Composition-Aware Fine-Grained Food Recognition for Dietary Analysis

Linghui Ye , Qingbing Sang , Zhiyong Xiao *

Posted Date: 30 January 2026

doi: 10.20944/preprints202601.2382.v1

Keywords: food recognition; fine-grained classification; vision transformer; key-region awareness; dietary analysis; feature fusion

 doi:10.20944/preprints202601.2382.v1

Article

Composition-Aware Fine-Grained Food Recognition for Dietary Analysis

Linghui Ye, Qingbing Sang and Zhiyong Xiao *

School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China
* Correspondence: zhiyong.xiao@jiangnan.edu.cn

Abstract

Reliable visual characterization of food composition is a fundamental prerequisite for image-based dietary assessment and health-oriented food analysis. In fine-grained food recognition, models often suffer from large intra-class variation and small inter-class differences, where visually similar dishes exhibit subtle yet discriminative differences in ingredient compositions, spatial distribution, and structural organization, which are closely associated with different nutritional characteristics and health relevance. Capturing such composition-related visual structures in a non-invasive manner remains challenging. In this work, we propose a fine-grained food classification framework that enhances spatial relation modeling and key-region awareness to improve discriminative feature representation. The proposed approach strengthens sensitivity to composition-related visual cues while effectively suppressing background interference. A lightweight multi-branch fusion strategy is further introduced for stable integration of heterogeneous features. Moreover, to support reliable classification under large intra-class variation, a token-aware subcenter-based classification head is designed. The proposed framework is evaluated on the public FoodX-251 and UEC Food-256 datasets, achieving accuracies of 82.28% and 82.64%, respectively. Beyond benchmark performance, the framework is designed to support practical image-based dietary analysis under real-world dining conditions, where variations in appearance, viewpoint, and background are common. By enabling stable recognition of the same food category across diverse acquisition conditions and accurate discrimination among visually similar dishes with different ingredient compositions, the proposed approach provides reliable food characterization for dietary interpretation, thereby supporting practical dietary monitoring and health-oriented food analysis applications.

Keywords: food recognition; fine-grained classification; vision transformer; key-region awareness; dietary analysis; feature fusion

1. Introduction

With the accelerating pace of modern life, dietary patterns have become increasingly diverse, accompanied by substantial visual variability in food presentation across regions, cooking styles, and consumption contexts. Such variability poses significant challenges for fine-grained food recognition, especially when visually similar food categories exhibit subtle yet discriminative differences in ingredient dominance and spatial organization, which in turn affect the reliability of image-based dietary assessment and food-related health analysis.

An ideal dietary analysis approach should be efficient, convenient, and accurate. However, traditional manual recording approaches are often tedious, error-prone, and highly dependent on users' subjective judgment, making them difficult to apply in daily life and limiting their scalability in practical food monitoring scenarios. Consequently, recent studies have incorporated computer vision and deep learning techniques into dietary image analysis, enabling tasks such as automatic food category recognition, in some cases, calorie-related analysis based on photographs captured in real-world dining environments [1]. These image-based approaches can lower the usage threshold

 doi:10.20944/preprints202601.2382.v1

[2] and improve the efficiency of dietary data acquisition, motivating further research on reliable fine-grained food recognition in health-related contexts. Against this background, fine-grained food recognition technology has gradually become a research hotspot in the field of intelligent nutrition analysis [3]. When dealing with diverse and fine-grained food image data, improving a model's ability to recognize visually similar food items has become an urgent problem to address [4].

Compared with traditional fine-grained tasks such as animals or vehicles, food images pose greater challenges due to significant intra-class variations and inter-class visual overlaps caused by ingredient composition, cooking processes, and presentation styles. Specifically, food images often exhibit intra-class diversity [5], where the same dish may appear with noticeable differences in color, shape, and background due to variations in region, cooking style, or shooting angle [6]. Meanwhile, there also exists inter-class similarity, as certain stir-fried dishes or desserts made from similar ingredients or arranged in similar plating styles present highly comparable visual appearances, making category distinction much more difficult [7], and resulting in dispersed intra-class distributions and small inter-class distances in the feature space.

A number of large-scale datasets dedicated to food recognition have been successively introduced [8], such as Food-101 [9], UEC Food-256, FoodX-251 [10]. The release of these datasets has not only accelerated the research on fine-grained food classification models but also revealed the complexity and challenges of this task in real-world scenarios, where complex backgrounds, diverse plating patterns, and fine-grained visual ambiguity are commonly observed [11]. In professional fields such as medical nutrition and health monitoring, accurate food recognition serves as an important prerequisite for subsequent dietary analysis and decision support. Therefore, developing models capable of accurately identifying and distinguishing fine-grained food categories is not merely a technical breakthrough, but also a key to improving the applicability, reliability, and trustworthiness of image-based dietary analysis systems [12].

Although mainstream models such as Swin-DR have demonstrated promising recognition capabilities in fine-grained food image classification, their fundamental architectures—based on local window modeling and multi-scale feature fusion—only partially improve feature representation [13], while significant limitations remain in spatial perception and key region modeling [14].

On the one hand, Swin Transformer has inherently limited spatial modeling capacity, making it difficult to effectively capture the complex variations in structure and arrangement within the same food category. For example, while both are "sushi" rolled sushi usually appears in a compact strip arrangement, whereas scattered sushi distributes rice and toppings freely across the plate. Such global layout differences are difficult to model adequately through local attention alone, thereby affecting classification stability and accuracy [15]. On the other hand, the model's positional awareness and attention mechanisms are relatively weak [16], often leading to overfitting on background areas or irrelevant edges while lacking precise localization and emphasis on the main object. Particularly when dealing with mixed dishes such as boiled fish with chili or spicy hot pot, recognition becomes challenging to focus accurately on the key food regions. As a result, existing models struggle to jointly model global–local spatial relations, fine-grained variations intra-class, and background suppression [17]. As can be seen from Figure 1 Food categories usually exhibit small inter-class differences, large intra-class variations, and high diversity in ingredients and plating forms. Therefore, models must not only finely model local structures but also maintain sensitivity to overall layout and positional relationships to break through existing performance bottlenecks.

Food images often require the joint consideration of heterogeneous visual cues, such as local texture patterns, global spatial layouts, and salient ingredient regions; however, existing modeling strategies tend to integrate these cues in a coarse manner, which can result in redundant representations, noise amplification, and inefficient use of discriminative information. Meanwhile, existing recognition pipelines often rely on overly simplified decision mechanisms, where global feature aggregation tends to discard important spatial distribution cues, and single-prototype representations are insufficient to accommodate the pronounced intra-class variability commonly observed within the same food

 doi:10.20944/preprints202601.2382.v1

category. These issues collectively constrain further performance improvement in fine-grained food recognition. The hybrid model integrates Convolutional Neural Networks with the Swin Transformer to simultaneously capture intricate local textures and global spatial dependencies. By embedding specialized convolutional blocks directly into the Transformer architecture, the framework significantly enhances the recognition of visually similar food items for precision dietary analysis [18].

Figure 1. Representative samples from the FoodX-251 and UEC-Food256 datasets, showing significant intra-class variations and complex background conditions in fine-grained food recognition.

To address these challenges, we propose a unified fine-grained food recognition framework that emphasizes spatial structure modeling and composition-aware feature representation, aiming to improve the robustness and reliability of food recognition in complex real-world scenarios. From a food analysis perspective, explicitly modeling spatial structure and ingredient-dominated regions enables more stable characterization of composition patterns within a dish, which is essential to distinguish visually similar foods with different underlying ingredient arrangements. Enhancing positional awareness of key food regions allows the model to focus on composition-related visual cues, such as ingredient concentration areas, while effectively suppressing background elements unrelated to food structure. Existing feature integration strategies often struggle to effectively leverage heterogeneous visual cues, leading to redundant representations, noise amplification, or inefficient information utilization [19], particularly in complex fine-grained food recognition scenarios [20]. Accordingly, an adaptive feature integration strategy is required to preserve complementary information while suppressing redundancy and noise in fine-grained food recognition. Previous studies indicate that adaptive normalization and gating mechanisms can help balance heterogeneous feature contributions and improve robustness in complex visual recognition tasks [21]. Low-rank interaction and attention-based enhancement have also been shown to improve fine-grained discrimination while maintaining computational efficiency [22,23]. At the decision stage, conventional classification designs often suffer from spatial information loss and limited ability to model intra-class diversity, which constrains performance in fine-grained food recognition. Preserving spatial distribution cues during pooling has been shown to improve recognition of food categories with subtle structural and compositional differences [24]. Such spatially aware pooling strategies help maintain sensitivity to ingredient distribution and local structure in visually similar food categories [25].

Experimental results demonstrate the effectiveness of the proposed framework for fine-grained food image recognition under complex visual conditions. The proposed approach achieves accuracies

 doi:10.20944/preprints202601.2382.v1

of 82.64% on Food-256 and 82.28% on FoodX-251, demonstrating strong discrimination among visually similar food categories, especially dishes sharing similar ingredients or culinary styles.

In the context of globally diversified food culture, a single food category often exhibits substantial visual variation across regions and preparation styles, resulting in a widespread "same-category, different appearance" phenomenon. Failure to capture variations in overall layout and local ingredient arrangements may lead to representation confusion, while insufficient positional awareness often causes attention to drift toward irrelevant regions (e.g., plates or tablecloths), weakening focus on key ingredients.

To address these challenges, we introduce a unified framework that jointly considers spatial structure and key food region awareness. By integrating global layout cues with local ingredient arrangements, the framework improves sensitivity to complex spatial structures and subtle appearance variations in food images. With enhanced positional awareness, the framework focuses on ingredient-dominant regions while suppressing background interference, thereby improving classification accuracy and robustness in food recognition. This approach provides reliable technical support for image-based dietary composition analysis and health-oriented food assessment pipelines.

The main contributions of this work are summarized as follows:

1. To handle complex spatial structures and subtle appearance variations caused by ingredient composition and preparation styles, we design AGRA (Adaptive Grouped Residual Attention) to jointly capture global layout and local ingredient arrangements in food images.
2. We introduce CEAG (Coordinate-Enhanced Adaptive Gating) to improve the model's ability to localize key regions while suppressing background distractions such as plates and tablecloths, thus enhancing classification accuracy and robustness.
3. We propose SGLR-Mixer, a soft-gated low-rank fusion strategy that adaptively integrates heterogeneous visual cues in food images while avoiding redundant representations and excessive computational overhead.
4. We design TAPSubCenterArcHead, which preserves spatial distribution cues during pooling and improves intra-class diversity modeling, enabling more reliable discrimination among visually confusing food categories.
5. We propose Swin-ACST, a fine-grained food classification framework that integrates spatial relationship modeling with key-region awareness, effectively enhancing feature discrimination and overall classification performance in complex food image analysis scenarios.

2. Related Works

2.1. Vision Transformers in Image Classification

In recent years, Vision Transformers (ViT) have demonstrated remarkable modeling capabilities in image classification, emerging as a mainstream architecture following convolutional neural networks (CNNs). ViT was the first to directly apply the Transformer architecture to sequences of image patches [26,27], leveraging a global self-attention mechanism to capture long-range dependencies and significantly improve classification performance. Subsequently, DeiT introduced efficient training strategies [28], enabling Vision Transformers to achieve competitive results even on medium-scale datasets.

To further reduce computational cost and enhance spatial modeling capability [29], Swin Transformer proposed a hierarchical architecture with a shifted-window mechanism [13], balancing global semantic modeling and local detail perception. It has achieved outstanding results in image recognition and object detection tasks. Despite these advancements, Vision Transformers still face challenges in spatial structure understanding [30] and fine-grained modeling [31]. For instance, window partitioning restricts the acquisition of full spatial context, and the attention mechanism often exhibits weak sensitivity to spatial distribution, making it difficult to precisely characterize subtle variations in key regions, an issue that becomes particularly evident in complex backgrounds or fine-grained classification tasks.

 doi:10.20944/preprints202601.2382.v1

2.2. Spatial and Attention Mechanisms

Spatial and attention mechanisms play a crucial role in enhancing a visual model's ability to capture key regions, enabling the network to selectively focus on informative parts of an image and thereby improve the effectiveness of feature representation. Early works such as SENet enhanced representational capacity through channel attention [32], while CBAM further integrated spatial and channel attention to strengthen the model's focus on salient regions [33].

In recent years, increasing research efforts have been devoted to the design of efficient spatial attention mechanisms, such as GENet, GALA, and Coordinate Attention. By incorporating spatial coordinate information [16], Coordinate Attention organically integrates positional encoding with channel attention [34], effectively improving the model's spatial localization capability. Existing spatial and attention mechanisms provide diverse strategies for enhancing the model's ability to identify key regions; however, achieving accurate modeling of both global and local spatial information while maintaining computational efficiency remains an open challenge.

Building upon these insights, this work designs a more efficient spatial perception module that strengthens the model's localization ability for key regions, enabling it to effectively focus on target food areas while suppressing background interference, thereby improving fine-grained regional discrimination in food image recognition.

2.3. Global Context Modeling Techniques

Global context modeling is essential for understanding complex visual scenes and plays a critical role in fine-grained image classification. Traditional convolutional networks, limited by their local receptive fields and inefficient long-range dependency capture, struggle to model global relationships effectively. To overcome these limitations, a variety of self-attention and non-local mechanisms, such as NLNet and GC-Net, have been proposed to strengthen the integration of global information within feature representations.

More recently, approaches like Grouped Residual Self-Attention have improved the efficiency of global–local feature interaction through grouped and residual structures [35], enhancing global spatial contextual sensitivity while balancing modeling capacity and computational cost. Global context modeling techniques not only enrich the semantic representation of inter-class differences but also improve the model's ability to understand complex spatial dependencies.

In this work, we further optimize the perception of global context, enhancing the model's capability to capture subtle differences and improving classification performance in fine-grained food image recognition.

2.4. Image-Based Food Recognition for Dietary and Health-Oriented Analysis

Image-based food recognition has been widely studied as a fundamental component of dietary assessment and health-oriented food analysis [36], offering a non-invasive alternative to traditional manual dietary recording [37]. By enabling automatic identification of food categories from images captured in daily dining scenarios, such approaches provide essential visual information for subsequent dietary interpretation and health-related analysis.

Compared with generic object recognition tasks, fine-grained food classification presents unique challenges due to large intra-class variation and small inter-class differences. Dishes belonging to the same category may differ substantially in appearance because of variations in ingredient dominance, cooking style, and presentation, while visually similar foods prepared with comparable ingredients often exhibit subtle structural differences. These characteristics make accurate and stable classification particularly important for reliable food characterization.

Recent research has increasingly emphasized the role of spatial structure modeling and region-aware feature representation in improving classification consistency for food images [38]. By enhancing sensitivity to composition-related visual cues and reducing background interference, such methods

 doi:10.20944/preprints202601.2382.v1

aim to provide more robust food category recognition under real-world conditions, thereby supporting image-based dietary analysis and health-oriented applications.

3. Methods

3.1. Overall Architecture of the Approach

The overall architecture of the proposed model is shown in Figure 2. The model mainly consists of three parts: the first part is the backbone network, the second part is the spatial perception and modeling enhancement module, and the third part is the fine-grained food classifier.

Among them, the backbone network is composed of multiple Swin Transformer modules. The spatial perception and modeling enhancement module sequentially integrates global spatial relation modeling enhancement and coordinate-aware spatial enhancement, combined with a DRConvBlock to further improve local feature representation ability. After multi-branch enhancement, to robustly integrate heterogeneous semantic information, SGLR-Mixer is introduced to adaptively integrate heterogeneous features from multiple enhancement paths, enabling stable fusion of complementary spatial and semantic information. Then, spatial selective amplification is applied to further strengthen the response of salient regions and suppress background interference.

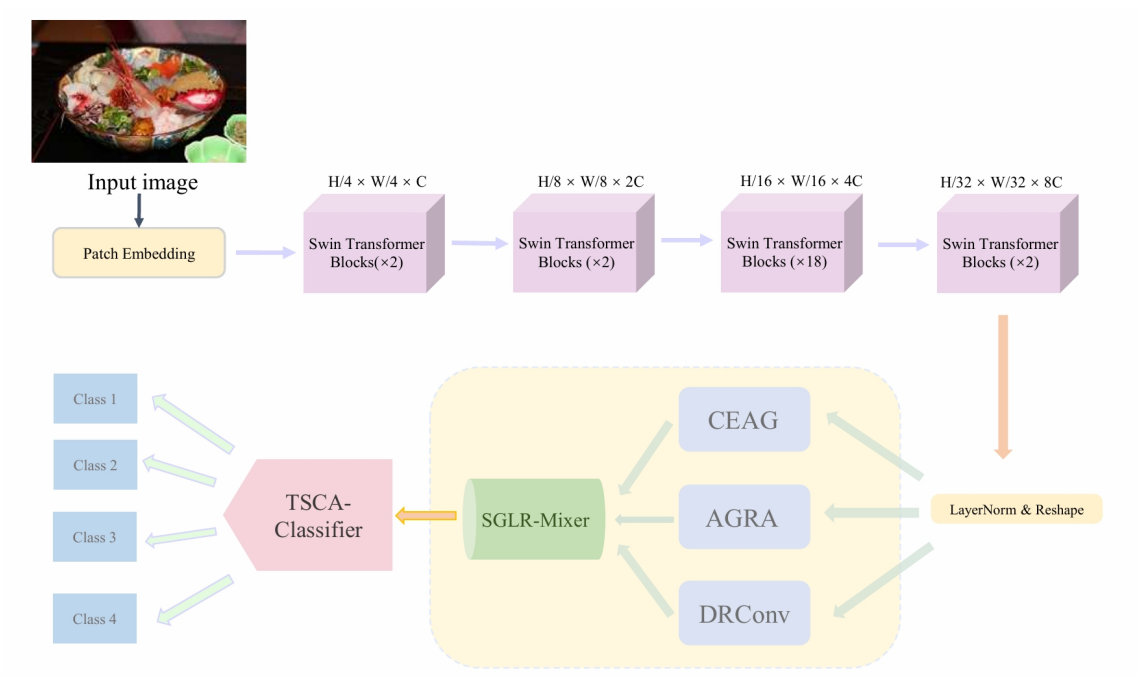

Figure 2. Overall architecture of the proposed Swin-ACST framework, illustrating the collaborative relationships among feature extraction, fusion, and classification stages.

The fine-grained food classifier no longer adopts a simple global average and single-center metric, but employs a combination of weighted aggregation preserving spatial distribution and multi-center angular metric, namely the TAPSubCenterArcHead, to preserve the spatial distribution of key food components and improve the modeling of intra-class appearance diversity.

The overall process is as follows: First, the three-channel RGB fine-grained food image is input into the backbone network, and the Swin Transformer extracts global feature representations. Then, the feature information is processed by AGRA for global and local relation modeling, which focuses on both overall structure and local arrangement to enhance spatial sensitivity and subtle-difference discrimination ability. Next, the CEAG module, based on coordinate awareness and spatial gating, improves the model's attention accuracy to target regions, effectively suppresses irrelevant background areas, and avoids background overfitting.After that, local feature enhancement is performed to improve edge and texture representations.

 doi:10.20944/preprints202601.2382.v1

For the features obtained from the above multi-branch enhancement, SGLR-Mixer performs path alignment and soft-gated low-rank fusion, highlighting salient regions and producing stable and more discriminative global representations. Finally, the enhanced features are fed into the TAPSubCenterArcHead, where weighted aggregation and multi-center angular metric are used to complete category prediction.

Overall, through multi-level spatial relation modeling, salient region localization, and robust fusion, the proposed model effectively enhances spatially aware feature representation and category discrimination for fine-grained food images, providing more stable and reliable predictions under diverse presentation conditions.

3.2. CEAG Block

Following the Swin-DR backbone, this section details the proposed CEAG Block to improve spatial awareness in fine-grained food image analysis. In food images, discriminative visual cues are often closely tied to the spatial distribution of ingredients and their relative positions within a dish, while irrelevant background elements such as plates or tablecloths may introduce significant interference. However, the window-based attention mechanism of Swin Transformer exhibits limited explicit positional sensitivity, making it difficult to consistently emphasize key food regions. The CEAG block addresses this issue by incorporating coordinate-aware attention to explicitly encode spatial location information into channel representations. This design enables the model to better align visual responses with ingredient-dominated regions while suppressing non-food background areas. By enhancing spatial localization and region-level discrimination, CEAG provides more reliable feature representations for fine-grained food recognition and downstream food composition analysis. The CEAG Block comprises two components: Coordinate Attention and Spatial Gating. Coordinate Attention encodes spatial locations and channel dependencies to provide strong coordinate guidance. It uses horizontal and vertical global average pooling to produce spatial embeddings:

$$z_c^h = \frac{1}{W}\sum_{i=1}^{W} x_c(h,i) \tag{1}$$

$$z_c^w = \frac{1}{H}\sum_{j=1}^{H} x_c(j,w) \tag{2}$$

These embeddings are concatenated and processed via 1×1 convolutions, BatchNorm2d, and h-swish to yield fused spatial features $\mathbf{f}$:

$$\mathbf{f} = \text{h-swish}\Big(\text{BN2d}\Big(\text{Conv2d}1 \times 1\Big([z_c^h, z_c^w]\Big)\Big)\Big) \tag{3}$$

$\mathbf{f}$ is then split into horizontal and vertical attention weights via separate 1×1 convolutions and sigmoid activation:

$$g^h = \sigma\Big(\text{Conv2d}1 \times 1^h(\mathbf{f})\Big), \quad g^w = \sigma(\text{Conv2d}1 \times 1^w(\mathbf{f})) \tag{4}$$

Input features $\mathbf{X}$ are recalibrated in space and across channels to produce $\mathbf{X}'$:

$$\mathbf{X}' = \mathbf{X} \odot g^h \odot g^w \tag{5}$$

where $\odot$ is element-wise multiplication. This encodes spatial coordinates explicitly, improving localization and suppressing background. The CEAG Block uses a dual gating mechanism. The first gate uses 1×1 convolution and sigmoid to produce feature gate $\mathbf{G}$, and modulates the recalibrated features via a residual path:

$$\mathbf{G} = \sigma\big(\text{Conv2d}1 \times 1(\mathbf{X}')\big) \tag{6}$$

$$\mathbf{Y} = \mathbf{X}' + \alpha \cdot \big(\mathbf{X}' \odot \mathbf{G}\big) \tag{7}$$

 doi:10.20944/preprints202601.2382.v1

Here, α is a scaling factor (set to 0.1), and $\mathbf{G} \in \mathbb{R}^{C\times H\times W}$ is the feature gate. This gate focuses on spatially relevant regions and reduces background contributions. The second gate produces spatial gate $\mathbf{M}$ via 1×1 convolution and sigmoid, modulating the coordinate-enhanced features. Results are merged with the input via a residual path to yield output $\mathbf{O}$:

$$\mathbf{M} = \sigma(\text{Conv2d}1 \times 1(\mathbf{Y})) \tag{8}$$

$$\mathbf{O} = \mathbf{Y} + \beta \cdot (\mathbf{Y} \odot \mathbf{M}) \tag{9}$$

Here, β is a scaling factor (set to 0.1), and $\mathbf{M} \in \mathbb{R}^{1\times H\times W}$ is the spatial gate. The dual gating refines features by focusing on relevant spatial regions and suppressing background to reduce overfitting. BatchNorm2d and nonlinearities (e.g., h-swish or ReLU) are applied after each convolution for stable, expressive distributions. The design enables progressive refinement: coordinate attention provides spatial awareness, and the dual gates refine representations. By integrating coordinate attention and dual gating into a single block, CEAG improves focus on relevant regions, suppresses background, reduces overfitting, and enhances recognition on complex scenes and fine-grained targets.

3.3. AGRA Block

The Swin Transformer employs window-based local self-attention, which limits its capacity for comprehensive spatial modeling. To solve it, we design a novel Adaptive Grouped Residual Attention(AGRA) module with lightweight adaptive weights. The core innovation of AGRA lies in its ability to simultaneously capture both global structural information and local spatial arrangements, thereby significantly enhancing the model's sensitivity to fine-grained spatial details. By explicitly modeling relationships between distant and nearby spatial positions, AGRA enables the network to better distinguish subtle structural differences arising from ingredient distribution and arrangement, which is crucial for fine-grained food recognition.

Given the input feature $X \in \mathbb{R}^{B\times N\times C}$, we first split the channel dimension into two equal groups, denoted as $X_1, X_2 \in \mathbb{R}^{B\times N\times \frac{C}{2}}$. For each group, we apply a grouped residual linear transformation to generate the query, key, and value representations, while introducing learnable adaptive weights α_1, α_2 to control the contribution of the residual branch. Specifically, the transformation for each group is formulated as

$$Q_1 = X_1 + \alpha_1 \cdot \text{Linear}_Q(X_1), \quad Q_2 = X_2 + \alpha_2 \cdot \text{Linear}_Q(X_2), \tag{10}$$

$$K_1 = X_1 + \alpha_1 \cdot \text{Linear}_K(X_1), \quad K_2 = X_2 + \alpha_2 \cdot \text{Linear}_K(X_2), \tag{11}$$

$$V_1 = X_1 + \alpha_1 \cdot \text{Linear}_V(X_1), \quad V_2 = X_2 + \alpha_2 \cdot \text{Linear}_V(X_2). \tag{12}$$

The grouped Q, K, and V are then concatenated and reshaped for multi-head attention computation. We normalize Q and K and compute the cosine similarity, scaled by a learnable parameter λ, to obtain the attention map. The attention calculation is given by

$$\text{Attn} = \text{Softmax}\left(\lambda \cdot \frac{Q}{\|Q\|} \cdot \left(\frac{K}{\|K\|}\right)^{\top} + B_{\text{ES-RPB}}\right), \tag{13}$$

where $B_{\text{ES-RPB}}$ denotes the exponential-space relative position bias. To further enhance the model's ability to capture spatial relationships and suppress background noise, we adopt an exponential mapping for the relative position coordinates. For any two tokens with relative coordinates $(\Delta X, \Delta Y)$, we compute

$$\Delta\hat{X} = \text{sign}(\Delta X) \cdot (1 - \exp(-|\Delta X|)), \tag{14}$$

$$\Delta\hat{Y} = \text{sign}(\Delta Y) \cdot (1 - \exp(-|\Delta Y|)), \tag{15}$$

$$B_{\text{ES-RPB}} = \text{MLP}([\Delta\hat{X}, \Delta\hat{Y}]). \tag{16}$$

 doi:10.20944/preprints202601.2382.v1

The output of the attention is then split into two groups, and each group is projected back to the original feature space using a grouped residual linear layer with the same adaptive weights:

$$[A_1, A_2] = \text{Split}(\text{Attn@}V), \tag{17}$$

$$O_1 = X_1 + \alpha_1 \cdot \text{Linear}_{\text{proj}}(A_1), \quad O_2 = X_2 + \alpha_2 \cdot \text{Linear}_{\text{proj}}(A_2), \tag{18}$$

$$O = \text{Concat}(O_1, O_2). \tag{19}$$

This design enables adaptive control of information flow across grouped attention branches, thereby enhancing the robustness and discriminative capability of spatial attention modeling. The introduction of exponential-space relative position bias further strengthens the modeling of spatial relevance across different structural regions, facilitating more consistent discrimination of fine-grained food patterns with complex spatial arrangements. All components are fully differentiable and can be seamlessly integrated into window-based self-attention frameworks such as Swin Transformer, enabling effective spatial relation modeling for fine-grained food classification under complex presentation variations.

3.4. SGLR-Mixer

In fine-grained food classification, multi-branch feature extraction is commonly adopted to capture complementary cues related to global layout, ingredient-dominant regions, and local structural details. However, in food images with complex composition and high visual variability, simple fusion strategies such as direct concatenation or uniform weighting often introduce feature redundancy, noise accumulation, and unnecessary computational overhead, which may lead to unstable category predictions under real-world conditions. To address this issue, we propose a Soft-Gated Low-Rank Mixer (SGLR-Mixer) that performs adaptive fusion of three heterogeneous feature streams in a lightweight and structured manner. By selectively integrating spatial structure–aware features and key-region–focused representations, SGLR-Mixer enhances the consistency and reliability of the final classification output, which is critical for composition-aware food recognition serving downstream dietary analysis applications. The framework diagram of SGLR-Mixer and its role within the overall multi-branch enhancement pipeline are illustrated in Figure 3.

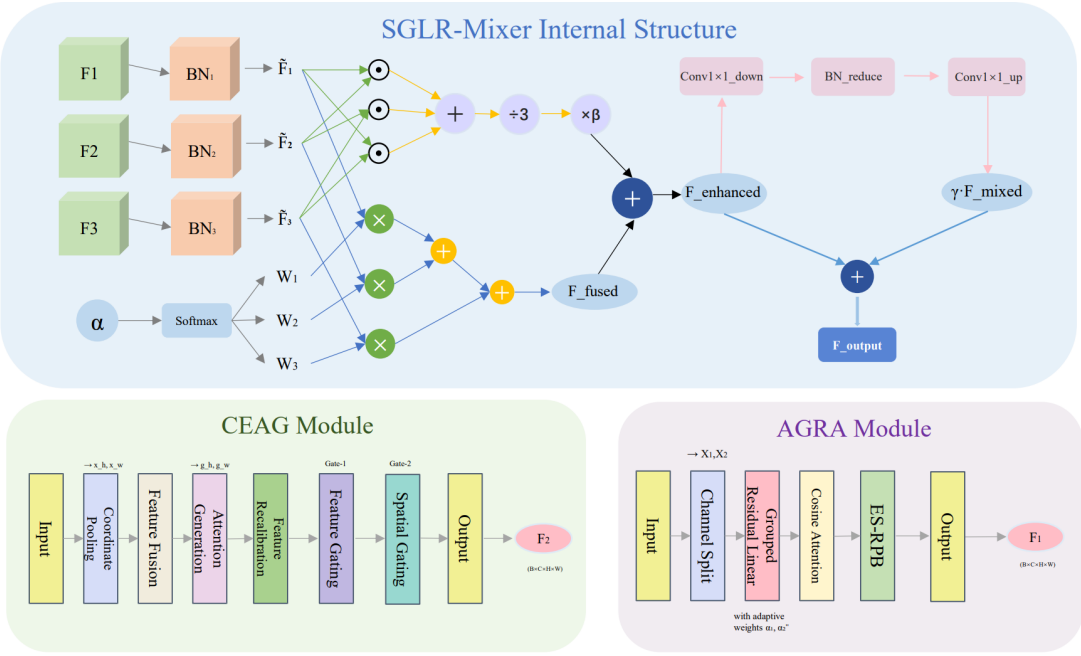

Figure 3. Structural illustration of the proposed AGRA, CEAG, and SGLR-Mixer modules, highlighting their internal mechanisms and information flow.

 doi:10.20944/preprints202601.2382.v1

Given three inputs $F_1, F_2, F_3 \in \mathbb{R}^{B\times C\times H\times W}$, we templately apply batch norm to align magnitude differences:

$$\tilde{F}1 = \mathrm{BN}_1(F_1), \quad \tilde{F}_2 = \mathrm{BN}_2(F_2), \quad \tilde{F}_3 = \mathrm{BN}_3(F_3) \tag{20}$$

Then fuse via learnable soft-gated weights $\alpha = [\alpha_1, \alpha_2, \alpha_3]$:

$$w = \mathrm{Softmax}(\alpha) = [w_1, w_2, w_3] \tag{21}$$

$$F_{\mathrm{fused}} = w_1 \cdot \tilde{F}1 + w_2 \cdot \tilde{F}_2 + w_3 \cdot \tilde{F}_3 \tag{22}$$

To capture complementary information, add a small second-order interaction term:

$$F_{\mathrm{enhanced}} = F_{\mathrm{fused}} + \beta \cdot \frac{\tilde{F}1 \odot \tilde{F}_2 + \tilde{F}_2 \odot \tilde{F}_3 + \tilde{F}_1 \odot \tilde{F}_3}{3} \tag{23}$$

where β is the interaction coefficient and $\odot$ is element-wise multiplication. A low-rank channel mixer performs lightweight cross-modal interaction. Reduce channels to a low-rank space:

$$F_{\mathrm{reduced}} = \mathrm{BNreduce}(\mathrm{Conv1} \times 1^{\mathrm{down}}(F_{\mathrm{enhanced}})) \tag{24}$$

then map back to the original channel space:

$$F_{\mathrm{mixed}} = \mathrm{Conv1} \times 1^{\mathrm{up}}(F\mathrm{reduced}) \tag{25}$$

Finally, add the mixed features back to the enhanced features via a residual:

$$F_{\mathrm{output}} = F_{\mathrm{enhanced}} + \gamma \cdot F_{\mathrm{mixed}} \tag{26}$$

where γ is a scaling factor controlling the residual contribution.

3.5. TSCA-Classifier

In fine-grained food classification, global average pooling and single-centroid classifiers tend to discard spatial distribution cues and inadequately model intra-class diversity, which are critical for distinguishing visually similar food categories with different composition-related characteristics. To address this limitation, we propose the TSCA-Classifier, as illustrated in Figure 4.

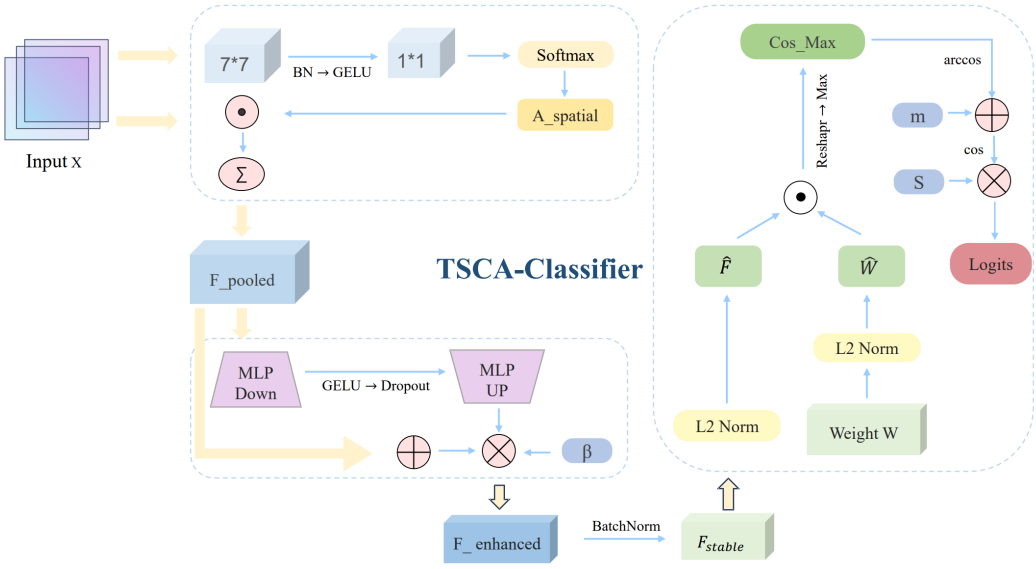

Figure 4. Architecture of the TSCA-Classifier, demonstrating its spatially weighted pooling and multi-center angular discrimination strategy.

 doi:10.20944/preprints202601.2382.v1

Given inputs $X \in \mathbb{R}^{B\times C\times H\times W}$, we use Token-aware Pooling (TAP) to aggregate spatial features.TAP uses two-layer convolutions to generate spatial attention:

$$A_{\text{spatial}} = \text{Softmax}(\text{Conv1} \times 1(\text{GELU}(\text{BN}(\text{Conv7} \times 7(X))))), \tag{27}$$

$$F_{\text{pooled}} = \sum_{i=1}^{H\times W} A_{\text{spatial}}^{(i)} \cdot X^{(i)} \tag{28}$$

with 7×7 to $C/16$ and 1×1 to single-channel. A lightweight enhancement follows:

$$F_{\text{enhanced}} = F_{\text{pooled}} + \beta \cdot \text{MLP}(\text{Dropout}(\text{GELU}(\text{MLP}(F_{\text{pooled}})))) \tag{29}$$

where $\beta = 0.008$; MLP maps $C \to C/32 \to C$; Dropout $= 0.7$. BatchNorm is applied:

$$F_{\text{stable}} = \text{BN}(F_{\text{enhanced}}) \tag{30}$$

SubCenter ArcFace keeps K sub-centers per class ($K = 3$), weight matrix $W \in \mathbb{R}^{(C\times K)\times N_{\text{classes}}}$. Normalize features and weights:

$$\hat{F} = \frac{F_{\text{stable}}}{||F_{\text{stable}}||2'}, \quad \hat{W} = \frac{W}{||W||_2} \tag{31}$$

Cosine similarity:

$$\cos = \hat{F} \cdot \hat{W}^T \tag{32}$$

Reshape to $[B, N_{\text{classes}}, K]$ and take the max over K sub-centers:

$$\cos_{\max} = \max_{j=1}^{K} \cos_{i,j} \tag{33}$$

Train with an angular margin m:

$$\theta = \arccos(\cos_{\max}) \tag{34}$$

The target logit for the ground-truth class is:

$$\cos_{\text{target}} = \cos(\theta + m) \tag{35}$$

Final logits are obtained by scaling with s:

$$\text{logits} = s \cdot \cos_{\text{final}} \tag{36}$$

4. Experiments and Results

4.1. Dataset

This paper employs two common fine-grained food image datasets to evaluate the proposed method. All datasets adhere to official standard partitions.

- **FoodX-251.** FoodX-251 (Kaur et al., 2019) comprises 251 visually similar fine-grained categories (e.g., cakes with varying decorations, sandwiches with distinct fillings, and pasta in diverse shapes) with 120,216 training images (raw web labels), 12,170 validation images, and 228,399 test images (human-verified labels for validation/test sets).
- **UEC FOOD-256.** UEC FOOD 256 (Kawano and Yanai, 2014) includes 256 categories of food images, each annotated with bounding boxes precisely localizing food regions. The dataset primarily features Japanese cuisine (e.g., tamagoyaki and takoyaki) alongside international dishes, where certain Japan-specific categories may present recognition challenges for non-native observers.

 doi:10.20944/preprints202601.2382.v1

4.2. Evaluation Metrics

To comprehensively assess the effectiveness of our proposed method in image classification, three widely accepted evaluation metrics were employed: Top-1 Accuracy, F1 Score, and Precision.

- Accuracy measures the proportion of correctly predicted samples over the total number of test instances, indicating overall performance.

- Precision reflects how many of the predicted positive results are actually correct.

- Recall represents the ability of the model to correctly identify all actual positive instances.

- F1 score is the harmonic mean of precision and recall, providing a balanced metric between the two.

The definitions of these metrics are given as:

$$\text{Acc} = \frac{1}{N}\sum_{i=0}^{N-1}(f(x_i) = y_i) \tag{37}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{38}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{39}$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{40}$$

Here, TP (True Positives) refers to correctly classified positive samples, FP(False Positives) are negative samples mistakenly predicted as positive, FN (False Negatives) are positive samples incorrectly classified as negative, and TN (True Negatives) are correctly classified negative samples.

All experiments were conducted under consistent settings and hardware configurations to ensure the reliability of the comparisons. In addition to benchmarking against state-of-the-art self-supervised learning methods, ablation studies were also performed to analyze the individual contributions of each module within the proposed architecture.

4.3. Performance Comparative Experiments

To comprehensively evaluate the effectiveness of the proposed method in fine-grained food recognition tasks, a series of systematic comparative experiments were conducted on two authoritative public datasets FoodX-251 and UEC FOOD-256. The compared methods include representative image recognition architectures, covering both conventional convolutional neural networks (e.g., ResNet, TResNet, EfficientNet) and a variety of recently developed Vision Transformer-based models (e.g., ViT, Swin Transformer, Twins, CaiT, ConvNeXt, CSwin, VOLO) [39,40], all of which have demonstrated strong benchmark performance in food image classification and fine-grained recognition tasks.

All experiments were performed on a single NVIDIA GeForce RTX 4090 GPU with mixed-precision training to improve computational efficiency. The batch size was set to 8, and the AdamW optimizer was employed. Both datasets were split into training and validation sets following their official protocols, without any manual intervention. To ensure fairness and reproducibility, three commonly used evaluation metrics, Top-1 Accuracy, F1 Score, and Precision were adopted to comprehensively assess the classification capability of different models under multi-class and imbalanced food image distributions.

As shown in Table 1, the performance differences among the compared models are evident across both datasets. Traditional convolutional networks exhibit certain advantages in local texture modeling, yet they often fail to establish global semantic consistency when facing food images with complex backgrounds or diverse structural forms. Transformer-based architectures, by contrast, improve the integration of global features through self-attention mechanisms, thereby achieving higher recognition

 doi:10.20944/preprints202601.2382.v1

accuracy. However, these methods still face limitations when handling food categories with large intra-class variations and scattered discriminative features.

Table 1. Swin-ACST compares results with other models on the FoodX-251 and UEC Food-256 datasets.

Methods	Epochs	Resolution	FoodX251			UECFood-256		
			Acc. (%)	Pre. (%)	F1. (%)	Acc. (%)	Pre. (%)	F1. (%)
ResNet-50	50	224 × 224	72.13	72.05	72.93	75.54	75.44	76.06
ResNet-101	50	224 × 224	73.11	73.10	74.00	75.65	75.57	76.30
TResNet-L	50	224 × 224	74.84	74.82	75.61	76.18	76.01	76.74
TResNet-XL	50	224 × 224	74.59	74.51	75.27	76.35	76.17	76.77
EfficientNet-b7	50	224 × 224	73.54	73.44	74.15	74.84	74.52	75.05
ConvNext-B	50	224 × 224	77.84	77.75	78.35	78.79	78.60	79.11
ConvNext-L	50	224 × 224	78.02	77.92	78.52	79.44	79.25	79.74
ViT-B	50	224 × 224	77.46	77.35	78.02	78.90	78.91	79.66
ViT-L	50	224 × 224	79.51	79.36	79.88	80.67	80.61	81.86
SwinT-B	50	224 × 224	78.57	78.55	79.17	80.81	80.74	81.61
SwinT-L	50	224 × 224	79.58	79.53	80.11	81.77	81.72	82.52
Swinv2-B	50	192 × 192	77.36	77.33	77.98	80.30	80.27	80.92
Swinv2-L	50	192 × 192	78.31	78.30	78.89	80.22	80.23	80.86
DeiT-S	50	224 × 224	72.62	72.63	73.39	75.28	75.28	75.98
DeiT-B	50	224 × 224	75.91	75.91	76.65	77.41	77.38	78.07
DeiTv2-B	50	224 × 224	75.52	75.41	76.03	78.05	78.06	78.70
DeiTv2-L	50	224 × 224	77.54	77.43	78.14	79.07	78.96	79.58
Twins-B	50	224 × 224	75.82	75.77	76.44	77.61	77.53	78.11
Twins-L	50	224 × 224	76.04	75.93	76.60	78.13	78.00	78.65
Cait-S	50	224 × 224	76.41	76.40	77.08	77.87	77.76	78.46
CSWin-L[†]	50	224 × 224	79.90	–	–	–	–	–
VOLO-D5[†]	50	224 × 224	79.51	–	–	–	–	–
Inception V3[§]	50	224 × 224	–	–	–	76.17	–	–
WRN[¶]	50	224 × 224	–	–	–	79.76	–	–
Swin-DR	50	224 × 224	81.07	80.98	81.48	82.15*	82.11	82.89
Swin-ACST	50	224 × 224	**82.28**	**82.21**	**82.76**	**82.64**	**82.60**	**83.39**

[†]Ródenas et al. (2022); [§]Hassannejad et al. (2016); [¶]Martinel et al. (2018); [*]Slight variation from the original paper, likely due to implementation environment. All models are compared fairly under our unified setup.

Our proposed model incorporates several optimization mechanisms tailored for fine-grained food imagery, effectively enhancing its capabilities in spatial modeling, region perception, and detail discrimination. On the FoodX-251 and UEC FOOD-256 datasets, our method achieves 82.28% and 82.64% Top-1 accuracy, respectively, confirming its superiority and robustness in complex food recognition scenarios.

The model centers around unified spatial relation modeling and composition-related salient region localization, strengthening spatial dependency representations across both global and local levels. Through robust alignment of heterogeneous semantic branches and lightweight low-rank fusion, the proposed architecture enables effective feature integration that preserves the spatial distribution of key regions while suppressing background noise and redundancy. Based on this, the classification head replaces the single-center paradigm with a multi-center angular metric, enhancing intra-class diversity representation and boundary robustness. With only marginal additional computational overhead, the entire pipeline achieves end-to-end optimization from feature extraction and cross-branch fusion to decision-space refinement, significantly improving recognition stability and discriminative capability in "same-class samples with different composition layouts" and cluttered background scenarios.

Notably, the proposed method demonstrates strong discriminative capability in handling visually similar food categories that belong to different semantic classes, a scenario that frequently occurs in real-world food image analysis. For example, soup-based dishes such as beef soup and pork bone soup often present highly comparable visual appearances in terms of color tone, liquid dominance, and serving context, despite differing in primary ingredients and dietary relevance. Conventional models tend to confuse such categories due to their reliance on coarse global representations. In contrast,

 doi:10.20944/preprints202601.2382.v1

the proposed framework can accurately attend to composition-related regions while modeling global spatial relationships, enabling more reliable differentiation between visually confusing yet semantically distinct food categories and maintaining stable classification performance under complex background conditions.

These results demonstrate that the proposed approach achieves systematic improvements in feature alignment, category boundary modeling, and local–global information fusion by integrating more efficient spatial dependency modeling and region-aware mechanisms. Consequently, it enhances the model's ability to discriminate subtle inter-class differences and greatly improves its classification robustness and stability in fine-grained food recognition. The above experimental results confirm that our method provides a more accurate and reliable solution for complex fine-grained food recognition tasks.

4.4. Ablation Analysis

To verify the effectiveness of the overall architectural design proposed in this study, we conducted a series of ablation experiments on two fine-grained food image datasets, FoodX-251 and UEC FOOD-256. Starting from the baseline backbone, the experiments follow the overarching principle of "enhancing spatial relation modeling — emphasizing key regions — robust multi-branch fusion and discrimination". Each capability is introduced progressively to assess its individual and cumulative contribution to the overall recognition performance, while avoiding isolated or fragmented evaluations of single modules.

Specifically, the original Swin-DR model without any structural enhancement is designated as the baseline. As shown in Table 2, Its Top-1 accuracies on the two datasets are 81.07% and 82.15%, respectively. After incorporating spatial relation modeling enhancement, the accuracies increase to 81.41% and 82.36%, indicating that such modeling facilitates better representation of global layouts and cross-region dependencies, yielding stable gains particularly on the UEC FOOD-256 dataset. Further introducing discriminative region focusing and guidance leads to accuracies of 81.67% and 82.48%. When these capabilities are jointly integrated under a unified framework with lightweight low-rank multi-branch fusion, the model achieves 81.97% and 82.54%, respectively. Finally, after convergence through a more robust multi-center discriminative space, the model attains its best performance of 82.28% on FoodX-251 and 82.64% on UEC FOOD-256. These results are clearly superior to those achieved by any single modification, validating the synergistic effect of the overall design in spatial representation, salient localization, and heterogeneous feature integration.

Table 2. Ablation study of different module combinations in Swin-ACST on FoodX-251 and UEC Food-256 datasets. **S**: Swin-DR, **C**: CEAG, **A**: AGRA ,**SG**: SGLR-Mixer, **T**: TSCA-Classifier.

Model Variant	S	C	A	SG	T	Acc. (%)	Pre. (%)	F1. (%)	Acc. (%)	Pre. (%)	F1. (%)
	Modules					FoodX-251			UEC Food-256		
S	✓	✗	✗	✗	✗	81.07	80.98	81.48	82.15	82.11	82.89
S+C	✓	✓	✗	✗	✗	81.41	81.52	82.09	82.36	82.28	83.16
S+A	✓	✗	✓	✗	✗	81.38	81.26	81.93	82.29	81.95	83.03
S+C+A	✓	✓	✓	✗	✗	81.67	81.47	82.12	82.48	81.99	83.13
S+C+A+SG	✓	✓	✓	✓	✗	81.97	81.64	82.44	82.54	82.17	83.22
S+C+A+SG+T (**Ours**)	✓	✓	✓	✓	✓	**82.28**	**82.21**	**82.76**	**82.64**	**82.60**	**83.39**

Taking visually similar soup-based dishes as an example, categories such as beef soup and chicken soup often exhibit highly comparable appearances, characterized by dominant liquid regions, similar color distributions, and overlapping serving contexts. Differences in primary ingredients are frequently reflected only in subtle local regions and spatial composition cues, making them prone to confusion under background interference and plating variations. The baseline model tends to be distracted by dominant background or container regions, resulting in unstable attention and inconsistent predictions.

 doi:10.20944/preprints202601.2382.v1

In contrast, when the proposed integrated strategy is enabled, the model can consistently emphasize ingredient-related regions within the global layout, preserve meaningful spatial distribution patterns, and suppress irrelevant background responses, thereby achieving more reliable discrimination between visually confusing yet semantically distinct food categories.

In summary, the ablation results indicate that the observed performance gains are not attributable to any single component in isolation, but arise from the coordinated optimization of spatial relation modeling, salient region localization, robust low-rank fusion, and discriminative decision modeling. This integrated design enhances the consistency of feature representation, stabilizes region-aware responses, and improves fine-grained alignment across complex visual conditions. These findings further validate the structural rationality of the proposed framework and its practical suitability for reliable fine-grained food image recognition in application-oriented settings.

5. Dietary Analysis Application

In recent years, increasing attention has been paid to dietary health and nutritional balance. However, in daily dining scenarios, many consumers lack accurate knowledge of food composition and ingredient structure, especially when visually similar dishes differ subtly in their main ingredients or preparation styles. This often leads to misunderstandings in dietary assessment and limits the effectiveness of image-based dietary analysis.

To address this problem, we develop an image-based dietary analysis application in which the proposed Swin-ACST framework serves as the core food recognition module. The application is designed to support reliable food category identification under real-world dining conditions and to provide users with consistent food-related information for subsequent dietary interpretation.

As illustrated in Figure 5, users capture food images using mobile devices in unconstrained environments, where variations in shooting angle, illumination, distance, and background are common. The captured images are uploaded through the application interface and transmitted to the inference server, where the trained Swin-ACST model performs fine-grained food recognition. Despite substantial visual variations in image appearance, the system produces stable and consistent recognition results for the same food category. This indicates that the proposed framework effectively captures food-related structural and compositional cues that remain reliable across different acquisition conditions, which is essential for practical dietary monitoring.

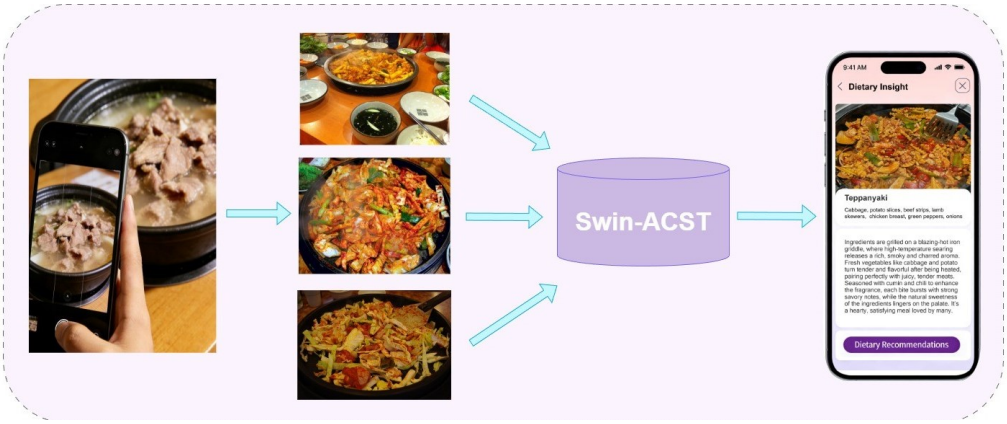

Figure 5. Despite significant variations in visual appearance caused by real-world dining environments, the system consistently identifies the food as the same category, demonstrating stable recognition performance and robustness for practical dietary monitoring

Beyond robustness to appearance variation, accurate discrimination among visually similar food categories is particularly important for dietary analysis. As shown in Figure 6, several dishes with highly similar visual appearances but different ingredient composition are evaluated. The visualization results demonstrate that the proposed system focuses on ingredient-dominant regions while suppressing irrelevant background areas, enabling more precise differentiation between visually

 doi:10.20944/preprints202601.2382.v1

confusing dishes. Such behavior is critical in dietary applications, where misclassification between similar-looking foods may lead to incorrect interpretation of nutritional characteristics.

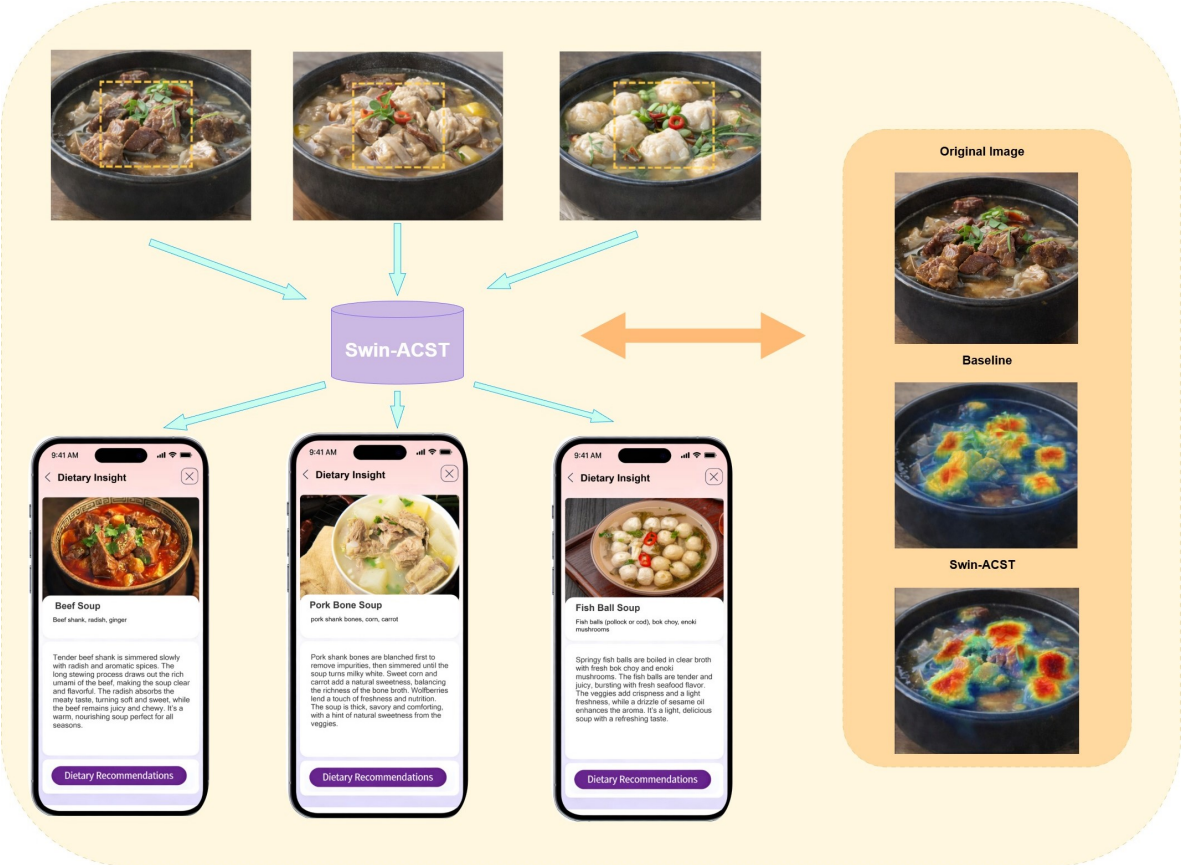

Figure 6. For food images with highly similar appearance, the proposed model highlights discriminative ingredient regions while suppressing irrelevant background areas. The system is able to focus on category-specific visual cues, enabling accurate differentiation between food items that are easily confused in real-world dietary scenarios.

After food recognition is completed, the system automatically retrieves corresponding food-related information and health-oriented references from the database according to the recognized category and returns the results to the application interface. By reliably distinguishing food categories with subtle compositional and structural differences, the application supports more meaningful interpretation of meal composition and provides users with clearer dietary insights. These results indicate that Swin-ACST is well suited for deployment in real-world image-based dietary analysis applications, providing a reliable visual foundation for intelligent food monitoring and health-oriented dietary assessment.

6. Conclusion

This study presents Swin-ACST, a fine-grained food classification framework designed to enhance the visual characterization of composition-related structures in food images. By jointly modeling global spatial layouts and local structural arrangements, the proposed approach effectively addresses key challenges in fine-grained food recognition, including large intra-class variation and small inter-class differences commonly observed in visually similar dishes.

The introduced spatial relation modeling and key-region awareness mechanisms enable the model to suppress background interference and maintain consistent attention to ingredient-dominant regions, improving both inter-class discrimination and intra-class consistency. In addition, the adaptive multi-branch fusion strategy and token-aware subcenter-based classification head further strengthen class separability by preserving spatial information and explicitly modeling intra-class diversity. Experimental results on FoodX-251 and UEC Food-256 demonstrate that Swin-ACST achieves superior

 doi:10.20944/preprints202601.2382.v1

performance compared with existing advanced methods, particularly for visually confusing food categories.

Beyond classification accuracy, the proposed framework provides a robust and non-invasive visual characterization basis for composition-aware food recognition in health-oriented applications. By reliably distinguishing food categories that differ subtly in ingredient dominance and spatial organization, Swin-ACST supports more consistent interpretation of meal composition and its potential dietary relevance. This work demonstrates that enhancing spatial structure modeling and key-region awareness is an effective strategy for bridging fine-grained food recognition with practical dietary analysis scenarios, offering a reliable visual foundation for intelligent food analysis systems.

Author Contributions: Investigation, L.Y.; methodology, L.Y. and Z.X.; data analysis, L.Y. and Q.S.; software, conceptualization, Z.X.; writing–original draft, L.Y.; supervision, Z.X.; project administration, Z.X.; funding acquisition, Z.X.; writing-review & editing, Q.S. and Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Jiangsu Province (CN) under Grant BK20190079.

Data Availability Statement: The datasets used in this study are all publicly available.UEC FOOD-256 is available at http://foodcam.mobi/dataset.html. FoodX-251 is available at https://www.selectdataset.com/dataset/439a25421e7c974dd41dc26a3e40e42c. All datasets can be used for academic purposes in accordance with their respective terms of use or open-source licenses.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Gao, X.; Xiao, Z.; Deng, Z. High accuracy food image classification via vision transformer with data augmentation and feature augmentation. *Journal of Food Engineering* **2024**, *365*, 111833. https://doi.org/https://doi.org/10.1016/j.jfoodeng.2023.111833.
2. Mansouri, M.; Benabdellah Chaouni, S.; Jai Andaloussi, S.; Ouchetto, O. Deep learning for food image recognition and nutrition analysis towards chronic diseases monitoring: A systematic review. *SN Computer Science* **2023**, *4*, 513.
3. Zhang, Y.; Deng, L.; Zhu, H.; Wang, W.; Ren, Z.; Zhou, Q.; Lu, S.; Sun, S.; Zhu, Z.; Gorriz, J.M.; et al. Deep learning in food category recognition. *Information Fusion* **2023**, *98*, 101859.
4. He, J.; Chen, J.N.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; Wang, C. Transfg: A transformer architecture for fine-grained recognition. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 852–860.
5. Boyd, L.; Nnamoko, N.; Lopes, R. Fine-grained food image recognition: A study on optimising convolutional neural networks for improved performance. *Journal of Imaging* **2024**, *10*, 126.
6. Ma, P.; Lau, C.P.; Yu, N.; Li, A.; Liu, P.; Wang, Q.; Sheng, J. Image-based nutrient estimation for Chinese dishes using deep learning. *Food Research International* **2021**, *147*, 110437.
7. Xiao, Z.; Diao, G.; Deng, Z. Fine grained food image recognition based on swin transformer. *Journal of Food Engineering* **2024**, *380*, 112134. https://doi.org/https://doi.org/10.1016/j.jfoodeng.2024.112134.
8. Min, W.; Wang, Z.; Liu, Y.; Luo, M.; Kang, L.; Wei, X.; Wei, X.; Jiang, S. Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**, *45*, 9932–9949.
9. Bossard, L.; Guillaumin, M.; Van Gool, L. Food-101–mining discriminative components with random forests. In Proceedings of the European Conference on Computer Vision. Springer, 2014, pp. 446–461.
10. Kaur, P.; Sikka, K.; Wang, W.; Belongie, S.; Divakaran, A. Foodx-251: a dataset for fine-grained food classification. *arXiv preprint arXiv:1907.06167* **2019**.
11. Qian, Y.; Xiao, Z.; Deng, Z. Fine-grained crop pest classification based on multi-scale feature fusion and mixed attention mechanisms. *Frontiers in Plant Science* **2025**, p. 1500571.
12. Xiao, Z.; Sun, Y.; Deng, Z. FGFoodNet: Ingredient-perceived fine-grained food recognition for dietary monitoring. *Journal of Food Measurement and Characterization* **2025** *19*, 7017–7033.

 doi:10.20944/preprints202601.2382.v1

13. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
14. Xiao, Z.; Zhang, Y.; Deng, Z.; Liu, F. Light3DHS: A lightweight 3D hippocampus segmentation method using multiscale convolution attention and vision transformer. *NeuroImage* **2024**, *292*, 120608. https://doi.org/https://doi.org/10.1016/j.neuroimage.2024.120608.
15. Xiao, Z.; Ling, R.; Deng, Z. FoodCSWin: A high-accuracy food image recognition model for dietary assessment. *Journal of Food Composition and Analysis* **2025**, *139*, 107110. https://doi.org/https://doi.org/10.1016/j.jfca.2024.107110.
16. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713–13722.
17. Zhuang, G.; Hu, Y.; Yan, T.; Gao, J. GCAM: Gaussian and causal-attention model of food fine-grained recognition. *Signal, Image and Video Processing* **2024**, *18*, 7171–7182.
18. Xiao, Z.; Diao, G.; Liu, C.; Deng, Z. Fine-grained food image recognition using a convolutional neural network and swin transformer hybrid model. *Journal of Food Composition and Analysis*, **2025**, *148*, 108395.
19. Ramzan, M.U.; Khaddim, W.; Rana, M.E.; Ali, U.; Ali, M.; ul Hassan, F.; Mehmood, F. Gated-attention feature-fusion based framework for poverty prediction. In Proceedings of the International Conference on Data Engineering and Communication Technology. Springer, 2024, pp. 415–426.
20. Li, Y.; Daho, M.E.H.; Conze, P.H.; Zeghlache, R.; Le Boité, H.; Tadayoni, R.; Cochener, B.; Lamard, M.; Quellec, G. A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine* **2024**, *177*, 108635.
21. Segu, M.; Tonioni, A.; Tombari, F. Batch normalization embeddings for deep domain generalization. *Pattern Recognition* **2023**, *135*, 109115.
22. Liu, X.; Zheng, G.; Beheshti, I.; Ji, S.; Gou, Z.; Cui, W. Low-Rank Tensor Fusion for Enhanced Deep Learning-Based Multimodal Brain Age Estimation. *Brain Sciences* **2024**, *14*, 1252.
23. Chen, S.; Sun, Q.; Li, C.; Zhang, J.; Zhang, Q. Attention-guided second-order pooling convolutional networks. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 2230–2234.
24. Wang, X.; Kang, M.; Chen, Y.; Jiang, W.; Wang, M.; Weise, T.; Tan, M.; Xu, L.; Li, X.; Zou, L.; et al. Adaptive local cross-channel vector pooling attention module for semantic segmentation of remote sensing imagery. *Remote Sensing* **2023**, *15*, 1980.
25. Marin, D.; Chang, J.H.R.; Ranjan, A.; Prabhu, A.; Rastegari, M.; Tuzel, O. Token pooling in vision transformers. *arXiv preprint arXiv:2110.03860* **2021**.
26. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
27. Xiao, Z.; Su, Y.; Deng, Z.; Zhang, W. Efficient combination of CNN and Transformer for dual-teacher uncertainty-guided semi-supervised medical image segmentation. *Computer Methods and Programs in Biomedicine* **2022**, *226*, 107099.
28. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning. PMLR, 2021, pp. 10347–10357.
29. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 22–31.
30. Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; Jégou, H. Going deeper with image transformers. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 32–42.
31. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems* **2021**, *34*, 9355–9366.
32. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
33. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

 doi:10.20944/preprints202601.2382.v1

34. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems* **2021**, *34*, 3965–3977.
35. Li, Y.; Deng, Z.; Cao, Y.; Liu, L. GRFormer: grouped residual self-attention for lightweight single image super-resolution. In Proceedings of the Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 9378–9386.
36. He, J.; Shao, Z.; Wright, J.; Kerr, D.; Boushey, C.; Zhu, F. Multi-task image-based dietary assessment for food recognition and portion size estimation. In Proceedings of the 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2020, pp. 49–54.
37. Dalakleidi, K.V.; Papadelli, M.; Kapolos, I.; Papadimitriou, K. Applying image-based food-recognition systems on dietary assessment: a systematic review. *Advances in Nutrition* **2022**, *13*, 2590–2619.
38. Tahir, G.A.; Loo, C.K. A comprehensive survey of image-based food recognition and volume estimation methods for dietary assessment. In Proceedings of the Healthcare. MDPI, 2021, Vol. 9, p. 1676.
39. Yuan, L.; Hou, Q.; Jiang, Z.; Feng, J.; Yan, S. Volo: Vision outlooker for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *45*, 6575–6586.
40. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12124–12134.