Advanced Intelligent Systems

ADVANCED INTELLIGENT SYSTEMS Open Access

www.advintellsyst.com

RESEARCH ARTICLE OPEN ACCESS

VALIANT: A Vision-Authenticity Language Framework Through Integrated Experts and Aligned Numerical-Textual Descriptors for *Citri Reticulatae Pericarpium*

Simon C. K. Chan | Bingxin Huang | Hannah H. Kim | Wentao Yu | Yinqian Luo | Terence T. W. Wong

Translational and Advanced Bioimaging Laboratory, Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

Correspondence: Terence T. W. Wong (ttwwong@ust.hk)

Received: 21 February 2026 | **Revised:** 26 March 2026 | **Accepted:** 8 April 2026

Keywords: authenticity verification | chenpi | citri reticulatae pericarpium (crp) | multimodal fusion | vision-language framework

ABSTRACT

Citri reticulatae pericarpium (CRP; Chenpi), the dried peel of *Citrus reticulata*, is a traditional Chinese medicine whose value increases with storage age. This price premium incentivizes fraudulent aging through cosmetic treatments, high-temperature fermentation, tea infusions, or chemical dyes. Expert sensory appraisal is subjective (approximately 83% accuracy), while vision-only deep models—effective in controlled settings (90%–98% accuracy)—often degrade under real-world smartphone conditions and against sophisticated counterfeits. We present VALIANT, a multimodal vision-language framework for smartphone images that enables age classification (1, 6, 10, 15, and 20 years) and authenticity verification against standard operating procedure (SOP) mimics. VALIANT integrates three modalities: (1) visual embeddings from a dual-expert system (Florence-2 for global context, and Swin Transformer for local detail), (2) numerical fingerprints, including CIE 1976 $L^*a^*b^*$ (CIELAB) color statistics, gray-level co-occurrence matrix (GLCM) texture properties, and Laplacian variance, and (3) textual attributes generated by GPT-4o-mini. A gated cross-attention module with contrastive alignment enhances semantic coherence and robustness to imaging noise. Evaluated on 942 smartphone images (761 genuine, 181 SOP mimics), VALIANT achieves 88.7% accuracy for genuine samples and 92.2% for mixed data, representing 4%–5% significant improvements over vision-transformer baselines. This framework enables portable, rapid, and point-of-purchase verification to improve market integrity and supplier credibility.

1 | Introduction

Citri reticulatae pericarpium (CRP), known as Chenpi, is a valued commodity in traditional Chinese medicine and gastronomy. CRP's pharmacological properties are reported to correlate with storage age, leading to substantial price premiums. For instance, market prices for products labeled as "10-year" Chenpi have been reported to reach RMB 2,000 per kilogram [1–3]. This economic incentive drives fraudulent practices, such as tea infusions to replicate aged coloration and aroma or dyes combined with mechanical abrasion to simulate surface features, thereby posing significant detection challenges [4]. Reliable and scalable methods are essential to protect consumers and uphold CRP market integrity.

Traditional sensory evaluation by experts remains common but demonstrates subjectivity and inconsistency, underscoring the need for objective and data-driven alternatives [3]. Instrumental approaches offer precision; for instance, gas chromatography-mass spectrometry identifies chemical profiles accurately but requires destructive sampling and specialized laboratory facilities [5]. Nondestructive options, such as near-infrared spectroscopy (NIRS) [6] and surface-enhanced Raman spectroscopy [7], achieve 90%–100% accuracy by targeting molecular markers like flavonoid

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Advanced Intelligent Systems* published by Wiley-VCH GmbH.

peaks, yet their reliance on laboratory equipment limits point-of-purchase accessibility.

Smartphones enable portable, image-based deep learning for CRP. Vision-based models such as ConvNeXt [8] and enhanced ResNets [4] can achieve 90%–98% accuracy under controlled conditions by leveraging visual cues such as oil-gland density, fibrous pith textures, and albedo features. However, in real-world applications, performance often degrades under uncontrolled lighting, background clutter, and device variability. Furthermore, models can be vulnerable to standard operating procedure (SOP)-produced mimics, leaving a gap between lab performance and field use [3].

Figure 1 shows the traditional and our proposed verification methods. Manual inspection (Figure 1a) relies on sensory steps, including looking (i.e., visual inspection), smelling, touching, and brewing, yielding subjective appraisals with variable confidence. In contrast, our proposed VALIANT pipeline (Figure 1b) uses smartphone image capture, cloud-based neural network inference, and authenticity screening to reject SOP mimics and predict age classes (e.g., 1-year, 6-year, 10-year, etc.), enabling rapid, cost-effective decisions at the point of purchase.

To overcome these challenges, we propose VALIANT, a multimodal vision-language fusion framework that prioritizes CRP age classification (primary task: five categories: 1, 6, 10, 15, and 20 years) while enabling mimic detection (secondary task). VALIANT integrates (i) high-fidelity visual features from a dual-expert system (Florence-2 [9] for global context fused with Swin Transformer [10] for local details via gating); (ii) quantitative numerical fingerprints, including CIE 1976 $L^*a^*b^*$ (CIELAB) color statistics to serve as a proxy for biochemical state, gray-level co-occurrence matrix (GLCM) properties to quantify textural surface changes, and Laplacian variance; and (iii) textual attribute descriptions generated by generative pretraining transformer (GPT-4o-mini) [11], a large language model utilized to encode expert botanical observations. A gated cross-attention module ensures balanced multimodal fusion, with contrastive alignment loss enhancing generalization and robustness by aligning visual-textual representations. This approach effectively exposes SOP mimics through numerical and linguistic inconsistencies often missed by vision-only methods [3].

Our key contributions include (1) a curated smartphone-image dataset of 942 CRP samples (761 genuine, 181 SOP mimics) from verified suppliers that captures real-world imaging variability and the distinctions between natural age-related biochemical changes (e.g., flavonoid accumulation) and artificial alterations (detailed in Section 3.1); and (2) VALIANT, a tailored multimodal framework that combines visual, numerical, and controlled textual cues for CRP age grading and authenticity verification, which outperforms vision-transformer baselines by 4%–6% on genuine samples and by over 2% on mixed data, achieving 88.7% age-classification accuracy for genuine peels and 92.2% for mixed data, supporting portable point-of-purchase verification.

The paper is structured as follows: Section 2 details the VALIANT methodology; Section 3 covers experiments, results, and analyses; Section 4 discusses limitations and future work; and Section 5 concludes.

2 | Methodology

2.1 | Design Rationale

While unimodal approaches capture surface appearances, integrating complementary modalities supports quantification and interpretation of subtle age-related features, thereby improving accuracy and robustness. For instance, numerical features offer quantitative precision: CIELAB color statistics serve as a quantitative proxy for the complex biochemical state of the peel, which is influenced by changes in its chemical composition during aging, including the transformation of flavonoids and other pigments; GLCM properties [12] (contrast and homogeneity) provide a statistical measure of image texture that is strongly correlated with the physical surface changes, such as coarsening, that result from dehydration; and Laplacian variance, a metric of global image sharpness, is included to capture fine topographical changes. The underlying hypothesis is that the presence of age-related structural features like scars and detailed pith texture increases the overall acutance of the peel's image, making this metric a useful proxy for such changes. These features refine visual interpretations by capturing shifts in chromatic entropy and textural uniformity, enabling improved performance over

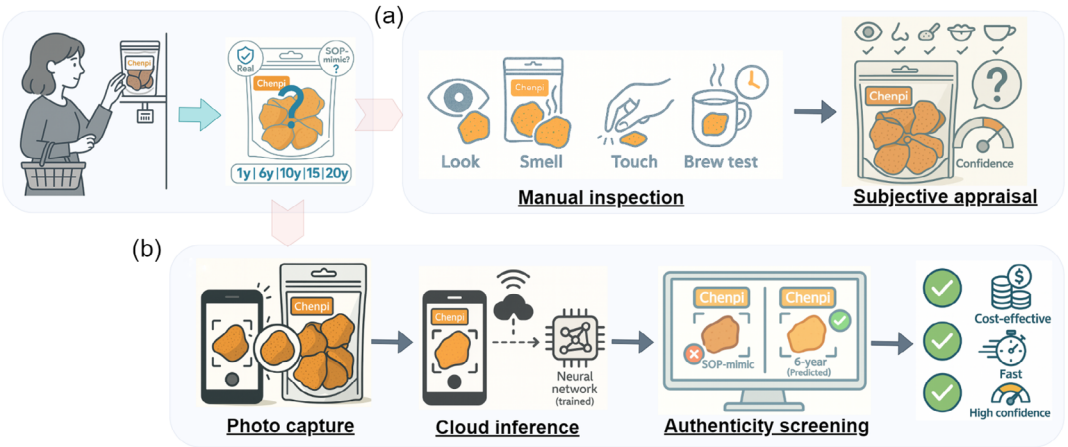

FIGURE 1 | Comparison of CRP verification methods: (a) the traditional manual inspection workflow and (b) our proposed smartphone-based pipeline for authenticity screening and age classification.

vision-only baselines (as demonstrated in Section 3.3), primarily through reduction of noise-related errors.

Textual descriptions provide qualitative, domain-specific insights for semantic reasoning. Generated by GPT-4o-mini, these descriptions encode expert observations (e.g., dense fibrous pith with subtle cicatrices) derived from botanical and traditional Chinese medicine knowledge. This modality resolves ambiguities, such as distinguishing dehydration-induced textures from environmental variations, complementing numerical metrics.

VALIANT's multimodal integration enhances resilience against SOP mimics by identifying inconsistencies (e.g., anomalous GLCM homogeneity or mismatched textual feature descriptions). Relative to vision-centric methods reported in the literature (e.g., ConvNeXt fusion attention module (CNFA) [8], achieving 97%–98% accuracy on authentic samples but with variable performance on mimics), VALIANT demonstrates enhanced generalization on mixed datasets, as shown in Section 3.4. The design addresses real-world smartphone imaging challenges, supporting practical, scalable point-of-purchase verification. These principles inform VALIANT's architecture, as detailed later in the text and illustrated in Figure 2.

2.2 | Architecture Overview

The VALIANT framework analyzes images of the albedo (inner peel) side of CRP samples captured by smartphones for these dual tasks. As shown in Figure 2, it employs a sequential pipeline: preprocessing, parallel modality encoding, gated cross-attention fusion, and classification. Preprocessing employs a fine-tuned YOLOv11 [13, 14] model for instance segmentation to isolate the peel from the background. YOLOv11 is used only to extract the peel region and is not the final classifier; age grading and authenticity verification are performed by the downstream multimodal fusion module. This approach focuses feature extraction on the region of interest despite background clutter or variable lighting. From the segmented peel, three feature sets are extracted: high-fidelity visual features, quantitative numerical metrics (GLCM texture properties and Laplacian variance computed using standard computer vision libraries) [15], and qualitative textual descriptions generated by GPT-4o-mini. Each modality is then encoded into a standardized 512-dimensional embedding for compatibility. The fusion stage applies gated cross-attention to integrate these embeddings adaptively, with a contrastive alignment loss [16, 17] during training to align visual and textual representations for better generalization. A lightweight classification head then outputs probabilities over six classes (five age categories and one mimic class), supporting efficient end-to-end training and mobile deployment.

2.3 | Modality Encoders

Each modality-specific encoder projects inputs into a standardized 512-dimensional embedding to enable effective fusion. The visual encoder processes albedo images with a dual-expert system. Florence-2 serves as the primary backbone for the full 384 × 384-pixel image, while the auxiliary Swin Transformer employs shifted-window attention to capture fine-grained details like scars and oil glands. A gating mechanism dynamically merges their outputs into the final visual embedding. The numerical encoder handles a 12-dimensional feature vector per image, comprising nine CIELAB color statistics, two GLCM texture properties (contrast and homogeneity), and one Laplacian variance sharpness metric

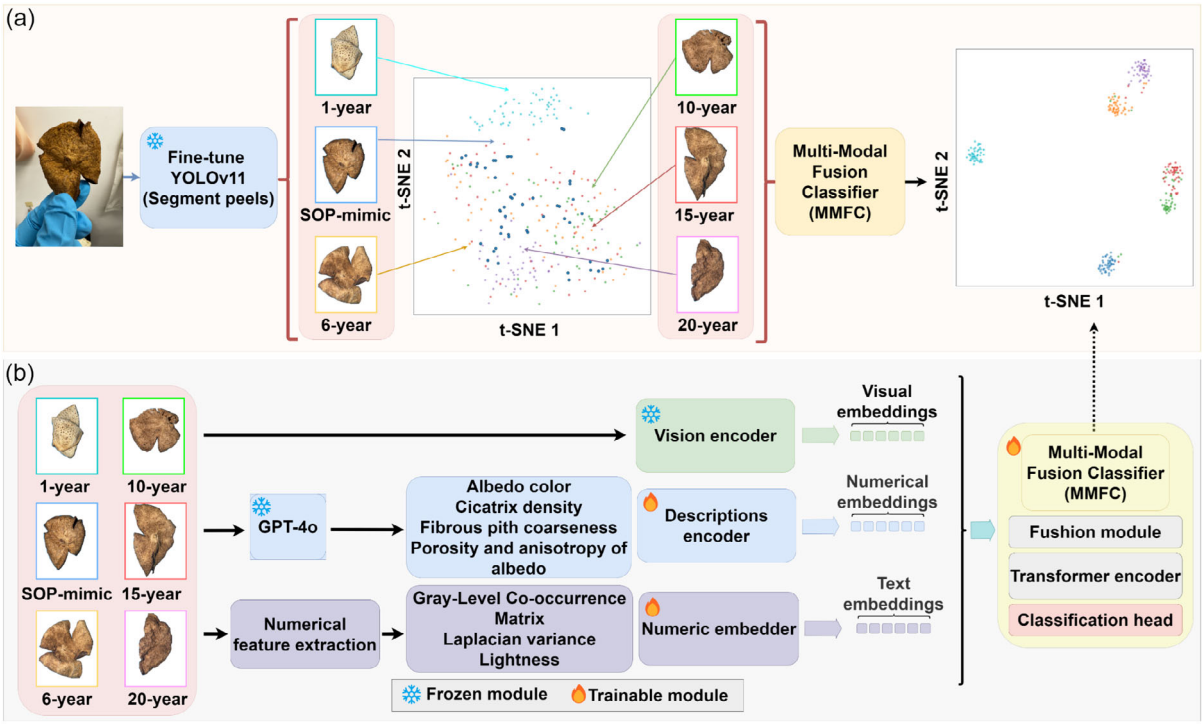

FIGURE 2 | Overview of the VALIANT framework. (a) Workflow from region-of-interest extraction to final prediction, with visualizations of the feature space. (b) The multimodal architecture details the visual, numerical, and text encoding streams. Snowflake icons denote frozen pretrained modules (weights fixed during training), whereas flame icons denote trainable modules updated on our dataset.

—selected for their correlation with citrus peel aging. A multilayer perceptron (MLP)-based weight generator with Gaussian error linear unit (GeLU) activations converts this vector into the numerical embedding. The text encoder processes qualitative descriptions of the peel's physical attributes. These descriptions are generated by GPT-4o-mini using a controlled prompt that focuses on observable features while avoiding direct age references (to prevent label leakage). Figure 3 provides examples of these generated attributes for an SOP-mimic sample (Figure 3a), a 6-year sample (Figure 3b), and a 20-year sample (Figure 3c), showcasing how the model describes distinct features like "Albedo color," "Fibrous pith coarseness," and "Cicatrix density." This process creates a rich, qualitative fingerprint for each sample without inferring its age directly. A pretrained all-MiniLM-L6-v2 [18] sentence transformer then encodes this text via mean pooling and linear projection to yield the final text embedding [19, 20].

2.4 | Fusion and Loss Functions

2.4.1 | Gated Cross-Attention Fusion and Classification

The fusion module integrates the three modality embeddings into a unified representation. Embeddings are normalized and scaled via modality-specific linear layers. A multi-head cross-attention mechanism facilitates intermodality attention for adaptive focus; for example, the visual embedding serves as query (Q), with numerical and textual embeddings as keys (K) and values (V):

$$Attention Q, K, V = softmax \frac{QK^T}{\sqrt{d_k}} V$$

where $(\cdot)^T$ denotes the matrix transpose operation and d_k is the key dimension. An MLP-based weight generator computes softmax-normalized coefficients for a weighted sum of attended features. The resulting fused vector feeds into a three-layer MLP classification head to produce logits for the six target classes. Residual connections stabilize training throughout.

2.4.2 | Loss Functions

The model optimizes a hybrid loss $L = L_{class} + \lambda L_{align}$, with balancing hyperparameter $\lambda = 0.1$. The classification loss L_{class}, combines focal loss for class imbalance and cross-entropy with label smoothing to mitigate overconfidence: $L_{\text{focal}} = -\alpha(1-p_t)^{\gamma} \log(p_t)$, where $\alpha = 0.25$ and $\gamma = 2$. The alignment loss L_{align} promotes semantic coherence between visual (E_v) and textual (E_t) embeddings via symmetric information noise contrastive estimation (InfoNCE) contrastive loss:

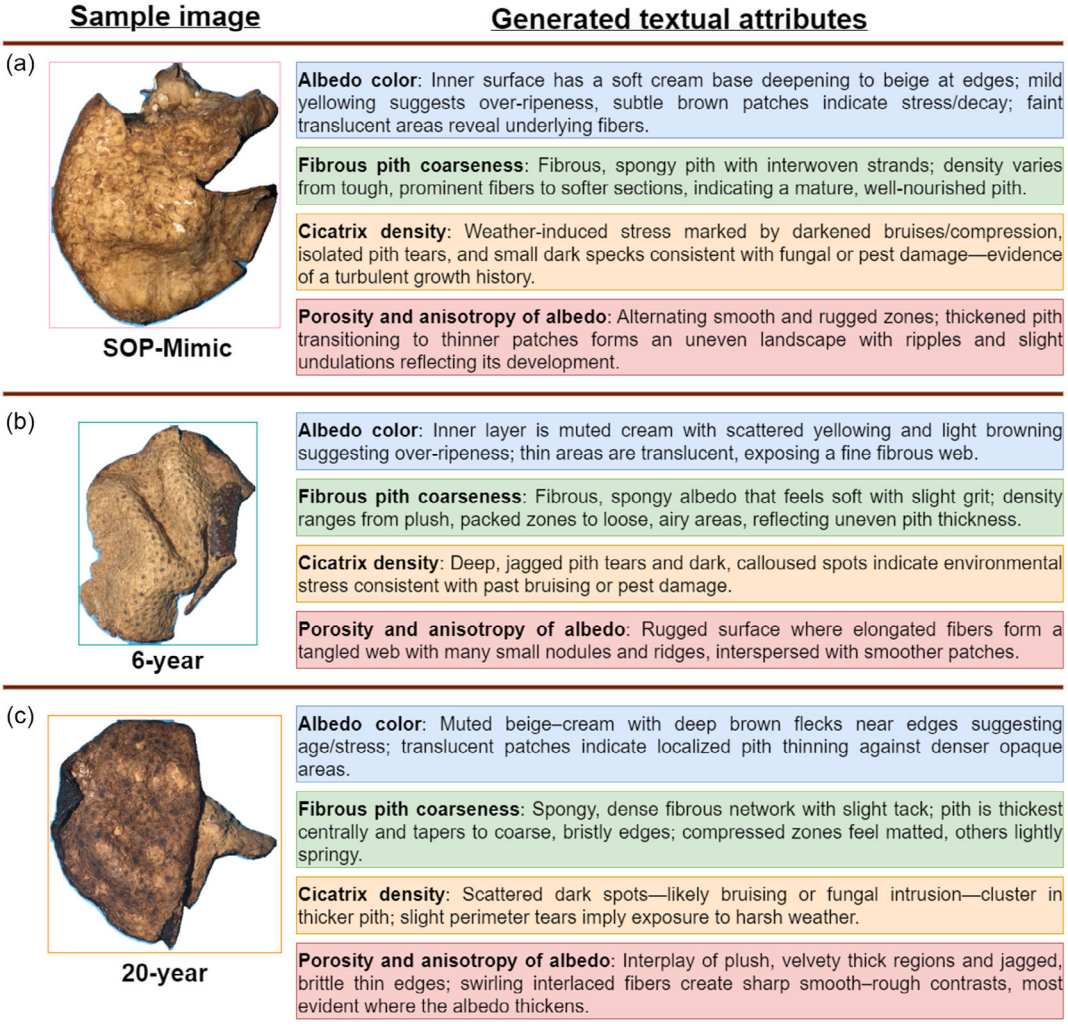

FIGURE 3 | Representative sample textual attributes generated by GPT-4o-mini for (a) an SOP-mimic sample, (b) a 6-year genuine sample, and (c) a 20-year genuine sample.

$$L_{\text{align}} = \frac{1}{2}\left[CE\left(\frac{\text{sim}\,(E_v, E_t)}{\tau}\right) + CE\left(\frac{\text{sim}\,(E_v, E_t)^T}{\tau}\right)\right]$$

where $sim(\cdot)$ is the batch-wise cosine similarity matrix, CE is the cross-entropy loss, and $\tau = 0.5$ is a temperature parameter. This loss enhances authenticity verification by penalizing visual-textual mismatches.

3 | Experiments and Results

3.1 | Dataset and Experimental Setup

To assess VALIANT's performance in both classification and verification tasks, we curated a dataset of genuine samples from Xinhui, Guangdong Province, China, and simulated SOP mimics. Genuine samples, sourced from authenticated suppliers, ensure varietal consistency with Xinhui chenpi, which is renowned for its aging potential. The genuine subset includes 761 images across five age categories: 63 for 1-year (fresh) peels, 177 for 6-year, 177 for 10-year, 181 for 15-year, and 163 for 20-year peels. These samples represent progressive biochemical changes during natural aging, including reduction of volatile oils, elevation of flavonoids (e.g., nobiletin and tangerine), and increases in phenolic acids and 5-hydroxymethylfurfural. Such changes boost pharmacological value, evident in darkening color, coarsened texture, and complex aromas [21].

For the secondary authenticity analysis, we generated 181 mimic images from 1-year genuine peels using an SOP designed to replicate common fraud tactics observed in the CRP market. This controlled simulation allows systematic evaluation of the model's ability to detect sophisticated counterfeits. The SOP mimics tea dyeing and chemical accelerated-aging methods through the following protocol: An aging solution is prepared by boiling 1 L water, steeping 50 g ripe Pu-erh tea (a common adulterant in CRP fraud) for 5–10 min, filtering, and adding 10 mL light soy sauce (for color darkening) and 2 g baking soda (to adjust pH and accelerate Maillard reactions) [22, 23]. Fresh 1-year peel batches undergo immersion for age-specific durations calibrated to approximate visual characteristics of target ages: 40 min for 6-year equivalents, 70 min for 10-year, 100 min for 15-year, and 120 min for 20-year. Peels are then incubated at 60°C for 4–6 h to simulate long-term oxidation and melanoidin formation [21, 24–26].

Experiments employed fivefold cross-validation for robustness. Implementations used PyTorch 2.4.0 on an NVIDIA RTX 3090 GPU (24 GB VRAM). For VALIANT, the Florence-2 backbone was fine-tuned with low-rank adaptation [27] (rank = 8, α = 16, and dropout = 0.1), using the AdamW optimizer, cosine annealing scheduler, batch size 8, and 80 epochs. Baselines (ConvNeXt [28], Vision Transformer (ViT) [29], MaxViT [30], Swin Transformer, and EfficientNet [31]) followed analogous protocols. Metrics—accuracy, precision, recall, F1 score, and area under the receiver operating characteristic (AUROC) curve—were averaged across folds, with standard deviations reported.

3.2 | Discriminative Capacity of Numerical Features

Figure 4a shows representative albedo-side images and enlarged local crops for each class. Younger CRP samples generally exhibit a lighter tone and a relatively loose pith structure, whereas older samples tend to show darker, denser, and more heterogeneous surface morphology. The enlarged local views further highlight differences in fibrous texture and local surface irregularity. At the same time, some adjacent age groups and SOP-mimic samples still display partial visual overlap, illustrating the practical difficulty of fine-grained discrimination from smartphone images alone.

To examine whether these visible trends can be captured by the numerical descriptors alone, we next evaluated the discriminative capacity of the numerical features using unsupervised clustering methods. As visualized through Principal Component Analysis (PCA) in Figure 4, these features demonstrate limited utility for independent classification. While the first two principal components capture substantial variance (PC1: 79.3%, PC2: 8.3%; 87.6% cumulative), the PCA projection displays considerable overlap among ground-truth age classes (Figure 4b) [32, 33]. The dominance of PC1 suggests that the descriptors primarily capture a coarse global trend rather than fine-grained age distinctions and that age-related variations are not well separated in a low-dimensional projection. To disentangle visual overlaps in 2D from true separability in the original feature space, we next examine whether an unsupervised algorithm can recover class-consistent structure when operating on the full 12D descriptors. Accordingly, we applied K-means clustering directly in the standardized 12D feature space with $K = 6$ and visualized the resulting assignments in the PCA plane for visualization only (Figure 4c). Whereas Figure 4b is colored by ground-truth labels, Figure 4c is colored by the raw K-means cluster assignments, with C0–C5 denoting arbitrary cluster indices (no semantic ordering and not aligned to age classes) and red crosses indicating the corresponding cluster centroids projected into the PCA coordinates. Qualitatively, K-means identifies groups in the numerical feature space, but these groups do not map cleanly onto the ground-truth classes, motivating a quantitative alignment analysis. Cluster quality metrics across K (Silhouette score and Davies–Bouldin index; Figure 4d) show no clear elbow or stable optimum; notably, the best Silhouette score occurs at $K = 2$, indicating that the numerical descriptors primarily support a coarse partition rather than a meaningful six-way separation consistent with the six classes. Quantitative metrics confirm this modest separability: for $K = 6$, the Silhouette score (0.256; higher is better) and Davies–Bouldin index (1.145; lower is better) indicate weak cohesion and separation. Agreement with the ground-truth labels is weak, with an adjusted rand index of 0.091 and normalized mutual information of 0.196. The class–cluster heatmap (Figure 4e) further reports a row-normalized contingency matrix; its columns are permuted via Hungarian matching solely to maximize diagonal mass for readability (cluster labels unchanged), so diagonal entries represent the fraction of each class captured by its best-matched cluster and reveal substantial fragmentation across clusters. Taken together, Figures 4c and 4e show that although clustering discovers structure in the 12D descriptors, that structure is largely orthogonal to the age-defined classes, and thus insufficient for standalone age discrimination. Physically, these features act as statistical proxies for global appearance changes—CIELAB summary statistics reflect overall lightness/chromatic shifts, GLCM properties capture coarse texture regularity, and Laplacian variance relates to high-frequency sharpness—but these cues are also sensitive to acquisition and surface-condition variability. However, when integrated within VALIANT's multimodal framework, these complementary

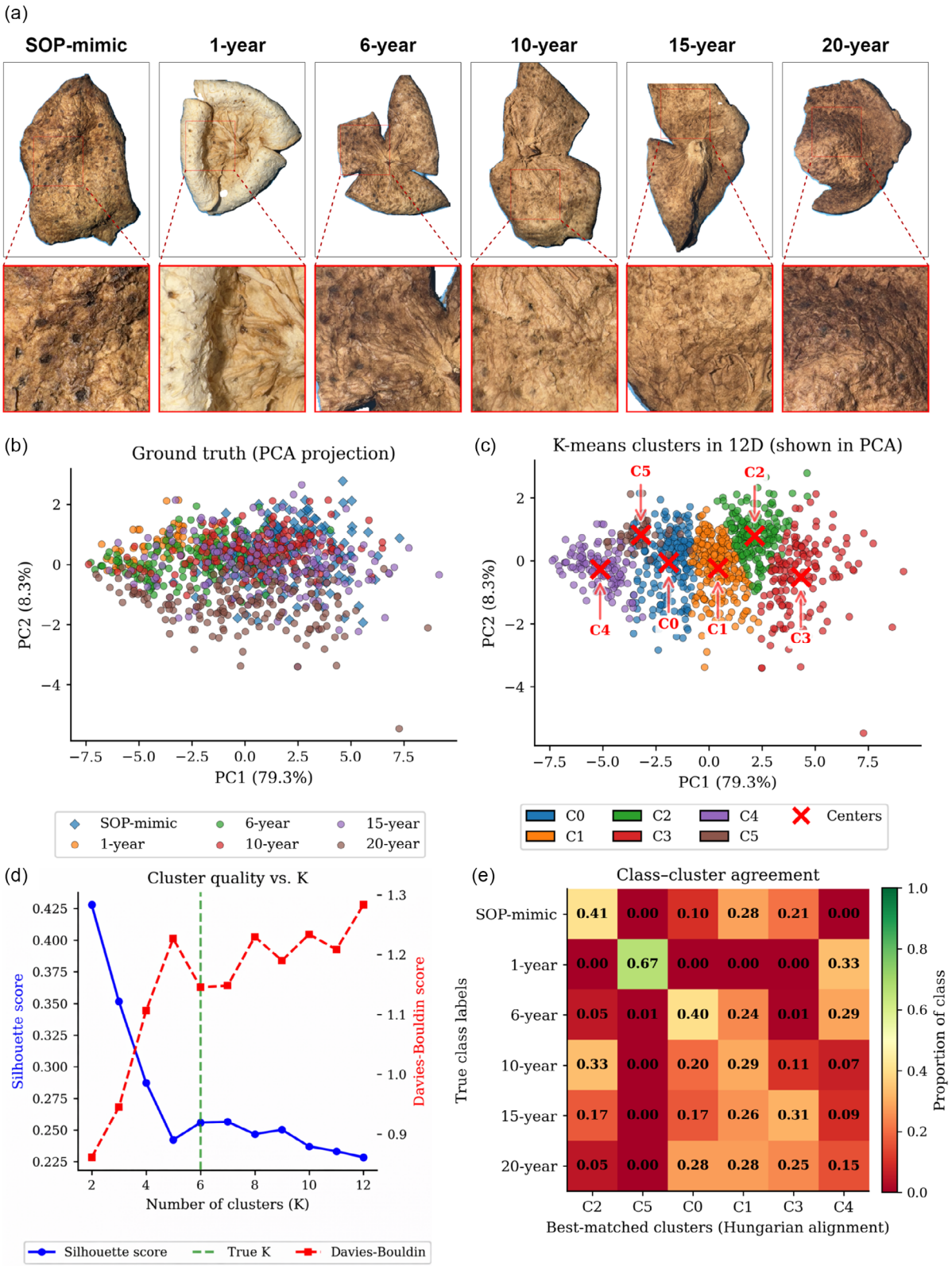

FIGURE 4 | Visual and numerical analysis of age-related differences in CRP. (a) Representative albedo-side images from each class, together with enlarged local crops highlighting finer surface details. (b) PCA projection of standardized numerical features colored by ground-truth classes. (c) K-means clustering performed in the full standardized 12D space ($K = 6$), visualized in the PCA plane with raw cluster IDs (C0–C5); red crosses indicate projected cluster centroids. (d) Silhouette score and Davies–Bouldin index versus K; dashed line indicates $K = 6$. (e) Row-normalized class–cluster contingency matrix; columns are permuted via Hungarian matching for readability.

TABLE 1 | Performance of VALIANT and vision-only baselines on the genuine CRP dataset. Metrics averaged over fivefold cross-validation (± standard deviations); bold indicates best performance.

Model	Accuracy	Precision	Recall	F1 score	AUROC
EfficientNet [31]	82.5 ± 2.4	82.5 ± 2.4	82.8 ± 2.4	82.4 ± 2.4	93.7 ± 1.6
ConvNeXt [28]	84.8 ± 2.8	85.2 ± 2.9	85.2 ± 2.7	84.6 ± 2.8	**95.0 ± 1.6**
MaxViT [30]	83.9 ± 3.3	84.5 ± 3.7	84.3 ± 3.3	83.8 ± 3.1	94.6 ± 1.0
ViT [29]	84.9 ± 1.7	85.2 ± 1.6	85.3 ± 1.6	84.8 ± 1.6	94.5 ± 0.9
Swin [10]	84.6 ± 2.0	84.9 ± 2.1	85.0 ± 2.0	84.5 ± 2.0	94.5 ± 0.8
VALIANT	**88.7 ± 1.7**	**89.2 ± 1.6**	**88.9 ± 1.7**	**88.6 ± 1.7**	94.7 ± 0.2

signals help reduce noise-related errors and improve performance over vision-only approaches (Section 3.3).

3.3 | Performance on Genuine Dataset

We benchmarked VALIANT against several prominent vision-based baselines—ConvNeXt, MaxViT, ViT, Swin Transformer, and EfficientNet—selected for their established efficacy in fine-grained classification tasks. As 1-year peels are visually distinct from aged samples and represent a trivial classification case, this experiment focused on the more difficult and commercially relevant challenge of distinguishing between the four aged classes (6–, 10-, 15–, and 20-year). All models underwent training and testing on the genuine dataset via fivefold cross-validation. Metrics are reported in Table 1.

VALIANT outperforms baselines across all metrics, achieving a mean accuracy of 88.7% (± 1.7%), precision of 89.2% (± 1.6%), recall of 88.9% (± 1.7%), F1 score of 88.6% (± 1.7%), and AUROC of 94.7% (± 0.2%). This represents 4%–6% accuracy improvements over top baselines such as ViT (84.9% ± 1.7%), attributable to multimodal integration. For qualitative insights, we visualize attention mechanisms on representative samples for 6-year (Figure 5a), 15-year (Figure 5b), and 20-year (Figure 5c) peels. The attention heatmaps (red = high attention, blue = low attention) reveal VALIANT's consistent focus on the central albedo region, where age-discriminative features such as oil glands, fibrous pith coarseness, and color gradients are most pronounced. In contrast, vision-based baselines like ViT display scattered attention across background areas and image edges, suggesting distraction by imaging artifacts or irrelevant features rather than true age markers.

3.4 | Performance on Mixed Dataset

Building on genuine dataset results, we assessed VALIANT's robustness on a mixed dataset that includes SOP mimics. This configuration tests both objectives jointly. Baselines and VALIANT were retrained on mixed data via fivefold cross-validation; metrics are listed in Table 2.

VALIANT demonstrates improved resilience, with an accuracy of 92.2% (± 2.1%), precision of 92.7% (± 2.0%), recall of 92.5% (± 2.1%), F1 score of 92.3% (± 2.1%), and AUROC of 97.7% (± 0.7%). It outperforms baselines by 2.2%–5.6% in accuracy. VALIANT's fusion, leveraging cross-attention on numerical and textual features, helps resolve ambiguities and reduce false positives for forgeries.

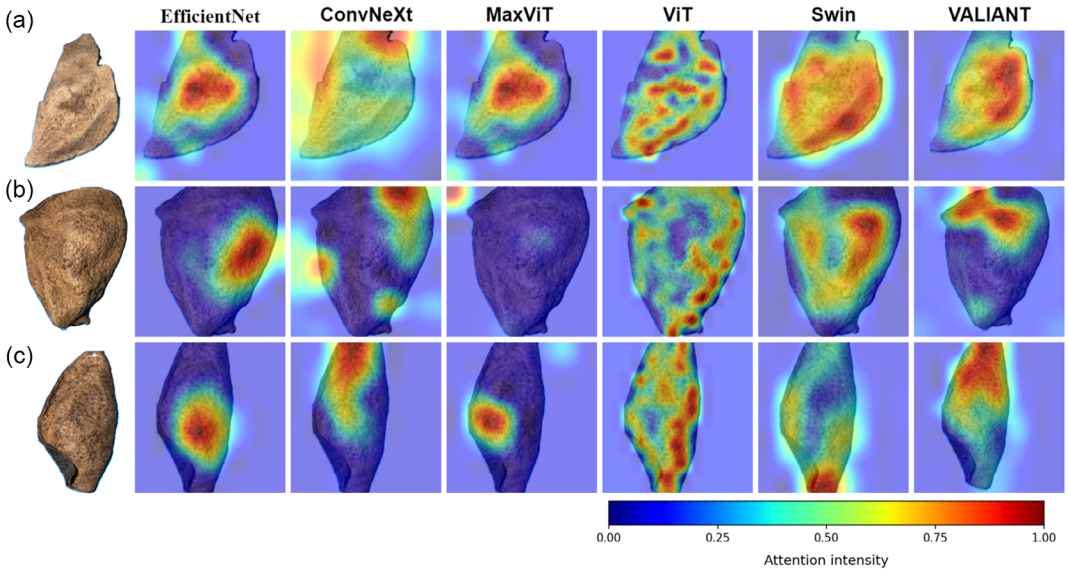

FIGURE 5 | Attention heatmaps from VALIANT and vision-only baselines for genuine CRP samples from three age classes: (a) 6-year, (b) 15-year, and (c) 20-year.

TABLE 2 | Performance of VALIANT and vision-only baselines on the mixed CRP dataset (genuine + SOP mimics). Metrics averaged over fivefold cross-validation (± standard deviations); bold indicates best performance.

Model	Accuracy	Precision	Recall	F1 score	AUROC
EfficientNet	86.6 ± 0.9	87.0 ± 1.0	87.1 ± 0.9	86.8 ± 0.9	96.7 ± 0.5
ConvNeXt	90.0 ± 1.6	90.4 ± 1.5	90.4 ± 1.5	90.3 ± 1.5	97.6 ± 0.3
MaxViT	90.0 ± 0.5	90.5 ± 0.5	90.4 ± 0.5	90.2 ± 0.5	97.6 ± 0.5
ViT	89.7 ± 0.5	90.0 ± 0.4	90.0 ± 0.5	89.8 ± 0.5	97.2 ± 0.9
Swin	90.0 ± 0.7	90.4 ± 0.9	90.4 ± 0.7	90.0 ± 0.8	97.3 ± 1.1
VALIANT	**92.2 ± 2.1**	**92.7 ± 2.0**	**92.5 ± 2.1**	**92.3 ± 2.1**	**97.7 ± 0.7**

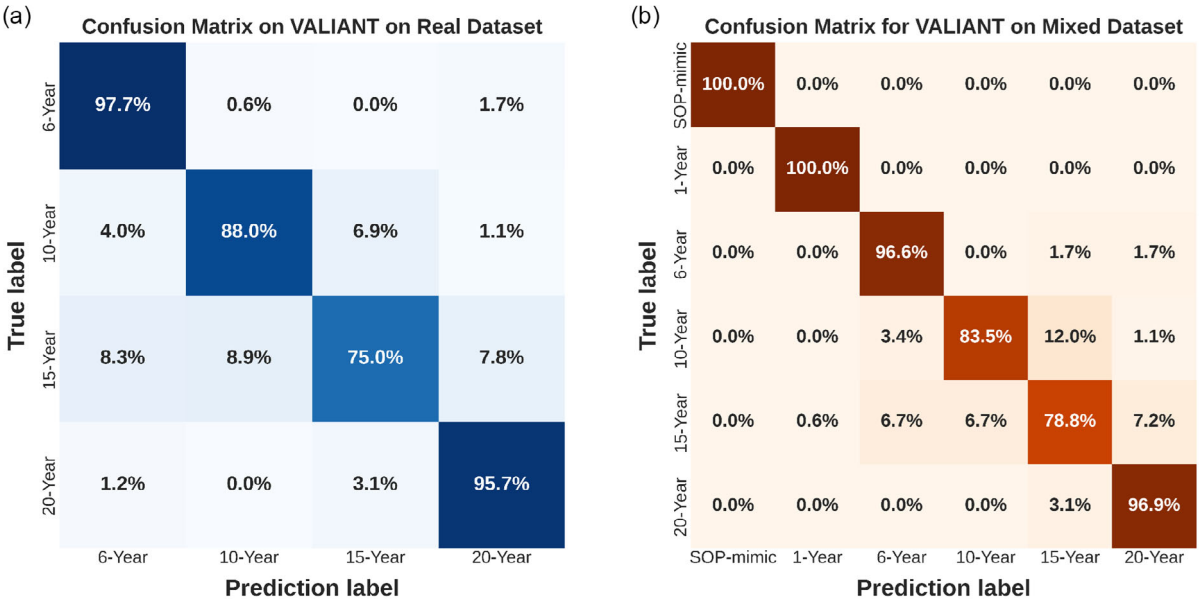

FIGURE 6 | VALIANT confusion matrices (row-normalized, %). (a) Genuine dataset (4 classes: 6–, 10-, 15–, and 20-year). (b) Mixed dataset (6 classes: SOP-mimic, 1–, 6-, 10–, 15-, and 20-year).

The confusion matrices in Figure 6 provide per-class insights into VALIANT's performance. For the genuine dataset (Figure 6a), the model demonstrates strong performance, with high accuracy for the 6-year (97.7%) and 20-year (95.7%) classes. Most misclassifications occur between the adjacent 10- and 15-year categories, which is expected given the subtle visual transitions during these aging periods. On the mixed dataset (Figure 6b), VALIANT correctly identifies all 181 SOP mimics, achieving 100% recall for that class. It is also noteworthy that the model achieves 100% accuracy on genuine 1-year samples, effectively discriminating them from the SOP mimics that were derived from them. Together, these results underscore VALIANT's dual capabilities in both fine-grained age classification and robust authenticity verification.

3.5 | Ablation Studies

We conducted an ablation study to assess each modality's contribution by progressively adding them to a visual-only baseline. Results are shown in Table 3.

The vision-only baseline delivers solid results but modest specificity (98.4%) for SOP-mimic detection. Incorporating numerical features produces incremental gains: accuracy rises from 92.1% to

TABLE 3 | Ablation study: Effects of numerical and textual modalities on VALIANT performance. Results are from a single split.

Model	Accuracy	Precision	Recall	F1 score	Specificity
Vision-only	92.1	92.3	92.5	92.2	98.4
Vision + numerical	92.6	92.7	92.9	92.7	98.5
Vision + text	93.1	93.2	93.4	93.2	98.6
Full	**95.1**	**95.1**	**95.3**	**95.1**	**99.0**

92.6%. Substituting numerical with textual features achieves similar accuracy (93.1%) alongside improved precision (93.2%). Full integration reveals synergies: accuracy attains 95.1% (a 2.0% gain over vision-plus-text), with precision (95.1%), recall (95.3%), and F1 (95.1%). SOP-mimic specificity reaches 99.0%, consistent with 100% recall (Figure 6b). These results demonstrate that numerical and textual modalities contribute complementary information, with their full integration yielding synergistic improvements in both tasks. VALIANT's multimodal design, rather than a simple concatenation of existing detection and text-generation tools, is what enables it to significantly outperform vision-only baselines.

4 | Discussion

VALIANT provides a sample preparation-free, smartphone-based approach for rapid CRP screening. With 92.2% accuracy on the mixed dataset, VALIANT approaches or exceeds reported expert accuracy (approximately 83%) and achieves performance comparable to laboratory-based methods, such as NIRS, though direct comparisons should be interpreted cautiously due to differences in evaluation protocols.

Despite its strong performance, this study has several limitations. First, the dataset consists solely of Xinhui CRP, which may limit generalizability to other geographical varieties and cultivars. Second, while the SOP mimics were designed to reflect common fraudulent methods, the model's robustness against novel or more sophisticated counterfeiting techniques remains to be validated. Third, the numerical features demonstrate sensitivity to variations in lighting and smartphone sensor characteristics; the absence of explicit color calibration may introduce systematic bias. Finally, the SOP mimics were all derived from 1-year peels using a single procedure, meaning the reported accuracy is specific to this type of forgery.

Future work should prioritize testing the model against a more diverse set of third-party forgeries to validate its real-world effectiveness. Additionally, the VALIANT pipeline could be extended to address other authenticity challenges, such as spice adulteration or wine aging [28].

5 | Conclusion

We present VALIANT, a smartphone-based multimodal framework for CRP classification and verification. By integrating visual, numerical, and textual features through gated cross-attention, VALIANT demonstrates improved performance over unimodal baselines and achieves high specificity for SOP-produced mimics. These results suggest VALIANT may serve as an accessible screening tool to complement expert appraisal and laboratory confirmation in CRP authentication workflows. Broader validation across geographic regions, device types, lighting conditions, and counterfeiting methods will be essential for large-scale deployment and to strengthen consumer protection in specialty agricultural markets.

Acknowledgments

The authors acknowledge funding support from the HKUST startup grant (R9421). The authors also acknowledge the discussion with Mr. Ping Wong, who revealed the value of the project in the early stage, and Prof. Karl W. K. Tsim at HKUST for providing a few CRPs from authenticated suppliers.

Funding

Hong Kong University of Science and Technology (R9421).

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. "新會陳皮造假, "I 廣西 「 工藝皮 」 冒充1個月速成扮 「 10年貨 」 售2000元," *Accessed: Dec* 17, 2025, [Online]. Available: https://www.hk01.com/%E5%8D%B3%E6%99%82%E4%B8%AD%E5%9C%8B/60303971/%E6%96%B0%E6%9C%83%E9%99%B3%E7%9A%AE%E9%80%A0%E5%81%87-%E5%BB%A3%E8%A5%BF-%E5%B7%A5%E8%97%9D%E7%9A%AE-%E5%86%92%E5%85%85-1%E5%80%8B%E6%9C%88%E9%80%9F%E6%88%90%E6%89%AE-10%E5%B9%B4%E8%B2%A8-%E5%94%AE2000%E5%85%83.

2. Y. Li, W. Qian, Z. Wen, et al., Recent Advances in the Authentication (geographical Origins, Varieties and Aging Time) of Tangerine Peel (Citri Reticulatae Pericarpium): A Review, *Food Chem* 442 (2024), https://doi.org/10.1016/j.foodchem.2024.138531.

3. Y. Guo, X. Yu, W. Hong, Y. Cai, W. Xu, and H. Gu, "On-Site Rapid Detection of Aging of Pericarpium Citri Reticulatae Using Multispectral Imaging," *Journal of Near Infrared Spectroscopy* 31, no. 5 (2023): 263–270, https://doi.org/10.1177/09670335231194737.

4. M. Zhang, C. Ma, J. Ma, Y. Yuan, J. Huang, and Y. Yan, "Vis/NIR Spectroscopy and Machine Learning Model for Counterfeit Citri Reticulatae Pericarpium Identification," *Journal of Food Composition and Analysis* 148 (2025), https://doi.org/10.1016/j.jfca.2025.108240.

5. Z. Chu, F. Li, D. Wang, S. Xu, C. Gao, and H. Bai, "Research on Identification Method of Tangerine Peel Year Based on Deep Learning," *Food Science and Technology* 42 (2022), https://doi.org/10.1590/fst.64722.

6. Y. Qin, Q. Zhao, D. Zhou, et al., "Application of Flash GC E-Nose and FT-NIR Combined with Deep Learning Algorithm in Preventing Age Fraud and Quality Evaluation of Pericarpium Citri Reticulatae," *Food Chem: X* 21 (2024), https://doi.org/10.1016/j.fochx.2024.101220.

7. S. Pan, X. Zhang, W. Xu, J. Yin, H. Gu, and X. Yu, "Rapid On-Site Identification of Geographical Origin and Storage Age of Tangerine Peel by Near-Infrared Spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 271 (2022), https://doi.org/10.1016/j.saa.2022.120936.

8. F. Deng, J. Li, L. Fu, et al., "CNFA: ConvNeXt Fusion Attention Module for Age Recognition of the Tangerine Peel," *Journal of Food Quality* 2024 (2024), https://doi.org/10.1155/2024/6439900.

9. B. Xiao, H. Wu, W. Xu, et al., "Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024): 4818–4829.

10. Z. Liu, H. Hu, Y. Lin, et al., "Swin Transformer V2: Scaling Up Capacity and Resolution," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022):12009–12019, http://arxiv.org/abs/2111.09883.

11. OpenAI, J. Achiam, S. Adler, S. Agarwal et al., "GPT-4 Technical Report," arXiv:2303.08774 (2024), http://arxiv.org/abs/2303.08774.

12. J. S. Nantongo, E. Serunkuma, G. Burgos, et al., "Color and Grey-Level Co-Occurrence Matrix Analysis for Predicting Sensory and Biochemical Traits in Sweet Potato and Potato," *International Journal of Food Science* 2024, no. 1 (2024), https://doi.org/10.1155/2024/1350090.

13. R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," *rXiv:2410.17725* (2024), http://arxiv.org/abs/2410.17725.

14. C. M. Badgujar, A. Poulose, and H. Gan, "Agricultural Object Detection With You Only Look Once (YOLO) Algorithm: A Bibliometric and Systematic Literature Review," *Computers and Electronics in Agriculture* 223 (2024), https://doi.org/10.1016/j.compag.2024.109090.

15. G. Vijayakumari, G. Prithipa, S. Rithvedha, and S. Santhiya, "Fruit Quality Analysis Using GLCM Algorithm in Image Processing," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering* 8, no. 2 (2019): https://doi.org/10.15662/IJAREEIE.2019.0802021.

16. A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning With Contrastive Predictive Coding," *arXiv:1807.03748* (2018), http://arxiv.org/abs/1807.03748.

17. K. E. Ak, J. Mohta, D. Dimitriadis, S. Manchanda, Y. Xu, and M. Shen, "Aligning Vision Language Models With Contrastive Learning," *in European Conference on Computer Vision – ECCV 2024 Workshops* (2025): 32–45. https://doi.org/10.1007/978-3-031-91672-4_3.

18. W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers," *Advances in Neural Information Processing Systems* (2020), http://arxiv.org/abs/2002.10957.

19. H. Zhu, S. Qin, M. Su, C. Lin, A. Li, and J. Gao, "Harnessing Large Vision and Language Models in Agriculture: A Review," *Frontiers in Plant Science* 16 (2025), https://doi.org/10.3389/fpls.2025.1579355.

20. Y. Xiao, Y. Li, G. Cui, H. Zhang, and W. Zhang, "A Systematic Review of Multimodal Fusion Technologies for Food Quality and Safety Assessment: Recent Advances and Future Trends," *Trends in Food Science & Technology* 164 (2025): 105277, https://doi.org/10.1016/j.tifs.2025.105277.

21. W. Zhang, X. Fu, Y. Zhang, et al., "Metabolome Comparison of Sichuan Dried Orange Peels (Chenpi) Aged for Different Years," *Horticulturae* 10, no. 4 (2024): 421, https://doi.org/10.3390/horticulturae10040421.

22. Y. Qi, W. Wang, T. Yang, W. Ding, and B. Xu, "Maillard Reaction in Flour Product Processing: Mechanism, Impact on Quality, and Mitigation Strategies of Harmful Products," *Multidisciplinary Digital Publishing Institute (MDPI)* 14, no. 15 (2025): 2721, https://doi.org/10.3390/foods14152721.

23. L. El Hosry, V. Elias, V. Chamoun, et al., "Maillard Reaction: Mechanism, Influencing Parameters, Advantages, Disadvantages, and Food Industrial Applications: A Review," *Multidisciplinary Digital Publishing Institute (MDPI)* 14, no. 11 (2025): 1881, https://doi.org/10.3390/foods14111881.

24. J. Liang, Y. Xi, J. Li, et al., "Metabolomic Profiling Reveals the Quality Variations in Citri Reticulatae Pericarpium (Citrus Reticulata Blanco Cv. Chachiensis) with Different Storage Ages in Response to Candidatus Liberibacter Asiaticus Infection," *Foods* 13, no. 6 (2024), https://doi.org/10.3390/foods13060827.

25. Y. Hu, D. Hu, L. Yin, et al., "Dynamic Profiling of Bioactive Compounds, Flavor Metabolites, and Quality-Related Microorganisms During the Freshening-Drying-Aging Process of Citri Reticulatae Pericarpium: Implications for Quality Formation Mechanisms," *Food Chemistry: X* 30 (2025), https://doi.org/10.1016/j.fochx.2025.102906.

26. TVB USA Official, "東張西望 | 平嘢扮貴嘢？陳皮大師教你分真偽！," Accessed 26, 2026, https://youtu.be/bLY2zVQDTBQ?si=iig04YON1MdNNO2s.

27. E. J. Hu, Y. Shen, P. Wallis, et al., "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations (ICLR)* (2022), http://arxiv.org/abs/2106.09685.

28. S. Woo, S. Debnath, R. Hu, et al., "ConvNeXt V2: Co-Designing and Scaling ConvNets With Masked Autoencoders," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), http://arxiv.org/abs/2301.00808.

29. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)* (2021), http://arxiv.org/abs/2010.11929.

30. Z. Tu, H. Talebi, H. Zhang et al, "MaxViT: Multi-Axis Vision Transformer," in *European Conference on Computer Vision (ECCV)* (2022). http://arxiv.org/abs/2204.01697.

31. M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," *in International Conference on Machine Learning* (2021):10096–10106, http://arxiv.org/abs/2104.00298.

32. F. A. Listya and N. Rokhman, "Classification of Tangerine (Citrus Reticulata Blanco) Quality Using Combination of GLCM, HSV, and K-NN," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems* 13, no. 4 (2019): 357, https://doi.org/10.22146/ijccs.47906.

33. P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics* 20 (1987): 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.