A two-stage approach to agricultural products authentication by mining subtle local features

Dat Tran Anh[1]

Abstract: Counterfeit agricultural products recognition poses a significant challenge due to the high visual similarity between genuine and fake items. Existing methods often struggle to capture the subtle details necessary for reliable differentiation. This paper presents Focus on Detail (FoD), a novel approach that emphasizes the automatic discovery of trustworthy 'authenticity cues' on the products. By employing custom-designed loss functions and a semi-supervised training strategy, FoD learns to suppress distracting background regions and focus exclusively on the most critical local features. Experimental results demonstrate that FoD achieves superior performance on standard benchmark datasets, establishing a new state-of-the-art in both accuracy and speed.

Keywords: Counterfeit Recognition, Local Feature Learning, Weakly-Supervised Learning.

1. Introduction

Counterfeit agricultural products recognition via Images (Mukhiddinov et al., 2022; Tran-Anh et al., 2024) is a critical task for verifying product authenticity, playing an essential role in consumer protection and brand reputation preservation. A major challenge in this task lies in the high visual similarity between genuine and counterfeit items, which often causes traditional methods relying on global features to fail. These methods typically overlook fine-grained spatial details, leading to perceptual aliasing-a phenomenon where visually similar objects become indistinguishable.

To address this issue, two-stage approaches-comprising coarse classification followed by refined verification-have been proposed to exploit local features. However, even the most advanced of these methods face three key limitations: (i) a lack of explicit modeling of discriminative regions, which are crucial for distinguishing genuine from fake products; (ii) insufficient fine-grained annotated data at the local level, which hinders effective supervision during the verification stage; and (iii) high computational overhead incurred by exhaustive matching of all local features, limiting real-world applicability.

This paper introduces Focus on Detail (FoD), an efficient two-stage framework specifically designed to overcome the above limitations. At its core, FoD automatically models authenticity markers-reliable and highly discriminative local regions indicative of product authenticity. We achieve this through custom-designed loss functions that guide the model to concentrate on informative details while suppressing noisy background regions. In addition, we propose a weakly-supervised training strategy to learn robust local representations without requiring dense annotations. Finally, FoD leverages these learned authenticity markers to guide a streamlined re-verification process that selectively matches only the most relevant regions, thereby improving both accuracy and efficiency.

Our contributions result in a powerful and practical framework that significantly enhances recognition performance. Experimental results show that FoD not only outperforms previous methods but also sets a new benchmark for balancing accuracy and computational efficiency in image-based product authentication.

2. Related works

This section provides an overview of image-based counterfeit products recognition methods, which can be broadly categorized into two main approaches: single-stage and two-stage methods.

2.1. Single-Stage Recognition Methods

Single-stage approaches treat the task as either an image classification or image retrieval problem. These methods typically generate a single global representation for each product image and perform similarity matching in that feature space. Common aggregation architectures such as NetVLAD (Arandjelovic et al., 2018), GeM (Goyal & Ferrara, 2018), and more recent methods based on Transformers and Vision Foundation Models (VFMs), such as DINOv2 (Wang et al., 2024), have been employed. For instance, SALAD (Izquierdo & Civera, 2023) redefined the soft assignment in NetVLAD (Arandjelovic et al., 2018) as an optimal transport problem to aggregate local features extracted from DINOv2.

However, a fundamental limitation of single-stage methods lies in their vulnerability to high visual similarity between genuine and counterfeit products. By relying solely on global information, these approaches often overlook subtle local cues essential for accurate identification, leading to misclassification when counterfeit products are intricately forged.

In contrast, our proposed FoD approach addresses this challenge by explicitly modeling and

[1] Thuyloi University

Received 8th Aug. 2025
Accepted 8th Sep. 2025
Publication date 31st Dec. 2025

leveraging spatially discriminative feature regions. This enables the model to attend more effectively to critical "authenticity cues," thereby reducing the confusion caused by global feature similarities. Moreover, we incorporate a validation step that performs local region matching to refine the initial classification output.

2.2. Two-Stage Recognition Methods

A recent trend in product recognition involves adopting a two-stage strategy. These methods first perform an initial ranking or classification based on global feature similarity, followed by a re-ranking step among the top-K candidates using spatially-aware local features. Two-stage methods typically achieve superior performance due to the incorporation of fine-grained spatial information.

Inspired by works in Visual Place Recognition (VPR), similar strategies have been applied in this domain. For example, Patch-NetVLAD (Hausler et al., 2021) proposes matching multi-scale features at the patch level, while R2Former introduces a unified framework for retrieval and re-ranking that considers feature correlation, attention values, and coordinates. More recently, SelaVPR (Lu et al., 2024) integrates trainable adapters into the DINOv2 architecture and performs re-ranking by introducing another adapter to generate local features.

Our proposed FoD method differs from existing two-stage approaches in two key innovations:

First, FoD explicitly models reliable feature regions using specially designed loss functions, rather than solely relying on the model's built-in attention mechanisms. These mined regions are then utilized to enable more effective and accurate local feature matching during the re-ranking phase;

Second, FoD introduces a semi-supervised local feature learning technique based on pseudo-correspondence, enhancing the precision of the verification stage and addressing the scarcity of fine-grained labeled data.

3. The proposed model

3.1. Problem Definition

The problem of Agricultural Product Authentication via Images can be formally defined as follows:

Given a reference database of images of verified genuine products, denoted as:

$$\mathrm{D}_{\mathrm{ref}} = \{I_1, I_2, \ldots, I_N\} \tag{1}$$

and a query image I_q of an unknown product, the goal is to learn a prediction function f that assigns a label $y_q \in \mathrm{Y} = \{\text{genuine}, \text{counterfeit}\}$ to the query:

$$y_q = f(I_q) \tag{2}$$

A key challenge in this task lies in the high visual similarity between authentic and counterfeit products, especially when counterfeits are skillfully replicated. Formally, if we define a distance function $d(\cdot,\cdot)$ in the feature space, it is possible that the distance between a counterfeit query image I_q (where $y_q = \text{counterfeit}$) and a genuine image $I_{\mathrm{ref}} \in \mathrm{D}_{\mathrm{ref}}$ is very small, potentially even smaller than the distance between two distinct genuine images in the database:

$$d(I_q, I_{\mathrm{ref}}) \approx 0, \quad \text{even though } y_q \neq y_{\mathrm{ref}} \tag{3}$$

To address this issue, our approach reformulates the problem into a two-stage process:

Stage 1: Retrieval and Coarse Classification

This stage aims to narrow the search space quickly. We learn a global feature extractor $g_{\text{global}}(\cdot)$ that generates a global representation vector D_g for each image. Based on D_g, a preliminary classification is performed to either identify the top-k most similar genuine candidates from $\mathrm{D}_{\mathrm{ref}}$, or to produce an initial prediction label.

$$D_g = g_{\text{global}}(I), \quad \forall I \in \mathrm{D}_{\mathrm{ref}} \cup \{I_q\} \tag{4}$$

Stage 2: Re-verification via Local Features

This stage conducts a deep inspection for final decision-making. We learn a local feature extractor $g_{\text{local}}(\cdot)$ that produces a set of fine-grained descriptors D_l representing discriminative regions on the product surface:

$$D_l = g_{\text{local}}(I) \tag{5}$$

A local matching process is then performed between the local features of the query image I_q and those of the top-k candidates retrieved from Stage 1. The final decision y_q is made based on the result of this fine-grained local feature matching, rather than relying solely on global similarity:

$$y_q = \text{Verify}_{\text{local}}(g_{\text{local}}(I_q), \{g_{\text{local}}(I_i)\}_{i=1}^{k}) \tag{6}$$

3.2. Model Architecture

The overall architecture of the proposed Focus on Detail (FoD) approach is illustrated in Figure 1. Our model comprises three main components: (1) a shared feature extractor, (2) a global feature aggregation branch, and (3) a local feature decoder branch.

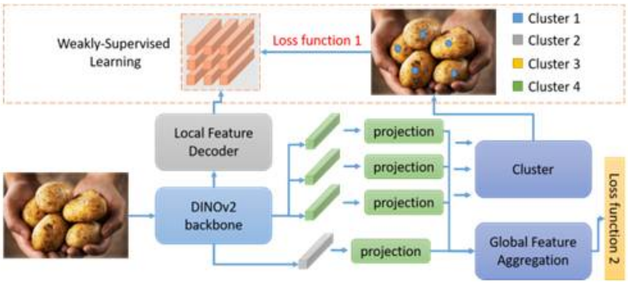

Figure 1. The proposed model

Following recent advancements in computer vision, we adopt a pre-trained Vision Foundation Model (VFM)-specifically, DINOv2-as the backbone to extract feature representations.

Given an input image I , the backbone outputs a set of feature tokens from its final layer, including:

- a [cls] token, and
- a sequence of patch tokens F , each representing a localized region of the input image.

Global Feature Aggregation branch is responsible for producing a compact global feature vector D_g , which encodes the overall visual content of the product image.

Inspired by recent clustering-based aggregation methods, we employ a learnable clustering mechanism. Specifically, the patch tokens F are first projected and then assigned to a set of M learnable cluster centers. The aggregated features from these clusters are then used to construct the global representation D_g .

To ensure that genuine products are embedded close together while counterfeit ones are pushed apart in the feature space, we apply a metric learning loss, such as the Multi-Similarity Loss, denoted as Loss Function 2 in Figure 1.

The goal of Local Feature Decoder branch is to generate high-resolution local feature maps D_l that preserve fine-grained details crucial for the re-verification stage.

To this end, we employ a lightweight decoder consisting of:

- two deconvolution layers, and
- a ReLU activation layer.

The decoder takes the patch tokens F as input and upsamples them to produce dense local descriptors. As illustrated in Figure 1, these features serve as the foundation for the Weakly-Supervised Learning mechanism, which is optimized using Loss Function 1—to be detailed in subsequent sections.

3.3. Loss function

To effectively train the FoD model, we adopt a composite loss function composed of multiple components aimed at optimizing both global and local representations. The total loss function $\mathrm{L}_{\mathrm{total}}$ is defined as:

$$\mathrm{L}_{\mathrm{total}} = \mathrm{L}_{\mathrm{MS}} + \mathrm{L}_{\mathrm{local}} + \alpha \mathrm{L}_{\mathrm{align}} + \beta \mathrm{L}_{\mathrm{PC}} \qquad (7)$$

where:

- L_{MS} : optimizes global feature representation,
- $\mathrm{L}_{\mathrm{local}}$ and $\mathrm{L}_{\mathrm{align}}$: model spatially discriminative regions,
- L_{PC} : supervises local feature learning via pseudo-correspondence, and α, β are balancing hyperparameters.

Global Feature Loss - L_{MS}

This corresponds to Loss Function 2 in Figure 1. To train the global feature branch, we adopt the Multi-Similarity Loss L_{MS} , a metric learning technique that structures the embedding space effectively. Specifically, it encourages the global representations D_g of genuine (positive) samples to be close to each other, while pushing them away from counterfeit (negative) samples by considering complex similarity relationships among sample pairs within a batch.

Losses for Local Feature Learning and Region Modeling

The following components form Loss Function 1 in Figure 1 and constitute the core of the FoD methodology.

(a) Alignment Loss - $\mathrm{L}_{\mathrm{align}}$

Inspired by the Extraction-Aggregation Spatial Alignment Loss (SAL), this component aligns attention between the feature extractor and the aggregation module. Specifically, we use Kullback-Leibler (KL) divergence to enforce consistency between:

- the attention map from the backbone (i.e., regions the backbone considers important), and
- the clustering map from the aggregation layer (i.e., regions retained to form global features).

This encourages the model to consistently attend to semantically meaningful regions.

(b) Local Contrastive Enhancement Loss - $\mathrm{L}_{\mathrm{local}}$

Inspired by Foreground-Background Contrast Enhancement Loss (CEL), we design a triplet-based loss to highlight "authenticity cues."

We define:

- a feature prototype for informative regions (foreground),
- and a background prototype for non-informative regions.

The loss encourages local regions from two genuine images (positive pair) to be similar, while ensuring both differ significantly from background prototypes.

(c) Pseudo-Correspondence Loss - L_{PC}

Due to the lack of pixel-level annotations for “authenticity cues,” we introduce a weakly-supervised learning strategy based on pseudo-correspondence, similar to the one proposed in prior work.

We assume that image patches from a positive pair (two images of the same genuine product) that are assigned to the same cluster in the global aggregation process form a pseudo-correspondence.

The loss L_{PC} is then defined to minimize the distance between the corresponding local features D_l at those pseudo-corresponding locations, weighted by a confidence score estimating the reliability of each pseudo-correspondence.

4. Experimental results

4.1. Evaluation indicators

To comprehensively assess the performance of the proposed model, we employ a set of widely accepted and standardized evaluation metrics. Core classification performance is measured using Accuracy, Precision, Recall, and F1-Score, which evaluate the model's ability to accurately distinguish between genuine and counterfeit products. Given the two-stage architecture, we also assess the effectiveness of the retrieval and re-verification components using Mean Average Precision (mAP) and Recall@k (with k=1, 5, 10). Finally, computational efficiency-a critical factor for real-world deployment-is evaluated by measuring the Inference Time per query and the computational cost in FLOPs (Floating Point Operations).

4.2. Experimental setup

To evaluate the effectiveness of the proposed FoD method, we primarily utilize Agri-TLU, a dataset for agricultural product recognition constructed by us for both training and testing purposes. To assess the model's robustness, we also incorporate COCO (Zhuo et al., 2025) as a source of distractor images, and leverage the representational power of pre-trained models on large-scale datasets such as ImageNet (Huang et al., 2024) by adopting the DINOv2 backbone. The model, based on a ViT-L backbone, is fine-tuned on the training set of Agri-TLU using a NVIDIA A100 GPU and the AdamW optimizer. Hyperparameters are selected based on the highest F1-Score achieved on the validation set of Agri-TLU to ensure optimal performance. The Agri-TLU dataset was compiled using images of genuine agricultural products sourced from certified local farms and counterfeit products obtained from market monitoring. All images were captured under controlled lighting conditions using a high-resolution camera, standardized to a 1024x1024 resolution.

4.3. Experimental results

Table 1 summarizes the comparison between FoD and other baseline methods. For Agri-TLU, we report the F1-Score (%) for the real/fake classification task. For ImageNet and COCO, we measure the Rejection Rate (%), which reflects the model's ability to correctly identify and discard images that do not belong to the target product category.

Table 1. The Agri-TLU column reports F1-Score (higher is better), while the ImageNet and COCO columns report Rejection Rate (higher is better). Best results are highlighted in bold

Method	Agri-TLU (F1-Score %)	ImageNet (Rejection Rate %)	COCO (Rejection Rate %)
Global-Classify (Huang et al., 2024)	92.8	85.4	88.2
Patch-Verify (Damm et al., 2025)	93.4	91.0	92.5
Archaeoscape (Perron et al., 2024)	95.1	94.6	94.3
FoD-global (ours)	95.8	95.2	95.5
FoD-verification (ours)	**96.5**	**97.3**	**96.8**

From the results, FoD-verification not only achieves the highest F1-Score of 96.5% on the primary dataset Agri-TLU, but also demonstrates outstanding robustness. With Rejection Rates of up to 97.3% on ImageNet and 96.8% on COCO, FoD shows its ability to focus on the distinctive characteristics of agricultural products and avoid being "fooled" by common objects-an essential factor for real-world deployment.

To demonstrate the critical role of using pre-trained weights from large-scale datasets such as ImageNet, we conducted an ablation experiment. The results in Table 2 show a significant performance drop when the model is trained from scratch.

Table 2. Impact of Using Pre-trained DINOv2 Weights

Method	F1-Score in Agri-TLU (%)
FoD (training from scratch)	92.5
FoD (initialize from DINOv2)	96.5

In terms of computational efficiency, Table 3 compares the inference time of FoD with other two-stage methods. FoD demonstrates significantly faster processing, especially in the verification stage, thanks to its mechanism of matching only on distinctive feature regions.

Table 3. Comparison of inference time across two-stage methods.

Method	Extract Time (ms)	Time of Verification (ms)	Total Time (ms)
Patch-Verify	18.5	150.2	168.7
Sela-Agri	23.0	71.0	94.0
FoD-verification (ours)	25.1	32.5	57.6

5. Conclusion and future work

In this paper, we introduced Focus on Detail (FoD), a novel two-stage approach that establishes a new state-of-the-art in agricultural product authentication by automatically discovering reliable "authenticity cues". Through custom loss functions and a weakly-supervised learning strategy, our method overcomes the challenge of high visual similarity between genuine and counterfeit items, demonstrating superior accuracy and efficiency.

While effective, we acknowledge that the model's performance depends on products having distinct surface features and may be less robust when dealing with low-quality images from real-world settings. Therefore, future work will focus on three key directions: enhancing model robustness and extending the framework to other domains such as pharmaceuticals and luxury goods; optimizing the model for efficient on-device deployment to enable in-situ authentication; and integrating Explainable AI (XAI) techniques to improve transparency and user trust.

Acknowledgments

This research is funded by Thuyloi University Foundation for Science and Technology under grant number TLU.STF.25-01.

References

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2018). *NetVLAD: CNN Architecture for Weakly Supervised Place Recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6), 1437–1451. https://doi.org/10.1109/TPAMI.2017.2711011

Damm, S., Laszkiewicz, M., Lederer, J., & Fischer, A. (2025). *AnomalyDINO: Boosting Patch-based Few-Shot Anomaly Detection with DINOv2.* Proceedings - 2025 IEEE Winter Conference on Applications of Computer Vision, WACV 2025, 1319–1329. https://doi.org/10.1109/WACV61041.2025.00136

Goyal, P., & Ferrara, E. (2018). *Graph embedding techniques, applications, and performance: A survey. Knowledge*-Based Systems, 151, 78–94. https://doi.org/10.1016/j.knosys.2018.03.022

Hausler, S., Garg, S., Xu, M., Milford, M., & Fischer, T. (2021). *Patch-NetVlad: Multi-scale fusion of locally-global descriptors for place recognition.* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 14136–14147. https://doi.org/10.1109/CVPR46437.2021.01392

Huang, Y., Zou, J., Meng, L., Yue, X., Zhao, Q., Li, J., Song, C., Jimenez, G., Li, S., & Fu, G. (2024). *Comparative Analysis of ImageNet Pre-Trained Deep Learning Models and DINOv2 in Medical Imaging Classification.* Proceedings - 2024 IEEE 48th Annual Computers, Software, and Applications Conference, COMPSAC 2024, 297–305. https://doi.org/10.1109/COMPSAC61105.2024.00049

Izquierdo, S., & Civera, J. (2023). Izquierdo, S., & Civera, J. (2023). *Optimal Transport Aggregation for Visual Place Recognition.* 17658–17668. https://doi.org/10.1109/CVPR52733.2024.01672Optimal Transport Aggregation for Visual Place Recognition. 17658–17668. http://arxiv.org/abs/2311.15937

Lu, F., Zhang, L., Lan, X., Dong, S., Wang, Y., & Yuan, C. (2024). *Towards Seamless Adaptation of Pre-Trained Models for Visual Place Recognition.* 12th International Conference on Learning Representations, ICLR 2024, 1–22.

Mukhiddinov, M., Muminov, A., & Cho, J. (2022). *Improved Classification Approach for Fruits and Vegetables Freshness Based on Deep Learning.* Sensors, 22(21). https://doi.org/10.3390/s22218192

Perron, Y., Sydorov, V., Wijker, A. P., Evans, D., Pottier, C., & Landrieu, L. (2024). *Archaeoscape: Bringing Aerial Laser Scanning Archaeology to the Deep Learning Era.* Advances in Neural Information Processing Systems, 37(NeurIPS), 1–25.

Tran-Anh, D., Vu, H. N., Bui-Quoc, B., & Dao Hoang, N. (2024). *LmGa: Combining label mapping method with graph attention network for agricultural recognition.* Knowledge and Information Systems. https://doi.org/10.1007/s10115-024-02234-z

Wang, H., Zhang, T., & Salzmann, M. (2024). *Sinder: Repairing the singular defects of dinov2.* European Conference on Computer Vision, 20–35.

Zhuo, W., Tang, Z., Xue, W., Ding, H., & Shen, L. (2025). *DINOv2-powered Few-Shot Semantic Segmentation: A Unified Framework via Cross-Model Distillation and 4D Correlation Mining.* https://arxiv.org/pdf/2504.15669