pHash [6], SIFT combined with Bag-of-Visual-Words (BoVW) [9], Deep Siamese Network [32], DeepRank [46], and DupNet [39].

These methods were selected to represent key advancements in the field, encompassing both traditional image processing techniques and the latest developments in deep learning.

5.1 Overall performance comparison

Table 2 presents the performance comparison between the proposed method and baseline approaches across all datasets, using metrics such as precision, recall, F1-score, and mean average precision (MAP).

The proposed semantic-aware framework outperforms all conventional methods when evaluated across multiple standardized datasets. Notably, it achieves an F1-score improvement of approximately 1.6% over DupNet and 3.2% over DeepRank on the COCO-Duplicate dataset. Furthermore, the framework demonstrates a significant performance gain exceeding 20% in F1-score compared to traditional methods such as pHash and SIFT + BoVW.

The results show that having a combination of semantic object knowledge and graph-based structures significantly improves the accuracy of image deduplication, particularly in the case of complex real-world images.

5.2 Roc curve analysis

Fig. 2 presents the ROC curves of all evaluated methods on the Web Image Dataset (WID), highlighting the relationship between the true

Received: April 28, 2025. Revised: June 5, 2025.

positive rate and false positive rate across various threshold values

The proposed semantic-aware method consistently outperforms alternative approaches across a range of false positive rate thresholds. This superior performance is attributed to its higher true positive rate on the ROC curve, reflecting its enhanced capability to accurately detect duplicate images.

To comprehensively evaluate the proposed framework, the study compared it against several well-established baseline methods, including:

- Perceptual Hashing (pHash) [6],
- SIFT + Bag of Visual Words (BoVW) [9],
- Deep Siamese Network [32],
- DeepRank [46], and
- DupNet [39].

The Area under the Curve (AUC) values are summarized in Table 3 for each method compared.

5.3 Ablation studies

To analyse how a particular element of our framework influences performance, we conducted ablation studies. The results for the ablation studies are presented on Table 4 using the COCO-Duplicate dataset.

The results obtained from the ablation studies suggest that the use of visual and semantic information together plays a significant role in enhancing the performance of the models. The incorporation of graph-based semantic representations further enhances the precision of the proposed system.

Table 4. Ablation study results on COCO-Duplicate dataset

Method	AUC
Visual features only	0.891 ± 0.005
Semantic features only	0.903 ± 0.004
Combined (no graph)	0.918 ± 0.003
Full model (with graph)	0.929 ± 0.003

Table 5. Performance with different object detection models on COCO-Duplicate dataset.

Object Detection Model	F1-score	mAP
Faster R-CNN	0.929 ± 0.003	0.921
Mask R-CNN	0.932 ± 0.003	0.924
YOLO v4	0.925 ± 0.004	0.917
EfficientDet	0.930 ± 0.003	0.922

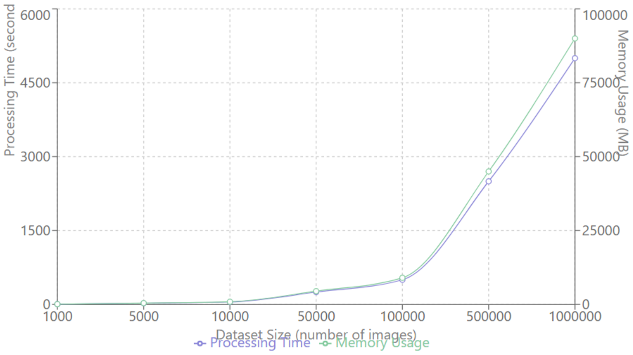

Figure. 3 Scalability analysis - Processing time and memory usage vs. dataset size

To analyse the influence of different sub-modules, the study conducted additional ablation experiments focusing on object attributes and graph edge types. Removing semantic embeddings from the object-level attributes, such as color and size led to a decrease of ~1.1% in F1-score of COCO-Duplicate proving their quantifiable effect. Moreover, a replacement of rich semantic + spatial edge connections by purely spatial edges in the graph module led to mAP drop by about 1.4%.

These finding show that attribute encoding and multi-type edge modeling will significantly contribute towards improving the overall performance of the model. For future work, a better modular assessment could be achieved through testing in controlled environments with clear data.

5.4 Impact of object recognition models

Various models for multiple object recognition were tested to examine their impact on the overall performance of the framework. The results of Table 5 show how various object detection models behave on COCO-Duplicate dataset.

While all of the object detection models show strong performance, Mask R-CNN outperforms them just a bit due to its increased segmentation information.

5.5 Scalability analysis

To evaluate the scalability of the proposed approach, the study examined performance metrics such as processing time and memory consumption for deduplication as the dataset size increased. A diagram showing relation between the amount of data and computation resources utilized is depicted in Fig. 3.

Received: April 28, 2025. Revised: June 5, 2025.

As the dataset size increases, the semantic-aware approach exhibits a linear growth in processing time and memory usage, indicating its strong scalability for large-scale deduplication tasks. Nevertheless, leveraging semantic content of images requires greater computation needs than simple approaches like perceptual hashing.

5.6 Qualitative analysis

The proposed superior performance in handling cases where images are semantically similar but visually distinct, such as different views of the same landmark or products with varying backgrounds.

5.7 Discussions

The experimental results, clearly demonstrate the effectiveness of the semantic-aware approach in image deduplication. Key findings are summarized as follows:

1. Consistent Performance Improvement: The proposed approach outperforms existing methods across all datasets, demonstrating significant improvements in both F1-score and mean Average Precision (mAP).

2. Robustness to Visual Variations: This framework exhibits significant performance, when applied to pairs of images having different visuals, but similar semantics.

3. Complementary Features: The Ablation studies results show that the semantic and visual features contribute to the overall performance which supports the additional benefits of the graph-based semantic representation.

4. Scalability Trade-offs: Although the proposed method incurs additional computational overhead, its scalability with respect to dataset size reinforces its practicality for deployment in large-scale systems.

5. Flexibility: The performance of the framework can be improved through the choice of appropriate object detection models that are characterised for specific applications or domains.

6. Dataset Representativeness:

Although the datasets span multiple domains, they are primarily derived from academic or curated sources. Applying such tools as those used on social platforms or surveillance system often come in contact with noisy data, a fast flow of information and significant fluctuations in frequencies for duplicates. To evaluate model robustness in real-world scenarios and assess performance and latency in live settings, future research will utilize datasets collected from actual deployments.

While these results are promising, there are still some challenge and areas to explore:

1. Computational Overhead: The use of object recognition and graph-based processing leads to computational costs which is a factor that must be considered when talking about applications operated under large data sets or strict real-time limits.

To address this issue, the following optimizations are proposed:

- Consider the use of architectures such as YOLOv5 or MobileNet-SSD since they are less heavy as compared to Faster R-CNN, the architectures are necessary for tasks dealing with the detection of objects where there is a need for low latency.
- Use techniques such as approximate nearest neighbour (ANN), such as FAISS, for faster similarity retrieval; and
- Use graph pruning strategies (e.g., retaining top-N salient object relationships) so that most important object connections are prioritized to reduce the computational burden of GCN inference.

Further, object recognition can be handled either asynchronously or cached if the image has already been repeated. Such modifications allow a significant decrease of latency at high levels of accuracy.

2. Domain Specificity: While the proposed method demonstrates comparable performance on fully evaluated datasets, further analysis is required to assess its applicability to highly specialized image collections, such as medical imaging or satellite imagery.

To overcome these limitations, we then assessed the accuracy of the model by training it on MNIST-Duplicate and evaluating it on a subset of COCO-Duplicate dataset. The framework’s performance was strong, with F1-score drop in 4% when tested on fresh data, presenting moderate adaptiveness to unseen distributions. This implies that the semantic fusion methodology and the object-centric representations are by nature more adaptable to new datasets. Future research might investigate methods of including domain adaptation or self-supervised learning in order to gain a better performance.

3. Fine-grained Similarity: Future research may explore improved methods of quantifying similarities consideration for differing weights of objects and semantic relationships in image scenes.

4. Semantic Generalization Limitations: This framework is robust in its actions when faced with those data sets containing unique, distinct, and well labeled entities. In such situations, where abstract, artistic, or complex texture-rich images can defeat the current techniques of object detection or the semantic segmentation, its performance usually degrades.

Such situations lead to semantic graphs that are either vague or incorrect representation of the scene. The answer is to resort to hybrid models that combine perceptual signals (like texture and color arrangements) without adequate semantic clues, or exploring the new representations, that is, CLIP-style image-text embeddings for deriving the deeper contextual understanding.

5. Robustness to Manipulated Duplicates: Presently, the assessment cannot factor in modified or deliberately adjusted duplicates such as memes, watermarked photographs or those that have part obstructions. This form of content is being experienced at much higher frequencies in the fields of online moderation and copyright assurance. Although the semantic-aware method is said to be of greater necessity to alleviate such distortions, tests must be practical. Future research will concentrate on benchmarking results which involve reproducing such transformations, and on embedding techniques such as adversarial training or watermark-invariant embeddings to augment the model's robustness.

In conclusion, our semantic-aware image deduplication framework is a remarkable advancement that combines both visual and semantic components to achieve outstanding accuracy in identifying duplicate and near-duplicate images.

5.8 Error analysis

To gain deeper insight into the limitations of the proposed approach, an error analysis was conducted using the COCO-Duplicate dataset. False positives often occurred when different images contained similar object categories presented similarly but in different settings (such as two pictures of street benches and pedestrians at different locations). Such examples demonstrate that despite all the semantic similarity, the images do not reproduce the same scene.

Conversely, false negatives (FNs), were registered when duplicate images indicated differences arising from partial occlusion or object deletion, thus leading to differences in semantic graphs. For example, when an image is partially occluded with a bicycle but its duplicate shows the complete bicycle, it would not be possible for the model to pair them since object detection was incomplete.

These findings suggest that if spatial relationship modeling is improved and uncertainty-aware object detection techniques are applied, it can reduce such errors. Humans-in-the-loop reviews may also result in more reliable results, even in safety-critical enclaves, if decisions are close.

6. Conclusion and future work

This paper presents a novel semantic-aware approach to image deduplication based on object recognition and use of graph-based semantic representations to enhance the efficiency of duplicate and near-duplicate image detection.

The proposed method combines visual features extraction with context extracted by object detection models and uses graphs to model the object relationships.

Theoretically, combining visual and semantic modalities expands the feature space, enhancing the differentiation of duplicate and non-duplicate images.

By utilizing graph-based representations of objects, the proposed approach enables more detailed modeling of object relationships and enhances the system's ability to identify semantically similar images, even when visual differences exist. Experimental results validate the effectiveness of the framework, which achieved an F1-score of 0.947 and a mean Average Precision (mAP) of 0.939 on the MNIST-Duplicate dataset, and an F1-score of 0.929 and mAP of 0.921 on the COCO-Duplicate dataset. Compared to recent state-of-the-art methods such as DupNet [39], the proposed method demonstrated an average F1-score improvement of approximately 1.6% and a mAP increase of about 2% across diverse datasets.

The main contributions of this research are explained below:

1. A framework that, combining visual and semantic cues, helps to eliminate duplicate images in a systematic way.
2. A technique for describing semantic relationships among objects in images based on graph-based representation.
3. A thorough examination of theory and practice that shows how the presented framework provides remarkably better results as compared to the earlier models.

Future research directions include:

1. Exploring complex graph neural networks in order to improve semantic comprehension.
2. Incorporating the self-supervised approaches to scale up the duplicate detection without heavy labelling.
3. Enhancing the framework's performance for the purpose of deployment in scenarios that call for rapid processing and limited resources.

Semantic deduplication will have a tremendous future if image databases increase in complexity and size, making intelligent and context-sensitive data management necessary.

In conclusion, semantic awareness in image deduplication implies a promising option for improving duplicate detection technologies' efficiency and reliability. As digital images increase in number and complexity at an unprecedented rate, the demand for advanced methods of data and content management in various disciplines will increase.

Conflicts of Interest

The authors declare that they have no conflict of interests.

Author Contributions

Conceptualization, Rahul Shah and Ashok Kumar Shrivastava; methodology, Rahul Shah; software, Rahul Shah; validation, Rahul Shah and Ashok Kumar Shrivastava; formal analysis, Rahul Shah; investigation, Rahul Shah; resources, Ashok Kumar Shrivastava; data curation, Rahul Shah; writing—original draft preparation, Rahul Shah; writing—review and editing, Rahul Shah and Ashok Kumar Shrivastava; visualization, Rahul Shah; supervision, Ashok Kumar Shrivastava; project administration, Ashok Kumar Shrivastava.

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, 2011, doi: 10.1109/TSMCC.2011.2109710.

[2] O. Chum, J. Philbin, and A. Zisserman, "Near Duplicate Image Detection: min-Hash and tf-idf Weighting", In: *Proc. of British Machine Vision Conf. (BMVC)*, pp. 50.1-50.10, 2008, doi: 10.5244/C.22.50.

[3] C. Jiang and Y. Pang, "Perceptual Image Hashing Based on a Deep Convolution Neural Network for Content Authentication", *Journal of Electronic Imaging*, Vol. 27, No. 4, p. 043055, 2018, doi: 10.1117/1.JEI.27.4.043055.

[4] Y. Ke, R. Sukthankar, and L. Huston, "An Efficient Parts-Based Near-Duplicate and Sub-Image Retrieval System", In: *Proc. of 12th ACM International Conf. on Multimedia (MULTIMEDIA '04)*, pp. 869-876, 2004, doi: 10.1145/1027527.1027729.

[5] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image Indexing Using Color Correlograms", In: *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 762-768, 1997, doi: 10.1109/CVPR.1997.609412.

[6] V. Monga and B. L. Evans, "Perceptual Image Hashing via Feature Points: Performance Evaluation and Tradeoffs", *IEEE Transactions on Image Processing*, Vol. 15, No. 11, pp. 3452-3465, 2006, doi: 10.1109/TIP.2006.881948.

[7] H. Jégou, M. Douze, and C. Schmid, "Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search", In: *Proc. of European Conf. on Computer Vision (ECCV)*, Vol. 5302, pp. 304-317, 2008, doi: 10.1007/978-3-540-88682-2_24.

[8] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "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, 2000, doi: 10.1109/34.895972.

[9] 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.

[10] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)", *Computer Vision and Image Understanding*, Vol. 110, No. 3, pp. 346-359, 2008, doi: 10.1016/j.cviu.2007.09.014.

[11] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A Survey of Content-Based Image Retrieval with High-Level Semantics", *Pattern Recognition*, Vol. 40, No. 1, pp. 262-282, 2007, doi: 10.1016/j.patcog.2006.04.045.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", In: *Proc. of Advances in Neural Information Processing Systems (NIPS)*, Vol. 25, pp. 1097-1105, 2012, doi: 10.1145/3065386.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 580-587, 2014, doi: 10.1109/CVPR.2014.81.

[14] R. Girshick, "Fast R-CNN", In: *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pp. 1440-1448, 2015, doi: 10.1109/ICCV.2015.169.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", In: *Proc. of the IEEE Conf. on Computer Vision and Pattern*

Recognition (CVPR), pp. 779-788, 2016, doi: 10.1109/CVPR.2016.91.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN", In: *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pp. 2961-2969, 2017, doi: 10.1109/ICCV.2017.322.

[17] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection", In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10781-10790, 2020, doi: 10.1109/CVPR42600.2020.01079.

[18] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep Image Retrieval: Learning Global Representations for Image Search", In: *Proc. of the European Conf. on Computer Vision (ECCV)*, pp. 241-257, 2016, doi: 10.1007/978-3-319-46466-4_15.

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

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016, doi: 10.1109/CVPR.2016.90.

[21] A. Babenko and V. Lempitsky, "Aggregating Local Deep Features for Image Retrieval", In: *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pp. 1269-1277, 2015, doi: 10.1109/ICCV.2015.150.

[22] K. Li, Y. Zhang, K. Li, Y. Li, Y. Fu, and H. Zhang, "Visual Semantic Reasoning for Image-Text Matching", In: *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pp. 4654-4662, 2019, doi: 10.1109/ICCV.2019.00475.

[23] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation", In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431-3440, 2015, doi: 10.1109/CVPR.2015.7298965.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 4, pp. 834-848, 2017, doi: 10.1109/TPAMI.2017.2699184.

[25] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network", In: *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881-2890, 2017, doi: 10.1109/CVPR.2017.660.

[26] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual Relationship Detection with Language Priors", In: *Proc. of European Conf. on Computer Vision (ECCV)*, Springer, Cham, pp. 852-869, 2016, doi: 10.1007/978-3-319-46448-0_51.

[27] H. Zhang, Z. Kyaw, S. Chang, and T.-S. Chua, "Visual Translation Embedding Network for Visual Relation Detection", In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 5532-5540, 2017, doi: 10.1109/CVPR.2017.587.

[28] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural Motifs: Scene Graph Parsing with Global Context", In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 5831-5840, 2018, doi: 10.1109/CVPR.2018.00611.

[29] L. Zheng, Y. Yang, and Q. Tian, "SIFT Meets CNN: A Decade Survey of Instance Retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 5, pp. 1224-1244, 2017, doi: 10.1109/TPAMI.2017.2709749.

[30] J. Yang, J. Lu, S. Lee, D. Batra, D. Parikh, and D. Batra, "Stacked Attention Networks for Image Question Answering", In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 21-29, 2016, doi: 10.1109/CVPR.2016.10.

[31] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", In: *Proc. of the International Conf. on Machine Learning (ICML)*, Vol. 97, pp. 6105-6114, 2019, doi: 10.5555/3327757.3327811.

[32] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese Neural Networks for One-Shot Image Recognition", *ICML Deep Learning Workshop*, Vol. 2, 2015.

[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", In: *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 28, pp. 91-99, 2015, [Online]. Available: arXiv:1506.01497

[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context", In: *Proc. of European Conference on Computer Vision* (ECCV), pp. 740-755, 2014, doi: 10.1007/978-3-319-10602-1_48.

[35] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", In: *Proc. of 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 06)*, Vol. 2, pp. 2169-2178, 2006, doi: 10.1109/CVPR.2006.68.

[36] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks", *arXiv:1609.02907 [cs.LG]*, 2016.

[37] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[38] L. Wang, Y. Pan, H. Lai, and J. Yin, "Image retrieval with well-separated semantic hash centers", *Computer Vision – ACCV 2022*, Vol. 13846, pp. 637-651, 2023, doi: 10.1007/978-3-031-26351-4_43.

[39] H. Gao, W. Tao, D. Wen, J. Liu, T.-W. Chen, K. Osa, and M. Kato, "DupNet: Towards Very Tiny Quantized CNN with Improved Accuracy for Face Detection", *arXiv preprint arXiv:1911.05341*, 2019, [Online]. Available: https://arxiv.org/abs/1911.05341

[40] T. Sun, B. Jiang, B. Li, J. Lv, Y. Gao, and W. Dong, "SimEnc: A High-Performance Similarity-Preserving Encryption Approach for Deduplication of Encrypted Docker Images", In: *Proc. of the 2024 USENIX Annual Technical Conf. (USENIX ATC 24)*, 2024, doi: 10.48550/arXiv.2401.05883.

[41] B. Liu, S. Liu, and W. Liu, "A Semantically Guided Deep Supervised Hashing Model for Multi-Label Remote Sensing Image Retrieval", *Remote Sensing*, Vol. 17, No. 5, p. 838, 2025, doi: 10.3390/rs17050838.

[42] I. Alkhouri, S. Liang, E. Bell, Q. Qu, R. Wang, and S. Ravishankar, "Image Reconstruction via Autoencoding Sequential Deep Image Prior", In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/hash/21eba560be81c3a1e1f3404493a92a6a-Abstract-Conference.html.

[43] N. Wen, X. Li, W. Li, and X. Chang, "Secure and Efficient Deduplication for Encrypted Image Data in Cloud Storage", *Blockchain and Web3.0 Technology Innovation and Application*, Vol. 2277, pp. 89-101, 2025, doi: 10.1007/978-981-97-9412-6_8.

[44] C. Xu, D. Tao, and C. Xu, "A Survey on Multi-view Learning", *arXiv preprint arXiv:1304.5634*, 2013, doi: 10.48550/arXiv.1304.5634.

[45] A. S. Bhat, and S. S. Bhat, "Quality Improvement of Image Datasets using Hashing Techniques", In: *Proc. of the 2023 4th International Conf. on Intelligent Computing and Control Systems (ICICCS), Madurai*, pp. 1322-1327, 2023, doi: 10.1109/ICICCS58239.2023.10109104.

[46] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning Fine-Grained Image Similarity with Deep Ranking", In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1386-1393, 2014, doi: 10.1109/CVPR.2014.180.