Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061086

ARTICLE

A Novel Approach Based on Graph Attention Networks for Fruit Recognition

Dat Tran-Anh[1] **and Hoai Nam Vu**[2,3,*]

[1]Faculty of Information Technology, Thuyloi University, Ha Noi, 100000, Viet Nam
[2]Faculty of Artificial Intelligence, Posts and Telecommunications Institute of Technology, Nguyen Trai, Ha Noi, 100000, Viet Nam
[3]Young Innovation Research Laboratory on Digital Technology (YIRLoDT), Posts and Telecommunications Institute of Technology, Nguyen Trai, Ha Noi, 100000, Viet Nam

*Corresponding Author: Hoai Nam Vu. Email: namvh@ptit.edu.vn

Received: 16 December 2024; Accepted: 02 January 2025; Published: 17 February 2025

ABSTRACT: Counterfeit agricultural products pose a significant challenge to global food security and economic stability, necessitating advanced detection mechanisms to ensure authenticity and quality. To address this pressing issue, we introduce iGFruit, an innovative model designed to enhance the detection of counterfeit agricultural products by integrating multimodal data processing. Our approach utilizes both image and text data for comprehensive feature extraction, employing advanced backbone models such as Vision Transformer (ViT), Normalizer-Free Network (NFNet), and Bidirectional Encoder Representations from Transformers (BERT). These extracted features are fused and processed using a Graph Attention Network (GAT) to capture intricate relationships within the multimodal data. The resulting fused representation is subsequently classified to detect counterfeit products with high precision. We validate the effectiveness of iGFruit through extensive experiments on two datasets: the publicly available MIT-States dataset and the proprietary TLU-States dataset, achieving state-of-the-art performance on both benchmarks. Specifically, iGFruit demonstrates an improvement of over 3% in average accuracy compared to baseline models, all while maintaining computational efficiency during inference. This work underscores the necessity and innovativeness of integrating graph-based feature learning to tackle the critical issue of counterfeit agricultural product detection.

KEYWORDS: Fruit recognition; graph attention network; multi-feature processing

1 Introduction

1.1 Background and Motivation

Counterfeit agricultural products represent one of the most serious threats to online agricultural business services today. The increasing complexity of business transactions exacerbates this issue facilitated through online payment applications (e.g., eBay and Alipay) and credit cards [1–3]. Fraudsters employ various tactics to counterfeit agricultural products, such as substituting them with cheaper alternatives. In this endeavor, we aim to identify counterfeit agricultural products using artificial intelligence on a real-world platform.

In the realm of commercial markets, counterfeit agricultural products stand out as a significant component of the potential risks faced by buyers. In our earlier detection systems, we noted that factors linked to specific regions, such as the geographical origin of crops and their distinctive colors, serve as pivotal indicators for identifying fraudulent activities. These factors are used to distill numerous patterns, numbering

in the hundreds. These are then integrated as attributes within a machine learning model or formulated as rules within a decision engine.

Nevertheless, expertly crafted by human specialists, the feature engineering process for these associative patterns remains effective only within a single step in the connected data structure. When considering steps beyond this immediate connection, human experts' task of uncovering and comprehending these patterns becomes notably inefficient. This inefficiency arises due to the considerable volume of derived characteristics generated through the cumulative aggregation of original attributes along the paths of risk propagation.

Distinctive traits of the graph for fruit recognition. Two critical attributes of identity graphs have been discerned [4,5] within a comparable application to detect counterfeit agricultural products. Initially, domains and their associated entities inherently constitute a graph comprising heterogeneous nodes. These fraudulent products tend to exhibit shared risk entities, like the crop's nomenclature and the geographical region of cultivation. Secondly, detecting counterfeit agricultural products within the identity graph necessitates addressing the issue of multi-feature, as products utilized by malevolent actors and authorized users frequently manifest resemblances (e.g., color, size, etc.).

Challenge 1: Integration of image representation features for agricultural objects. *Can we rely solely on information extracted from a fundamental feature of the agricultural product's characteristics?* Concerning the issue of counterfeit agricultural products, the image representation features—each delineating distinct attributes of the image into features—encompass various characteristics, each bearing its unique significance. Two crucial attributes come to the fore: (i) Image similarity attribute; and (ii) Text similarity attribute. Intuitively, the feature patterns exhibited by agricultural objects find interconnections within attribute data, encompassing images, text, and the interplay between the two. Consequently, the present study aims to combine these attributes to discern counterfeit agricultural products effectively.

Challenge 2: Graph inference efficiency. Within the realm of graph inference, each connecting edge serves distinct objectives. Specific scenarios, like the Dalat potato recognition procedure encompassing potato procurement and post-purchase evaluation for detecting tampered agricultural products, might accommodate relatively extended latency periods. Nonetheless, the prompt identification of counterfeit agricultural products necessitates swift responsiveness, given the heightened user sensitivity to latency. In general, models employing Graph Attention Networks (GAT) aggregate information from neighbors spanning at least two graph steps; consequently, each inference operation often demands hundreds of milliseconds or even seconds, rendering it inadequate to fulfill internal requisites and user-side anticipations regarding system latency.

1.2 Main Contributions

To address the aforementioned challenges, we propose the iGFruit (Integration of image representation features and Graph attention networks for FRUIT recognition) framework.

- We propose a novel Graph Attention Networks (GAT) transformation technique designed for feature-centric forgery identification graphs, resulting in a dual-stage Navigated Graph (TD) storage approach. This architecture is attuned to features and holds the potential for instantaneous alerts regarding counterfeit fruit products.
- Based on the features, we introduce a Neural Network (NN) architecture that supports multi-feature and real-time processing. The Neural Network architecture leverages snapshot feature synthesis. We also improved the efficiency of real-time inference using the Lambda architecture.

- Our tests show that iGFruit outperforms basic models by more than 33% in average accuracy. In addition, detecting fake agricultural products in real time is computationally efficient. iGFruit reduces end-to-end inference latency by more than 3% (including neighborhood query and graph inference), compared to traditional GNN inference frameworks. Our speedup is 0.58 on average for the inference phase compared to traditional GNN.

1.3 Organization

The structure of this paper is organized as follows: Section 2 provides a comprehensive review of related works on fruit recognition and explores the potential of multi-feature approaches combined with GAT. In Section 3, we define the primary problem and introduce the proposed framework. Section 4 details the experimental setup and presents the corresponding results. Lastly, Section 5 concludes the paper with a discussion of the findings and their implications.

2 Related Works

Numerous feature types are employed in various applications, including image features [6], text features [7], shape, color, and others. Ahmed et al. [8] proposed a methodology integrating three distinct feature extraction techniques: the Bag of Words (BoW) model, intensity graph, and gray-level co-occurrence matrix (GLCM). In 2022, Ge et al. [9] enhanced SVM-based classification by leveraging a crop segmentation classifier and Multi-features Fusion. Their approach incorporated features from Red Green Blue (RGB), Hue Saturation Value (HSV), Curvatures, Fast Point Feature Histogram (FPFH), and Spin Image, significantly improving both the accuracy and processing speed of the classifier. Similarly, Vishnoi et al. [10] employed a combination of the GLCM, Gabor feature extraction algorithm, and k-means clustering segmentation to extract features such as contrast, correlation, homogeneity, entropy, shape, color, texture, and intensity. These image-processing techniques, combined with computational intelligence or soft computing methods, have demonstrated their potential to assist farmers in detecting diseases more effectively. This integration of algorithms enabled more efficient clustering at hierarchical levels, leading to improved outcomes in accuracy and processing speed.

In our previous works, specifically involving the Graph Convolutional Network (GCN) [11] and GAT [12] methodologies, we undertook an in-depth exploration of the complexities inherent in dynamic graph modeling for the realm of image recognition. To elucidate further, GCN [11] was conceptualized primarily to enhance image recognition capabilities by deploying a dynamic heterogeneous graph neural network. Within the GAT framework, two distinct classifications of subgraphs are delineated: firstly, the structural subgraph corresponding to each timestamp, which serves to explicate the intricate interdependencies among diverse entity categories; secondly, a temporal subgraph that establishes connections between these structural subgraphs utilizing feature edges. Consequently, temporal dimensions seamlessly integrate within an expansive graph, coexisting with the structural linkages. The assimilation of this histogram technique equips GAT with the prowess to outperform elementary benchmarks such as GCN and GNN [13] in the domain of image recognition.

GAT [12], constructed as an extension of GCN [11], has undergone examination for the integration of feature embeddings while upholding substructure representations. This approach investigates the substitution of elementary convolutional layers with more intricate alternatives, such as the heterogeneous transformer layer [14]. Furthermore, we delve into the configuration of histogram architecture and data distribution across multiple features, along with their consequential impacts on the dynamics inherent in the GNN model.

However, we notice three issues when deploying these prototype models in production:

- **(I1)** The utilization of a two-dimensional graph structure involving domains and objects has the propensity to consume significant GPU memory resources, particularly when the graph expands to encompass millions of vertices and edges.
- **(I2)** In many instances, the use of aggregation functions like mean and max in Graph Neural Networks (GNNs) poses challenges for the model in distinguishing between distinct structures, as these functions are inherently non-injective.
- **(I3)** Multi-feature of image and text, GCN and GNN exhibit proximity query latencies exceeding several hundred milliseconds or even seconds, which falls short of being ideal for a real-time responsive system.

3 Research Methodology

In this work, we propose iGFruit (Integration of image representation features and Graph attention networks for FRUIT recognition) to tackle these remaining issues when deploying GAT in production.

- To address the issue stated as **(I1)**, we opt to preserve the structural diagrams of the Graph Attention Network (GAT). Subsequently, we partition these diagrams to facilitate domain-specific learning as well as object-centric learning, all performed on a part-by-part basis.
- To solve **(I2)**, we strictly control using only feature information to predict fake agricultural products. Here, we also combine the features into a generic feature of the forgery product that helps support GAT in making more accurate decisions.
- In order to mitigate the expense associated with querying neighbors as outlined in **(I3)**, we implement incorporating batch inference capabilities, clustering features, and housing feature embeddings within a key-value store. Subsequently, these stored feature embeddings are leveraged during inference processes to curtail inference latency effectively.

In this section, we introduce and explain the multi-feature Convolutional Neural Network (Section 3.2), illustrate the graph transformation module GAT (Section 3.2) and experiments as well as a detailed discussion of the components of iGFruit (Section 4.4).

3.1 Problem Definition

3.1.1 Notations and Basic Assumptions

The notations utilized in this manuscript are succinctly summarized in Table 1 and are further elaborated in the subsequent sections detailing our methodology and technical framework.

Table 1: Summary of notations

Notation	Description
x	The input image
$\mathcal{D}$	The domain set includes product origins
$\mathcal{O}$	The object set includes types of product
$\mathcal{Y}_s$	The seen labels
$\mathcal{Y}_u$	The unseen labels
y	The label belonging to one of the seen labels $\mathcal{Y}_s$
p_c^u	representation vector (Prototype) for class c in S^u
ω	Weights of embedding network
$d(a, b)$	Euclidean distance of vector a and vector b

This study addresses the challenge of identifying counterfeit agricultural products in scenarios where a novel crop emerges abruptly within an agricultural system lacking prior crop-specific data. For example, the case of Dalat potatoes illustrates such a scenario, where data collection is confined to potatoes cultivated in the Da Lat region. Additionally, we consider a setting involving multiple localized agricultural hubs, each capable of collecting farmer-specific data samples. However, these centers are reluctant to share data to protect proprietary product information. Our approach enables classification and detection without relying on a unified, extensive dataset.

To achieve this, we first collect labeled data on agricultural products from localities with sufficient data availability. This dataset, referred to as the base set, is denoted as $\mathcal{A} = (x, y) \mid x \in \mathcal{X}, y \in \mathcal{Y}$, where x represents an image in the RGB space $\mathcal{X}$, and y is its label from the label set $\mathcal{Y}$. Each label is structured as a tuple $y = (d, o)$, where $d \in \mathcal{D}$ is the domain and $o \in \mathcal{O}$ is the object, with $\mathcal{D}$ and $\mathcal{O}$ representing the sets of domains and objects, respectively. Training separate classifiers for domains and objects risks overfitting due to their interdependence; for example, visual features of a sliced domain differ significantly depending on whether the object is an apple or a potato. To address this, we propose jointly modeling domains and objects using a compatibility function $\xi : \mathcal{X} \times \mathcal{D} \times \mathcal{O} \to \mathbb{R}$, which evaluates the compatibility between an image, domain, and object. Given an input image x, the label $y = (d, o)$ is determined by identifying the domain-object pair with the highest compatibility score. This integrated modeling approach reduces overfitting risks and effectively captures the intertwined nature of the problem.

Our approach leverages the Graph Attention Network (GAT) framework [12], which utilizes the base set $(\mathcal{X}, \mathcal{Y})$ to train a classifier. GAT addresses the complexities of dynamic node relationships over extended distances and reduces the reliance on the individual graph structure of nodes. It employs an attention mechanism to aggregate feature vectors from neighboring nodes, allowing each node to integrate information from all its neighbors. This is achieved by assigning coefficients to neighboring features while accounting for the influence of distant nodes. Although this approach enhances feature integration, it has increased computational and memory demands.

We consider the input image as x and the pair of input texts as 'domain-object' as y. We employ N_V backbones (denoted as $f_V^1, \ldots, f_V^{N_V}$) to extract image features and N_T backbones (denoted as $f_V^1, \ldots, f_V^{N_T}$) to extract text features. Additionally, we utilize a fusion function f_F to merge these features, resulting in N_K-dimensional vectors for each of the $N_V + N_T + 1$ extracted features. In this paper, we exclusively use six features inspired by the principles of image analysis [15], the BERT language model [16], multimodal inputs [17], and TF-IDF representation [18].

$$e_{ij} = \mathcal{A}(\mathcal{W}v_i, \mathcal{W}v_j), \tag{1}$$

where e_{ij} are computed for nodes $j \in \mathcal{N}(i)$, and $\mathcal{N}(i)$ is the neighborhood of node i in the graph. Then, the coefficients are normalized across all j using the softmax function [15]:

$$\alpha_{ij} = softmax_j(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} exp(e_{ik})}. \tag{2}$$

3.1.2 Problem Formulation

Definition 3.1 (Transform the visual and text features into dimensional vectors with the same dimension) For a given dataset consisting of x as input images and y as labels representing agricultural produce, the output comprises the features extracted from both x and y.

Definition 3.2 (Fruit Recognition with GAT) The input consists of the features of the input image and the labels representing agricultural produce. The output is a GAT (Graph Attention Network) graph that recognizes agricultural produce.

The input to our GAT layer is a set of transformed subgraphs $\mathcal{G}$, $\mathcal{G} = \{g_1, g_2, \ldots, g_N\}$, $g_i \in \mathbb{R}^{N_F}$. The directed graph is structured to simplify the visualization of the relationship mode from the perspective of target features, making it easier to partition the graph into various crop classification charts. Here, $\mathcal{G}$ represents a subgraph with undirected edges, each weighted with feature-specific values. The partitioning of the target features is introduced to distinguish between feature roles, as a feature can serve as either an extraction or a reference feature. All reference feature nodes share common entities stored alongside target features in a partition. In the graph, each feature is assigned only one role, either as a target or reference feature.

$$g_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} g_j^{(l)} \mathcal{W}^{(l)}\right), \tag{3}$$

where $g_j^{(l)}$ is the current input feature of layer l; $g_i^{(l+1)}$ is the output feature; σ is an activation function.

3.2 iGFruit Framework

The ongoing competition offers a captivating avenue for exploration due to its utilization of multimodal data. This intriguing aspect introduces a convergence of different data types, fostering an environment for intricate analysis and innovative solutions. The underlying pipeline designed for this competition is a testament to the complexity of the task.

The iGFruit framework, in Fig. 1, unfolds in two distinctive stages, each catering to a specific facet of the multimodal data challenge: (i) In the initial stage (Stage Training), the focus lies on the training of metric learning models. These models operate meticulously to derive cosine similarity measurements from various data sources, including images, textual content, and the fusion of both image and text data. This critical step establishes a foundation for understanding the intrinsic relationships within and between these data modalities; (ii) Moving forward to the second stage (Stage Testing), the emphasis shifts towards training "meta" models. These models, having been primed with the cosine similarity data, are responsible for classifying whether a given pair of items belongs to the same label group. This classification task is a pivotal juncture in the competition, demanding a higher level of abstraction and synthesis of the learned similarities.

In Stage Training, three core modules are highlighted: image feature extraction, text feature extraction, and feature fusion. The image feature extraction module utilizes NFNet and ViT architectures to independently process input images, generating high-dimensional embeddings that encode visual semantics. Simultaneously, the text feature extraction module employs multiple transformer-based models, including English BERT, Multilingual BERT, and XLM, to encode the textual labels (e.g., "Dalat Potato") into embeddings that capture semantic information across languages. These embeddings are then combined in the fusion block, where cross-modal similarity metrics, such as cosine similarity, are calculated. These fused features and similarity scores are fed into a graph-based module (GAT) and a gradient-boosting classifier (LightGBM), enabling the model to learn inter-modal and intra-modal relationships. Optimization during this stage is guided by two losses: a fusion loss for aligning embeddings across modalities and a classification loss for improving downstream task performance.

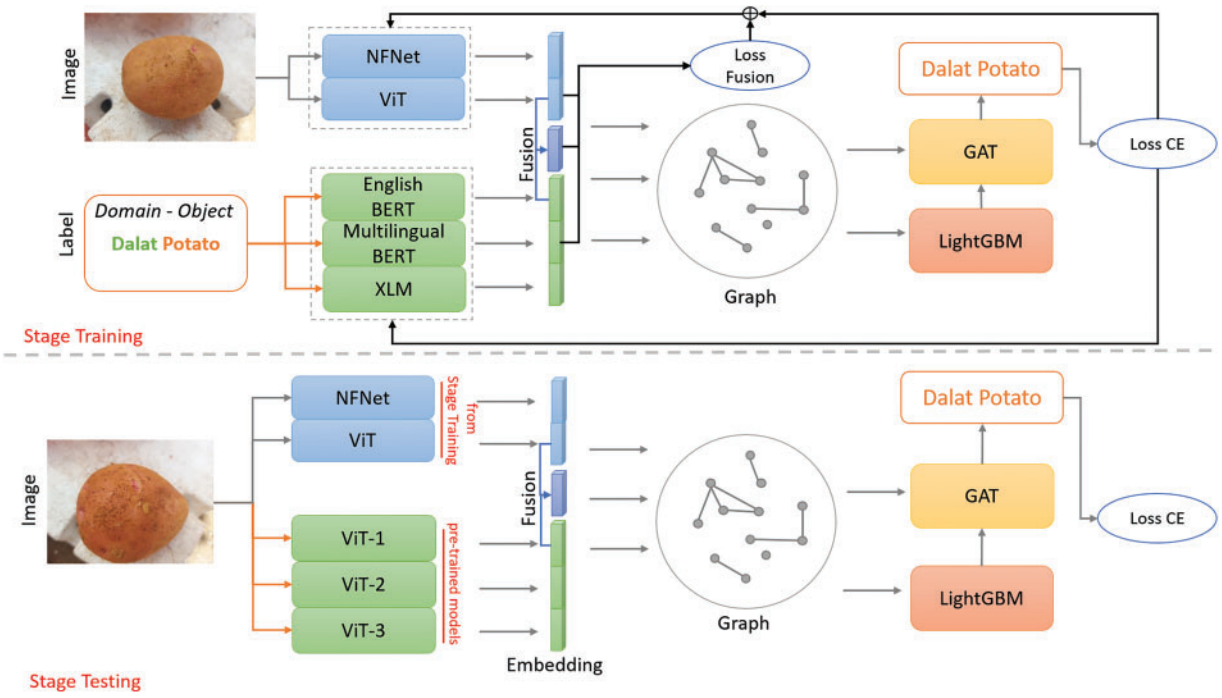

Figure 1: The general design pipeline for iGFruit framework

In the testing stage, the model utilizes pre-trained components from Stage Training, with additional enhancements to improve prediction robustness. Multiple variants of ViT (e.g., ViT-1, ViT-2, ViT-3) are employed to extract diverse embeddings from input images, enabling ensemble-style feature integration. These embeddings and pre-trained text embeddings are concatenated and passed through the GAT and LightGBM modules to classify whether the input pair belongs to the same label group. This hierarchical flow ensures that inter-modal and intra-modal relationships learned during training are preserved and leveraged effectively for inference.

We have incorporated distinct visual elements to ensure Fig. 1 effectively communicates the framework. Arrows indicating data flow (e.g., image embeddings, text embeddings, similarity scores) are labeled explicitly to represent their roles. Additionally, color coding has been introduced to differentiate key components: blue for image processing, green for text processing, and orange for graph-based classification. This refined illustration complements the textual explanation, offering a comprehensive and intuitive understanding of the iGFruit framework's design and functionality.

The selection of ViT, NFNet, and BERT as backbone networks for the iGFruit model is driven by their state-of-the-art performance and complementary strengths in handling multimodal data. ViT is chosen for its ability to effectively capture local and global visual features through self-attention mechanisms, making it particularly suitable for distinguishing subtle differences in agricultural product images. NFNet complements this by providing stable and efficient training without batch normalization, ensuring robust performance on high-variance visual data typical of agricultural products. Meanwhile, BERT is utilized for its strength in processing and understanding textual data, enabling the model to analyze product descriptions, labels, or packaging text for inconsistencies indicative of counterfeiting. These networks create a comprehensive feature representation by combining robust image processing with contextual text understanding. This multimodal synergy ensures high accuracy, computational efficiency, and scalability,

aligning with iGFruit's goal of delivering an effective and practical solution for counterfeit agricultural product detection.

A strategic selection of tools and techniques has been employed to power this intricate pipeline. LightGBM [19], a gradient boosting framework, lends its prowess to aid in making sense of the intricate patterns present within the data. Additionally, incorporating GAT underscores the importance of harnessing relational information, especially in cases where the data exhibits inherent graph-like structures. These tools synergistically complement each other, contributing to a comprehensive approach to tackling the multimodal data challenge.

In summary, this competition is an engaging platform for scientific inquiry, inviting participants to navigate the intricacies of multimodal data analysis. The pipeline's dual stages and the adept utilization of LightGBM and GAT showcase a well-rounded strategy that seeks to extract meaningful insights and solutions from the amalgamation of diverse data sources. As the competition unfolds, it promises to unravel new dimensions of understanding and innovation in the realm of multimodal data analysis.

3.2.1 Metric Learning

As illustrated in Fig. 1, our proposed iGFruit initially conducts a projection of both image and text features to acquire transformed representations, denoted as s_V and s_T. Simultaneously, the index constraints operate within and between methodologies, while the method classifiers confine the discriminative and methodology-independent learned subspace representations. Concretely, we deconstruct the process of subspace learning into two distinct loss components: (1) Cross-Entropy Loss: This is a common loss function for text and image classification tasks; (2) Discriminatory Loss: This term emphasizes the similarity of models within the same method category while ensuring that the acquired representations remain discriminative;

With (2), to uphold intrinsic discrimination within the data after feature projection, a classifier is implemented to predict the semantic labels of the projected items in the shared subspace. For this purpose, a feedforward network is activated by softmax, and it is added to the head of each embedded subspace neural network. This classifier takes the anticipated features of instances composed of concatenated image and text data as training input and generates a probability distribution of semantic categories for each item as output. Assuming y to be the underlying ground truth label of each representation, represented as a one-hot vector, and the predicted probability distribution from the classifier's output as p_i the intra-method objective function can be formulated as follows, irrespective of the representation of the object being transformed from any method.

$$\begin{aligned} L_d = {} & k_N L_N + k_V L_V + k_F L_F + k_I L_I + k_M L_M + k_X L_X \\ & - \frac{1}{N}\sum_{i}^{N}(y_i \cdot (k_N log p_i(f_N(\mathcal{I}_i)) + k_V log p_i(f_V(\mathcal{I}_i)) + k_F log p_i(f_F(\mathcal{I}_i, y_i)) \\ & + k_I log p_i(f_I(y_i)) + k_M log p_i(f_M(y_i)) + k_X log p_i(f_X(y_i)))), \end{aligned} \tag{4}$$

where $k_i, i \in \{N, V, F, I, M, X\}$ are the associated parameters that signify the contribution of each respective loss. It's important to mention that the values of k_i can be fine-tuned through empirical training experiments, which will be elaborated upon in the subsequent section.

3.2.2 Joint feature fusion

The study incorporates five distinct input sources for classification tasks related to individual patient nodes within the graph network. These input sources encompass NFNet [20] image feature data denoted as $x_{N,i} \in \mathcal{X}_N$, image feature data extracted from Vision Transformer (ViT) [21] labeled as $x_{V,i} \in \mathcal{X}_V$, fused

image feature data merged with ViT features termed as $x_{F,i} \in \mathcal{X}_{\mathcal{F}}$, text feature data from English BERT represented as $x_{I,i} \in \mathcal{X}_{\mathcal{I}}$, text feature data from Multilingual BERT denoted as $x_{M,i} \in \mathcal{X}_{\mathcal{M}}$ and text feature data from XML denoted as $x_{X,i} \in \mathcal{X}_{\mathcal{X}}$. These input sources collectively facilitate three separate graph network methodologies for the classification task. Particularly, the inclusion of XLM metadata can provide supplementary independent information augmenting the two other text-based contributions. A linear transformation is applied to standardize the object representation size across all methods to ensure equitable consideration of each methodology during processing. Subsequently, these standardized representations are subjected to an aggregate function σ, yielding the corresponding associative representation $g_{c,i}$, which is subsequently employed within the graph network framework.

$$g_{c,i} = \sigma(\omega(\mathcal{W}_N x_{N,i}), \omega(\mathcal{W}_V x_{V,i}), \omega(\mathcal{W}_F x_{F,i}), \omega(\mathcal{W}_I x_{I,i}), \omega(\mathcal{W}_M x_{M,i}), \omega(\mathcal{W}_X x_{X,i})). \tag{5}$$

Possible approaches for δ include concatenation, averaging, pooling, or the use of attention mechanisms. The activation function ω is a non-linear operator, and the learnable linear transformation matrices $\mathcal{W}_N \in \mathbb{R}^{F_N \times F_c}$, $\mathcal{W}_V \in \mathbb{R}^{F_V \times F_c}$, $\mathcal{W}_F \in \mathbb{R}^{F_F \times F_c}$, $\mathcal{W}_I \in \mathbb{R}^{F_I \times F_c}$, $\mathcal{W}_M \in \mathbb{R}^{F_M \times F_c}$, and $\mathcal{W}_X \in \mathbb{R}^{F_X \times F_c}$ are employed to map the input feature dimensions to a common dimensionality F_c.

The multi-head output features generated by the joint neuron attention module are subsequently combined through a gated fusion layer. This layer selectively filters and integrates the most relevant information, producing a unified representation of both image and text data. As illustrated in Fig. 2, the gated fusion layer accepts two input vectors, x_N and x_I, and outputs a fused representation that encapsulates the essential features from both modalities.

$$x_F = x_N \oplus x_I. \tag{6}$$

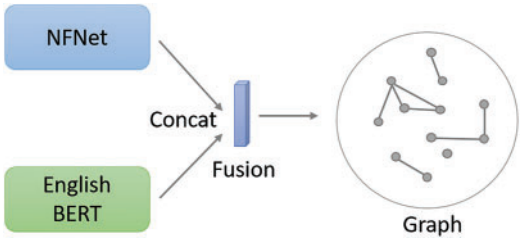

Figure 2: The fusion process for fruit recognition

To facilitate the training of feature extraction models within the training pipeline, the loss value is computed from six different sources, including cross-entropy loss from Eq. (4) and Huber loss. It is important to emphasize that the reason for opting for Huber loss over mean square error (MSE) or mean absolute error (MAE) lies in the fact that Huber loss combines both the characteristics of MAE and MSE. The challenge associated with using MAE in training CNN and BERT-based models stems from its consistently steep gradient, which can lead to difficulties in converging to the minimum point at the end of training. In contrast, MSE exhibits a more precise gradient as the loss approaches its local minimum, making it more advantageous in this regard. However, MSE loss is less robust to outliers, while Huber loss has been demonstrated to not only be less sensitive to outliers in the data compared to MSE but also possesses differentiability at zero. Consequently, Huber loss combines robust properties from both MSE and MAE. In our research, the Huber

loss between the six features is expressed as follows:

$$L_{huber} = \begin{cases} \frac{1}{2}(\hat{y})^2 & if\,|(\hat{y})| < \mathcal{T} \\ \mathcal{T}\left(\hat{y} - \frac{1}{2}\mathcal{T}\right) & otherwise \end{cases}, \tag{7}$$

where $\mathcal{T}$ is the threshold which is empirically set to 1 and $\hat{y}$ is the difference between three feature vectors from six auto-encoders, defined according to the following formula:

$$\begin{aligned} \hat{y} = & \frac{1}{2}(mean(f_I(y), f_M(y), f_X(y)) - mean(f_N(\mathcal{I}), f_V(\mathcal{I}))) \\ & + \frac{1}{2}(f_F(\mathcal{I}, y) - mean(f_N(\mathcal{I}), f_V(\mathcal{I}))). \end{aligned} \tag{8}$$

3.2.3 Meta Model with GAT

The graph processing in our proposed method is based on the Graph Attention Network (GAT) [12], which integrates efficient neighborhood processing with the ability to generalize to unseen data samples. This is achieved while maintaining localized filtering and low computational complexity. Each GAT layer processes 1-hop neighborhood regions within the graph $\mathcal{G}(V, E)$ for each vertex e_i. The representation of each vertex is updated by considering not only its own feature vector but also those of its neighbors, which are transformed through the network layer.

In GAT, the significance of each neighbor is determined by learned attention coefficients, which evaluate the contribution of each feature vector to the update of the target vertex's representation. The updated feature representation, $g_{c,i}^{att}$, is computed using the following equation:

$$g_{c,i}^{att} = \|_{p=1}^{P} \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{p} \mathcal{W}^{p} g_{c,j}\right). \tag{9}$$

Here, α_{ij}^{p} is the learned attention coefficient for attention head p, encoding the significance of the neighboring feature vector $g_{c,j}$ in the 1-hop neighborhood $\mathcal{N}(i)$ for updating $g_{c,i}$. The term $\mathcal{W}^{p}$ denotes a learned linear transformation, and α_{ij}^{p} includes self-attention, enabling the model to consider a node's own representation in the update process.

The graph processing approach leverages attention mechanisms to effectively integrate meta-data, $\mathcal{X}$, into the learning process. The meta-data is structured into the graph by creating neighborhoods $\mathcal{N}(i)$ based on feature similarity, enabling the incorporation of features from neighbors $g_{c,j} \in \mathcal{N}(i)$ in the transformation of $g_{c,j}$. Consequently, each transformed representation depends not only on its own features but also on weighted contributions from its neighbors. This mechanism stabilizes predictions for instances where initial representations of a class are indistinct but are located within well-defined data clusters, improving classification accuracy and robustness.

The GAT synthesizes representations of neighboring nodes along dependent paths. However, this process does not consider all dependency relationships, potentially leading to the loss of crucial dependent information. Intuitively, nodes with different dependency relationships can have varying influences. We propose an extension of the original GAT by incorporating six features from both images and text, including image, vit, fusion image and text feature, bert, multimodal, and TF-IDF similarities. We utilize these features as intelligent gates for relationships to control the flow of information from neighboring nodes, specifically

applying graph attention to "other edges connected to the node connected to the target edge" within the neighborhood. Additionally, we preprocess the graph by: (1) Ensuring that label groups form a clique; (2) Bridging two cliques with abundant edges; (3) Assigning higher betweenness centrality to such edges; and (4) Recursively removing such edges until there are none left or connected components become smaller than a predefined threshold. The overall architecture of this method is depicted in Fig. 3.

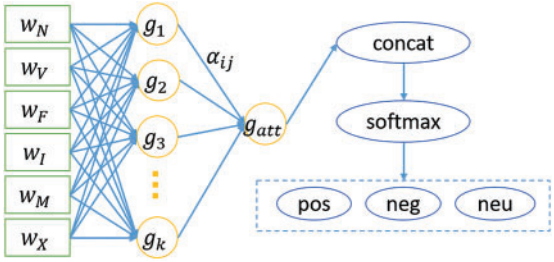

Figure 3: Structure of the proposed graph attention network (GAT)

In addition to using Huber loss, we aim to train iGFruit to achieve optimal performance in the activity recognition task, meaning that the embedding vectors can also be utilized for real-world action recognition. Consequently, we introduce another cross-entropy loss similar to the one in Eq. (4). The final loss of the validation stream is backpropagated to the feature selection layer as follows:

$$L_{iGFruit} = k_a L_d + k_b L_{huber} + k_c L_{CE}, \tag{10}$$

where k_a, k_b, and k_c are the associated parameters that signify the contribution of each respective loss. It's important to mention that the values of k_a, k_b and k_c can be fine-tuned through empirical training experiments, which will be elaborated upon in the subsequent section.

4 Experiments and Discussions

4.1 Data Description

To assess the efficacy of our approach, we analyze two distinct datasets: MIT-States [22] and TLU-States (our dataset). Each of these datasets is accompanied by authentic annotations. The MIT-States dataset comprises a total of 53,000 images, out of which 10,000 are designated as query images. For each query image, there exists a meticulously curated list of relevant images that serve as the ground truth. Concurrently, the TLU-States dataset encompasses 9000 images, organized into 21 clusters. Within this dataset, every image is associated with a set of three corresponding images that have been established as part of the ground truth information.

This ground truth information in both datasets is vital for evaluating the performance of our approach in a controlled and reliable manner. Specifically, the "list of relevant images" in the MIT-States dataset refers to images that share similar semantic and visual attributes with the query image, carefully curated through domain-specific knowledge and validated annotations. Similarly, in the TLU-States dataset, our research team developed the associated set of three relevant images for each image through a meticulous annotation process. The creation of TLU-States involved a structured pipeline, including data collection from diverse sources, manual annotation of images with attributes such as geographic origin, object type, and cluster labels, and a rigorous quality control process to ensure the integrity and consistency of the annotations. The curated ground truth pairs in TLU-States were designed to reflect real-world semantic and visual

relationships, enabling robust evaluation of retrieval and classification tasks. The TLU-States dataset offers a novel contribution to the field by addressing gaps in existing datasets, providing a rich diversity of high-quality annotated images organized into clear clusters, and serving as a benchmark for evaluating advanced machine learning models. This addition complements the MIT-States dataset, allowing for a comprehensive analysis of our approach across different scales and domains.

The TLU-States dataset is organized into 21 prominent clusters, each corresponding to a specific product type and geographic origin. These clusters include, for instance, the American apple (386 images), American grape (251 images), American orange (81 images), Australian orange (43 images), China apple (64 images), and China potato (280 images). Additionally, the dataset features key clusters such as the Dalat potato (1,290 images), japan apple (383 images), and Vietnam grape (256 images), among others. Each cluster is designed to reflect unique semantic and visual characteristics, providing a nuanced evaluation ground for machine learning models. The diversity within the dataset is exemplified by its inclusion of rare combinations, such as the South African grape (147 images) and New Zealand apple (666 images), alongside more common clusters like Vietnam orange (59 images).

By carefully balancing cluster sizes and ensuring diversity in image composition, the TLU-States dataset serves as a robust benchmark for tasks requiring both inter-cluster and intra-cluster discrimination. The explicit annotations and ground truth lists further enhance the dataset's utility for retrieval, classification, and evaluation tasks, making it an invaluable resource for the community.

4.2 Metrics

The evaluation of the model's performance on primitives, specifically in object and state classification, is conducted across various bias factors associated with unseen compositions. The results are quantified using multiple metrics to ensure a comprehensive assessment. First, the Best Seen metric measures the highest accuracy achieved exclusively on images belonging to seen compositions, reflecting the model's performance on familiar data. Conversely, the Best Unseen metric evaluates the highest accuracy attained on images of unseen compositions, highlighting the model's generalization capability. The Best Harmonic Mean (Best HM) is reported to balance these two aspects, representing the optimal harmonic mean of the seen and unseen accuracies. Additionally, the Area Under the Curve (AUC) is computed to provide a holistic view of the model's performance by considering seen and unseen accuracies across a range of bias values. These metrics collectively offer a rigorous analysis of the model's ability to generalize while maintaining robust performance on known compositions.

4.3 Baselines

We compare our proposed method-iGFruit with state-of-the-art methods:

- *SymNet:* SymNet introduces the previously overlooked principle of symmetry in attribute-object transformations. SymNet enhances attribute modeling by considering coupling and decoupling operations, guided by symmetry and group theory principles. It trains these networks in an end-to-end paradigm, optimizing them based on group axioms and the symmetry property.
- *LabelEmbed+:* LabelEmbed+ embeds attributes, object vectors, and image features into a semantic space which allows us to represent these components in a shared, meaningful way. It optimizes the input representations, adapting them to capture the relationships between attributes and objects better.
- *Attribute as Operators:* Attributes as Operators creates a semantic embedding that explicitly separates attributes from their accompanying objects
- *Task-Modular Neural Networks (TMN):* TMN adopts a joint processing approach that considers the input image, object, and attributes together, which enhances its understanding of visual information. The key

innovation lies in the use of gatings that depend on the specific object-attribute pair in the input. These gatings dynamically control the flow of information, allowing TMN to adaptively compose modules for different tasks

4.4 Experiments and Discussions

This section begins with an overview of the pre-training details, followed by an in-depth experimental analysis of our proposed method. The experimental analysis is structured to address the following key research questions (RQ):

- *RQ1*. Effectiveness of iGFruit: How does our proposed iGFruit method perform compared to state-of-the-art approaches in the domain?
- *RQ2*. Impact of iGFruit on Classifier Performance: To what extent can iGFruit enhance the accuracy and overall performance of the classifier, as demonstrated through ablation studies?

4.4.1 Experimental Setup

Data preprocessing. The data preprocessing pipeline for our approach consisted of several steps to ensure data quality and enhance model performance. First, all raw images underwent a cleaning process to remove duplicates, filter out low-quality images, and normalize labels for consistency. Each image was then resized to a fixed resolution of 224×224 pixels, ensuring compatibility with the input layer of deep learning models. Data augmentation techniques such as random cropping, flipping, brightness adjustment, and Gaussian noise addition were applied to increase dataset diversity and improve model generalization. Additionally, pixel values were normalized to the range $[0, 1]$ and standardized using the mean and standard deviation of the ImageNet dataset. The dataset was stratified into training, validation, and test sets with a ratio of 70:15:15, maintaining balanced representation across clusters.

Image extractor. In iGFruit, our image feature extraction process centers around the utilization of two prominent architectures: the NFNet-F0 [20] and a variant of the Vision Transformer (ViT) [21] known as ViT-B/16. For the NFNet-F0, we perform pre-training on the ImageNet dataset, utilizing a batch size of 4096 and a training duration of 360 epochs. Similarly, the ViT-B/16 model is also pre-trained, but for a shorter span of 20 epochs, with a batch size of 2880. We employ the AdamW [23] optimizer with a weight decay of 0.05. To facilitate effective training, the learning rate undergoes a warm-up phase, reaching a peak of $3e$-4, and is subsequently decayed linearly with a rate of 0.85. Notably, the ViT-B/16 model's pre-training is conducted using the ImageNet dataset. The subsequent step involves concatenating the output features of both the NFNet-F0 and ViT models, enabling us to harness the complementary strengths of these architectures for improved feature representation.

Text extractor. Our approach is rooted in the fusion of pre-trained embeddings from diverse sources. Specifically, we initialize the word embeddings by concatenating pre-trained English-BERT, Multilingual-BERT [24], and Paraphrase-xlm-r-multilingual-v1. This amalgamation ensures a rich representation of language semantics and context.

Graph attention network. To incorporate contextual relationships between different instances in our iGFruit framework, we employ a Graph Attention Network (GAT) [12] layer. The GAT layer facilitates information propagation across the graph structure, effectively integrating image and text features. In this layer, nodes represent instances (image-text pairs), and edges signify relationships between these instances.

LightGBM. As the final component of our iGFruit framework, we employ the LightGBM [19] algorithm for the task. LightGBM is a gradient-boosting framework that excels in handling tabular data and has

been widely used for various machine-learning tasks, including classification. It is capable of handling large datasets efficiently and provides strong predictive performance.

Training. We employ a combination of data augmentation techniques, starting with transformations derived from the learned Auto Augmentation policy, applied to each individual image. Standard augmentations, including random size cropping, random horizontal flipping, color jittering, and lighting adjustments complement these. The optimization is performed using the AdamW optimizer [23] with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a peak learning rate of 10^{-3}. A batch size of 64 and a high weight decay of 0.1 are used. The training process spans 200 epochs, with the first 5 epochs designated as warmup periods, utilizing a cosine learning rate schedule for decay. To further refine optimization, we incorporate the Exponential Moving Average (EMA) method with a decay rate of 0.9995. Our implementation is based on the Timm library, and all experiments are conducted on two NVIDIA RTX 2080 GPUs. The code for our approach will be released upon acceptance.

4.4.2 Comparison with State-of-the-Art Methods (RQ1)

Table 2 presents a comparison of iGFruit with state-of-the-art methods on the MIT-States and TLU-States datasets. On MIT-States, iGFruit demonstrates superior performance across all evaluated metrics. Notably, it achieves a significant improvement in AUC, recording 4.2 compared to 3.0 from SymNet [25], despite obtaining a comparable best harmonic mean. When compared to the closest competitor, LabelEmbed+ (LE+) [26], iGFruit exhibits clear advantages across all metrics, with AUC increasing from 2 to 4.2 and the best harmonic mean improving from 10.7% to 16.3%.

Table 2: Results on MIT states and TLU states. We measure states (Sta.) and objects (Obj.) accuracy on the primitives, best seen (S) and unseen accuracy (U), best harmonic mean (HM), and area under the curve (auc)

Method	MIT-States						TLU-States					
	Sta.	**Obj.**	**S**	**U**	**HM**	**auc**	**Sta.**	**Obj.**	**S**	**U**	**HM**	**auc**
AoP	21.1	23.6	14.3	17.4	9.9	1.6	38.9	66.8	54.5	50.3	36.2	21.7
LE+	23.5	26.3	15.0	20.1	10.7	2.0	41.7	70.5	51.8	60.1	39.8	24.5
TMN	23.3	26.5	20.2	20.1	13.0	2.9	40.2	68.2	57.3	58.6	43.5	28.1
SymNet	26.3	28.3	24.2	**25.2**	16.1	3.0	40.8	66.9	49.1	56.7	38.8	22.3
iGFruit (ours)	**26.9**	**30.8**	**24.8**	24.9	**16.3**	4.2	**45.3**	**74.2**	**60.1**	**61.5**	**43.8**	**30.2**

Similarly, iGFruit maintains its dominance on the TLU-States dataset. It outperforms all competing methods, including Attribute as Operators (AoP) [27], LE+, Task-Modular Neural Networks (TMN) [28], and SymNet. iGFruit achieves a state accuracy of 45.3% and an object accuracy of 74.2%, both of which are significantly higher than those achieved by alternative approaches. Furthermore, iGFruit's best harmonic mean of 43.8% surpasses all other methods, while its AUC of 30.2 represents a remarkable improvement over the nearest competitors, solidifying its effectiveness and robustness in these tasks.

Our model (iGFruit) is designed to be adaptable to a wide range of agricultural products by leveraging generalizable feature extraction techniques, such as NFNet-F0 and ViT-B/16, pre-trained on large-scale datasets like ImageNet. The model demonstrates effective performance across these product types by fine-tuning on the iGFruit dataset, which includes diverse categories such as fruits (e.g., apples, grapes, oranges) and vegetables (e.g., potatoes, strawberries). Due to their unique texture, color, and shape characteristics, it performs most robustly on products with visually distinctive features, such as apples and grapes. Conversely, products with more uniform or ambiguous features, such as potatoes, exhibit slightly lower but still

satisfactory performance. The model's ability to distinguish fine-grained differences between similar varieties of the same product further underscores its effectiveness for high-precision tasks.

Despite its versatility, the model has certain limitations. Its performance is highly dependent on the availability of high-resolution images with minimal noise, which may not always be feasible in real-world scenarios involving poor lighting or occlusions. Additionally, for rare or underrepresented products, the limited variability in training data can lead to reduced accuracy. Domain shifts, such as environmental differences affecting product appearance across geographic regions, may also pose challenges to generalization without further adaptation. Moreover, scalability to large-scale farm environments or dynamic conditions might require additional domain-specific fine-tuning and integration. Addressing these limitations in future work will involve incorporating more diverse data sources, exploring domain adaptation techniques, and optimizing the model for real-world deployment. This discussion highlights the model's strengths and areas for improvement, offering a foundation for broader applicability in agricultural domains.

4.4.3 Ablation Studies (RQ2)

Image encoder. In this section of our research, we delve into the impact of various vision encoders on the performance of our current architecture. The experimental investigation focuses on a comprehensive set of two state-of-the-art models, including Vision Transformer (ViT) [21] and NFNet [20]. In addition, we investigate the model's performance when combining the features of both the ViT and NFNet models through concatenation. Table 3 presents the experimental results using the ViT model, the NFNet model, and their combination on the MIT-States and TLU-States datasets. Notably, the synergistic utilization of the ViT and NFNet models as visual feature extractors yielded a substantial enhancement in the AUC metric, surpassing the AUC values of ViT and NFNet individually by 10.53% and 20% on the MIT-States dataset, respectively. Similarly, on the TLU dataset, the AUC values improved by 3.78% and 11.03%, respectively. This robust improvement underscores the advantages of concatenating features from both models, facilitating the holistic model in capturing a diverse spectrum of image features. This amalgamation significantly contributes to the overall framework's performance.

Table 3: Compare performance with different image encoders on two datasets MIT-States and TLU-States

Image encoder	MIT-States			TLU-States		
	Sta.	**Obj.**	**auc**	**Sta.**	**Obj.**	**auc**
ViT	20.2	23.2	3.8	38.8	60.9	29.1
NFNet	18.5	19.8	3.5	35.8	57.9	27.2
ViT + NFNet (ours)	**26.9**	**30.8**	**4.2**	**45.3**	**74.2**	**30.2**

Text encoder. The Table 4 provides a comprehensive overview of the performance of various text encoders on two benchmark datasets, MIT-States and TLU-States. Our model combines all three text encoders and achieves Sta. of 26.9, Obj. of 30.8, and auc of 4.2 on MIT-States. On TLU-States, our model excels with Sta. at 45.3, Obj. at 74.2, and auc at 30.2. These results demonstrate that the combination of different text encoders in the model significantly improved performance compared to individual text encoders.

Table 4: Compare performance with different text encoders on two datasets MIT-States and TLU-States

Text encoder	MIT-States			TLU-States		
	Sta.	Obj.	AUC	Sta.	Obj	AUC
English BERT	22.3	29.7	3.4	42.4	72.9	28.3
Multilingual BERT	20.2	28.4	3.6	43.2	71.2	30.0
XLM	18.3	29.0	3.1	41.8	66.6	28.1
English BERT + Multilingual BERT	24.8	27.8	3.4	43.5	73.3	28.8
English BERT + XLM	23.7	30.4	3.5	42.7	72.2	27.8
Multilingual BERT + XLM	21.6	29.1	4.0	42.2	71.9	25.8
English BERT + Multilingual BERT + XLM (ours)	**26.9**	**30.8**	**4.2**	**45.3**	**74.2**	**30.2**

The impact of different graph networks on performance. In this section we ablate our iGFruit framework with respect to the graph depth, and graph architecture variants. We extensively investigate the architectural nuances of the graph across varying depths, ranging from 1 to 6 layers, with the intention of quantifying the extent of knowledge propagation essential for achieving optimal performance. Our focal point is the comparison between Graph Attention Networks (GAT) [12] and Graph Convolutional Networks (GCN) [11] in this evaluation. We aim to underscore the superior efficacy of GAT over GCN in our findings. From the insights garnered in Figs. 4 and 5, we discern a remarkable trend. With GAT, a shallower architecture with only 1 layer attains the most impressive Area Under the Curve (AUC) of 4.2% on the MIT-States dataset and 30.2% on the TLU-States dataset, outperforming the configurations with a greater depth. We explore a graph convolution formulation known as GCN to investigate whether our success is inherently tied to a relatively superficial representation. Our investigation reveals that while GCN displays improved resilience against the smoothing challenge and sustains performance at deeper architectures, it only achieves an AUC of 3.5 on the MIT-States dataset and 26.5 on the TLU-States dataset in the optimal model setting.

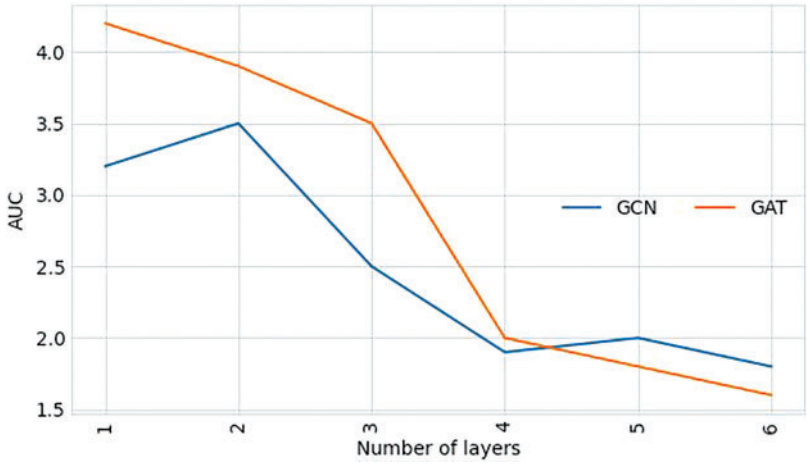

Figure 4: Comparison at various depths of the GCN network and GAT network on the validation set of MIT-States

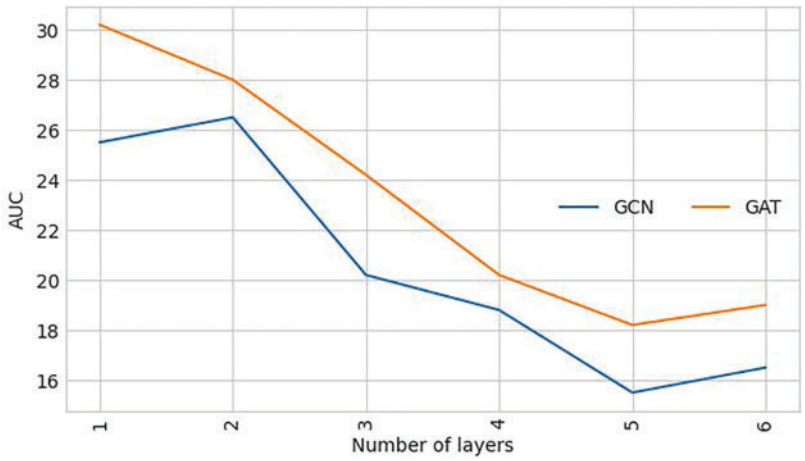

Figure 5: Comparison at various depths of the GCN network and GAT network on the validation set of TLU-States

Loss function. To determine the optimal values for the coefficients k_a, k_b, and k_c in Eq. (10), we conducted a grid search over a predefined range of hyperparameters, training the model with various combinations of these values. Performance was evaluated on a validation dataset to identify the configuration that maximized validation accuracy and minimized loss. Our experimental results highlighted the critical impact of these coefficients on model performance. After an exhaustive search, the optimal values were identified as $k_a = 0.4$, $k_b = 0.1$, and $k_c = 0.5$. This configuration achieved the highest validation accuracy and the lowest loss for our task, emphasizing the importance of fine-tuning the $L_{iGFruit}$ loss function to enhance the model's effectiveness.

5 Conclusion

Our proposed (iGFruit) approach demonstrates the effectiveness of leveraging both image and text data for counterfeit agricultural product detection. By combining features extracted from different modalities using multiple backbone models and modeling their relationships through a Graph Attention Network, we enhance the discriminative power of our model. The achieved state-of-the-art performance on both the MIT-States and TLU-States datasets underscores the potential of our method in practical applications. However, several areas remain for future research. First, further refinement of the model can be achieved by incorporating advanced techniques such as self-supervised learning and domain adaptation to address the domain shift challenges inherent in agricultural datasets. Second, exploring additional data sources, such as environmental data (e.g., weather conditions or soil quality) and multi-temporal imagery, may improve the model's ability to handle complex, real-world scenarios.

Additionally, while our method excels in accuracy, its scalability to larger datasets and real-time applications needs further exploration. Research into model compression techniques and efficient deployment strategies, such as edge computing, could address these challenges. Another promising avenue is enhancing the interpretability of the model's decision-making process, enabling end-users to understand and trust its predictions. This is particularly critical in high-stakes applications like counterfeit detection in the agricultural sector. Finally, collaborations with domain experts can guide the development of tailored solutions, ensuring the model's relevance and usability in diverse agricultural contexts. By addressing these recommendations, future research can build upon our work to create robust, scalable, and interpretable solutions for counterfeit agricultural product detection and beyond.

Acknowledgement: Dat Tran-Anh was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2024.TS.070. Hoai Nam Vu was funded by the Postdoctoral Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2024.STS.39.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Dat Tran-Anh, Hoai Nam Vu; data collection: Dat Tran-Anh; analysis and interpretation of results: Dat Tran-Anh, Hoai Nam Vu; draft manuscript preparation: Dat Tran-Anh, Hoai Nam Vu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Hoai Nam Vu, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Nomenclature

ViT	Vision Transformer
GAT	Graph Attention Network
BoW	Bag of Words
GLCM	Gray-Level Co-occurrence Matrix
SVM	Support Vector Machine
RGB	Red Green Blue
HSV	Hue Saturation Value
FPFH	Fast Point Feature Histogram
GCN	Graph Convolutional Network
MSE	Mean Square Error
MAE	Mean Absolute Error
HM	Harmonic Mean
AUC	Area Under the Curve
TMN	Task-Modular Neural Network
EMA	Exponential Moving Average

References

1. Wabeke T, Moura GCM, Franken N, Hesselman C. Counterfighting counterfeit: detecting and taking down fraudulent webshops at a ccTLD. In: Sperotto A, Dainotti A, Stiller B, editors. Passive and active measurement. Springer; 2020. p. 158–74.
2. Zhao S, Ge D, Zhao J, Xiang W. Fingerprint pre-processing and feature engineering to enhance agricultural products categorization. Future Gener Comput Syst. 2021;125:944–8. doi:10.1016/j.future.2021.07.005.
3. Zhang X, Han Y, Xu W, Wang Q. HOBA: a novel feature engineering methodology for credit card fraud detection with a deep learning architecture. Inf Sci. 2021;557:302–16. doi:10.1016/j.ins.2019.05.023.
4. Torky M, Hassanein AE. Integrating blockchain and the internet of things in precision agriculture: analysis, opportunities, and challenges. Comput Electron Agric. 2020;178:105476. doi:10.1016/j.compag.2020.105476.
5. Guan M, Cai X, Shang J, Hao F, Liu D, Jiao X, et al. HMSG: heterogeneous graph neural network based on Metapath SubGraph learning. Knowl Based Syst. 2023;279:110930. doi:10.1016/j.knosys.2023.110930.
6. Yu L, Xiong J, Fang X, Yang Z, Chen Y, Lin X, et al. A litchi fruit recognition method in a natural environment using RGB-D images. Biosyst Eng. 2021;204:50–63. doi:10.1016/j.biosystemseng.2021.01.015.

7. Zhao G, Zhang C, Shang H, Wang Y, Zhu L, Qian X. Generative label fused network for image-text matching. Knowl Based Syst. 2023;263:110280. doi:10.1016/j.knosys.2023.110280.
8. Ahmed KT, Ummesafi S, Iqbal A. Content based image retrieval using image features information fusion. Inf Fusion. 2019;51(1):76–99. doi:10.1016/j.inffus.2018.11.004.
9. Ge L, Zou K, Zhou H, Yu X, Tan Y, Zhang C, et al. Three dimensional apple tree organs classification and yield estimation algorithm based on multi-features fusion and support vector machine. Inf Process Agric. 2022;9(3):431–42. doi:10.1016/j.inpa.2021.04.011.
10. Vishnoi VK, Kumar K, Kumar B. A comprehensive study of feature extraction techniques for plant leaf disease detection. Multimed Tools Appl. 2022;81(1):1–53. doi:10.1007/s11042-021-11375-0.
11. Chen ZM, Wei XS, Wang P, Guo Y. Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019; Long Beach, CA, USA: IEEE.
12. Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. In: Proceedings of the International Conference on Learning Representations (ICLR); 2018; Vancouver, BC, Canada.
13. Lin X, Ding C, Zhan Y, Li Z, Tao D. HL-Net: heterophily learning network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022; New Orleans, LA, USA: IEEE. p. 19476–85.
14. Gufran D, Tiku S, Pasricha S. Heterogeneous device resilient indoor localization using vision transformer neural networks. In: Tiku S, Pasricha S, editors. Machine learning for indoor localization and navigation. Springer; 2023. p. 357–75.
15. Barisoni L, Lafata KJ, Hewitt SM, Madabhushi A, Balis UG. Digital pathology and computational image analysis in nephropathology. Nat Rev Nephrol. 2020;16(11):669–85. doi:10.1038/s41581-020-0321-6.
16. Lee JS, Hsiang J. Patent classification by fine-tuning BERT language model. World Pat Inf. 2020;61:101965. doi:10.1016/j.wpi.2020.101965.
17. Khattak MU, Rasheed H, Maaz M, Khan S, Khan FS. MaPLe: multimodal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023; Vancouver, BC, Canada: IEEE. p. 19113–22.
18. Kim D, Seo D, Cho S, Kang P. Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec. Inf Sci. 2019;477:15–29. doi:10.1016/j.ins.2018.10.006.
19. Yalcin Kuzu S. Evaluation of gradient boosting and deep learning algorithms in dimuon production. J Mol Struct. 2023;1277:134834. doi:10.1016/j.molstruc.2022.134834.
20. Brock A, De S, Smith SL, Simonyan K. High-performance large-scale image recognition without normalization. In: Proceedings of the International Conference on Machine Learning (ICML); 2021. p. 1059–71.
21. Yin H, Vahdat A, Alvarez JM, Mallya A, Kautz J, Molchanov P. A-ViT: adaptive tokens for efficient vision transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022; New Orleans, LA, USA: IEEE. p. 10809–18.
22. Naeem MF, Xian Y, Tombari F, Akata Z. Learning graph embeddings for compositional zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021; Nashville, TN, USA: IEEE. p. 953–62.
23. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: Proceedings of the International Conference on Learning Representations (ICLR); 2018; Vancouver, BC, Canada.
24. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019; Minneapolis, MN, USA. p. 4171–86.
25. Li YL, Xu Y, Mao X, Lu C. Symmetry and group in attribute-object compositions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020; Seattle, WA, USA: IEEE; p. 11316–25.
26. Misra I, Gupta A, Hebert M. From red wine to red tomato: composition with context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; Honolulu, HI, USA: IEEE. p. 1792–801.

27. Nagarajan T, Grauman K. Attributes as operators: factorizing unseen attribute-object compositions. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018; Munich, Germany: Springer. p. 169–85.
28. Purushwalkam S, Nickel M, Gupta A, Ranzato MA. Task-driven modular networks for zero-shot compositional learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR); 2019; Long Beach, CA, USA: IEEE. p. 3593–3602.