2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Knowledge-Guided Graph Convolutional Network for Multi-Label Image Classification

Christine Dewi[12*], Dhananjay Thiruvady[1], Stephen Abednego Philemon[2], and Nayyar Zaidi[1]

***Abstract*— Multi-label image classification is a significant challenge in computer vision due to the presence of multiple interconnected objects in a single image. Traditional convolutional neural networks (CNN) often fail to capture semantic dependencies between labels, limiting performance in complex scenes. To address this issue, we propose a novel framework that combines Knowledge-Guided Graph Convolutional Network (KGGCN) with Darknet53 backbone to improve label dependency modelling. Our method fuses external semantic information from ConceptNet5, which allows the model to learn contextual relationships between labels. Our work evaluate this approach on two benchmark datasets, VOC 2007 and COCO, and obtain state-of-the-art results. KGGCN achieves an Average Precision (mAP) of 96.24% on VOC 2007 and 85.25% on COCO, outperforming existing methods in most categories. Moreover, ablation studies further highlight the benefits of external knowledge integration contributing to higher mAP scores. Finally, our proposed method KGGCN demonstrates the effectiveness of combining deep visual features with structured semantic knowledge for multi-label image classification.**

I. INTRODUCTION

In computer vision, multi-label image recognition enables models to identify multiple objects within a single image simultaneously. This has become a prominent research area because of its wide range of practical applications, including human attribute detection [1], medical diagnosis recognition [2], scene recognition [3], and multi-object recognition [4][5]. Unlike single-label classification, multi-label image classification requires the model not only to recognize several objects at once but also to understand the relationships among them within a shared visual scene.

In the literature, multi-label classification methodologies can generally be divided into two primary categories. The first category, often referred to as direct methods, develops a single Deep Neural Network (DNN) to handle multiple binary classification tasks without explicitly incorporating prior knowledge into the model architecture [6]. Although direct methods have demonstrated strong performance, as reported in [7], they often require deeper and more complex network structures to achieve optimal results. This increases memory consumption and limits their applicability in memory-constrained environments. In contrast, the second category, known as indirect methods, leverages prior knowledge about the relationships among objects or labels within an image [8]. Such approaches are especially useful for multi-label recognition, where the presence of one object often provides contextual cues for the existence of other related objects.

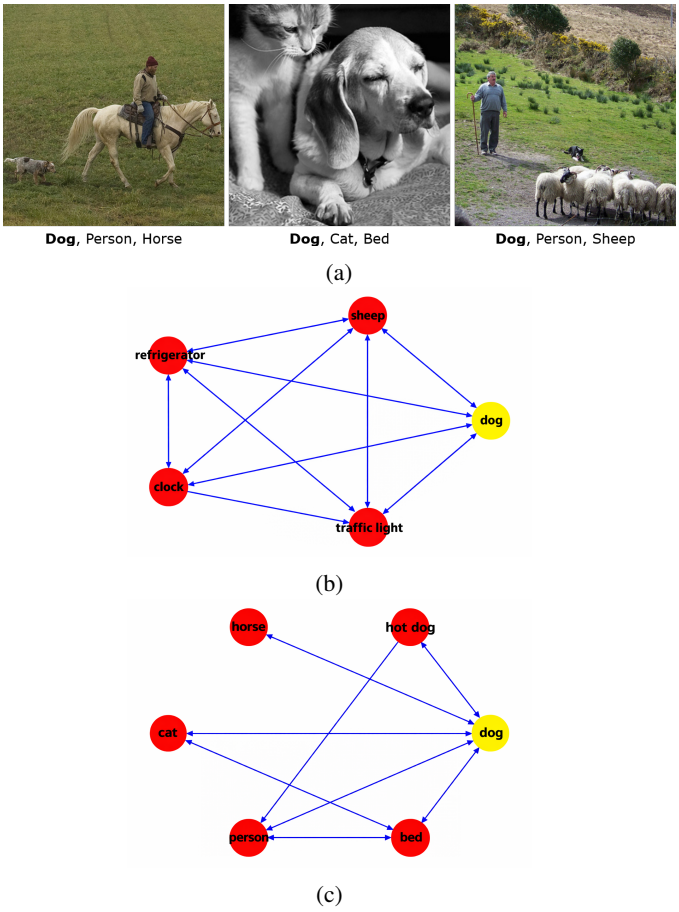


Fig. 1: (a) Image samples with multiple labels. (b) Graph based on label co-occurrence. (c) Graph utilizing the external knowledge from ConceptNet5.

Graph Convolutional Networks (GCNs) are effective for modelling structured relationships in graph-based data. In multi-label image classification, they represent object categories as nodes and semantic or co-occurrence relationships as edges, allowing information to propagate between related labels. This helps capture contextual dependencies that are often missed by conventional convolutional backbones. However, standard GCNs mainly rely on the predefined graph topology and learned node representations, which may not be sufficient to fully capture richer external semantic knowledge. To address this limitation, Knowledge-Guided Graph Convolutional Networks (KGGCNs) can be introduced as

*Corresponding Author: Christine Dewi (c.dewi@deakin.edu.au)
[1]Faculty of Sci Eng & Built Env, Deakin University, VIC 3220, Australia.
[2]Department of Information Technology, and Satya Wacana Christian University, 52-60 Diponegoro Rd, Salatiga City, 50711, Indonesia.

a modification of conventional GCNs. KGGCN extends the original GCN framework by integrating external knowledge, such as semantic similarity, category co-occurrence, or prior domain relationships, into the graph learning process. By guiding message passing with this additional knowledge, KGGCN can learn more informative and discriminative label representations, thereby improving the reasoning capability of the model for complex multi-label recognition tasks.

In particular, Darknet53 has shown excellent performance in real-time detection and image representation learning. Nevertheless, such backbones generally predict labels independently and do not explicitly model semantic dependencies among object categories. This limits their ability to exploit contextual relationships that are crucial in multi-label image classification. Motivated by this, our approach combines Darknet53 and KGGCN in a unified framework. Darknet53 extracts visual features, while KGGCN models knowledge-guided label dependencies. By integrating visual representation learning with graph reasoning, the proposed method captures both appearance and contextual relationships to improve multi-label classification performance.

Figure 1 shows the motivation for using external knowledge in multi-label classification. In Figure 1a, images with labels like dog, person, and horse illustrate frequent co-occurrence of related objects. Figure 1b depicts graph-based models linking label nodes via co-occurrence statistics. However, such data-driven approaches can introduce noisy or irrelevant associations—for example, spurious edges between dog and unrelated objects like clock or traffic light—due to their reliance solely on frequency rather than meaning. To solve this limitation, we propose a knowledge-guided graph representation leveraging structured semantic embeddings from ConceptNet5, as depicted in Figure 1c.

ConceptNet5 is a public semantic knowledge graph that captures meaningful label relationships—e.g., dog, cat, person, bed—enabling the GCN to model correlations, reduce noise, and generalize effectively [9]. Figures 1b and 1c show that adding commonsense knowledge creates more context-aware graphs, boosting multi-label classification. KGGCN builds a label graph with nodes as object categories and edges from prior co-occurrence or semantic relations, enabling the GCN to share context and improve classification accuracy. Our approach combines the visual strength of Darknet53 with a KGGCN that models high-level label semantics, integrating bottom-up visual features with top-down semantic cues for coherent multi-label classification.

This paper has three main contributions:

1) We propose a hybrid architecture, KGGCN, that combines the powerful visual feature extraction of Darknet53 with a knowledge-guided graph convolutional network. This design enables joint learning of spatial and semantic features for improved label correlation modelling.
2) We propose KGGCN, leveraging ConceptNet5 and ConceptNet Numberbatch embeddings to build a semantically rich label graph and better capture label relationships.
3) We conduct extensive experiments on the VOC 2007 [10] and COCO [11] datasets, demonstrating state-of-the-art performance. The results of our study provide additional evidence that KGGCN surpasses prior methods in both classification performance and model robustness.

II. RELATED WORK

A. Graph Convolution Networks (GCN)

GCN have garnered considerable interest for their capacity to model graph data structure, including social networks, citation graphs, and scene graphs.

Kipf and Welling [12] proposed a scalable semi-supervised node classification method using spectral-based convolution. Subsequent GCN variants address over-smoothing with residual connections and attention mechanisms. For example, Graph Attention Transformer (GAT) [13] employs masked self-attention to allocate varying weights to adjacent nodes, hence enhancing expressiveness in heterogeneous graphs. GraphSAGE [14] introduced an inductive architecture that learns aggregation functions, enhancing its applicability for dynamic graphs. Wang et al. [15] proposed a knowledge-aware GCN for multi-label picture classification by constructing a graph based on semantic labels, enabling the model to infer label co-occurrence and dependencies. Chen et al. [16] suggested a structured inference neural network that captures label correlations via message passing on a label graph. Recently, knowledge-guided and context-aware GCN have been employed to integrate external knowledge bases, such as WordNet, into visual recognition frameworks [16]. These methodologies improve generalization by representing higher-order semantic linkages, even without direct label associations [17].

In contrast to conventional convolutions that operate on local Euclidean regions of an image, GCNs aim to learn a function $f(.,.)$ on a graph G, which takes feature descriptions $H^l \in \mathbb{R}^{n\times C}$ and the corresponding correlation matrix $A \in \mathbb{R}^{n\times n}$ as inputs, where n represents the total number of nodes and C denotes the feature dimension of each node. The graph convolution then updates the node features as $H^{l+1} \in \mathbb{R}^{n\times C_1}$. Every GCN layer can be written as a non-linear function by

$$H^{l+1} = f(H^l, A). \tag{1}$$

where l denotes the layer index of the graph convolution network, H^l represents the node features at the l-th layer, and H^{l+1} denotes the updated node features at the next layer. Following the application of the convolutional operation from [12], $f(.,.)$ can be expressed as

$$H^{l+1} = h(\hat{A}H^lW^l), \tag{2}$$

where $W^l \in \mathbb{R}^{C\times C'}$ is a transformation matrix to be learned and $\hat{A} \in \mathbb{R}^{n\times n'}$ is the normalized version of correlation matrix A, and $h(.)$ denotes a non-linear operation, which is acted by LeakyReLU [18] in our experiments. Consequently, we can learn and model the intricate interrelationships of the nodes by stacking many GCN layers. For further

information, we direct interested readers to read research by [12].

B. Multi-Label Graph Convolutional Networks (ML−GCN)

Unlike the original GCN for instance classification, our approach treats each node as a label classifier. Lacking a predefined graph, we construct the correlation matrix directly from the training data to capture label relationships for multi-label image recognition. ML−GCN [16] was one of the pioneers in employing GCN for multi-label image classification to simulate label correlations. This architecture consists of two primary branches. The initial branch comprises a conventional image representation learning network. The authors utilized ResNet101 [19] to extract discriminative image characteristics from the input image. ML−GCN utilize word embedding reconstructions as input node characteristics, referred to as F_W. The node features are produced using GloVe [20]. Consequently, in this work, the initial node feature matrix is defined as $H^0 = F_W$, where F_W denotes the word-embedding feature matrix used for initialization. In addition, a reweighting strategy is introduced, where a threshold τ is used to filter noisy edges, resulting in:

$$A_{ij} = \begin{cases} 0, & \text{if } P_{ij} < \tau \\ 1, & \text{if } P_{ij} \geq \tau \end{cases} \tag{3}$$

where P_{ij} =$P(L_j|L_i)$ is the probability of the occurrence of an object label L_j in an image provided that the label L_i is already present.

III. Proposed framework

A. Semantic Consistency and Embedding Method

Semantic consistency describes how well concepts, objects, or labels are logically connected and contextually aligned, often reflecting the likelihood of their coherent co-occurrence within an image. Fundamentally, knowledge is symbolic and logical in nature, while most modern machine learning algorithms rely on numerical or subsymbolic representations [21]. Bridging this gap requires translating structured knowledge into a form that can be integrated into visual models, particularly for handling novel contexts in images. To this end, we adopt the approach proposed by [22], where semantic consistency between pairs of concepts is quantified numerically. A high degree of semantic consistency indicates that two concepts are likely to co-occur within the same image, providing a useful signal for improving multi-label recognition. GloVe [20] and ConceptNet Numberbatch [9] are word embedding methods; GloVe learns 300-dimensional vectors from co-occurrence statistics, while ConceptNet Numberbatch integrates distributional data with commonsense knowledge to capture richer semantics.

Algorithm 1 Knowledge Learning Update via Semantic Consistency

Input:
- $X_{\text{gcn}} \in \mathbb{R}^{n \times C}$ ▷ Graph Representation
- $S \in \mathbb{R}^{n \times n}$ ▷ Semantic Consistency: ConceptNet5
- lk ▷ Number of Top-lk Neighbors
- i, n ▷ Number of Iterations, Number of Nodes
- $\varepsilon \in [0, 1]$ ▷ Trade-off Parameter

Output:
Updated Graph Representation $X'_{\text{gcn}} \in \mathbb{R}^{n \times C}$

Compute Top-lk of Semantic Consistency Matrix
for each $i \in \{1, 2, \ldots, n\}$ **do**
  Get Top-lk indices T_i from S_i:

$$T_i \subseteq \{1, 2, \ldots, n\}, \quad |T_i| = lk$$

  Set all $S_{i,j}^{(lk)} = 0$ for $j \notin T_i$
  Normalize the values:

$$S_i^{(lk)} \leftarrow \frac{S_i^{(lk)}}{\sum_{j=1}^{n} S_{i,j}^{(lk)} + 1e{-}6}$$

end for
Update Graph Representation Using External Knowledge
Initialize $X \leftarrow (X_{\text{gcn}})^\top$
if $n = 1$ **then**
  Initialize $X'_{\text{gcn}} \leftarrow X_{\text{gcn}}$
else
  Initialize $X'_{\text{gcn}} \leftarrow \frac{1}{n} \cdot \text{ones\_like}((X_{\text{gcn}})^\top) \in \mathbb{R}^{C \times n}$
  for each iter $\in \{1, 2, \ldots, i\}$ **do**
    Compute weighted update:

$$w \leftarrow X'_{\text{gcn}} \cdot (S^{(lk)})^\top$$

    Apply smoothing and update:

$$X'_{\text{gcn}} \leftarrow (1 - \varepsilon) \cdot \left(\frac{w}{\sum_{j=1}^{n} S_{i,j}^{(lk)} + 1e{-}6} \right) + \varepsilon \cdot X, \quad \forall i \in \{1, \ldots, n\}$$

    Normalize updated X'_{gcn}:

$$X'_{\text{gcn}} \leftarrow \frac{X'_{\text{gcn}}}{\sum_{i=1}^{n} X'_{\text{gcn}}(i)}$$

  end for
end if
return $(X'_{\text{gcn}})^\top$

B. Knowledge-Guided Graph Convolutional Network (KGGCN)

Figure 2 illustrates KGGCN with a Darknet53 backbone, combining a CNN for feature extraction and a knowledge-enhanced GCN for semantic label modelling. The CNN processes 448×448 input images through multiple convolutional stages. This stage extracts spatial features and reduces resolution through stacked convolutional and residual blocks. The final feature representation is generated by a global max pooling layer, which produces a D-dimensional visual feature vector with a fixed length. The GCN component processes semantic label representations using word embeddings from ConceptNet Numberbatch, which encode commonsense relationships. These embeddings pass through two GCN layers, where an adjacency matrix—constructed

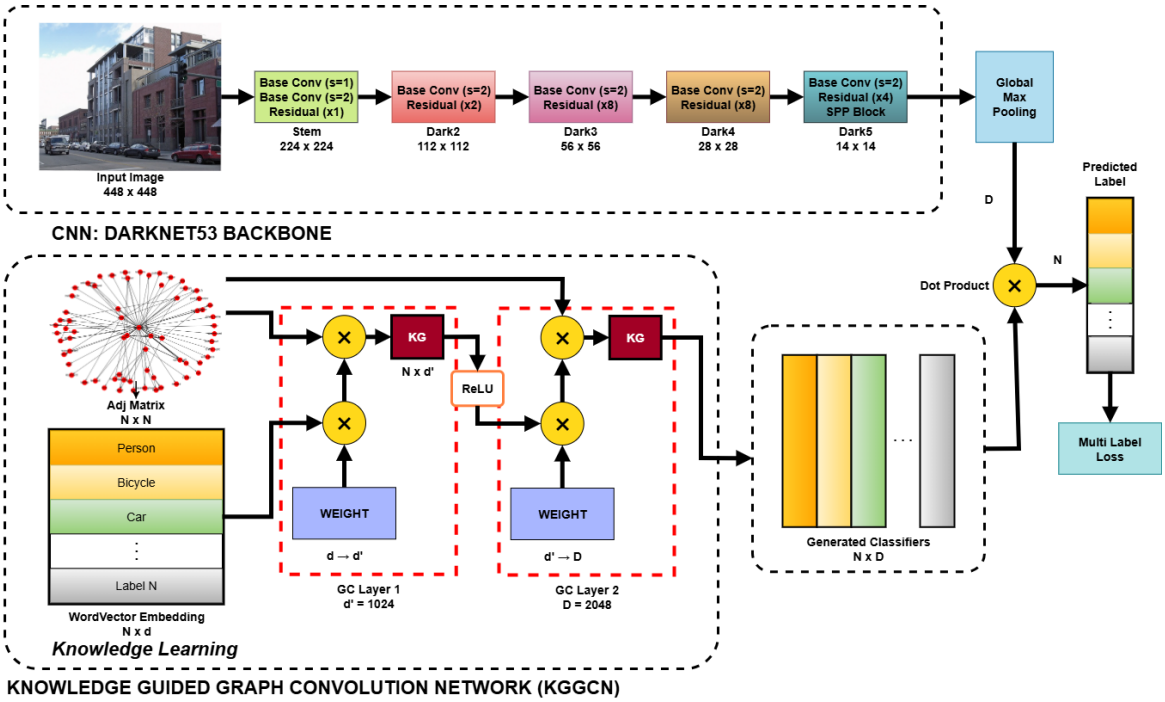


Fig. 2: Overall framework of our KGGCN model for multi-label image recognition. The object labels are represented by semantic representation from ConceptNet5. The generated classifier is a new embedding derived from the KG that represents the semantic connections between labels.

from external knowledge—models label correlations. The KG modules further improve message passing by incorporating semantic consistency into the GCN layers. GCN learns label classifiers and integrates them with CNN features. A multi-label loss aligns predictions with ground-truth, enabling joint optimization of spatial and semantic information for higher accuracy in complex tasks. Our research develops a KG to represent semantic consistency with ConceptNet5 [9]. To measure semantic consistency in a KG, we utilize random walk with restart by [23], with the derived probability value serving as an indicator of semantic consistency. A random walk is defined as a sequence of nodes $(v_0, v_1, v_2, \ldots, v_t)$, and $p(v_t = l'|v_0 = l)$ is the probability of reaching concept l' in t steps that we start from l. This probability can be utilized to establish semantic consistency, indicating that a higher likelihood from l to l' suggests greater semantic alignment.

The Algorithm 1 describes a method for updating a graph representation using semantic consistency derived from external knowledge sources. The input includes a graph representation matrix $X_{\text{gcn}} \in \mathbb{R}^{n \times C}$, a semantic consistency matrix $S \in \mathbb{R}^{n \times n}$, the number lk indicating how many top-related nodes to retain, the number of update iterations i, and a trade-off parameter $\varepsilon \in [0, 1]$ controlling the balance between the original features and knowledge-based updates. In Algorithm 1 line 1 sparsifies the semantic consistency matrix by selecting only the $Top - lk$ most semantically similar nodes for each node i. All other entries in the corresponding row are set to zero, and the remaining values are normalized to ensure they form a probability distribution (with a small constant $1e-6$ added for numerical stability). In Algorithm 1 line 7 updates the graph representation by iteratively applying a knowledge-guided smoothing procedure. It first transposes the original representation for column-wise operations. If there is only one node $(n = 1)$, the representation remains unchanged. Otherwise, it initializes an average feature matrix and performs i iterations of updates. In each iteration, it computes a weighted average of neighboring node features using the normalized semantic matrix $(S^{(lk)})$, applies smoothing by blending this average with the original features via the parameter ϵ, and normalizes the result across nodes. The updated representation is reshaped to its original form, aligning node features with semantically similar nodes and embedding external knowledge into the graph.

IV. Experiments

A. Dataset and Evaluation Metrics

The COCO dataset [11] is a large-scale benchmark with 122218 images spanning 80 object categories, supporting tasks like detection, segmentation, and captioning. The VOC 2007 dataset [10] is a standard benchmark for multi-label recognition, containing 9963 images across 20 classes.

Following [15], we train on the combined training and validation sets and evaluate on the official test set. In line with standard protocols, we report mean Average Precision (mAP), class-wise Precision (CP) and Overall Precision (OP), class-wise Recall (CR) and Overall Recall (OR), as well

as class-wise and Overall F1-scores (CF1, OF). For fair comparison, we also report Top-3 label performance [16], [24].

Our experiment trained on Google Colab Pro+ with NVIDIA V100/A100 GPUs, up to 52 GB RAM. Input resolution was 448×448, trained for 40 epochs with batch size 36 using SGD (momentum 0.9, base learning rate 0.01), and a multi-step learning rate schedule at epochs 10, 20, and 30. No warm-up was used, the random seed was fixed to 1, and validation used a 0.8 confidence threshold. Data loading used 8 workers, with TensorBoard enabled. The backbone was Darknet53 with ConceptNet Numberbatch embeddings and a label adjacency graph. For dataset-specific settings, the backbone learning rate was scaled by a factor lrp, set to 0.1 for COCO and 0.01 for VOC2007, with dataset paths configured accordingly. Here, lr is the base learning rate, while lrp adjusts the backbone learning rate.

B. Results and Discussions

1) Results on COCO: Table I compares KGGCN with state-of-the-art methods on the COCO dataset. Traditional baselines such as CNN−RNN [26] and RLSD [27] yield relatively low performance, while stronger backbones like ResNet101 [19] and graph-based methods (ML-GCN [16], SSGRL [8], KGGR [29]) achieve notable improvements in mAP and F1-scores. Among these, ASL [7] and ML-AGCN [6] deliver competitive results with mAP values of 86.6% and 86.0%, respectively. In contrast, KGGCN achieves 85.25% mAP, which is comparable to the top-performing methods, while significantly outperforming others in overall precision (96.1%) and class precision (92.8%). Importantly, KGGCN attains these results with a moderate parameter count of 53.4 M, demonstrating a strong balance between accuracy and efficiency. These results highlight the effectiveness of integrating external semantic knowledge with graph reasoning to enhance multi-label recognition on complex datasets such as COCO.

2) Results on VOC 2007: Table II shows KGGCN outperforming all methods on VOC 2007 with a mAP of 96.24%, achieving top AP in categories like bike (98.8%), boat (99%), bus (98.7%), and person (99.3%). Furthermore, compared to recent competitive approaches such as ML−AGCN (95%) [6], SSRGL (95%) [8], and CFMIC (94.7%) [36], our proposed model demonstrates superior robustness. The performance gains primarily arise from integrating external semantic knowledge through graph convolution with Darknet53 features. The proposed model achieves strong results on high-variability classes such as bus, car, and person by effectively capturing both semantic dependencies and visual representations, thereby establishing a new benchmark on VOC 2007 for multi-label classification.

Figure 3 shows a comparison of image classification results using two models: (a) Darknet53 and (b) KGGCN with VOC dataset, showing that KGGCN predicts more complete and accurate labels by leveraging external knowledge. Examples include detecting chair, bus, person and tvmonitor missed by the baseline (Darknet53), demonstrating improved inference of related objects.

C. Ablation Studies

Table III demonstrates that while the baseline achieves 95.64% mAP, adding GCN increases accuracy to 95.99% mAP. The full integration of KG, GCN, and ConceptNet embeddings yields the best performance (96.24% mAP, 90.5% OF1, 89.3% CF1) and consistent gains across precision, and recall, with only a slight rise in parameters without increasing complexity (53.48 M params, 213.93 MB), highlighting the effectiveness of semantic knowledge integration for multi-label classification. Table IV provides per-class AP comparisons, showing KGGCN with ConceptNet Numberbatch often outperforming others in challenging categories like bike, boat, car, and cat, where semantic context is crucial. The full model KGGCN with KG, GCN, and embedding leads in 14 of 20 classes, confirming its robustness and the complementary benefits of knowledge-guided representations and graph-based reasoning in reducing implicit label linkages and visual dependencies beyond what conventional CNN or text-only embeddings achieve.

Table V shows that adding GCN improves performance over the baseline, while combining KG, GCN, and embeddings achieves the best results (85.25% mAP) with minimal parameter increase, confirming the benefit of integrating semantic knowledge for multi-label recognition. Results show that ConceptNet Numberbatch embeddings improve multi-label classification, especially on COCO with high label co-occurrence, with gains stemming from knowledge integration rather than increased model complexity.

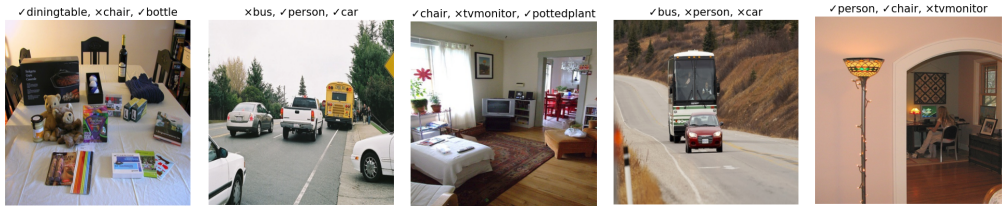
We vary the values of lk in Algorithm 1 to compute the semantic consistency matrix and present the results in Table VI. The parameter lk defines how many of the most confident predicted labels are used per image to construct the semantic consistency matrix. As shown, the optimal value of lk is dataset-dependent: on VOC, $lk = 5$ achieves the best results with the highest mAP (96.24%) and CF1 (89.3%), while on COCO, $lk = 4$ yields the best performance with mAP (85.25%) and OF1 (80.5%). The difference arises from dataset characteristics: VOC has fewer classes and simpler co-occurrence patterns, thus benefiting from a larger neighborhood, whereas COCO is larger and noisier, where a moderate neighborhood size strikes a better balance between capturing semantic relations and avoiding noise. Table VII reports the impact of varying the trade-off parameter ε. We observe that performance is relatively stable across settings, but the best results consistently occur at $\varepsilon = 0.7$. On VOC, this setting yields the highest mAP (96.24%) and CF1(89.30%), while on COCO it achieves the best mAP (85.25%) and OF1 (80.52%). The consistent optimal value across both datasets indicates that $\varepsilon = 0.7$ provides the most balanced weighting between supervised and semantic consistency losses, ensuring stable gains and good generalization.

TABLE I: Comparisons with state-of-the-art methods on the COCO dataset.

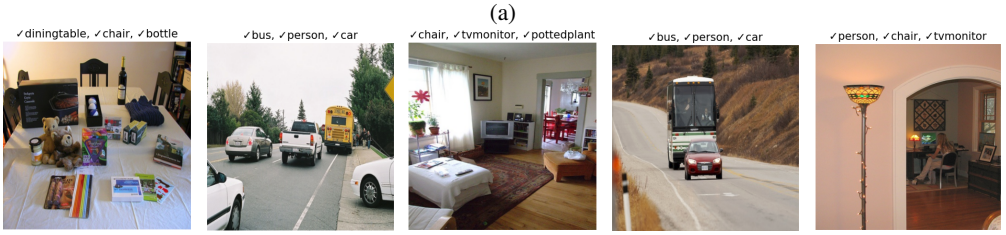
Model	mAP	OP	OR	OF1	CP	CR	CF1	Total Params (M)
Multi-Evidence [25]	-	85.2	72.5	78.4	80.4	70.2	74.9	$\sim$ 47.0
CNN-RNN [26]	61.2	-	-	-	-	-	-	66.2
RLSD [27]	68.2	70.1	63.4	66.5	67.6	57.2	62.0	–
SRN [24]	77.1	81.6	65.4	71.2	82.7	69.9	75.8	$\sim$ 48.0
ResNet101 [19]	77.3	80.2	66.7	72.8	83.9	70.8	76.8	44.5
ML-GCN [16]	83	85.1	72	78	85.8	75.4	80.3	44.9
SSGRL [8]	83.8	89.9	68.5	76.8	91.3	70.8	79.7	92.2
FLNet [28]	84.1	85.5	77.4	81.1	84.9	73.9	79.0	–
KGGR [29]	84.3	85.6	72.7	78.6	87.1	75.6	80.9	$\sim$ 45.0
C-Tran [30]	85.1	86.3	74.3	79.9	87.7	76.5	81.7	120.0
ASL [7]	**86.6**	87.4	76.4	**81.4**	88.1	79.2	81.8	53.8
ML-AGCN [6]	**86.6**	78.8	**82.6**	80.2	79	**85.1**	**81.9**	31.5
KGGCN	85.25	**96.1**	69.3	80.5	**92.8**	64.7	74.6	53.4

TABLE II: Comparisons of AP and mAP with state-of-the-art methods on the VOC 2007 dataset.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
CNN+RNN [26]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70	92.4	91.7	84.2	93.7	59.8	93.2	75.3	**99.7**	78.6	84
RLSD [27]	96.4	92.7	93.8	94.1	71.2	92.5	94.2	95.7	74.3	90	74.2	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.5	87.6	88.5
VeryDeep [31]	98.9	95	96.8	95.4	69.7	90.4	93.5	96	74.2	86.6	87.8	96	96.3	93.1	97.2	70	92.1	80.3	98.1	87	89.7
ResNet-101 [19]	99.5	97.7	97.8	96.4	65.7	91.8	96.1	97.6	74.2	80.9	85	98.4	96.5	95.9	98.4	70.1	88.3	80.2	98.9	89.2	89.9
FeV+LV [32]	97.9	97	96.6	94.6	73.6	93.9	96.5	95.5	73.7	90.3	82.8	95.4	97.7	95.9	98.6	77.6	88.7	78	98.3	89	90.6
HCP [33]	98.6	97.1	98	95.6	75.3	94.7	95.8	97.3	73.1	90.2	80	97.3	96.1	94.9	96.3	78.3	94.7	76.2	97.9	91.5	90.9
RNN-Attention [15]	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92	87.7	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
Atten-Reinforce [34]	98.6	97.1	97.1	95.5	75.6	92.8	96.8	97.3	78.3	92.2	87.6	96.9	96.5	93.6	98.5	81.6	93.1	83.2	98.5	89.3	92
ML-GCN [16]	99.5	98.5	98.6	98.1	80.8	97.2	97.2	98.2	82.3	95.7	86.4	98.2	98.4	96.7	99	84.7	96.7	84.3	98.9	93.7	94
F-GCN [35]	99.5	98.5	98.7	98.2	80.9	94.8	97.3	98.3	82.5	95.7	86.6	98.2	98.4	96.7	99	84.8	96.7	84.4	99	93.7	94.1
FLNet [28]	99.6	98.1	**98.9**	97.9	84.6	95.3	96.2	96.5	85.6	96.1	87.2	97.7	98.6	97	98.1	86.5	97.4	86.5	98.8	90.8	94.4
CFMIC [36]	99.7	98.5	98.8	98.3	83.9	96.5	97.5	98.8	83.1	96.1	87.4	98.6	98.9	97.2	99	85.4	97.1	84.9	99.2	94.2	94.7
MCAR [37]	99.7	99	98.5	98.2	85.4	96.9	97.4	**98.9**	83.7	95.5	88.8	99.1	98.2	95.1	99.1	84.8	97.1	87.8	98.3	94.8	94.8
SSGRL [8]	99.7	98.4	98	97.6	**85.7**	96.2	98.2	98.8	82	98.1	89.7	98.8	98.7	97	99	86.9	98.1	85.8	99	93.7	95
ML-AGCN [6]	**99.9**	98	98.5	98	81.6	96.8	96.6	98.2	85.6	**99.4**	88.2	**99.2**	99	96.5	98.8	84.8	**99.5**	88.1	98.9	94.5	95
KGGCN	99.8	**98.8**	98.7	**99**	84.8	**98.7**	**98.8**	98.2	**89.4**	98.7	**92**	98.4	**99**	**98.4**	**99.3**	**88.3**	98.4	**90.7**	99.5	**96**	**96.24**



(a)



(b)

Fig. 3: Image classification results with (a) Darknet53 and (b) KGGCN.

TABLE III: Ablation study on the impact of knowledge graph and embedding methods integration on the VOC 2007 dataset.

KG	GCN	Embedding Method	All							Top-3						Total Params	Params Size (MB)
			mAP	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1		
X	X	X	95.64	**97.9**	80.9	88.6	**97.4**	79.6	86.8	49.6	96.5	65.5	49.8	95.9	64.7	52,662,004	210.65
X	✓	✓	95.99	97.3	84.4	90.4	96.5	**83.4**	89.1	**49.7**	96.6	**65.6**	**52**	96	66.5	53,483,232	213.93
✓	✓	✓	**96.24**	97.6	**84.4**	**90.5**	97.1	83.3	**89.3**	49.6	**96.7**	65.5	50	**96**	**66.5**	53,483,232	213.93

V. Conclusion

We introduced KGGCN, a knowledge-guided graph convolutional network integrated with Darknet53 backbone for multi-label image classification. By leveraging ConceptNet Numberbatch embeddings, the model effectively captures label dependencies and semantic context that conventional CNNs often overlook. Comprehensive experiments on VOC 2007 and COCO demonstrated its superiority, achieving 96.24% and 85.25% mAP, respectively, and outperforming state-of-the-art methods across most metrics. Ablation studies further confirmed that ConceptNet embeddings significantly enhance mAP, precision, and F1-score without increasing complexity. Overall, KGGCN provides a scalable, knowledge-enhanced framework that advances

TABLE IV: Ablation study evaluating the impact of integrating knowledge graphs and embedding methods on AP and mAP on the VOC 2007 dataset

KG	GCN	Embedding Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
X	X	X	99.8	98.6	98.2	98.7	83.8	**98.8**	98.6	97.9	88.1	**99**	90.9	98.1	98.8	98.1	99.1	86.4	97.5	88.2	**99.6**	94.8	95.64
X	✓	✓	**99.9**	98.8	**98.8**	98.7	83.5	98.1	98.8	98.1	89.1	98.6	92	98.4	98.9	**98.5**	99.3	87.1	97.8	90.2	99.5	95.7	95.99
✓	✓	✓	99.8	**98.8**	98.7	**99**	**84.8**	98.7	**98.8**	**98.2**	**89.4**	98.7	**92**	**98.4**	**99**	98.4	**99.3**	**88.3**	**98.4**	**90.7**	99.5	**96**	**96.24**

TABLE V: Ablation study on the impact of knowledge graph and embedding methods integration on the COCO dataset.

KG	GCN	Embedding Method	All							Top-3						Total Params	Params Size (MB)
			mAP	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1		
x	x	x	81.03	**97.3**	56.7	71.6	**93.3**	49.4	60.6	68.7	70.9	69.8	65.9	66.1	64.4	52,692,784	210.77
x	✓	✓	83.80	96.6	65.2	77.9	93.0	60.1	71.0	69.9	72.2	71.0	67.8	68.5	66.7	53,483,232	213.93
✓	✓	✓	**85.25**	96.1	**69.3**	**80.5**	92.8	**64.7**	**74.6**	**70.4**	**72.7**	**71.6**	**69.3**	**69.2**	**68.0**	53,483,232	213.93

TABLE VI: Accuracy comparisons with different values of lk.

lk	VOC					COCO				
	All			Top-3		All			Top-3	
	mAP	OF1	CF1	OF1	CF1	mAP	OF1	CF1	OF1	CF1
1	96.227	90.41	89.16	65.55	65.01	85.232	80.51	74.56	71.56	67.94
2	96.237	90.49	89.24	65.55	65.01	85.230	80.51	74.57	71.58	67.96
3	96.198	90.29	88.96	65.61	**65.24**	85.079	80.31	74.43	71.51	67.70
4	96.231	90.46	89.22	65.56	65.10	**85.249**	**80.52**	**74.62**	**71.58**	**67.95**
5	**96.242**	**90.53**	**89.30**	**65.62**	65.05	85.234	80.49	74.58	71.56	67.89

TABLE VII: Accuracy comparisons with different values of Trade-off Parameter ε.

ε	VOC					COCO				
	All			Top-3		All			Top-3	
	mAP	OF1	CF1	OF1	CF1	mAP	OF1	CF1	OF1	CF1
0.1	96.1430	90.33	89.09	65.53	63.78	83.282	78.35	71.51	70.67	66.35
0.3	96.2422	90.44	89.19	65.53	64.82	84.861	80.08	73.88	71.48	67.60
0.5	96.2402	90.52	89.27	**65.54**	65.02	85.202	80.48	74.54	71.55	67.91
0.7	**96.2424**	**90.53**	**89.30**	65.53	**65.05**	**85.249**	**80.52**	**74.62**	**71.58**	**67.95**
0.9	96.2418	90.53	89.28	65.51	65.04	85.228	80.47	74.55	71.59	67.96

robust and interpretable multi-label recognition. Future work will extend this approach to temporal visual tasks, such as video classification, and explore dynamic graph learning to adaptively capture semantic associations. To encourage reproducibility and further research, we will release the source code and pretrained models upon acceptance.

ACKNOWLEDGMENT

This research is supported by Faculty of Sci Eng & Built Env, Deakin University, VIC 3220, Australia. OpenAI (ChatGPT) was used solely to assist with language editing and grammar refinement. All technical content, results, and conclusions were produced and verified by the authors.

REFERENCES

[1] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pp. 684–700, Springer, 2016.

[2] R. Arora, I. Saini, and N. Sood, "Multi-label classification of thoracic diseases using structure deep learning framework," *Biomedical Engineering: Applications, Basis and Communications*, vol. 36, no. 02, p. 2450006, 2024.

[3] K. Kang, W. Ouyang, H. Li, and X. Wang, "Object detection from video tubelets with convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 817–825, 2016.

[4] Y. Luo, J. Qin, X. Xiang, and Y. Tan, "Coverless image steganography based on multi-object recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 7, pp. 2779–2791, 2020.

[5] C. Dewi, R. Amirzadeh, D. Thiruvady, and N. Zaidi, "Knowledge-guided object detection via bayesian networks and knowledge graphs (kgbncnet)," *Expert Systems with Applications*, p. 129385, 2025.

[6] I. P. Singh, E. Ghorbel, O. Oyedotun, and D. Aouada, "Multi-label image classification using adaptive graph convolutional networks: from a single domain to multiple domains," *Computer Vision and Image Understanding*, vol. 247, p. 104062, 2024.

[7] T. Ridnik, H. Lawen, A. Noy, E. Ben Baruch, G. Sharir, and I. Friedman, "Tresnet: High performance gpu-dedicated architecture," in *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1400–1409, 2021.

[8] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 522–531, 2019.

[9] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[11] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.

[12] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[13] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[14] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[15] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *Proceedings of the IEEE international conference on computer vision*, pp. 464–472, 2017.

[16] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5177–5186, 2019.

[17] P. Kapanipathi, V. Thost, S. S. Patel, S. Whitehead, I. Abdelaziz, A. Balakrishnan, M. Chang, K. Fadnis, C. Gunasekara, B. Makni, *et al.*, "Infusing knowledge into the textual entailment task using graph convolutional networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 8074–8081, 2020.

[18] A. L. Maas, A. Y. Hannun, A. Y. Ng, *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, p. 3, Atlanta, GA, 2013.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[20] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[21] J. Liu, C. Fan, F. Zhou, and H. Xu, "Complete feature learning and consistent relation modeling for few-shot knowledge graph completion," *Expert Systems with Applications*, vol. 238, p. 121725, 2024.

[22] Y. Fang, K. Kuan, J. Lin, C. Tan, and V. Chandrasekhar, "Object de-

tection meets knowledge graphs," in -, International Joint Conferences on Artificial Intelligence, 2017.

[23] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in *Sixth international conference on data mining (ICDM'06)*, pp. 613–622, IEEE, 2006.

[24] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5513–5522, 2017.

[25] W. Ge, S. Yang, and Y. Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1277–1286, 2018.

[26] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2285–2294, 2016.

[27] J. Zhang, Q. Wu, C. Shen, J. Zhang, and J. Lu, "Multilabel image classification with regional latent semantic dependencies," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2801–2813, 2018.

[28] D. Sun, L. Ma, Z. Ding, and B. Luo, "An attention-driven multi-label image classification with semantic embedding and graph convolutional networks," *Cognitive Computation*, pp. 1–12, 2023.

[29] T. Chen, L. Lin, R. Chen, X. Hui, and H. Wu, "Knowledge-guided multi-label few-shot learning for general image recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1371–1384, 2020.

[30] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16478–16488, 2021.

[31] S.-F. Chen, Y.-C. Chen, C.-K. Yeh, and Y.-C. Wang, "Order-free rnn with visual attention for multi-label classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[32] H. Yang, J. Tianyi Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai, "Exploit bounding box annotations for multi-label object recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 280–288, 2016.

[33] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "Hcp: A flexible cnn framework for multi-label image classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1901–1907, 2015.

[34] T. Chen, Z. Wang, G. Li, and L. Lin, "Recurrent attentional reinforcement learning for multi-label image recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[35] Y. Wang, Y. Xie, Y. Liu, K. Zhou, and X. Li, "Fast graph convolution network based multi-label image recognition via cross-modal fusion," in *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 1575–1584, 2020.

[36] Y. Wang, Y. Xie, J. Zeng, H. Wang, L. Fan, and Y. Song, "Cross-modal fusion for multi-label image classification with attention mechanism," *Computers and Electrical Engineering*, vol. 101, p. 108002, 2022.

[37] B.-B. Gao and H.-Y. Zhou, "Learning to discover multi-class attentional regions for multi-label image recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 5920–5932, 2021.