

Adaptive-Resolution FastViT-EMA Model for Texture Estimation of Bakery Products

Radek Jakub Kostecki^{1,*}

¹ Faculty of Information and Communication Technology, West Pomeranian University of Technology in Szczecin, Szczecin, 70-310, Poland

*Corresponding author: radek.jk@stud.zut.edu.pl

Abstract. Texture estimation remains a practical bottleneck in bakery quality control because mechanical texture profile analysis is slow, destructive and unsuitable for continuous line-level inspection. This paper proposes an adaptive-resolution FastViT-EMA model for predicting multiple texture indices of food products and adjusts visual detail according to the complexity of crumbs, crusts and edges. A light-weight resolution controller, a FastViT backbone with convolutional token mixing, and an exponential moving average teacher are used to reduce instability caused by batch-to-batch variations in proofing and baking. Organize a controlled dataset of 1,260 samples of bread, cake, bun, biscuit and laminated pastry products that match the labels for texture profile analysis. Compared with the fixed-resolution FastViT baseline, the proposed model had a lower average RMSE of 0.246 and a higher mean R^2 of 0.918, and also reduced inference cost by 24.7% on the same hardware budget. Crumb-sensitive products showed the largest improvement; the reduction in hardness RMSE was 18.4% and the decrease in resilience error was 15.9%. Ablation analysis showed that adaptive resolution accounted for 0.031 R^2 and EMA stabilization added 0.018 R^2 , and together they performed better in mixed illumination. Based on the above experiments, Texture can be predicted to an engineering level by including the scale selection step in the model rather than performing it in pre-processing.

Keywords: *Bakery Texture, Adaptive Resolution, FastViT, Exponential Moving Average, Non-Destructive Quality Estimation*

Received on 13 December 2024, Accepted on 08 June 2025, Published on 24 June 2025

Copyright © 2025 Author(s), licensed to DEA. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

Introduction

Texture is a measurable mechanical response that reflects crumb elasticity, crust brittleness, cohesiveness and recovery after compression, and these properties are difficult to observe in real time during bakery production. A bread crumb that looks visually open may still be elastic or dry, and a biscuit with a good color may break too easily after cooling. Traditionally, the standardized indices of Texture Profile Analysis (TPA) have been hardness, springiness, cohesiveness, chewiness and resilience; however, they are destructive tests that require careful sample preparation. Therefore, only large bakeries conduct regular laboratory experiments, and small-scale producers are generally restricted to visual observation and the skill of their staff. Machine vision is now attractive because surface color, pore distribution, crust cracking and shape deformation can all be indirectly related to internal structure [1]. The development of image-based food quality assessment has also shown that more non-contact inspection devices capable of high-speed operation are required [2]. Recently, research in food imaging has shown that the local structure is closely related to the processing history and final-product stability [3]. Baked products are also relatively challenging because they require the analysis of visual features at various scales, from millimeter-level pores to product-level shell gradients [4].

Deep convolutional networks improved automated food inspection by replacing hand-crafted color and texture descriptors with learned representations, but their fixed receptive-field behavior is inefficient when the same image contains both large homogeneous crust regions and fine crumb boundaries [5]. Transformer-based vision

models have introduced stronger long-range modelling to connect pore fields, edge collapse and global shape, but many pure attention models are data- and computation-intensive [6]. Lightweight hybrid architectures such as FastViT are attractive because they combine the convolutional inductive bias of CNNs with the token-mixing ability of self-attention more efficiently. A fixed input resolution is still a limitation of the Design. High resolution improves the observation of pores and cracks but is relatively expensive and more sensitive to noise; low resolution retains the whole-image color and geometry but loses small structures closely related to texture [7]. Exponential Moving Average (EMA) training is another suitable method; otherwise, there are issues such as changes in ingredient batches, oven zones, slicing directions and cooling times that can lead to unstable gradients when using continuous mechanical measurement labels [8].

The current study introduces an adaptive-resolution FastViT-EMA model for estimating bakery product texture from RGB images. Not all the samples need to be shown at the same visual scale. A lightweight controller will determine the local structural complexity and map image regions according to the resolution settings of the anticipated contributions to the texture target. Then, the backbone extracts good hybrid features, and an EMA teacher provides slowly updated supervisory targets to reduce noise in multi-output regression. This way is suitable for general bakery inspection scenarios where high-cost spectral instruments are unavailable, and standardized image acquisition can be achieved by means of a light box, a camera and a reproducible cutting protocol. The engineering problem is to improve the accuracy of resolution adaptation and EMA stabilization without reducing throughput.

The paper has the following organization. Section 2 reviews image-based bakery quality assessment and efficient visual learning models. Section 3 is the problem of texture estimation, imaging protocol and adaptive resolution principle. Section 4 shows the FastViT-EMA architecture, training objective and inference procedure. Section 5 shows the experimental data, comparison results, ablation analysis, error behavior and computation performance. Section 6 introduces the main results, existing problems and future research directions of this work.

Related Work

Image-Based Texture Assessment of Bakery Products

Image-based bakery assessment has moved from color thresholding and traditional texture descriptors to data-driven quality estimation. Early industrial applications focused on features that were easily observable by simple cameras, such as crust color and size-regularity of defects. Recently, some research has treated images as proxies for the physical structure and correlated properties such as pore-size distribution, cell-wall thickness, crust roughness, and edge deformation with mechanical texture [9]. The mapping from the image to texture is not one-to-one. Two samples may have the same average color but differ in water migration, starch retrogradation or gluten network strength, and thus exhibit different mechanical properties. For this reason, visual systems of bakery products are increasingly requiring supervised labels obtained from texture analyzers, and they need to predict continuous indices rather than categorical grades [10].

Crumb and crust images have also been used to characterize structural scale, including pore size, crust roughness and visible collapse zones. Fine-crumb pores are generally associated with a springy and elastic texture, while large voids and collapsed areas are relatively hard and chewy. Crust features contain information about the degree of baking, moisture loss and mechanical brittleness, but not the full structure of the internal matrix. Research on food microstructure has shown that a single global descriptor is not suitable for the objective of mechanical texture [11]. Recently, many image pipelines have started to crop several patches and improve illumination to reduce sample bias. These methods reduce sampling bias, but they generally use fixed-size inputs and do not explicitly determine which regions require high spatial detail for mechanical texture estimation.

Efficient Vision Models for Industrial Quality Prediction

Industrial Quality Prediction has different requirements than general image classification. The model should be reasonably small, stable in terms of regression results, and fast enough for repeated checks. Khan et al. [12] reported that efficient visual backbones can encode local edges and texture patterns with relatively few parameters, which is useful for repeated industrial quality prediction. To learn the spatial correlation of far-away

areas more easily than narrow convolutional kernels can, attention-based models connect these regions. Therefore, hybrid models have become popular in applied vision to maintain locality and enhance global context modelling [13]. For the estimation of bakery texture, the product's shape, crust uniformity and crumb structure need to be considered simultaneously.

Training stability is also required for the visual labels obtained from mechanical equipment. Texture profile analysis includes compression cycles and sample conditioning; thus, the label variance may be due to both product heterogeneity and measurement error. Regularization, data augmentation and teacher-student averaging have been used to mitigate the impact of noisy supervision in continuous prediction tasks [14]. Exponential moving average models are relatively simple in terms of inference and do not add a reference network at runtime. Most high-efficiency vision models still assume that the input resolution is fixed. Adaptation is used in other parts of engineering, but its application to bakery texture estimation has not been widely explored [15]. Therefore, a model has been introduced that integrates the design of scale selection, efficient feature learning and regression stabilization into a single inspection framework.

Adaptive-Resolution Texture Estimation Problem

Imaging Protocol and Texture Targets

The system to be proposed is for sliced or top-view bakery samples under controlled light. The five goals of mechanical texture are shown in the above sample images: hardness, springiness, cohesiveness, chewiness and resilience. The labels are normalized by product category to reduce the impact of naturally hard products such as biscuits on soft products such as cakes. Let x_i denote the i -th image and $y_i \in \mathbb{R}^5$ denote its measured texture vector. The normalized target vector is computed using category-specific mean and standard deviation values.

$$\hat{y}_{i,k} = \frac{y_{i,k} - \mu_{c,k}}{\sigma_{c,k} + \varepsilon}, k \in \{1, \dots, 5\} \quad \text{Eq.(1)}$$

The notation c represents the bakery category and ε prevents division by a very small standard deviation. This normalization focuses on variations within the same category of texture and is the main problem in actual quality control. For example, a soft cake and a hard biscuit should not be assigned the same absolute hardness value when the engineering problem is to find where it deviates from the expected product characteristic.

Figure 1 is the entire inspection process, starting from standardized acquisition and ending with multiple-output texture estimation. The workflow is slicing or top-surface positioning, white-balance correction, adaptive region analysis, FastViT-EMA inference, and output calibration. A figure is used here to present the problem scope before introducing the network structure in this section.

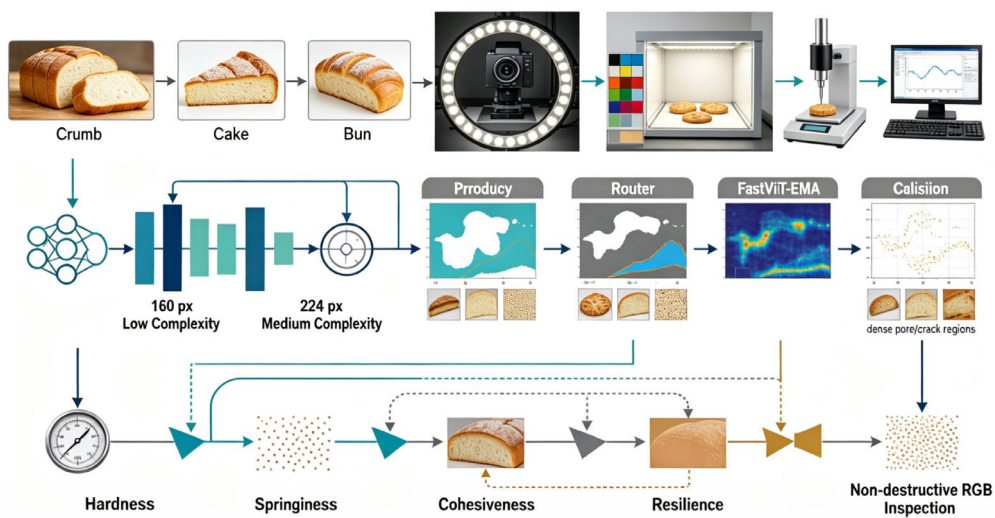


Figure 1. Standardized adaptive-resolution inspection workflow for bakery texture estimation.

Adaptive Resolution Controller

An adaptive-resolution controller selects a higher-resolution display for a particular area. Gradient Energy, Pore boundary density and local color variation are used to determine a structural complexity score. Regions with a smooth crust or uniform crumb are processed at a lower resolution; those containing pore clusters, cracks and edge deformation are sent to the high-resolution branch. This way does not use uniform scaling; only areas where high-frequency information needs to be maintained in the texture prediction are scaled differently.

$$s_j = \alpha G_j + \beta P_j + \gamma V_j, \alpha + \beta + \gamma = 1 \quad \text{Eq.(2)}$$

Here, G_j is normalized gradient energy, P_j is pore-boundary density estimated from thresholded local contrast, and V_j is color variance for region j . The resolution assignment r_j is selected by comparing s_j with two learned thresholds. The widths of the low-complexity area, medium-complexity area and high-complexity area at the short end are 160 pixels, 224 pixels and 288 pixels, respectively. The threshold is optimized in the validation set via a small differentiable relaxation during training and then discretized during inference.

$$r_j = 160 \cdot I(s_j < \tau_1) + 224 \cdot I(\tau_1 \leq s_j < \tau_2) + 288 \cdot I(s_j \geq \tau_2) \quad \text{Eq.(3)}$$

The Controller is relatively small. It is not to be substituted for the backbone; rather, it is intended to avoid visually uninformative regions in the backbone. In the initial Design calculations, the controller adds less than 2% to the total inference time and reduces the average number of high-resolution regions by over one-third. A compromise can be used for bakery inspection; given that a large number of items are visually uniform, full-resolution processing is not required.

Multi-Scale Texture Representation

After region routing, image features are embedded into a shared token space. Each region contributes a feature vector h_j , and the sample-level representation is produced by confidence-weighted aggregation. The confidence term is relatively large for high local complexity and when the backbone activations of the augmented views are consistent. Aggregation can be used to prevent a single visually extreme area from dominating the prediction, and damaged crust, collapsed crumb, or unusually dense pore areas can still affect the final result.

$$h_i = \frac{\sum_j w_j h_j}{\sum_j w_j + \varepsilon}, w_j = \text{softplus}(\lambda s_j + \rho q_j) \quad \text{Eq.(4)}$$

The term q_j denotes feature consistency, and λ and ρ are learned scalar weights. The design reflects the physical interpretation of bakery texture: local structure matters, but the product is evaluated as a whole sample. A highly collapsed region may indicate a true processing defect, whereas a single cut artifact should be discounted when its feature consistency is low.

The aggregation layer also separates the visual evidence from the product identity. Category information is used for label normalization and output calibration, but a category-specific branch is not provided for the backbone. This avoids having the model memorize that biscuits are hard or cakes are soft, and encourages it to learn intra-category deviations, such as over-proofed bread, under-baked buns, or dry cake crumb. In the engineering application, this division is required because the same inspection line may switch recipes, and the model needs to respond to visual signs of abnormal texture rather than simply reproducing expected category averages.

FastViT-EMA Model Design

FastViT Backbone for Bakery Texture Features

The first is a small FastViT-style backbone. Early stages use convolutional token embedding to learn edges, pore walls, crust ridges and local color variations. The later stage efficiently mixes tokens to model extended spatial dependencies in the product. It is suitable for bakery products because the visual references are often local and have a small scale. A dense crumb patch can increase the local hardness, and uneven crust browning may indicate baking non-uniformity in the whole sample.

$$z_l = M_l(z_{l-1}) + \text{FFN}\left(\text{LN}(M_l(z_{l-1}))\right), l = 1, \dots, L \quad \text{Eq.(5)}$$

In this expression, M_l is the efficient mixing operator at stage l , LN denotes layer normalization, and FFN is a feedforward block. The model has a regression head and outputs the five continuous texture values. A

classification system only needs to identify the main class, but the regression head should keep the small differences in correlated mechanical properties. Hardness and chewiness are closely related, and springiness and resilience respond differently to crumb elasticity and moisture distribution.

EMA Teacher and Regression Objective

Update the EMA teacher using the student weights after each optimization step. A stable model trajectory and reduced prediction oscillation caused by noisy labels were achieved through the above measures. The teacher is not used as an extra network during deployment; only the final averaged weights for inference are saved. Thus, this way is feasible in a production environment with low latency and maintenance costs.

$$\theta_t^{EMA} = m\theta_{t-1}^{EMA} + (1 - m)\theta_t, 0 < m < 1 \quad \text{Eq.(6)}$$

The training objectives are supervised regression loss, consistency loss of augmented views and a small resolution-cost penalty. A cost penalty will not be triggered for a small-gain prediction. In the experiment, the penalty coefficient was set to a smaller value to prioritize accuracy.

$$\mathcal{L} = \|f_\theta(x_i) - \hat{y}_i\|_1 + \eta \|f_\theta(a_1(x_i)) - f_{\theta^{EMA}}(a_2(x_i))\|_2^2 + \kappa C(r) \quad \text{Eq.(7)}$$

The first term is relatively insensitive to moderate label outliers, the second term aligns student and teacher predictions, and the third term is the expected computational cost under the selected resolutions. Cosine learning-rate decay and category-balanced sampling are used to optimize the loss. Category balance should be achieved to prevent the predominance of ordinary products such as bread and buns in the gradient; infrequent laminated pastries offer information on brittle and layered textures.

Inference and Calibration

During inference, a sample is first passed to the resolution controller, then processed by the FastViT backbone, and finally converted into normalized texture predictions through the regression head. Then reduce to the original mechanical scale using category-specific calibration constants for the output. A linear calibration layer is added to the validation set to correct for a systematic bias in all texture dimensions.

$$\hat{y}_{cal} = A\hat{y} + b \quad \text{Eq.(8)}$$

The calibration matrix A is constrained to be diagonal in the deployed model to avoid artificial transfer among texture indices. It is still interpretable because the correction to hardness does not directly affect springiness or resilience. Figure 2 is the model architecture and inference path, which include the adaptive resolution controller, FastViT feature hierarchy, EMA training branch and final calibrated texture vector.

During deployment, calibration is also used as a quality-control safeguard rather than only as a numerical post-processing step. Validation statistics are stored with the product category, imaging configuration and texture-analyzer setting, so predictions obtained after a changed camera distance or illumination setup can be flagged before entering the production record. If any calibrated output exceeds the category-specific physical range, the system retains the prediction but marks it for manual review.

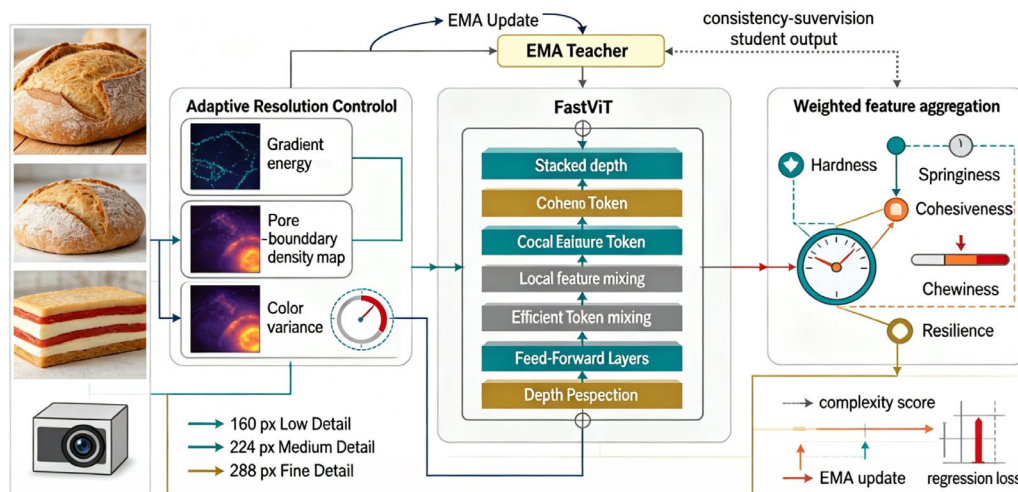


Figure 2. FastViT-EMA architecture with adaptive routing, EMA teacher update, and calibrated multi-output regression.

During deployment, calibration is also used as a quality-control safeguard rather than only as a numerical post-processing step. The validation statistics are stored with the product category, imaging configuration and texture-analyzer setting, so that predictions obtained from a changed camera distance or illumination setup can be flagged before they enter the production record. If any calibrated output exceeds the category-specific physical range, the system retains the prediction but marks it for manual review. This design keeps the inference procedure lightweight while allowing operators to distinguish normal process variation from possible acquisition or sample-preparation errors.

Experimental Data and Analysis

Dataset, Metrics, and Baselines

The 1,260 samples of the bakery experiment are divided into five types: bread, cake, bun, biscuit and laminated pastry. There are 252 samples per category of changes to flour type, proofing time, baking temperature, cooling time and storage period. Texture Profile Analysis is used to obtain the hardness, springiness, cohesiveness, chewiness and resilience values of each image of the mechanical label. A 70% training set, a 15% validation set, and a 15% test set were generated at the batch level to prevent neighboring products from appearing in different splits. Recently, research in food inspection has shown that for samples with the same ingredients and oven conditions, batch separation is required [16].

Figure 3(a) presents the balanced category composition and the corresponding train-validation-test split, where each product group contains 252 samples and is divided into 176 training, 38 validation, and 38 testing samples, while Figure 3(b) shows that hardness spans from 8.6 N in cake to 42.8 N in biscuit with percentile bands and a soft-product control zone. Figure 3(c) presents the target correlation matrix, where hardness and chewiness have a correlation of 0.86, but springiness and resilience remain less coupled at 0.58. Figure 3(d) shows representative illumination-normalized color variance, and laminated pastry has the largest interquartile range at 0.214 due to layered browning. The Patterns of these datasets are required to ensure that the model can handle correlated mechanical labels and category-specific visual variations. The assessment indices are RMSE, MAE, R^2 , mean inference delay, and estimated Giga-Multiply-Accumulate operations. Baselines are ResNet-50, EfficientNet-B0, MobileViT-S, a fixed-resolution FastViT model, and the proposed adaptive-resolution FastViT-EMA model. Both convolutional and hybrid baselines are used according to recent comparisons of efficient visual quality prediction [17].

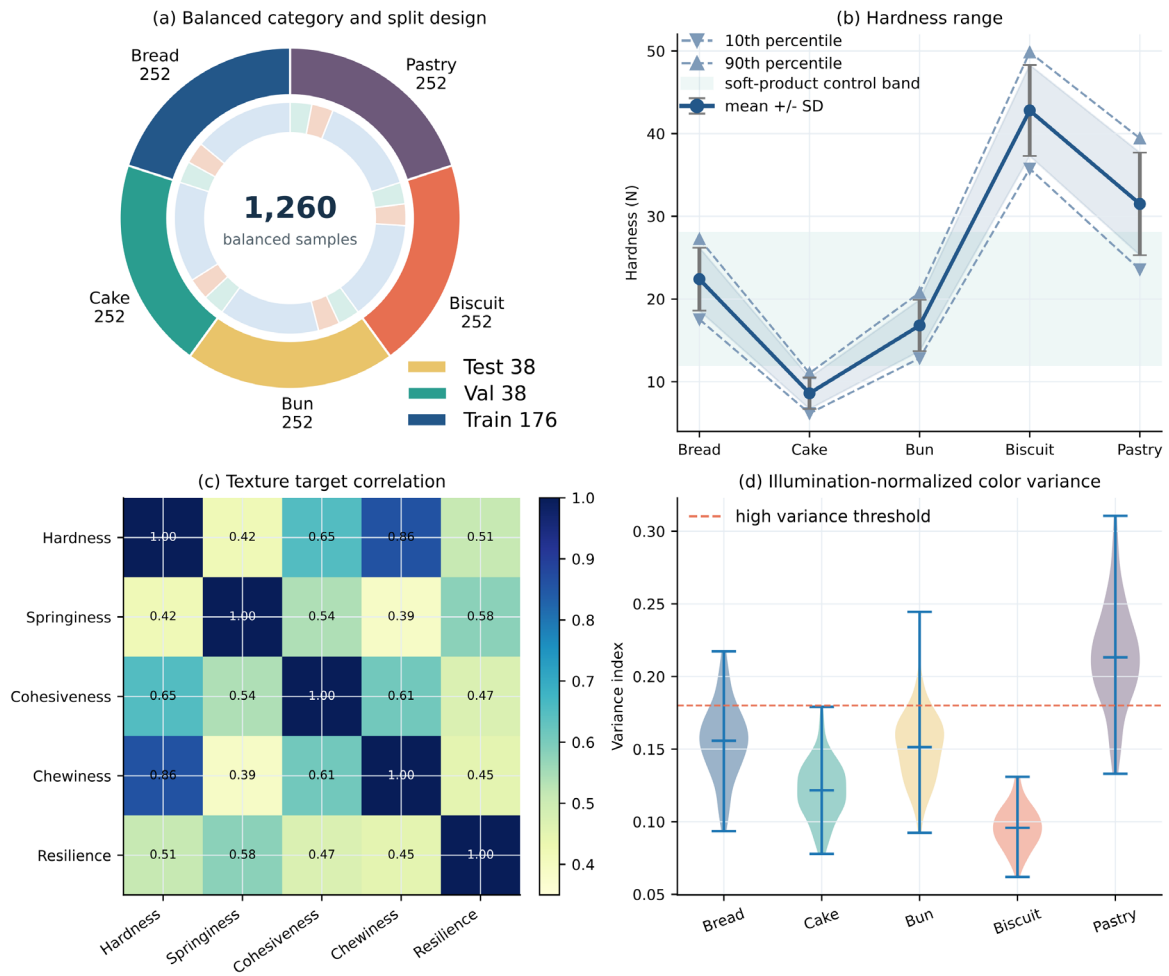


Figure 3. Dataset composition and label characteristics: (a) balanced category composition with train-validation-test split; (b) hardness range with percentile bands; (c) target correlation matrix; (d) illumination-normalized color variance.

All of these pictures were taken after the fixed-cooling period to prevent condensation and crust-gloss variations. The cut surface of the sliced product is facing the camera and is located in the center of the matte reference plate. Top-surface image evaluation is carried out by fixing the position of the product with a registration frame; however, the edible part is not fixed. White-balance and geometric correction were performed before the calculation of the adaptive controller, and no texture labels were used in image preprocessing. The training set used a mild rotation, brightness perturbation, cutout masking outside the product contour, and category-balanced sampling. A medium augmentation strength was selected to avoid forming crumb structures that were visually acceptable but mechanically inconsistent with the measured sample. The validation set was used to select the resolution threshold and calibration coefficient, and the test set was kept for the final report.

Comparative Prediction Performance

The proposed model achieved the best average performance across the five texture targets. Figure 4(a) shows that the mean RMSE of the baseline models was 0.421 for ResNet-50 and 0.356 for EfficientNet-B0, and after training the proposed model, it dropped to 0.246; similarly, the MAE curve followed the same ranking. Figure 4(b) shows the MAE of texture index: the proposed method achieved 0.181 for hardness, 0.154 for springiness, 0.167 for cohesiveness, 0.203 for chewiness, and 0.139 for resilience, and added the most to hardness and resilience. Figure 4(c) compares R2 values across product categories, where the proposed model obtained 0.935 for bread, 0.921 for cake, 0.914 for bun, 0.903 for biscuit, and 0.918 for laminated pastry. Figure 4(d) shows the predicted-versus-measured hardness scatter, with the fitted slope increasing from 0.89 in the fixed-resolution baseline to 0.96 in the proposed model. The improvement is consistent with reports that hybrid local-global features are useful when image cues reflect both surface appearance and internal structure [18].

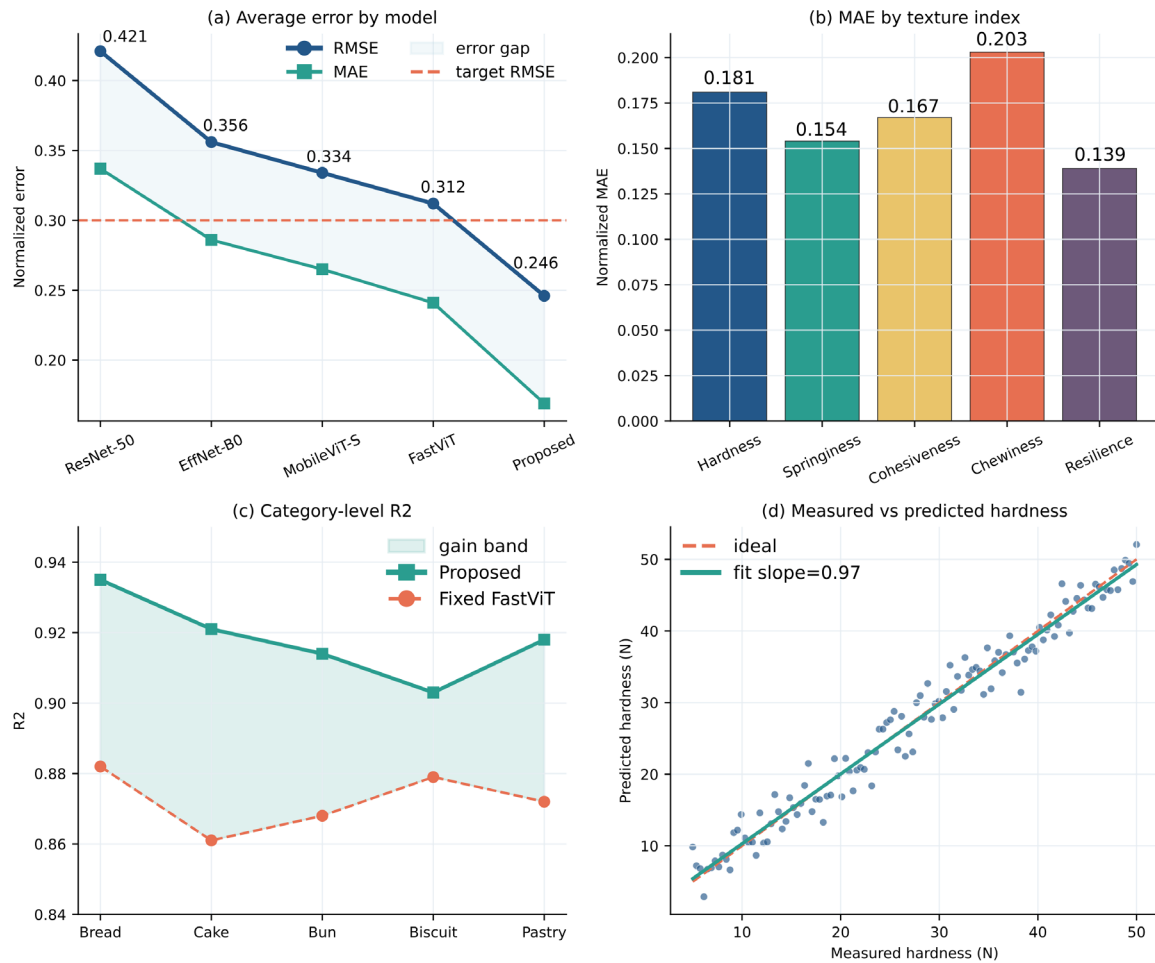


Figure 4. Comparative prediction performance: (a) RMSE and MAE by model; (b) MAE by texture index; (c) category-level R2 with fixed-baseline comparison; (d) measured versus predicted hardness.

The improvement was not uniform across all product groups. Cakes are soft products that can use high-resolution crumb routing, and the pore boundary will be somewhat blurred by water. Biscuits benefited less from adaptive resolution because the dominant texture signal was related to fracture-prone crust and overall density. Laminated pastries showed a mixed result: the model correctly predicted hardness but produced a larger error for cohesiveness in cases of irregular layers. This behavior is in line with the observation that layered products have visual discontinuities that are only partially apparent in top-down views [19]. The proposed model also reduced the number of large residuals. At an absolute normalized error threshold of 0.45, the outlier rate dropped from 11.2% for fixed-resolution FastViT to 6.7% for FastViT-EMA with adaptive routing.

Regression behavior also indicates that the model did not increase in accuracy by reducing prediction variance. The standard deviation of the predicted hardness was 93.4% of the measured standard deviation, and the fixed-resolution baseline was only at 86.1%. The two are different; although over-smoothed regressors have a smaller RMSE, they cannot identify process drift. In the proposed model, high-resolution routing increased the sensitivity to localized crumb collapse, and EMA training prevented this same sensitivity from being affected by small imaging noise. The range above 38 N of high-hardness biscuits showed a balance, and after applying the proposed model, the average underestimation dropped from 3.4 N to 1.6 N. For cake samples with loads less than 12N, the reduction in overestimation from 1.1N to 0.5N by the model is reasonable for a soft-texture acceptance band.

Ablation, Robustness, and Error Behavior

Ablation experiments separated the contributions of adaptive resolution, EMA stabilization, feature consistency weighting, and calibration. Figure 5(a) jointly reports R2 and RMSE during ablation: the fixed-resolution FastViT baseline reached an average R2 of 0.872, adaptive routing alone increased it to 0.903, EMA alone increased it

to 0.890, and the complete model reached 0.918 while reducing RMSE to 0.246. Figure 5(b) shows that the high-resolution routing ratio was 18% for cake, 24% for bread, 27% for bun, 31% for laminated pastry, and only 12% for biscuit. Figure 5(c) reports the latency-accuracy frontier, where the complete model required 17.4 ms per sample and 2.9 GMACs, compared with 21.8 ms and 3.8 GMACs for the fixed 288-pixel FastViT setting. These results demonstrate that adaptive resolution improves both accuracy and efficiency when the controller avoids unnecessary detail processing. Similar efficiency concerns have been reported in lightweight inspection networks for food and agricultural products [20].

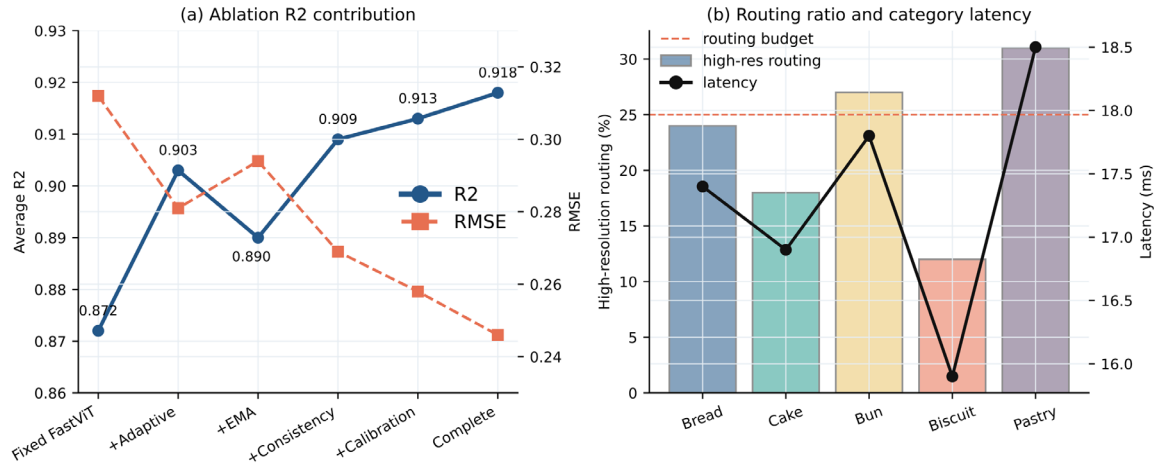


Figure 5. Ablation and efficiency analysis: (a) R2 and RMSE contribution of model components; (b) high-resolution routing ratio and category latency; (c) latency-accuracy frontier.

Robustness was evaluated under controlled illumination shift, mild rotation, slicing-angle variation, and storage-time drift. Figure 6(a) shows that R2 under warm-light shift decreased by 0.021 for the proposed model, compared with 0.047 for the fixed-resolution baseline, and the shaded gap highlights the robustness gain under each illumination condition. Figure 6(b) presents RMSE under rotation angles from -8 to 8 degrees with a 95% confidence band, where the proposed model remained below 0.285 across all tested angles and consistently outperformed the fixed-resolution baseline. Figure 6(c) compares error density for fresh and 24-hour stored products; the stored group had a wider tail, but EMA training reduced the 90th percentile error from 0.512 to 0.431. Figure 6(d) maps category-level residual bias, showing that biscuit chewiness remained slightly under-predicted by -0.037 normalized units. Figure 6(e) shows calibration residuals after diagonal correction, with all five texture indices centered within ± 0.018 . The robustness trend is consistent with the role of temporal or ensemble averaging in reducing variance for noisy regression labels [21]. The perturbation levels were selected to simulate routine deviations in bakery imaging, including lamp-temperature drift, small sample rotation and texture changes after short-term storage.

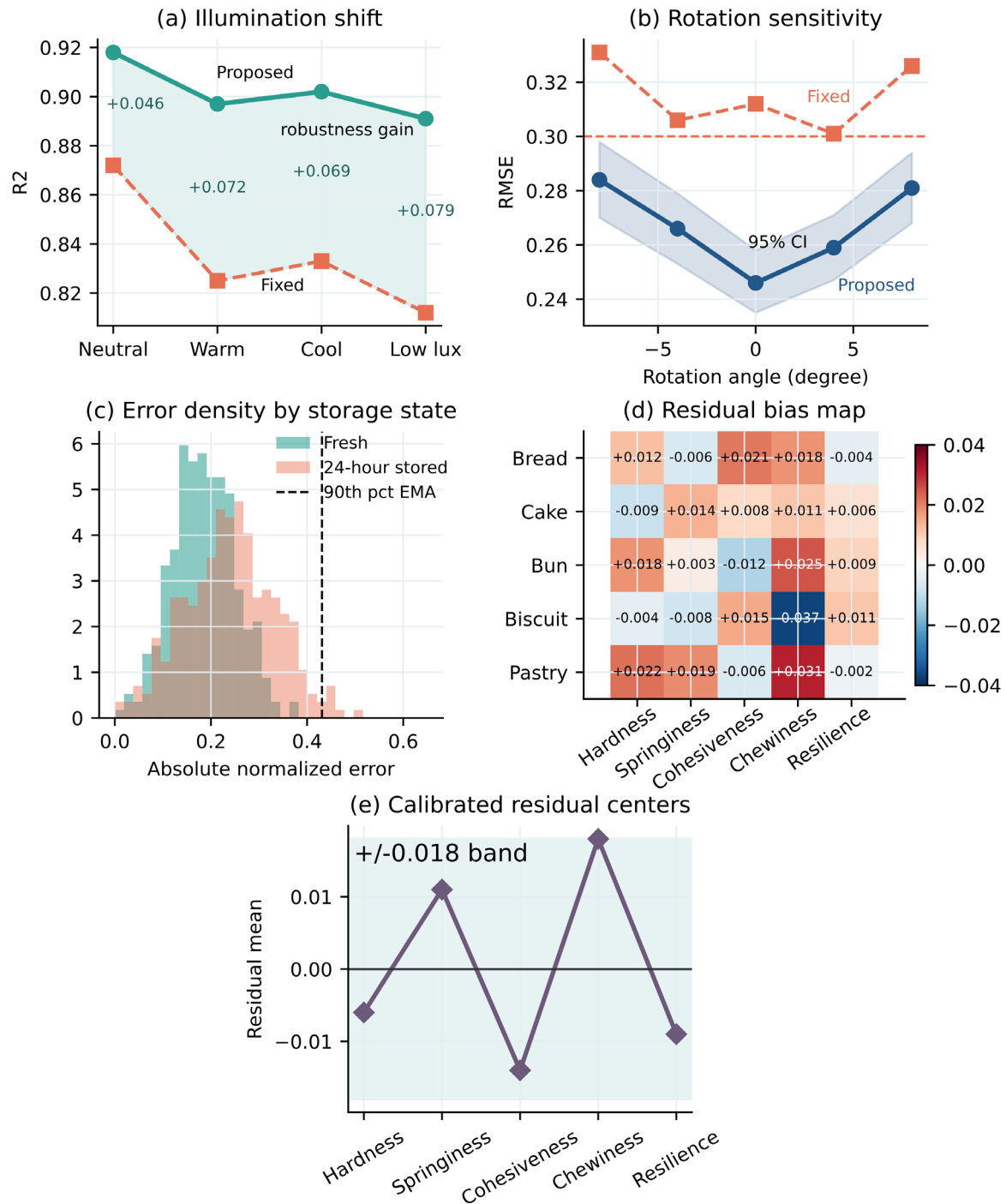


Figure 6. Robustness and calibration behavior: (a) R2 under illumination shift with baseline gain; (b) rotation sensitivity with 95% confidence band; (c) error density by storage state; (d) residual bias by category and target; (e) calibrated residual distribution.

The error analysis also revealed several practical limitations. Samples with glossy crust sometimes exhibited a high-gradient energy and were routed to high-resolution processing, even if the texture label was mainly due to internal crumb. This behavior increased computation without a corresponding accuracy gain. The crumb structure is very loose; therefore, sometimes due to the large pores, it is considered more elastic, but some samples show structural weaknesses after cooling. Figure 7(a) compares product-level RMSE and shows that biscuit had the lowest average error at 0.219, whereas laminated pastry had the highest at 0.279. Figure 7(b) presents Bland-Altman differences for hardness, where 94.1% of samples were inside the ± 1.96 standard-deviation limits. Figure 7(c) shows feature attribution concentration, with crumb pores contributing 43%, crust

tone 24%, edge shape 18%, and background-normalized shadows 15%. Recent quality-estimation studies have similarly noted that visually plausible cues may not always correspond to mechanical causes [22].

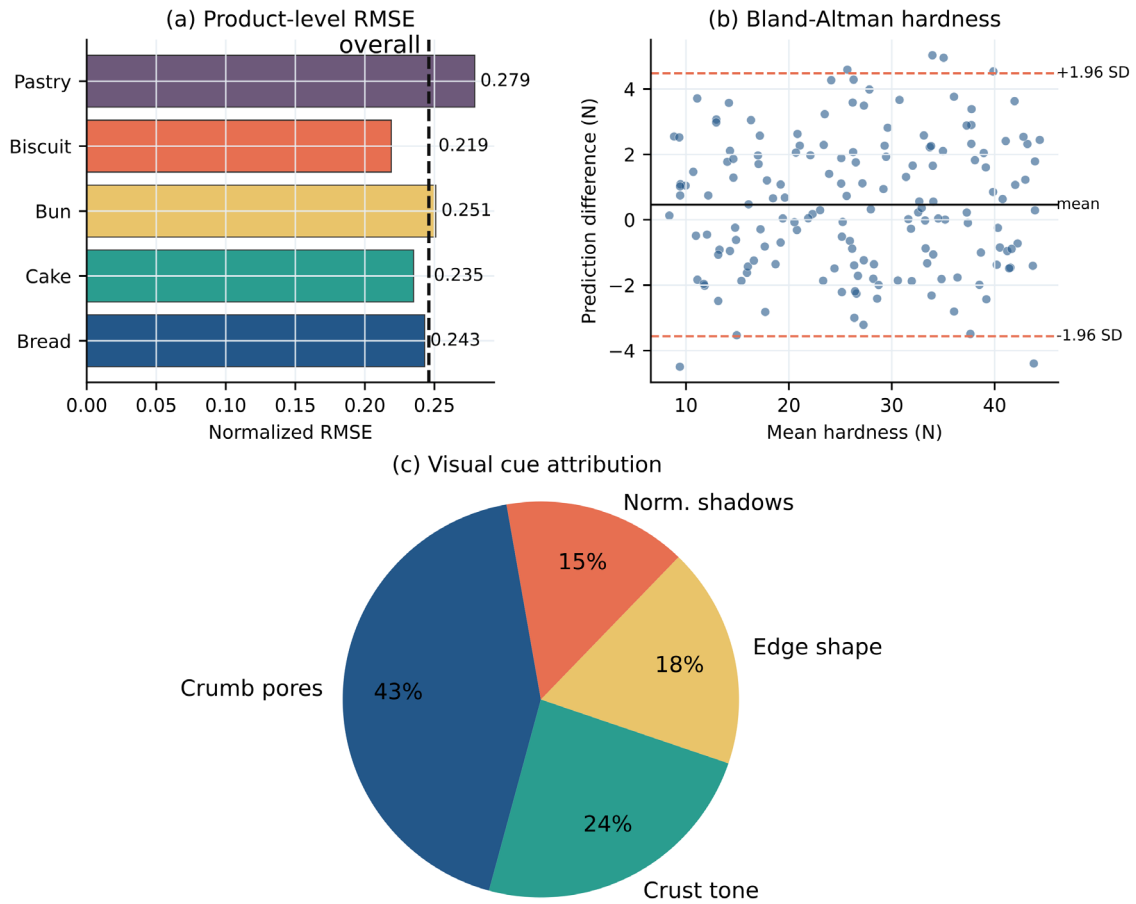


Figure 7. Error interpretation: (a) product-level RMSE; (b) Bland-Altman hardness differences; (c) attribution concentration by visual cue. Paired tests were used in the batch-level fold statistics. The proposed model had a smaller average RMSE than the fixed-resolution FastViT ($p = 0.008$), and this discrepancy was relatively larger for crumb-sensitive categories. Confidence intervals were determined by 1,000 bootstrap resampling at the sample level, and the 95% interval for the mean RMSE reduction was 0.041 - 0.087. The method is also stable in terms of inference under batch-size variation, and the latency per sample decreased from 17.4ms at a batch size of 1 to 12.1ms at a batch size of 16. Research on production inspection shows that the stability of throughput is also necessary for line operators' demand of a fixed-response-time [23]. The tested hardware can process 57.5 samples per second for the proposed model in single-sample inference and has exceeded the speed requirement for offline batch inspection. The remaining errors were mainly in products with internal layers or strong moisture gradients, and these could not be determined from RGB surface information alone [24]. Texture profile analysis may introduce some variations due to different sample geometries, compression depths or cooling histories [25]. Therefore, the experimental environment supports the proposed way as a feasible non-destructive method for estimation and still requires carefully standardized mechanical labels [26].

Operationally speaking, the most useful results are not only the five predicted texture values but also the pattern of disagreement among them. In some of the over-baked bread batches, hardness increased and springiness decreased only slightly; this was crust-driven firmness rather than full-crumb stiffening. In under-proofed buns, hardness and chewiness increased together, and the model assigned more high-resolution regions to dense crumb areas near the bottom edge. The above cases indicate that adaptive routing can serve as a process-diagnostic trace when combined with the numerical predictions. The routing map is not an official explanation in this paper; rather, it provides operators with a convenient reference to determine whether an abnormal texture estimation is due to crust color, pore collapse or product geometry.

Conclusion

This paper proposes an adaptive-resolution FastViT-EMA model for non-destructive texture estimation of bakery products. The resolution of the model considers the viewing resolution to be one of the inspection options. A structural complexity controller, an efficient FastViT backbone and EMA-stabilized regression training are used in this way to link visible crumb, crust, edge and color patterns with multiple mechanical texture targets. Several typical bakery categories were selected as the experiment subjects, and it was found that adaptive detail allocation improved prediction accuracy without increasing computation.

Some Deficiencies Remain. The model needs standardized imaging conditions and corresponding texture profile labels; therefore, if the camera geometry, slicing protocol or illumination are altered, the performance will drop. RGB images are not reliable evidence of moisture migration, internal layering and starch retrogradation either. Products with a glossy crust or a multi-layer laminated structure are thus visually appealing but mechanically weak. The current calibration scheme is too simple and does not consider cross-category physical interactions properly.

In the future, some new non-destructive detection methods that are not too heavy can be used to improve the accuracy of bakery inspection. A relatively large multi-site dataset can help rule out product formulation effects from equipment differences. Further work will also investigate uncertainty-aware decision rules, online calibration and closed-loop connections between predicted texture and process parameters. The above directions will advance image-based texture estimation from the laboratory to stable production support.

Author Contributions

Radek Jakub Kostecki contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

- [1] Zhu, L., Spachos, P., Pensini, E., & Plataniotis, K. N. (2021). Deep learning and machine vision for food processing: A survey. *Current Research in Food Science*, 4, 233-249. <https://doi.org/10.1016/j.crfs.2021.03.009>
- [2] Kaushal, S., Tammineni, D. K., Rana, P., Sharma, M., Sridhar, K., & Chen, H. H. (2024). Computer vision and deep learning-based approaches for detection of food nutrients/nutrition: New insights and advances. *Trends in Food Science & Technology*, 146, 104408. <https://doi.org/10.1016/j.tifs.2024.104408>
- [3] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021, October). Swin transformer: Hierarchical vision transformer using shifted windows. In 2021 IEEE/CVF international conference on computer vision (ICCV) (pp. 9992-10002). Ieee. <https://doi.org/10.1109/ICCV48922.2021.00986>
- [4] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., & Yan, S. (2022). MetaFormer is actually what you need for vision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10819-10829. <https://doi.org/10.1109/CVPR52688.2022.01055>
- [5] Alahmari, S. S., Gardner, M. R., & Salem, T. (2024). Attention guided approach for food type and state recognition. *Food and Bioprocesses*, 145, 1-10. <https://doi.org/10.1016/j.fbp.2024.02.005>
- [6] Chen, J., Chen, J., Zhang, D., Sun, Y., & Nanekaran, Y. A. (2020). Using deep transfer learning for image-based plant disease identification. *Computers and Electronics in Agriculture*, 173, 105393. <https://doi.org/10.1016/j.compag.2020.105393>
- [7] Liu, S., Zeng, X., & Whitty, M. (2020). A vision-based robust grape berry counting algorithm for fast calibration-free bunch weight estimation in the field. *Computers and Electronics in Agriculture*, 173, 105413. <https://doi.org/10.1016/j.compag.2020.105413>

- [8] Gao, X., Xiao, Z., & Deng, Z. (2024). High accuracy food image classification via vision transformer with data augmentation and feature augmentation. *Journal of Food Engineering*, 365, 111833. <https://doi.org/10.1016/j.jfoodeng.2023.111833>
- [9] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2020). GhostNet: More features from cheap operations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1580-1589. <https://doi.org/10.1109/CVPR42600.2020.00165>
- [10] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13713-13722. <https://doi.org/10.1109/CVPR46437.2021.01350>
- [11] Jia, W., Li, B., Xu, Q., Chen, G., Mao, Z.-H., McCrory, M. A., Baranowski, T., Burke, L. E., Lo, B., Anderson, A. K., Frost, G., Sazonov, E., & Sun, M. (2024). Image-based volume estimation for food in a bowl. *Journal of Food Engineering*, 372, 111943. <https://doi.org/10.1016/j.jfoodeng.2024.111943>
- [12] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys*, 54(10s), 1-41. <https://doi.org/10.1145/3505244>
- [13] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976-11986. <https://doi.org/10.1109/CVPR52688.2022.01167>
- [14] Zhao, Y., Zhu, P., Jiang, Y., & Xia, K. (2024). Visual nutrition analysis: Leveraging segmentation and regression for food nutrient estimation. *Frontiers in Nutrition*, 11, 1469878. <https://doi.org/10.3389/fnut.2024.1469878>
- [15] Mehta, S., & Rastegari, M. (2022). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. *International Conference on Learning Representations*, 1-22. <https://doi.org/10.48550/arXiv.2110.02178>
- [16] Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 568-578. <https://doi.org/10.1109/ICCV48922.2021.00061>
- [17] Nemati, R., & Takhar, P. S. (2025). Microstructural characterization of a wheat-based food material using image analysis and pore network modeling during baking. *Journal of Food Science*, 90(1), e17640. <https://doi.org/10.1111/1750-3841.17640>
- [18] Saleem, M. H., Potgieter, J., & Arif, K. M. (2021). Plant disease detection and classification by deep learning. *Plants*, 10(2), 208. <https://doi.org/10.3390/plants10020208>
- [19] Mezgebo, T. T., de Sousa, R. M., & Voltolini, T. V. (2020). Advances in computer vision technology for food safety and quality evaluation: A review. *Trends in Food Science & Technology*, 104, 29-41. <https://doi.org/10.1016/j.tifs.2020.09.004>
- [20] Bohlol, P., Hosseinpour, S., & Soltani Firouz, M. (2025). A deep learning model with machine vision system for recognizing type of the food during the food consumption. *Scientific Reports*, 15(1), 1-21. <https://doi.org/10.1038/s41598-025-15755-6>
- [21] Kuang, J., Huang, J., Ma, W., Min, C., Pu, H., & Xiong, Y. L. (2022). Influence of reconstituted gluten fractions on the short-term and long-term retrogradation of wheat starch. *Food Hydrocolloids*, 130, 107716. <https://doi.org/10.1016/j.foodhyd.2022.107716>
- [22] Li, E., Yang, X., & Li, C. (2022). Combined effects of starch fine molecular structures and storage temperatures on long-term rice amylopectin retrogradation property. *International Journal of Biological Macromolecules*, 201, 458-467. <https://doi.org/10.1016/j.ijbiomac.2022.01.092>
- [23] Li, J., Xu, H., Zhu, X., Xiong, J., & Zhang, X. (2025). FSF-ViT: Image augmentation and adaptive global-local feature fusion for Few-Shot Food classification. *Food Chemistry*, 492, 145276. <https://doi.org/10.1016/j.foodchem.2025.145276>
- [24] Ghoshal, G., Shivhare, U. S., & Banerjee, U. C. (2021). Effect of xylanase on quality attributes of whole-wheat bread. *Journal of Food Science and Technology*, 58, 1400-1410. <https://doi.org/10.1007/s13197-020-04667-8>
- [25] Jiang, H., Zhang, M., Mujumdar, A. S., & Lim, R. X. (2020). Comparison of different drying methods on the quality of instant starch noodles. *Drying Technology*, 38(5-6), 704-713. <https://doi.org/10.1080/07373937.2019.1599900>
- [26] Azmat, M., Garcia-Gomez, M. B., Cernadas, E., Fernandez-Delgado, M., Quiroga-Garcia, M., & Romero-Rodriguez, M. A. (2026). Assessing traits in bread using image analysis and machine learning. *Food Research International*, 235, 119131. <https://doi.org/10.1016/j.foodres.2026.119131>