Journal of Information Hiding and Multimedia Signal Processing ©2026 ISSN 2073-4212
Ubiquitous International Volume **17**, Number **1**, March **2026**

Multi-Label Image Retrieval Based on Semantic-Aware Representation Learning and Graph Attention Networks

Thanh Nguyen Van[1], Hai Do Van[2,*], Son Nguyen Thanh[3], Long Phan Phuoc[3], Sang Ha Van[3]

[1]Lecturer at Academy of Finance; Ph.D. Student at Thuyloi University, Hanoi, Vietnam
nguyenvanthanh@hvtc.edu.vn
[2]Thuyloi University, 175 Tay Son, Hanoi, Vietnam, Hanoi, Vietnam
haidv@tlu.edu.vn
[3]Academy of Finance, 58 Le Van Hien, Dong Ngac, Hanoi, Vietnam
sonnt@hvtc.edu.vn, phanphuoclong@hvtc.edu.vn, sanghv@hvtc.edu.vn

*Corresponding author: Hai Do Van

Received August 14, 2025, revised November 24, 2025, accepted December 1, 2025.

Abstract. *Multi-label image retrieval (MLIR) is more challenging than its single-label counterpart, as it must capture complex spatial and semantic dependencies among multiple co-occurring objects. Existing approaches often rely on global CNN features or static label graphs, which fail to explicitly model object-level relations and adapt to image-specific contexts, leading to suboptimal retrieval performance. To overcome these limitations, we propose a novel method for* ***M****ulti-****L****abel* ***I****mage* ***R****etrieval based on* ***S****emantic-****A****ware* ***R****epresentation* ***L****earning and* ***G****raph* ***At****tention Networks, called* ***MLIR-SARL-GAT****. Object-level visual features are extracted and paired with corresponding label embeddings to form semantic nodes, while edges are adaptively computed to reflect image-specific relationships. A Graph Attention Network (GAT) is then employed to weigh inter-object connections, enabling the model to selectively focus on the most informative semantic and spatial dependencies, that something fixed-graph or purely CNN-based methods cannot achieve. Trained in a multi-label classification setting, the learned graph-level embeddings are directly reused for retrieval without extra supervision. Experiments on MS-COCO and PASCAL VOC show that MLIR-SARL-GAT consistently outperforms state-of-the-art methods in both classification and retrieval, particularly for images with multiple, overlapping object categories.*

Keywords: multi-label image retrieval, graph attention network, semantic representation, object-level reasoning

1. **Introduction.** Multi-label images often depict complex scenes in which multiple objects co-exist, varying significantly in number, location, size, color, and visual appearance. Retrieving relevant images from a multi-label image database, referred to as multi-label image retrieval (MLIR), requires matching a query image with database images that share semantically similar content, particularly in terms of object co-occurrence and contextual relationships.

As a foundational task in computer vision, MLIR plays a crucial role in enabling higher-level applications such as image localization, segmentation, attribute recognition, and recommendation systems. With the advancement of deep learning, convolutional neural networks (CNNs) have achieved notable progress in image representation and retrieval. However, the semantic complexity and label entanglement in multi-label images present significant challenges for traditional CNN-based retrieval systems. Specifically, these models struggle to capture the rich contextual dependencies among co-occurring objects, particularly when object appearances vary or overlap. Moreover, global feature representations tend to overlook object-level semantics and inter-object relations, which are critical for accurate retrieval in complex scenes. For example, two images might both contain "person" and "bicycle," but differ significantly in their contextual relationships or background scenes, which global features struggle to distinguish. In [1], Hend A. Elsayed et al. proposed a new method to analyze multiple multimedia elements, including images, texts and graphics to enhance the accuracy in detecting small objects. Recent studies have sought to enhance image representation through attention mechanisms and graph-based modeling. For instance, self-attention has proven effective in emphasizing discriminative regions within an image [2, 3, 4]. Wang et al. [3] proposed a method to recurrently detect attention regions using RNNs, while Guo et al. [4] introduced visual attention consistency to reinforce important regions across image transformations. However, these methods primarily focus on visual saliency, without explicitly modeling the inter-label dependencies that are vital for multi-label understanding. To address this, Chen et al. [5] introduced ML-GCN, a graph convolutional network where labels are embedded into a graph structure and relational dependencies are modeled through label co-occurrence. Building upon this, Ye et al. [6] proposed ADD-GCN, which dynamically constructs image-specific graphs and integrates object detection, enabling better context modeling. More recently, approaches such as Patch-GAT and semantic-optimal transport methods [7] have sought to align visual features with semantic embeddings, enhancing semantic consistency in classification and retrieval tasks. Although the above methods have achieved remarkable improvements in the multi-label image retrieval task, these methods still have limitations such as: object structure modeling is underutilized, as many models rely on global features or static label graphs, failing to adapt to the unique object compositions of each image; label relations are typically static, derived from global co-occurrence statistics, which are insufficient for rare or novel object combinations in real-world scenes; visual-semantic alignment remains incomplete, due to the inherent gap between CNN feature spaces and semantic embeddings (e.g., word vectors), which can hinder accurate semantic matching during retrieval.

To overcome these limitations, we propose a unified framework called MLIR-SARL-GAT. Our method first applies an object detector (e.g., YOLOv8) to extract object-level visual features and their associated label embeddings. We then construct a dynamic semantic graph for each image, where nodes combine visual and semantic information, edges are adaptively learned through a Graph Attention Network (GAT) to model contextual relationships among objects. The resulting graph-level embedding is jointly optimized for multi-label classification and subsequently reused as a semantic descriptor for image retrieval. Multi-label image features learned through classification are used to measure similarity for retrieval.

Our main contributions include:

- A semantic-aware representation framework that leverages GAT for object-level reasoning, capturing both spatial and semantic dependencies in multi-label images.
- A dynamic graph construction mechanism that builds a semantic graph per image, instead of relying on fixed co-occurrence statistics.
- Demonstration that classification embeddings can directly support retrieval, eliminating the need for separate feature learning for CBIR, and yielding competitive performance in both tasks.

The rest of the paper is presented as follows: Section 2 we present related studies on multi-label image retrieval using representation learning and graph attention networks. Section 3 presents our proposed method. Finally the experiments and results are described in section 4. Conclusions are given in section 5.

2. **Related Works.** The task of multi-label image analysis has garnered substantial attention due to the complexity of modeling images containing multiple co-occurring objects. Early approaches attempted to model semantic connection-based learning aimed at creating separate classification spaces for each label [8], combining multi-label learning with contrastive learning to multi-labl image recognition [9] or model label dependencies using recurrent neural networks (RNNs), treating label prediction as a sequential process [10]. These methods learn to refine label predictions over time by capturing contextual dependencies, but they struggle with modeling unordered label sets and complex visual interactions among objects.

With the success of Graph Convolutional Networks (GCNs) in non-Euclidean data learning [11], numerous studies have proposed building label graphs to encode semantic or statistical correlations between object categories. Chen et al. [5] introduced ML-GCN, which constructs a global label graph where nodes are initialized with semantic embeddings (e.g., word vectors), and edges are derived from co-occurrence statistics. Although effective, such static graphs fail to capture image-specific object relationships and may overlook rare or coarsely labeled objects.

To address this limitation, several studies have proposed dynamic graph construction, where the label or object graph is tailored per image. Ye et al. [6] introduced ADD-GCN, an attention-driven framework that dynamically generates a graph based on detected objects and contextual relevance, significantly improving classification robustness. Yuan et al. [7] extended this idea by aligning patch-level visual features with semantic label embeddings using optimal transport and graph attention transformers, leading to improved multi-label classification and retrieval.

Parallel to these efforts, research on semantic-aware representation learning has grown rapidly. Instead of relying solely on global CNN embeddings, semantic-aware methods aim to bridge the gap between low-level visual features and high-level semantic concepts. Works such as PatchCT [8] and TDRG [12] emphasize aligning features from image patches with corresponding label embeddings, leveraging semantic consistency to enhance representation quality. These approaches often integrate natural language embeddings (e.g., GloVe or BERT) with vision features, enabling cross-modal understanding and more precise label prediction.

The Transformer architecture [13], initially developed for sequence modeling, has also demonstrated strong potential in multi-label image classification. For example, DETR [14] redefined object detection as a set prediction task using self-attention. TDRG [12] leverages Transformer encoders to capture long-range dependencies and incorporate dual relation graphs—one for visual relations and another for semantic label relations—leading to enhanced semantic-awareness in feature learning. Despite these advances, existing methods still exhibit limitations: (i) static label graphs lack flexibility to adapt to scene-specific object interactions; (ii) global CNN features do not capture fine-grained spatial structure among objects; and (iii) visual-semantic embedding misalignment limits the discriminability of learned features, especially for retrieval.

Recently, Graph Attention Networks (GATs) have emerged as a promising alternative, enabling the model to assign attention weights dynamically to neighbor nodes during message propagation [15]. This flexibility allows GAT-based models to capture contextual relationships more accurately and to learn task-adaptive representations. When applied to multi-label learning, GATs can model interactions between detected objects or between visual and semantic embeddings—opening new opportunities for object-centric and label-aware image retrieval systems. In this paper, we propose a novel framework that integrates these insights by combining object detection, semantic-aware graph construction, and GAT-based reasoning. Unlike previous works that use fixed label graphs or rely solely on visual cues, our model constructs a per-image dynamic graph where each node represents a fusion of object-level features and label embeddings. The use of GAT enables our model to adaptively learn contextual relations, thereby improving both classification and retrieval performance.

3. Proposed Method.

We propose a unified framework called MLIR-SARL-GAT, which jointly tackles multi-label image classification and retrieval by learning semantic-aware representations. Our method involves two main stages: SARL_GAT, a graph-based feature learning module that models object-level semantics using attention mechanisms, and MLIR, a retrieval process that reuses the learned embeddings to identify similar images. In this section, we will present each component in detail.

3.1. Semantic-Aware Representation Learning with Graph Attention.

As shown in the first frame, top of Figure 1, the first step in our semantic-aware representation learning pipeline involves detecting salient objects in each input image. This is crucial for transforming the image from a global grid of pixels into a structured set of meaningful entities, each of which becomes a node in the later graph construction phase. To achieve this, we use a state-of-the-art object detector, we choose the YOLOv8 [16] over prior detectors (Faster R-CNN [17], YOLOv3 [18], DETR [14]) due to YOLOv8 is an anchor-free detector that predicts bounding boxes, object class probabilities, and confidence scores in a single forward pass through the network. Its streamlined architecture and decoupled head design enable fast inference and robust localization of multiple object instances, which is essential in multi-label scenes with overlapping or small-scale objects. And it has demonstrated competitive performance in both speed and accuracy on benchmarks such as MS-COCO and PASCAL VOC.

For each image $\mathcal{I} \in \mathcal{R}^{\mathrm{H}\times\mathrm{W}\times 3}$, YOLOv8 outputs a set of $\mathcal{N}$ object detections: $\{(b_i, c_i, s_i)\}_{i=1}^{N}$ where:

- $b_i = (x_i, y_i, w_i, h_i)$ denotes the bounding box coordinates of the i^{th} object (center, width, height)
- $c_i \in \{1, ..., C\}$ is the predicted class label among C predefined categories
- $s_i \in [0, 1]$ is the confidence score for the detection.

We apply non-maximum suppression (NMS) to remove redundant overlapping boxes. Detections with confidence scores below a threshold τ (e.g., 0.25) are filtered out to reduce noise. Each retained object region, the bounding box coordinates are used to extract local features from the global feature F, the object labels will be mapped to a semantic embedding vector through a pre-trained text model (e.g.

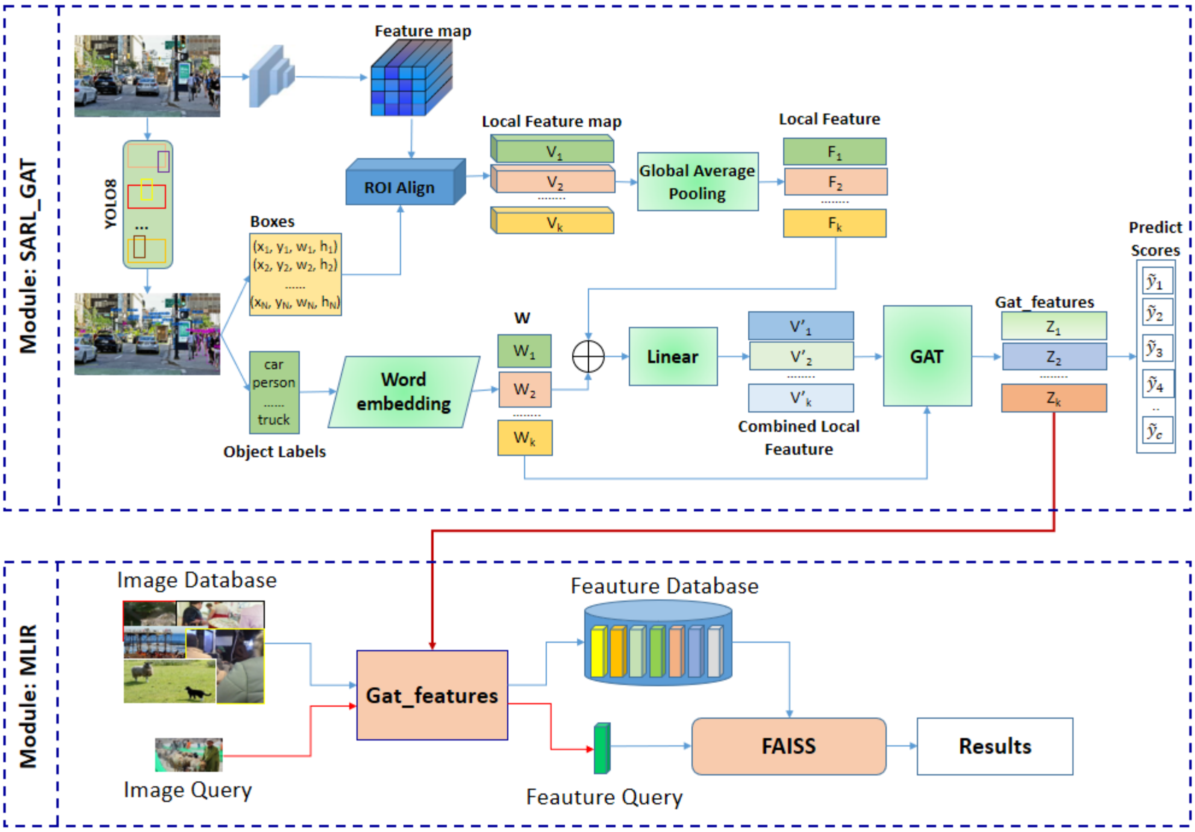

Figure 1. General diagram of the MLIR-SARL-GAT method

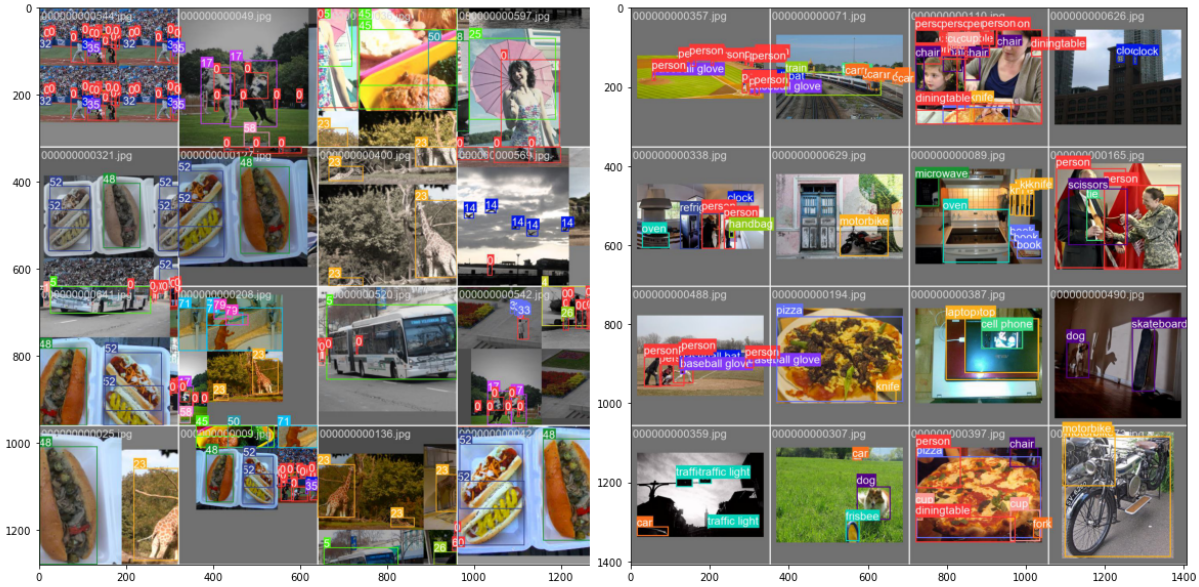

Figure 2. Visualization of object detection results using YOLOv8 on the MS COCO dataset

GloVe, BERT). 2 illustrates the visualization of the process of using YOLOv8 to detect objects with corresponding boxes, ids and labels on 16 random images taken from the MS COCO dataset.

To enable object-centric reasoning in multi-label scenes, we extract both global contextual features and localized visual features corresponding to detected objects. This dual representation ensures that the model benefits from high-level scene understanding as well as fine-grained region-specific details

3.1.1. *Global and local features representation:* Global feature representation: Each input image $\mathcal{I}$ is adjusted to a resolution of 448 × 448 and passed through a backbone convolutional neural network (CNN) for deep visual encoding. Specifically, we use ResNet101 [19] to extract the global feature map. We remove the linear layer of ResNet101 and obtain the output of "conv5_x", generating $2048 \times 14 \times 14$ feature maps:

$$F = \text{ResNet101}(\mathcal{I}) \in \mathbb{R}^{C \times H \times W}.$$

where $C = 2048$, and H, W denote the spatial dimensions 14×14. This tensor captures high-level semantic abstraction while preserving spatial localization. The global feature map F serves as a reference for deriving localized object-specific features

Following the object detection step using YOLOv8 [16], we obtain a set of bounding boxes $\{b_i\}_{i=1}^{N}$, each associated with a predicted class label c_i. To extract visual descriptors aligned with each object proposal, we apply RoIAlign [20] over the global feature map. This operation extracts a fixed-size region (e.g., $2048 \times 7 \times 7$) corresponding to the spatial extent of the detected object:

$$\mathrm{V}_i = \mathrm{RoIAlign}(\mathrm{F}, \mathrm{b}_i).$$

We then apply global average pooling on each V_i to obtain a compact local visual feature vector:

$$\mathrm{F}_i = \mathrm{GAP}(\mathrm{V}_i).$$

This feature encapsulates the visual content within the object's bounding box, contextualized by the global scene representation learned through the CNN.

3.1.2. *Semantic embedding and node vector construction.* Each detected object is assigned a predicted label c_i, which is mapped to a semantic embedding vector $\mathrm{w}_i \in R^300$ using pre-trained GloVe embeddings [21]. These embeddings provide semantic understanding of class labels based on large-scale text corpora. The visual and semantic features are concatenated to form an intermediate feature:

$$\hat{\mathrm{F}}_\mathrm{i} = [\mathrm{F}_i \parallel \mathrm{w}_i] \in R^{2348}. \tag{1}$$

To harmonize the concatenated feature and project it into a shared latent space, we apply a linear transformation followed by a non-linear activation, yielding the final node representation:

$$\mathrm{V}_i^{'} = \sigma(\mathrm{W}\hat{\mathrm{F}}_\mathrm{i} + \mathrm{b}) \in R^d \tag{2}$$

where $\sigma(\cdot)$ is an activation function (e.g., ReLU or Sigmoid), and d is a tunable dimension is a tunable hyperparameter (e.g., d=512 or d=1024) This transformation ensures that the resulting vector $V_i^{'}$ not only encodes both appearance and label semantics, but is also normalized and aligned in a space suitable for attention-based message passing in the Graph Attention Network

3.1.3. *Semantic Graph Construction via Optimal Transport.* After extracting object-level descriptors $\mathrm{V}_\mathrm{i}^{'} \in R^d$, for each detected object (which combine local visual features F_i and label embeddings w_i), we construct a semantic-aware graph to model contextual relationships among objects. This graph serves as the foundation for learning attention-based representations using Graph Attention Networks. We construct a semantic-aware graph $G = (V, E)$:

- **Nodes** $V = \{u_1, u_2, , ...u_N\}$: each node corresponds to an object and contains its fused visual-semantic representation
- **Edges** $E \subseteq V \times V$: srepresent semantic relationships between objects, dynamically computed using optimal transport.

The use of word embeddings, such as GloVe vectors [21], allows the graph to encode semantic proximity among object classes. For example, the semantic distance between "dog" and "cat" is smaller than between "dog" and "car," allowing the model to attend to semantically coherent groupings during message passing. This design embeds a priori semantic structure into the graph and helps guide attention to meaningful object interactions. To compute the adjacency matrix $A \in \mathcal{R}^{NN}$, we adopt an Optimal Transport (OT)-based strategy, inspired by recent works [6, 7], where each visual feature F_i is matched against all class label embeddings $\{w_j\}_{j=1}^{C}$, where C is the total number of known classes (e.g., 80 for COCO dataset). This process defines a semantic cost matrix:

$$C_{ij} = 1 - \cos(F_i, w_j) \tag{3}$$

where cos(.,.) denotes cosine similarity. OT then finds a minimal-cost mapping between detected object features and the label space, effectively modeling how closely the visual appearance of an object aligns with known semantic categories.

Edges A_{ij} between detected objects u_i and u_j are defined based on the alignment of their respective visual features with shared or related label embeddings. This mechanism ensures that the graph is dynamically adapted to the semantic content of each image.

We construct node embeddings: $u_1 = \phi([F_1 \| w_1])$, $u_2 = \phi([F_2 \| w_2])$, where $\phi(.)$ is a linear projection followed by an activation function (e.g., ReLU or Sigmoid). Next, we compute edge weights A_{12} based on the similarity between F_1 and all 80 class embeddings:

$$C_1 = [1 - cos(F_1, w_j)]_{j=1}^{8}0 \tag{4}$$

OT assigns weights to minimize transport cost across this space, and edges in A are built accordingly. Only detected objects are instantiated as graph nodes, but all labels are considered in computing OT-based similarity, ensuring that the edge weights reflect semantic awareness beyond co-occurrence statistics. This modular graph design provides a rich, semantically structured representation, well-suited for GAT-based attention modeling, which we describe in the next section.

3.1.4. *GAT-based Representation Learning.* Once the semantic graph G=(V,E) is constructed, we employ a Graph Attention Network (GAT) [15] to perform contextual reasoning over object-level nodes. GAT enables the model to capture dynamic and context-aware relationships among detected objects, refining their features by aggregating information from semantically and visually related neighbors.

Each input node $V_i^{'}$ is first linearly projected into a latent space: $z_i = W_G V_i^{'}$ where $W_G \in \mathcal{R}^{d' \times d}$ is a trainable weight matrix. The transformed vector $z_i \in \mathcal{R}^{d'}$ is used in subsequent attention computations. For each edge $(i, j) \in E$, an attention score e_{ij} is computed as:

$$e_{ij} = LeakyReLU(a^T [z_i \| z_j]) \tag{5}$$

where $a \in \mathcal{R}^{2d'}$ is a learned parameter vector, $\|$ denotes vector concatenation, and LeakyReLU is used as activation (typically with $slope = 0.2$). The attention scores are normalized using a softmax operation over the neighbors of node i:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})} \tag{6}$$

This yields attention weights $\alpha_{ij} \in [0, 1]$ that determine the importance of each neighboring node. The output feature for node iii is computed as a weighted sum of its neighbors' features:

$$\mathrm{h}_i^{'} = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathrm{z}_j \right) \tag{7}$$

where σ is an activation function (e.g., ReLU or ELU). This process allows each node to incorporate contextual information from its semantically relevant surroundings.

To improve robustness and expressivity, we adopt multi-head attention, where K independent attention mechanisms are run in parallel:

$$\mathrm{h}_i^{'} = \|_{k=1}^{K} \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(k)} \mathrm{z}_j^{(k)} \right) \tag{8}$$

The outputs of each head are concatenated, resulting in a final embedding of size $K.d'$ After applying GAT over the graph, we obtain the updated node features $\{h_1^{'}, h_2^{'}, \ldots, h_N^{'}\}$. These are aggregated into a single image-level embedding $H \in \mathcal{R}^{K.d'}$, using average pooling or attention-based pooling: $H = AvgPool([h_1^{'}, h_2^{'}, \ldots, h_N^{'}])$. This final embedding H is then used in multi-label classification via a sigmoid-activated linear layer over C classes; or image retrieval by computing cosine similarity between embeddings from different images.

The ultimate goal of the representation learning module is to generate an expressive, semantic-aware embedding that captures both visual and relational dependencies across multiple objects in an image. After refining the node features via a GAT as described in above section, we aggregate these features to form a global image-level representation suitable for multi-label classification. Aggregation to Global Embedding Given a set of updated node embeddings $\{h_1^{'}, h_2^{'}, \ldots, h_N^{'}\} \in \mathcal{R}^{d'}$, where each $h_i^{'}$ encodes both the semantic and contextual information of a detected object, we apply an aggregation function $\mathcal{A}$ to obtain the final image representation: $\mathcal{H} = \mathcal{A}(h_1^{'}, h_2^{'}, \ldots, h_N^{'})$. Common aggregation strategies include Average Pooling:

$$\mathcal{H} = \frac{1}{N} \sum_{i=1}^{N} h_i^{'} \tag{9}$$

To predict the presence of multiple object categories in the image, we use a fully connected output layer followed by a sigmoid activation function:

$$\hat{y} = \sigma(\mathcal{W}_{cls} H + b) \tag{10}$$

where: $\mathcal{W}_{cls} \in \mathcal{R}^{C \times d'}$ and $b \in \mathcal{R}^{C}$ are learnable parameters, C is the number of possible classes (e.g., 80 for COCO), $\hat{y} \in [0,1]^C$ is the predicted probability vector for all labels, $\sigma(.)$ is the element-wise sigmoid function, producing independent predictions per label.

3.2. **Multi-label Image Retrieval using Learned Features.** In this section, we describe how the semantic-aware representations learned through our classification model are repurposed for the task of multi-label image retrieval (MLIR module). Unlike traditional CBIR systems that often require separate feature extraction pipelines or fine-tuning specific to retrieval, our method enables direct retrieval based on the graph-level embeddings obtained during classification training, thus ensuring task alignment and computational efficiency. Multi-label image retrieval aims to return images from a database that share semantically similar object compositions with a given query image. This task is particularly challenging due to:

- The presence of multiple objects per image, often with complex inter-label dependencies.
- The need to reflect not just label overlap, but also the contextual arrangement and co-occurrence semantics of objects.

The graph-level embedding H learned via our GAT-based module inherently encodes such relational and semantic information, making it a strong candidate for retrieval representation. As shown in Figure 1, in the second frame (bottom frame) is the general pipeline for the image retrieval task. The retrieval process proceeds as follows:

- **Embedding Extraction:** For both query and dataset images, we reuse the same feature extraction pipeline from section 3.1 to obtain the final graph embeddings. Each image $\mathcal{I}$ is mapped to an embedding vector $\mathcal{H}_{\mathcal{I}} \in \mathcal{R}^{d'}$.
- **Similarity Search and Ranking Using FAISS:** To efficiently perform similarity-based retrieval on large-scale multi-label image datasets, we leverage $FAISS$ (Facebook AI Similarity Search) [24], a library optimized for high-speed similarity search and clustering of dense vectors. Given that our framework encodes each image into a compact fixed-length vector $\mathcal{H} \in \mathcal{R}^{d'}$, we can directly index these embeddings using $FAISS$ for fast nearest-neighbor retrieval. After training, the image database (gallery) is preprocessed by computing the graph-based embeddings $\{\mathcal{H}_{I_1}, \mathcal{H}_{I_2}, ...\mathcal{H}_{I_M}\}$ for all images. These embeddings are stored in a $FAISS$ index, which supports fast search using either exact or approximate nearest neighbor algorithms. When a query image is submitted, we extract its graph embedding $\mathcal{H}_q$, normalize it, and search for top-K nearest neighbors.

3.3. **Loss functions.** To effectively optimize the proposed MLIR-SARL-GAT framework for both multi-label classification and discriminative feature learning, we design a composite loss function that integrates three complementary components: a multi-label classification loss, a metric-based embedding loss, and a graph consistency regularization. The total loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \cdot \mathcal{L}_{\text{triplet}} + \lambda_2 \cdot \mathcal{L}_{\text{graph}} + \lambda_3 \cdot \mathcal{L}_{\text{reg}} \tag{11}$$

where $\lambda_1, \lambda_2, \lambda_3$ are weighting coefficients that balance each component.

- **Multi-Label Classification Loss $\mathcal{L}_{cls}$**
  We adopt Asymmetric Loss (ASL) [25], a variant of Binary Cross-Entropy tailored for multi-label tasks with class imbalance. ASL down-weights the gradient contributions of easy negatives and applies focal-like modulation to focus learning on hard examples. For each class $c \in \{1, ..., C\}$, the classification loss is:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{C} \sum_{c=1}^{C} \left[y_c \log(\hat{y}_c) + (1 - y_c)\, \hat{y}_c^{\gamma} \log\left(1 - \hat{y}_c\right) \right] \tag{12}$$

  where $y_c \in \{0,1\}$ is the ground truth label, $\hat{y}_c \in (0,1)$ is the predicted probability, and γ controls the focus on hard negatives.

- **Triplet Margin Loss $\mathcal{L}_{triplet}$**
  To encourage separable feature embeddings for retrieval, we apply a Triplet Margin Loss on the final graph-level representations. Given an anchor image **a**, a positive image **p** sharing at least one label, and a negative image **n** with no overlapping labels, the loss is:

$$\mathcal{L}_{\text{triplet}} = \max\left(0, \|\mathcal{H}_a - \mathcal{H}_p\|_2^2 - \|\mathcal{H}_a - \mathcal{H}_n\|_2^2 + \alpha\right) \tag{13}$$

where $\mathcal{H}$ denotes the graph embedding, and α is the margin hyperparameter. Triplets can be dynamically sampled during training using semi-hard negative mining to improve convergence [25].

- **Graph Consistency Loss** $\mathcal{L}_{graph}$
  To regularize node embeddings within the semantic graph and promote semantic smoothness, we introduce a graph consistency loss inspired by [11]. This loss encourages connected nodes in the attention-based graph to have similar representations, and is defined as:

$$\mathcal{L}_{\text{graph}} = \sum_{i,j} A_{ij} \cdot \|h_i - h_j\|_2^2 \tag{14}$$

  where A_{ij} is the attention weight (or similarity score from Optimal Transport) between nodes i and j, and h_i, h_j is the node embedding after GAT propagation.

- **Regularization Term** $\mathcal{L}_{reg}$
  To prevent overfitting, we apply an L_2 regularization term on all trainable parameters:

$$L_{\text{reg}} = \|\theta\|_2^2 \tag{15}$$

  where θ denotes all parameters in the network. This composite loss ensures that the learned representation is not only effective for multi-label classification but also discriminative and structurally consistent, which is crucial for accurate and semantically aware image retrieval

4. **Experiments.** To validate the effectiveness of our proposed MLIR-SARL-GAT framework, we conduct extensive experiments on two benchmark datasets commonly used in multi-label image classification and retrieval. We compare our method with various state-of-the-art baselines in both classification accuracy and retrieval performance

4.1. **Dataset.** We evaluate our proposed MLIR-SARL-GAT framework on two widely used benchmark datasets for multi-label image classification and retrieval: MS-COCO 2017 [22] and PASCAL VOC 2012 [23]. MS-COCO 2017 dataset contains 118,287 training images, 5,000 validation images, and 40,670 test images, annotated with 80 object categories. Each image contains an average of 2.9 labels, making it a challenging benchmark for modeling co-occurring object relationships. PASCAL VOC 2012 dataset comprises 17,125 images with 20 object categories. Following the standard protocol, we use the trainval set (5,717 images) for training and the test set (5,823 images) for evaluation. Each image may contain multiple object labels with significant variations in scale, appearance, and spatial arrangement.

All images are resized to 448×448 pixels for both training and inference. Standard data augmentation is applied during training, including random horizontal flipping, random cropping, and color jittering. No augmentation is applied at inference time. For both datasets, we follow the standard multi-label classification and retrieval protocols as used in prior works [5, 6].

4.2. **Experiment setup.**

- **Backbone and Feature Extraction:** We use YOLOv8 [16] as the object detector to localize bounding boxes and predict object classes. ResNet-101 pretrained on ImageNet is employed to extract global and region-level visual features F, from which local ROI features v_i are obtained using RoIAlign.
- **Semantic Embedding:** Semantic features are obtained by mapping each detected object's class label to a 300-dimensional GloVe embedding [21] (trained on 6B tokens). The local visual feature F_i and the corresponding word embedding w_i are concatenated, passed through a fully connected layer with a sigmoid activation, producing the node feature $V_i^{'}$ for graph construction
- **Graph Construction:** Each image builds a dynamic semantic graph, where nodes $V_i^{'}$ are formed by concatenating visual and semantic features. dge weights computed using an Optimal Transport (OT) cost matrix C_{ij} based on similarity between object visual features F_i and word embeddings w_j of all possible labels. This allows semantically related objects (e.g., "dog" and "cat") to have higher connection weights than unrelated objects (e.g., "dog" and "car"), even if they are not co-occurring in the training set, ensuring contextual relevance in the graph structure.
- **Graph Attention Network:** We use a two-layer Graph Attention Network (GAT) [15] with 8 attention heads per layer to learn adaptive weights for each edge in the semantic graph. The GAT outputs a final global embedding Z via global average pooling.
- **Training:** We utilize the AdamW optimizer [27] for training the $SARL_GAT$ module network, with a batch size of 32, the learning rate at 0.0001, weight decay is 0.00001. We implement early stopping at epoch 55 to avoid overftting.

- **Retrieval Setup:** During inference, the learned embedding Z for each image is stored in a $FAISS$ [24] index using IndexFlatIP for cosine similarity search. Retrieval is performed by querying this index with the embedding of a query image.

Our experiments were conducted to address two key research questions (RQs):

- **RQ1:** How well does the proposed MLIR-SARL-GAT model perform compared to state-of-the-art models in both tasks of multi-label image classification and multi-label image retrieval?
- **RQ2:** How does the MLIR-SARL-GAT model work in the classification task and retrieval?

4.3. **Performance Compare (RQ1).** To answer RQ1, we compared MLIR-SARL-GAT with several strong baselines on MS-COCO 2017 and PASCAL VOC 2012 datasets in terms of classification and retrieval performance. For multi-label classification, we report: mean average precision (mAP); per-class: precision (CP), recall (CR), and F1-score (CF1); overall: precision (OP), recall (OR), and F1-score (OF1)

Table 1. Classification Results on MS-COCO 2017

Method	**mAP**	**CP**	**CR**	**CF1**	**OP**	**OR**	**OF1**
ResNet-101 [19]	81.2	77.5	69.8	73.4	78.3	70.4	74.2
ML-GCN [5]	82.9	79.1	71.6	75.2	80.2	72.0	75.9
ADD-GCN [6]	84.1	80.6	72.8	76.5	81.4	73.2	77.1
PatchGAT [7]	85.4	82.0	73.9	77.7	82.8	74.3	78.3
TDRG [12]	81.2	77.5	69.8	73.4	78.3	70.4	74.2
MLIR-SARL-GAT(Ours)	**87.1**	**84.2**	**76.2**	**79.5**	**84.2**	**76.7**	**80.4**

Table 2. Classification Results on PASCAL VOC 2012

Method	**mAP**	**CP**	**CR**	**CF1**	**OP**	**OR**	**OF1**
ResNet-101 [19]	91.4	88.6	85.3	86.9	89.1	85.8	87.4
ML-GCN [5]	92.3	89.8	86.5	88.1	90.1	86.9	88.4
ADD-GCN [6]	93.1	90.4	87.2	88.8	90.8	87.5	89.1
PatchGAT [7]	93.8	91.0	87.8	89.4	91.4	88.2	89.8
TDRG [12]	94.0	91.3	88.0	89.6	91.7	88.4	90.0
MLIR-SARL-GAT(Ours)	**95.2**	**92.5**	**89.4**	**90.9**	**93.0**	**89.8**	**91.4**

For multi-label image retrieval, we report: precision@K ($P@K$) for $K \in 1, 5, 10$, mean average precision (mAP), $top - K$ recall. Retrieval similarity is computed via cosine similarity between the learned graph embeddings, using FAISS for efficient large-scale nearest neighbor search.

Table 3. Retrieval Results on MS-COCO 2017

Method	**mAP**	**P@1**	**P@5**	**P@10**
ResNet-101 [19]	64.7	70.5	65.3	61.2
ML-GCN [5]	66.1	72.3	66.9	62.4
ADD-GCN [6]	68.4	74.1	68.5	63.7
PatchGAT [7]	70.5	75.6	70.1	64.8
TDRG [12]	71.2	76.0	70.5	65.3
MLIR-SARL-GAT(Ours)	**73.6**	**78.4**	**72.3**	**67.0**

Tables 1, 2, 3 and 5 show the classification and retrieval results on MS-COCO 2017 and PASCAL VOC 2012, respectively. Our MLIR-SARL-GAT achieves the highest mAP in both tasks, with particularly strong improvements in retrieval metrics such as P@1 and mAP, indicating the effectiveness of semantic-aware representation learning with GAT in capturing inter-object relationships.

To better understand the contribution of each component in the proposed MLIR-SARL-GAT, we conduct an ablation study on the MS-COCO 2017 dataset. Specifically, we compare the full model with three variants obtained by (i) replacing the graph attention layers with graph convolution (w/o Graph Attention), (ii) substituting the OT-based adjacency with cosine-similarity adjacency (w/o OT adjacency),

TABLE 4. Retrieval Results on PASCAL VOC 2012

Method	**mAP**	**P@1**	**P@5**	**P@10**
ResNet-101 [19]	75.6	80.2	76.1	72.0
ML-GCN [5]	77.3	81.8	77.4	73.2
ADD-GCN [6]	78.6	82.9	78.5	74.0
PatchGAT [7]	80.1	84.5	79.6	75.2
TDRG [12]	80.7	85.0	80.0	75.6
MLIR-SARL-GAT(Ours)	**82.9**	**87.3**	**82.1**	**77.4**

and (iii) removing label embeddings so that only visual features are used (w/o Label Embeddings). As reported in Table 5, all ablated variants exhibit noticeable performance drops in both multi-label classification and retrieval metrics, which confirms that each component is beneficial and that their combination is crucial for achieving the best overall results.

TABLE 5. Ablation Study on MS-COCO 2017

Variant	**mAP (Cls)**	**mAP (Ret)**	**P@1**	**P@5**
Full model (MLIR-SARL-GAT)	**87.1**	**73.6**	**78.4**	**72.3**
w/o Graph Attention (GCN instead)	85.1	70.8	75.9	70.1
w/o OT adjacency (cosine similarity)	85.7	71.4	76.5	70.8
w/o label embeddings (visual only)	84.3	69.9	74.7	69.2

As shown in Table 5, the full MLIR-SARL-GAT model outperforms all ablated variants on every metric. Replacing graph attention with graph convolution reduces performance, confirming the benefit of attention for modeling label dependencies. Similarly, removing the OT-based adjacency harms results, indicating that the learned transport-based graph is more informative than a simple cosine graph. The largest drop occurs when label embeddings are removed, highlighting the crucial role of explicit semantic label information. These findings show that all components contribute and that their combination is necessary to achieve the best performance.

4.4. **Qualitative Study (RQ2).** To answer RQ2, we visualize the model's predictions in both classification and retrieval tasks.

Figure 3 shows example predictions from MLIR-SARL-GAT on five randomly selected images from MS-COCO 2017. The red labels are the actual labels present in the image but not detected by the model. The predicted labels closely match the ground truth, even in complex scenes containing multiple overlapping objects, demonstrating the model's ability to exploit both spatial and semantic dependencies.

Figure 4 presents qualitative retrieval results. For each query image, the top-5 retrieved images are shown. The red labels are the actual labels present in the image but not detected by the model. The dark blue labels are the labels that the model detects that are present in the result image but not in the query image. The retrieved results not only share the same object categories but also exhibit similar contextual arrangements, indicating that the learned graph embeddings capture high-level semantic alignment beyond simple object presence.

These visualizations confirm that MLIR-SARL-GAT successfully integrates object-level semantics and inter-object relationships into a unified representation, leading to robust performance across both classification and retrieval tasks.

5. **Conclusions.** In this paper, we proposed MLIR-SARL-GAT, a multi-label image retrieval framework that combines semantic-aware representation learning with Graph Attention Networks (GAT). Instead of relying solely on global image descriptors, the model integrates fine-grained object-level visual features with label embeddings and constructs dynamic semantic graphs to capture both spatial and contextual relationships among objects. By performing attention-based message passing over these graphs, MLIR-SARL-GAT explicitly models inter-label dependencies and produces unified embeddings that can be used for both multi-label classification and image retrieval.

Extensive experiments on two widely used benchmarks, MS-COCO 2017 and PASCAL VOC 2012, demonstrate the effectiveness of the proposed approach. Our method consistently outperforms strong

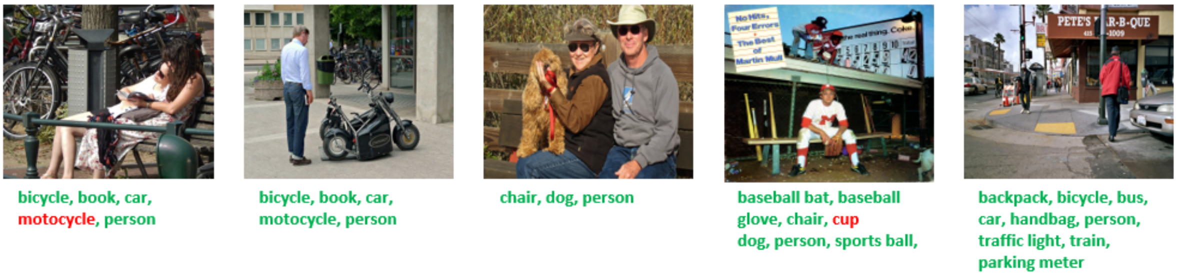

FIGURE 3. Visualization of multi-label image classification results on COCO 2017 dataset.

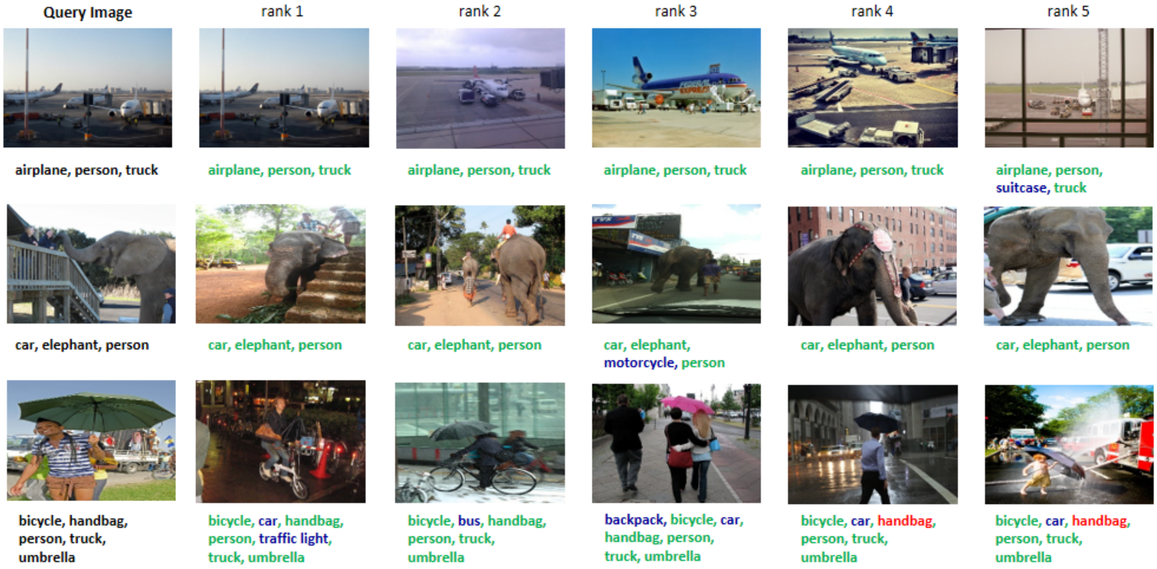

FIGURE 4. Visualization of multi-label image retrieval results on the COCO 2017 dataset

baselines in terms of both classification and retrieval metrics. In addition, the ablation study confirms that each key component—graph attention, OT-based adjacency construction, and label embeddings—contributes positively to the final performance, and that their combination yields the best results. Qualitative visualizations further show that the learned representations capture not only shared object categories but also higher-level contextual alignment between images.

Despite these promising results, the proposed approach still has several limitations. First, the model assumes a fixed label set with supervised annotations; noisy or incomplete labels may deteriorate the learned semantic graph and harm retrieval quality, and the current design does not explicitly address open-set or unseen labels. Second, the OT-based graph construction and attention-based message passing introduce additional computational and memory costs, which may limit scalability to very large label vocabularies or real-time deployment scenarios. Third, our evaluation is restricted to two benchmark datasets and a specific backbone architecture; further experiments on more diverse data and backbones are needed to fully assess robustness and generality. Finally, MLIR-SARL-GAT mainly leverages label co-occurrence and visual features, without exploiting richer side information such as textual descriptions, hierarchical label structures, or user feedback, which are directions that we plan to explore in future work to further enhance semantic representation and retrieval performance.

Acknowledgment. This research is funded by Academy of Finance, Hanoi, Vietnam.

REFERENCES

[1] Elsayed, Hend A., Mustafa F. Hameed, and Mohammed M. El Sherbiny. ”Image Classification using Decision Tree Classifier and Features Extraction using Hough transform and Genetic Algorithm.” Journal of Information Hiding and Multimedia Signal Processing 16.1 (2025): 438-452.

[2] Zhu, Feng, et al. ”Learning spatial regularization with image-level supervisions for multi-label image classification.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. 5513–5522.

[3] Wang, Zhouxia, et al. ”Multi-label image recognition by recurrently discovering attentional regions.” Proceedings of the IEEE International Conference on Computer Vision. 2017. 464–472.

[4] Guo, Hao, et al. ”Visual attention consistency under image transforms for multi-label image classification.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. 729–739.

[5] Chen, Zhao-Min, et al. ”Multi-label image recognition with graph convolutional networks.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. 5177–5186.

[6] Ye, Jin, et al. ”Attention-driven dynamic graph convolutional network for multi-label image recognition.” European Conference on Computer Vision. Springer, 2020. 649–665.

[7] Yuan, Jin, et al. ”Graph Attention Transformer Network for Multi-label Image Classification.” ACM Transactions on Multimedia Computing, Communications, and Applications 19.4 (2024).

[8] Ha, Duong Manh, et al. ”Semantic Connection-Based Learning for Dragonfruit Disease Classification.” Journal of Information Hiding and Multimedia Signal Processing 15.4 (2024): 281–291.

[9] Dao Thi Thuy Quynh, et al. ”Multi-modal tooth decay recognition based on Contrastive Learning and Multi-label.” Journal of Information Hiding and Multimedia Signal Processing 15.4 (2024): 271–280.

[10] Yazici, Vacit Oguz, et al. ”Orderless recurrent models for multi-label classification.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. 13440–13449.

[11] Kipf, Thomas N., and Max Welling. ”Semi-supervised classification with graph convolutional networks.” arXiv preprint arXiv:1609.02907 (2016).

[12] Zhao, J., et al. ”Transformer-based dual relation graph for multi-label image recognition.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. 163–172.

[13] Vaswani, Ashish, et al. ”Attention is All You Need.” Advances in Neural Information Processing Systems (NIPS). 2017.

[14] Carion, Nicolas, et al. ”End-to-end object detection with transformers.” Computer Vision – ECCV 2020.

[15] Velickovic, Petar, et al. ”Graph Attention Networks.” ICLR Conference. 2018.

[16] Jocher, Glenn, Ayush Chaurasia, and Jing Qiu. ”YOLOv8: Ultralytics Documentation.”, .https://docs.ultralytics.com/models/yolov8

[17] Ren, Shaoqing, et al. ”Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017.

[18] Redmon, Joseph, and Ali Farhadi. ”YOLOv3: An Incremental Improvement.” arXiv preprint arXiv:1804.02767 (2018).

[19] He, Kaiming, et al. ”Deep residual learning for image recognition.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. 770–778.

[20] He, Kaiming, et al. ”Mask R-CNN.” IEEE International Conference on Computer Vision (ICCV). 2017, .https://arxiv.org/abs/1703.06870.

[21] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. ”GloVe: Global Vectors for Word Representation.” Empirical Methods in Natural Language Processing (EMNLP). 2014, .https://nlp.stanford.edu/pubs/glove.pdf.

[22] Lin, Tsung-Yi, et al. ”Microsoft COCO: Common objects in context.” European Conference on Computer Vision. 2017. 740–755.

[23] Everingham, M., et al. ”The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.”, .http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html (2012).

[24] Johnson, Jeff, Matthijs Douze, and Hervé Jégou. ”Billion-scale similarity search with GPUs.” IEEE Transactions on Big Data 7.3 (2017), .https://doi.org/10.48550/arXiv.1702.08734.

[25] Ridnik, Tal, et al. ”Asymmetric Loss For Multi-Label Classification.” CVPR 2020.

[26] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. ”FaceNet: A Unified Embedding for Face Recognition and Clustering.” CVPR 2015.

[27] Llugsi, Ricardo, et al. "Comparison between Adam, AdaMax and Adam W optimizers to implement a Weather Forecast based on Neural Networks for the Andean city of Quito." IEEE Fifth Ecuador Technical Chapters Meeting (ETCM). IEEE, 2021.