

Food Image Ingredient Recognition Based on VMamba and Cross-layer Attention

Ping-Han Chen,^{1,2} Ting-Ting Chen,^{3*} and Chin-Feng Lai^{1**}

¹Department of Engineering Science, National Cheng Kung University, Tainan 701, Taiwan

²Internet of Things Laboratory, Telecommunication Laboratories, Chunghwa Telecom Co., Ltd.,
Taoyuan 326, Taiwan

³Department of Leisure Industry and Health Promotion, National Ilan University, Yilan 260, Taiwan

(Received April 10, 2026; accepted September 25, 2026)

Keywords: food image analysis, multitask learning, VMamba, lightweight networks, cross-layer attention

Food image analysis remains challenging owing to substantial visual variations, multiscale ingredient appearances, and the computational demands of global visual modeling. In practical dietary sensing systems, food images are acquired by vision sensors and processed on edge devices, placing additional requirements on both recognition accuracy and computational efficiency. To address these issues, we propose a lightweight multitask framework for joint dish classification and ingredient recognition. The framework employs VMamba-Tiny to provide efficient long-range visual modeling and introduces a cross-layer attention block to integrate fine-grained texture cues with high-level semantic information for multiscale ingredient representation. In addition, a decoupled feature integration strategy (DFIS) adopts a forward-shared and backward-decoupled design, preserving ingredient representations for dish classification while restricting the backward effect of the dish-classification loss on ingredient-specific representations. Experiments on Vireo Food-172 and ISIA Food-200 demonstrate that the proposed model achieves Micro-F1 scores of 81.41 and 72.78%, respectively, with only 7.1 giga floating-point operations. Further analysis shows that the improvement introduced by DFIS becomes increasingly pronounced for low-frequency ingredients, with gains on Tail classes approximately 2.4–3.2× larger than those on Head classes. Moreover, the model achieves 25.8 frames per second on an NVIDIA Jetson Orin Nano with a batch size of 1, supporting its use as the inference component of a vision-sensor-based edge dietary monitoring system.

1. Introduction

As public health awareness continues to rise, the public has an increasing demand for physical fitness management and dietary control. Previous studies have highlighted the importance of nutritional knowledge and appropriate dietary intake in health and sports-related applications.⁽¹⁾ However, traditional dietary assessment methods, such as food records,^(2,3) 24 h dietary recalls, and food frequency questionnaires,⁽⁴⁾ often require substantial user involvement or professional assessment, resulting in considerable time and labor costs. These methods are inherently time-

*Corresponding author: e-mail: ttchen@niu.edu.tw

**Corresponding author: e-mail: cinfon@ncku.edu.tw

<https://doi.org/10.18494/SAM6371>

consuming and incapable of delivering real-time feedback. Although photographic food logging has improved the convenience of dietary tracking, users may still have difficulty accurately identifying the ingredient composition of complex meals. Consequently, the adoption of computer vision technologies offers objectivity, high accuracy, and practicality in daily dietary tracking.⁽⁵⁾ From a sensing perspective, such dietary monitoring systems rely on vision sensors to acquire food images at the point of consumption, after which the captured visual data are analyzed to infer dish categories and ingredient composition. Deploying the recognition model close to the sensing device can further reduce dependence on remote computation and support responsive dietary monitoring, making computational efficiency an important requirement for practical edge sensing. Similar vision-sensor systems have been investigated in other application domains, including facial skin analysis for skincare recommendation and IoT-based nighttime vehicle detection, demonstrating the feasibility of integrating image acquisition, visual analysis, and application-level decision making within practical sensing systems.^(6,7)

Chinese dishes typically consist of multiple mixed ingredients, making the recognition of individual components inherently challenging, while diverse cooking methods and regional variations introduce substantial visual heterogeneity that further complicates accurate recognition. Traditionally, food image analysis has relied on convolutional neural networks (CNNs) for feature extraction.^(8–10) While CNNs are effective at capturing local textures and spatial patterns, their locality makes long-range dependence modeling less direct. This limitation can be restrictive for food images containing complex spatial layouts and substantial appearance variations. To address this, vision transformer (ViT) models,⁽¹¹⁾ which offer a global receptive field, have been introduced and have demonstrated strong performance in food recognition tasks.⁽¹²⁾ However, the self-attention mechanism of ViT introduces a quadratic computational complexity $O(N^2)$, leading to substantial computational costs and memory consumption.

To support such vision-sensor-based dietary monitoring, an efficient recognition model should retain global visual modeling capability while maintaining low computational complexity for edge deployment. We therefore adopt VMamba⁽¹³⁾ as the backbone. Through 2D-Selective-Scan (SS2D), VMamba models long-range visual dependencies with linear computational complexity $O(N)$, substantially reducing the computational overhead associated with self-attention-based architectures.

Beyond computational efficiency, ingredient recognition also requires visual representations across different spatial scales and semantic levels. Shallow features retain fine-grained texture and spatial details, whereas deeper features encode stronger semantic information. Relying on a single feature level may therefore be insufficient for recognizing ingredients with diverse appearances and scales. Moreover, although dish classification and ingredient recognition share the same visual input, their optimization objectives are not fully aligned. Dish classification emphasizes category-discriminative representations, whereas ingredient recognition requires fine-grained multilabel representations. Under fully coupled multitask optimization, the dish-classification objective may therefore interfere with the learning of ingredient-specific representations.

To address these challenges, we propose a lightweight multitask food recognition framework consisting of VMamba-Tiny, a cross-layer attention block (CLAB), and a decoupled feature integration strategy (DFIS). VMamba-Tiny provides efficient global visual modeling, while CLAB integrates shallow texture cues with deep semantic features to construct multiscale ingredient representations. DFIS further adopts a forward-shared and backward-decoupled design: ingredient representations remain available to the dish-classification branch during forward propagation, while stop-gradient prevents the dish-classification loss from directly updating the ingredient-specific representation through the feature-fusion path.

Extensive experiments on Vireo Food-172 and ISIA Food-200 demonstrate that the proposed framework achieves competitive performance on both ingredient recognition and dish classification with only 7.1 G floating-point operations (FLOPs). Further long-tail analysis shows that the benefit of DFIS becomes increasingly pronounced as ingredient frequency decreases, with substantially larger improvements on Tail classes. In addition, the complete model achieves 25.8 frames per second (fps) on an NVIDIA Jetson Orin Nano with a batch size of 1, validating its feasibility for practical edge deployment.

2. Methods

2.1 Overall architecture

The proposed architecture jointly performs ingredient recognition and dish classification while maintaining low computational complexity. As illustrated in Fig. 1, the framework consists of three main components: (1) a VMamba-Tiny backbone for hierarchical feature extraction and efficient global modeling, (2) CLAB for multiscale ingredient representation learning, and (3) DFIS for controlling cross-task gradient propagation.

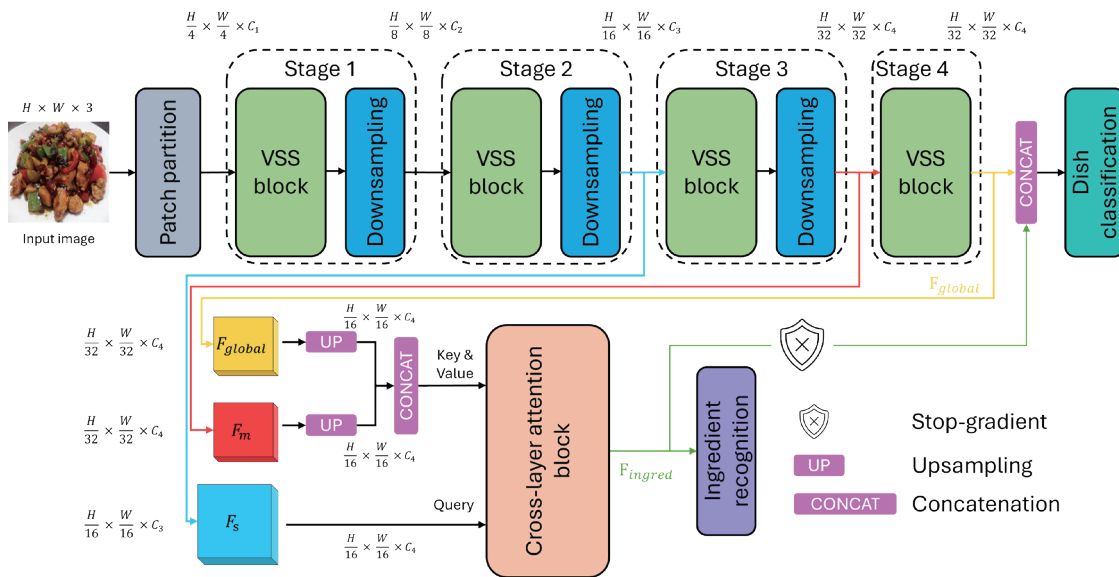


Fig. 1. (Color online) Overall architecture of proposed multitask framework.

Given an input image, VMamba-Tiny first extracts hierarchical visual features at multiple stages. CLAB then integrates fine-grained texture and spatial cues from shallower features with high-level semantic information from deeper features to construct the ingredient representation for multilabel ingredient recognition. For dish classification, this ingredient representation is combined with the global backbone feature through DFIS. Specifically, DFIS preserves the ingredient representation during forward propagation while applying stop-gradient to its feature-fusion path, thereby restricting the backward effect of the dish-classification loss on the ingredient-specific representation. The resulting fused representation is subsequently used for dish classification.

2.2 VMamba-Tiny backbone and SS2D

To achieve a global receptive field under limited computational resources, we employ the VMamba architecture as our feature extraction backbone. Whereas ViT relies on self-attention and suffers from a quadratic computational complexity of $O(N^2)$, VMamba uses the state space model and SS2D to reduce this complexity to linear $O(N)$.

In the VMamba architecture, the input image undergoes hierarchical feature extraction across four main stages. Each stage contains multiple visual state space (VSS) blocks, with the SS2D module at their core. The SS2D first sequentially unfolds the 2D image along four directions. These sequences are then fed in parallel into the internal selective state space model for processing, after which the features are merged. Through this process, the SS2D overcomes the lack of an inherent sequence in 2D images and effectively establishes long-range dependencies between features. Assuming the input feature is x_t , this continuous state-space transition can be discretized into the form of Eq. (1):

$$\begin{aligned} h_t &= \bar{A}h_{t-1} + \bar{B}x_t, \\ y_t &= Ch_t + Dx_t, \end{aligned} \quad (1)$$

where $\bar{A}$ and $\bar{B}$ represent the discretized state transition matrices, h_t is the hidden state, and y_t is the output feature. Through this state-space modeling approach, VMamba significantly reduces computational resource consumption compared with ViTs while simultaneously yielding features with a robust global receptive field.

Recent studies have demonstrated the adaptability of Mamba-based visual architectures across diverse image analysis tasks. Representative applications include GroupMamba with auxiliary Fourier features for waste classification,⁽¹⁴⁾ spatial Mamba for facial expression recognition,⁽¹⁵⁾ and frequency-aware U-Mamba for breast tumor segmentation.⁽¹⁶⁾ These developments further support the effectiveness of state-space-based modeling for both visual recognition and dense prediction tasks.

2.3 Multiscale ingredient recognition via CLAB

Ingredients within food imagery frequently exhibit diverse spatial scales. To capture these multiscale characteristics, we extract feature representations from the last three stages of the

VMamba-Tiny backbone. These outputs are denoted as the shallow feature F_s , intermediate feature F_m , and global feature F_{global} . To perform cross-layer attention, we introduce multihead attention. Within this module, the input features are linearly mapped into query (Q), key (K), and value (V) representations. The inputs are partitioned into h distinct attention heads and projected using h sets of learned weight matrices (W_i^Q, W_i^K, W_i^V). The scaled dot-product attention for each independent head is then computed according to Eq. (2).

$$Attention(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \quad (2)$$

In this formulation, the term $Q_i K_i^T$ computes the similarity matrix between queries and keys, whereas d_k denotes the embedding dimension. Scaling by $1/\sqrt{d_k}$ controls the magnitude of the dot products and stabilizes the subsequent softmax operation. These attention weights are subsequently aggregated with the value matrix V_i . The outputs from all h heads are then concatenated and subjected to a final linear projection via the weight matrix W^O , as defined in Eq. (3).

$$MultiHeadAttention(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O \quad (3)$$

In the proposed architecture, the shallow feature F_s , which preserves fine-grained textural and spatial details, is used as the Query. The intermediate feature F_m and the global feature F_{global} are first upsampled to match the spatial resolution of F_s and are then concatenated and projected to form the Key and Value. Through this query-driven interaction, shallow textural cues selectively aggregate semantic information from deeper representations, thereby facilitating cross-scale feature integration for ingredient recognition. The cross-layer attention operates on a 14×14 feature grid, using eight attention heads with embedding dimensions of 768, i.e., 96 dimensions per head. No additional positional encoding is applied.

2.4 Decoupled feature integration strategy

The proposed framework jointly performs dish classification and ingredient recognition within a multitask learning setting. Although the two tasks share the same visual input and backbone features, their optimization objectives are not fully consistent. Dish classification focuses on learning discriminative representations for category prediction, whereas ingredient recognition requires fine-grained representations for multilabel ingredient recognition. Therefore, under fully coupled end-to-end optimization, the gradient produced by the dish-classification objective may interfere with the learning of ingredient-specific representations. However, ingredient representations also provide useful semantic information for dish classification. Completely separating the two tasks would therefore reduce the information available to the dish-classification branch.

To address this issue, we introduce DFIS, which is based on a forward-shared and backward-decoupled design. DFIS preserves the ingredient representation as an input to the dish classifier during forward propagation, while restricting the backward effect of the dish-classification loss on the ingredient-specific representation.

Let F_{ingred} denote the ingredient representation generated by CLAB, and let F_{global} denote the global feature extracted by the VMamba backbone. The feature used for dish classification is defined as

$$F_{dish} = \text{Concat}(F_{global}, SG(F_{ingred})), \quad (4)$$

where $SG(\cdot)$ denotes the stop-gradient operation. During forward propagation, the operation does not change F_{ingred} ; therefore, the ingredient representation remains available to the dish-classification branch. During backpropagation, its gradient with respect to F_{ingred} is

$$\frac{\partial SG(F_{ingred})}{\partial F_{ingred}} = 0. \quad (5)$$

Accordingly, the dish-classification loss does not directly update F_{ingred} through the feature-fusion path, whereas the ingredient-recognition loss continues to optimize the ingredient representation normally. This design preserves forward feature sharing between the two tasks while reducing direct interference from the dish-classification objective on the ingredient-specific representation.

2.5 Edge sensor prototype implementation

To bridge the gap between theoretical algorithms and practical dietary assessment, we designed and implemented a customized sensor prototype, as illustrated in Fig. 2. The hardware architecture consists of a multiview camera array, an edge computing unit, and an interactive display. The array is designed to use three high-resolution image sensors, positioned at the left, center, and right of the tray. This arrangement is intended to address the visual occlusion caused by stacked or overlapping ingredients in complex dishes.

In the current implementation, only the central sensor is connected to the inference pipeline; the left and right positions are reserved as mounting points for the planned multiview extension. Each captured 2560×1440 frame is split into multiple regions of interest, and each region is resized to 224×224 before passing to the model. Using a high-resolution sensor preserves more details after cropping. The recognized ingredients and the resulting dietary record are then presented to the user through the interactive display.

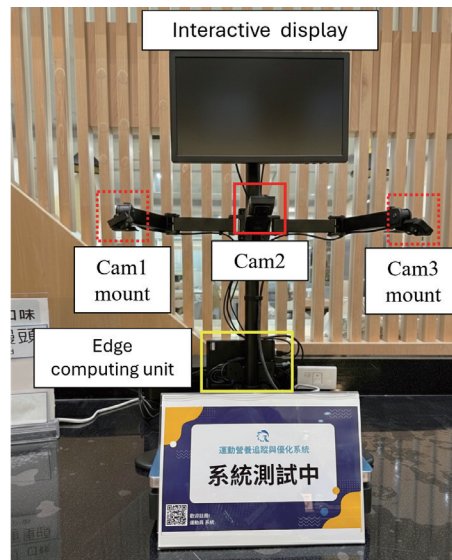


Fig. 2. (Color online) Customized dietary sensing prototype. Note that the red boxes labeled Cam1 and Cam3 indicate reserved mounting positions for future camera expansions; only Cam2 is installed and utilized in the current configuration.

3. Experimental Procedure

3.1 Dataset

To comprehensively evaluate the effectiveness of the proposed architecture, we conducted experiments on two widely adopted benchmark datasets: Vireo Food-172⁽⁸⁾ and ISIA Food-200.⁽⁹⁾

The first dataset utilized in this research, Vireo Food-172, was established by Chen and Ngo⁽⁸⁾ and consists of 110,241 images covering 172 Chinese dish categories and 353 distinct ingredient labels, with an average of three visible ingredients annotated per image. The images were predominantly crawled from the Baidu and Google image search engines and subsequently subjected to rigorous manual filtering to ensure high data quality. Strictly following the official dataset splits, the dataset is partitioned into training, validation, and testing subsets at a standard split ratio of 6:1:3.

The ISIA Food-200 dataset, constructed by Min *et al.*,⁽⁹⁾ serves as a large-scale benchmark designed specifically for real-world food recognition and dietary analysis. It encompasses a total of 197,323 images distributed across 200 food categories and 319 ingredient labels, ensuring a minimum of 500 images per category. Candidate images were scraped utilizing search engines such as Google and Bing, followed by a manual curation process to eliminate irrelevant or noisy data. Consistent with the official evaluation protocol, the dataset is partitioned into training, validation, and testing sets at a standard split ratio of 6:1:3.

3.2 Experimental setup

The proposed architecture is fundamentally driven by a VMamba backbone and CLAB. Following the settings of VMamba-Tiny,⁽¹³⁾ the backbone is initialized with ImageNet-1K pretrained weights, and the input image size is set to 224×224 . Regarding the loss functions, cross-entropy loss is utilized for the dish classification task, while binary cross-entropy loss is employed for the ingredient recognition task. During training, the batch size is set to 64, coupled with 8 steps of gradient accumulation. We adopt the AdamW optimizer with an initial learning rate of 1×10^{-4} , paired with a cosine learning rate scheduler. Additionally, the exponential moving average (EMA) technique is applied to the model weights during training to enhance stability and generalization. All experiments were conducted on a single NVIDIA RTX 5090. Our training phase employs a standard data augmentation pipeline. The process initiates with random resized cropping utilizing bicubic interpolation. This is sequentially followed by random horizontal flipping and the application of RandAugment. Subsequently, the input images undergo normalization on the basis of ImageNet statistics and are finally subjected to random erasing. To ensure reliability, all reported results represent the average of three independent experiments.

FLOPs and parameter counts are reported for a single 224×224 input with a batch size of 1 in inference mode, measured with the fvcare counter under the multiply–accumulate convention, in which one fused multiply–add is counted as one FLOP, the same convention adopted by Swin Transformer,⁽¹⁷⁾ and VMamba. As fvcare cannot trace the fused selective-scan CUDA kernel, its cost is accounted for using the analytic handle provided by the official VMamba implementation.⁽¹³⁾ During multilabel inference, ingredient labels with sigmoid probabilities greater than 0.5 are regarded as positive predictions.

Among the benchmark methods evaluated in this study, ViT,⁽¹¹⁾ MobileNetV3,⁽¹⁸⁾ GhostNetV2,⁽¹⁹⁾ and EfficientNet⁽²⁰⁾ were retrained under the same experimental settings and hyperparameters as the proposed method, each with the same dual-head output configuration; the reported parameter counts therefore correspond to this dual-task setting. The results for ASL, DSDL, MGTN, SGTN, CACNet, SLCI-T, and SLCI-B were adopted from the comparative results reported by Liu *et al.*⁽¹²⁾

3.3 Experiments on ingredient recognition task

To evaluate the effectiveness and applicability of the proposed architecture, we benchmark it against eleven representative ingredient recognition methods. To explicitly assess the computational efficiency of our lightweight design, the number of model parameters (#Params) and FLOPs are also reported.

Table 1 shows a comparison of the performance of the proposed method with those of the existing methods for food ingredient recognition. To provide a more comprehensive evaluation of the trade-off between recognition performance and computational efficiency, methods operating at different input resolutions are included in the comparison. In general, higher input resolutions preserve finer-grained visual details and can lead to notable improvements in

Table 1

Performance comparison of different methods on the multilabel food ingredient recognition task. Best and second-best results are shown in bold and underline, respectively.

Method	Input size	#Params	FLOPs	Vireo Food-172		ISIA Food-200	
				Micro-F1	Macro-F1	Micro-F1	Macro-F1
ASL ⁽²²⁾	448 × 448	25.5M	16.5G	74.38	60.79	62.74	64.20
DSDL ⁽²³⁾	448 × 448	42.5M	31.4G	66.85	49.44	63.93	62.46
MGTN ⁽²⁴⁾	448 × 448	45.9M	34.3G	79.46	58.48	61.87	60.28
SGTN ⁽²⁵⁾	448 × 448	22.9M	17.1G	76.83	55.27	58.97	55.21
CACLNet ⁽¹⁰⁾	448 × 448	42.9M	62.9G	79.28	65.71	64.85	62.94
SLCI-T ⁽¹²⁾	224 × 224	46.1M	11.6G	80.96	<u>68.73</u>	70.16	69.58
SLCI-B ⁽¹²⁾	384 × 384	121.0M	75.2G	<u>81.34</u>	69.22	<u>70.76</u>	<u>70.64</u>
ViT-Base-p16 ⁽¹¹⁾	224 × 224	94.4M	17.4G	80.00	66.83	70.48	70.22
MobileNetV3-Large ⁽¹⁸⁾	224 × 224	12.2M	0.51G	78.14	64.45	65.59	64.80
GhostNetV2 ⁽¹⁹⁾	224 × 224	15.0M	1.3G	76.34	62.85	63.73	62.69
EfficientNet-Lite4 ⁽²⁰⁾	224 × 224	22.0M	2.6G	78.95	65.85	69.07	68.40
Proposed	224 × 224	41.5M	7.1G	81.41	68.46	72.78	71.48

recognition performance, particularly for small objects or datasets exhibiting long-tailed distributions.⁽²¹⁾ In contrast, lower-resolution inputs impose substantially lower computational and memory demands, making them more suitable for edge computing environments.

As shown in Table 1, the proposed method achieves strong recognition performance and computational efficiency on both Vireo Food-172 and ISIA Food-200 despite using a relatively low input resolution and substantially fewer FLOPs. Specifically, on ISIA Food-200, our method achieves the best Micro-F1 and Macro-F1 scores among all compared methods. It outperforms the higher-resolution CACLNet, MGNTN, and ASL models while also requiring fewer parameters and a lower computational cost than SLCI-B and ViT-Base despite achieving superior F1 scores.

On Vireo Food-172, the proposed method achieves the highest Micro-F1 score while maintaining a highly competitive Macro-F1 score. Although SLCI-B achieves a slightly higher Macro-F1 score, it requires approximately 2.9× more parameters and 10.6× more FLOPs than the proposed method, imposing a considerably greater computational burden for practical deployment. Compared with SLCI-T, our method exhibits only a 0.27% decrease in Macro-F1 on Vireo Food-172 while outperforming SLCI-T in all other reported ingredient recognition metrics. These results demonstrate that the proposed architecture achieves a favorable balance between recognition performance and computational efficiency.

Furthermore, compared with lightweight architectures such as MobileNetV3-Large, GhostNetV2, and EfficientNet-Lite4, the proposed method requires moderately more parameters and computation but delivers substantial improvements in both dish classification and ingredient recognition performance. Overall, these results demonstrate that the proposed architecture effectively alleviates the feature representation limitations commonly observed in lightweight networks without incurring excessive computational overhead. This combination of high recognition performance and moderate computational complexity highlights its strong potential for accurate food ingredient recognition on edge devices.

3.4 Experiments on dish classification task

Table 2 shows a comparison of the recognition performance and computational cost of the proposed method with those of the representative models on the dish classification task. Consistent with the results on ingredient recognition, our method maintains competitive dish classification performance with substantially lower computational complexity, indicating that the multitask framework does not compromise dish-level discriminability.

On Vireo Food-172, the proposed method achieves a Top-1 accuracy of 91.22%. Compared with SLCI-T using the same 224×224 input resolution, our method improves Top-1 and Top-5 accuracies by 0.52 and 0.39 percentage points, respectively, while reducing the FLOPs from 11.6 to 7.1 GFLOPs. Even compared with SLCI-B using a larger 384×384 input, the Top-1 accuracy differs by only 0.36 percentage points, while our method requires only approximately 9.4% of its FLOPs and 34.3% of its parameters. On ISIA Food-200, our method achieves a Top-1 accuracy of 70.43% and outperforms SLCI-T in both Top-1 and Top-5 accuracies. Compared with SLCI-B, the Top-1 accuracy decreases by only 0.49 percentage points, while the computational cost is reduced from 75.2 to 7.1 GFLOPs. Overall, the consistent results across both datasets demonstrate that the proposed architecture preserves strong dish-level representations at a substantially lower computational cost, achieving a favorable balance between recognition performance and computational efficiency.

3.5 Ablation study

We conduct ablation experiments to evaluate the effects of EMA and the DFIS on dish classification and food ingredient recognition. We also implemented Micro-mAP (mean average precision) and Macro-mAP for ingredient recognition to complement the threshold-dependent F1 metrics.

Table 2

Performance comparison of different methods on the dish classification task. Best and second-best results are shown in bold and underline, respectively.

Method	Input Size	#Params	FLOPs	Vireo Food-172		ISIA Food-200	
				Top-1 accuracy	Top-5 accuracy	Top-1 accuracy	Top-5 accuracy
VGG ⁽²⁶⁾	224×224	143.6M	85.8G	80.41	94.59	59.05	86.53
DenseNet161 ⁽²⁷⁾	224×224	60.2M	46.4G	86.93	97.17	62.62	88.28
Swin-B ⁽¹⁷⁾	224×224	88.5M	15.4G	89.15	98.02	—	—
VOLO-D3 ⁽²⁸⁾	224×224	86.0M	19.3G	90.95	<u>98.70</u>	—	—
IG-CMAN ⁽⁹⁾	224×224	389.2M	124.4G	90.63	98.40	67.47	91.75
SLCI-T ⁽¹²⁾	224×224	46.1M	11.6G	90.70	98.05	70.16	92.34
SLCI-B ⁽¹²⁾	384×384	121.0M	75.2G	91.58	98.75	70.92	93.21
ViT-Base-p16 ⁽¹¹⁾	224×224	94.4M	17.4G	88.99	97.75	68.00	91.02
MobileNetV3-Large ⁽¹⁸⁾	224×224	12.19M	0.5G	87.40	97.10	63.04	88.74
GhostNetV2 ⁽¹⁹⁾	224×224	15.02M	1.3G	87.69	97.36	63.50	89.09
EfficientNet-Lite4 ⁽²⁰⁾	224×224	22.0M	2.6G	87.93	97.35	66.22	90.40
Proposed	224×224	41.5M	7.1G	<u>91.22</u>	98.44	<u>70.43</u>	<u>92.47</u>

As shown in Table 3, applying EMA alone improves most evaluation metrics on both datasets, although the gains are relatively modest. Introducing DFIS results in a slight decrease in Top-1 accuracy for dish classification while consistently improving both F1 and mAP metrics for ingredient recognition. In particular, the gains in Micro-mAP and Macro-mAP range from 2.46 to 5.51 percentage points. This trend is consistent with the intended behavior of DFIS, which restricts the backward effect of the dish-classification loss on the ingredient-specific representation. When EMA and DFIS are jointly employed, the model achieves the best performance across all evaluation metrics on both datasets. This suggests that the improved training stability provided by EMA complements the backward-decoupled design of DFIS, yielding a better overall balance between the two tasks.

3.6 Edge deployment and prototype validation

To further evaluate the deployment feasibility of the proposed architecture in practical edge-computing environments, we integrate the complete recognition pipeline into a prototype dietary sensing system. As illustrated in Fig. 3, tray images are first captured by the image sensor and preprocessed before being fed into the proposed lightweight model, which predicts both the dish

Table 3
Ablation study on Vireo Food-172 and ISIA Food-200.

EMA	DFIS	Top-1 accuracy	Top-5 accuracy	Micro-F1	Macro-F1	Micro-mAP	Macro-mAP
Vireo Food-172							
V	V	90.16	97.88	77.31	60.95	82.40	65.60
		90.39	97.99	77.55	61.34	84.63	65.69
		88.49	97.52	78.55	64.28	85.14	68.06
V	V	91.22	98.44	81.41	68.46	88.31	71.86
ISIA Food-200							
V	V	66.60	89.41	64.47	63.84	69.72	67.19
		68.22	90.51	66.37	66.02	71.79	69.04
		66.23	90.15	68.29	66.86	75.23	72.19
V	V	70.43	92.47	72.78	71.48	80.80	77.40

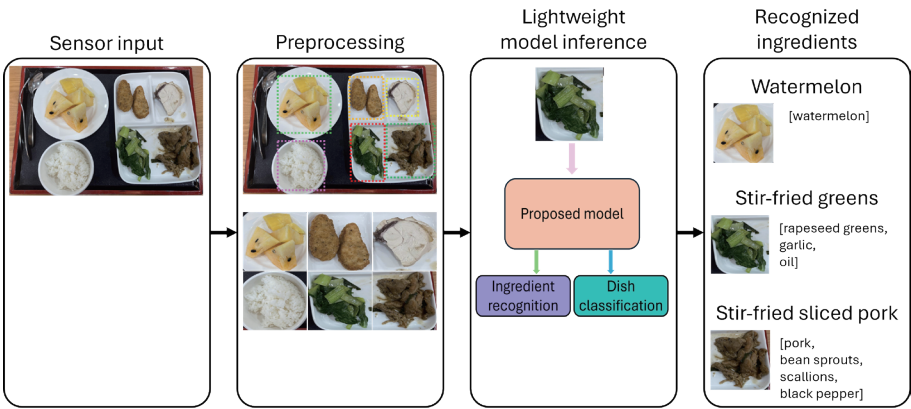


Fig. 3. (Color online) Operational pipeline of dietary sensing system.

category and the ingredient composition. This establishes an end-to-end pipeline from image sensing and recognition to dietary information output. The prototype therefore extends the evaluation beyond offline benchmarks and enables the direct assessment of the model on practical edge hardware.

We further benchmark the inference efficiency of the proposed model on an NVIDIA Jetson Orin Nano using PyTorch and the official SS2D implementation. To isolate the computational efficiency of the model itself, the reported measurements exclude image preprocessing and other system components. With a batch size of 1, corresponding to the interactive inference setting, the proposed model achieves 25.8 fps; when the batch size is increased to 32, the throughput reaches 52.95 fps. Since the target application performs event-driven recognition after tray placement rather than continuous high-frame-rate video analysis, the batch-1 inference performance is sufficient to support responsive interactive dietary recognition. These results demonstrate the practical feasibility of deploying the proposed architecture on low-power edge devices while maintaining competitive recognition performance.

4. Discussion

4.1 Analysis of long-tail robustness in ingredient recognition

To further analyze the recognition capability of the proposed method under long-tailed ingredient distributions, we group the ingredient labels according to their occurrence frequencies in the Vireo Food-172 training set. Unlike ISIA Food-200, which provides ingredient annotations at the dish-category level, Vireo Food-172 annotates the ingredients actually present in each image and therefore better reflects image-level ingredient distributions. As shown in Fig. 4, the 353 ingredient classes exhibit a pronounced long-tailed distribution. On the basis of K-means clustering, the ingredient classes are partitioned into Head, Medium, and Tail groups containing 59, 186, and 108 classes, respectively. Although the Head group accounts for only 16.7% of all ingredient classes, it contains 66.85% of the annotated instances. In contrast, the Tail group comprises 30.6% of the classes but only 1.68% of the annotated instances, with the rarest

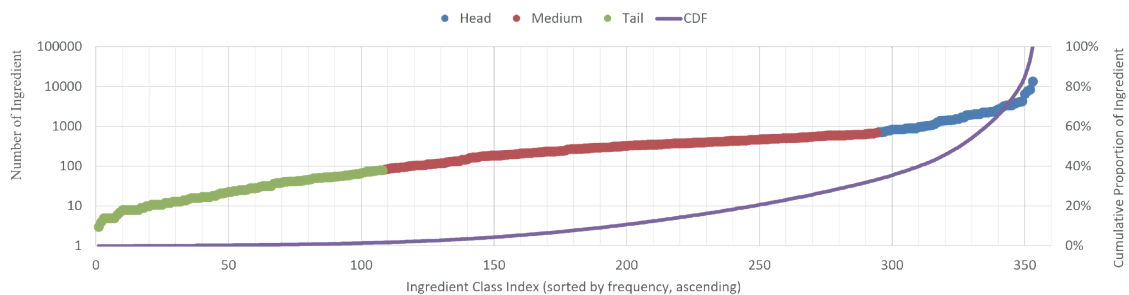


Fig. 4. (Color online) Long-tailed ingredient distribution in Vireo Food-172. CDF: cumulative distribution function.

ingredient appearing in only three training images. These statistics confirm the severe long-tailed nature of ingredient recognition in Vireo Food-172.

As shown in Table 4, all evaluated settings exhibit a clear performance degradation on Tail classes, indicating that low-frequency ingredients remain substantially more challenging to recognize. Notably, the gains introduced by DFIS increase as ingredient frequency decreases. On Vireo Food-172, incorporating DFIS improves Micro-mAP by 3.12, 4.19, and 9.96 percentage points for the Head, Medium, and Tail groups, respectively, while the corresponding Macro-mAP gains are 3.77, 5.18, and 8.98 percentage points. The improvement on Tail classes is therefore approximately 2.4–3.2× larger than that on Head classes, indicating that DFIS provides greater benefits for data-scarce ingredients.

This behavior is consistent with the design of DFIS. Through its forward-shared and backward-decoupled formulation, DFIS preserves the semantic contribution of ingredient representations to dish classification during forward propagation while using stop-gradient to reduce interference from the dish-classification loss on ingredient-specific representations during backpropagation. This decoupling is particularly beneficial for Tail classes, whose feature representations are more difficult to learn because of limited training samples. Table 4 also shows improvements across all ingredient-frequency groups on ISIA Food-200, with the largest Micro-mAP gain observed for Tail classes. However, the increase in gain toward Tail classes is less pronounced than that on Vireo Food-172.

4.2 Multitask weighting and gradient decoupling

To further examine the relationship between multitask weighting and DFIS, we compare fixed weighting, homoscedastic uncertainty weighting (UW)⁽²⁹⁾ and GradNorm⁽³⁰⁾ with and without DFIS. As shown in Table 5, on Vireo Food-172, UW without DFIS achieves the highest overall Micro-mAP, whereas its advantage is concentrated in the Head classes: on Tail classes, it falls below both DFIS-equipped counterparts. On ISIA Food-200, the aggregate ordering differs, but the Tail-class trend is consistent.

These results indicate that aggregate metrics may obscure degradation on low-frequency ingredients. Unlike task weighting, which adjusts the relative contribution of task losses, DFIS directly constrains the backward effect of the dish-classification loss on ingredient-specific representations. Accordingly, its benefits are more evident on Tail classes. Notably, incorporating DFIS consistently improves Tail-class performance across different task-weighting strategies, indicating that its effectiveness is complementary to specific weighting schemes.

Table 4
Long-tail ingredient recognition performance with and without DFIS.

Dataset	DFIS	Micro-mAP				Macro-mAP			
		Head	Medium	Tail	Total	Head	Medium	Tail	Total
Vireo Food-172	V	84.64	86.53	35.00	84.63	81.85	78.43	34.91	65.69
		87.76	90.72	44.96	88.31	85.62	83.61	43.89	71.86
ISIA Food-200	V	73.21	72.60	69.03	71.79	71.93	69.64	67.98	69.04
		82.02	80.87	78.41	80.80	80.91	77.70	76.30	77.40

Table 5

Effects of task-weighting strategies and DFIS on ingredient recognition. Best and second-best results are shown in bold and underline, respectively.

Strategy	DFIS	Micro-mAP				Macro-mAP			
		Head	Medium	Tail	Total	Head	Medium	Tail	Total
Vireo Food-172									
UW	V	88.45	<u>90.78</u>	43.73	88.70	86.22	83.39	42.18	71.25
		87.80	90.45	44.00	88.16	85.60	83.33	42.82	71.32
GradNorm	V	<u>88.22</u>	90.34	41.66	88.37	85.83	82.69	40.46	70.29
		87.94	90.91	45.27	<u>88.43</u>	<u>85.85</u>	83.83	44.33	72.08
Fixed (proposed)	V	84.64	86.53	35.00	84.63	81.85	78.43	34.91	65.69
		87.76	90.72	<u>44.96</u>	88.31	85.62	<u>83.61</u>	<u>43.89</u>	<u>71.86</u>
ISIA Food-200									
UW	V	81.81	80.73	78.29	80.68	80.75	77.62	76.09	77.19
		82.57	81.39	79.12	81.36	81.42	78.15	77.08	77.97
GradNorm	V	77.21	75.48	72.08	75.5	76.05	72.66	70.87	72.12
		81.87	80.69	78.15	80.62	80.71	77.49	75.96	77.06
Fixed (proposed)	V	73.21	72.6	69.03	71.79	71.93	69.64	67.98	69.04
		<u>82.02</u>	<u>80.87</u>	<u>78.41</u>	<u>80.80</u>	<u>80.91</u>	<u>77.70</u>	<u>76.30</u>	<u>77.40</u>

DFIS also narrows the spread of overall Micro-mAP across the three strategies from 4.07 to 0.27 points on Vireo Food-172 and from 8.89 to 0.74 on ISIA Food-200. Within this band, the best setting is dataset-dependent, and fixed weighting with DFIS remains within 0.8 points of the best result in every reported metric. We therefore adopt it for its consistent cross-dataset behavior and because it requires no additional parameters or per-dataset tuning.

5. Conclusions

In this work, we propose a lightweight multitask food recognition framework based on VMamba-Tiny for joint dish classification and ingredient recognition. CLAB integrates texture and semantic information across different feature levels to construct multiscale ingredient representations, while DFIS adopts a forward-shared and backward-decoupled design to preserve cross-task feature sharing while restricting interference from the dish-classification loss on ingredient-specific representations. Experimental results on Vireo Food-172 and ISIA Food-200 demonstrate that the proposed method maintains competitive performance on both tasks with only 7.1 GFLOPs, achieving a favorable balance between recognition performance and computational efficiency.

Further long-tail analysis shows that the gains introduced by DFIS become more pronounced as ingredient frequency decreases, with particularly substantial improvements on Tail classes. The multitask weighting experiments further indicate that task weighting and gradient decoupling are complementary rather than interchangeable, as they operate at different levels of the optimization process. Finally, the proposed model achieves 25.8 fps on an NVIDIA Jetson Orin Nano with a batch size of 1, supporting its practical use as the inference component of a vision-sensor-based dietary monitoring system on resource-constrained edge platforms. Future work will extend the sensing prototype to multiview image acquisition to mitigate occlusion in complex dishes.

Acknowledgments

This work was supported by the Taiwan Institute of Sports Science (Grant Nos. TISS-2024-002 and TISS-2025-009).

References

- 1 S. Heaney, H. O'Connor, S. Michael, J. Gifford, and G. Naughton: *Int. J. Sport Nutr. Exerc. Metab.* **21** (2011) 248. <https://doi.org/10.1123/ijsnem.21.3.248>
- 2 T. Osadchiy, I. Poliakov, P. Olivier, M. Rowland, and E. Foster: *J. Med. Internet Res.* **22** (2020) e13266. <https://doi.org/10.2196/13266>
- 3 M. K. Esquivel and C. P. Lozano: *Am. J. Lifestyle Med.* **18** (2024) 475. <https://doi.org/10.1177/15598276241240690>
- 4 J. S. Shim, K. Oh, and H. C. Kim: *Epidemiol. Health* **36** (2014) e2014009. <https://doi.org/10.4178/epih/e2014009>
- 5 P. Ma, C. P. Lau, N. Yu, A. Li, P. Liu, Q. Wang, and J. Sheng: *Food Res. Int.* **147** (2021) 110437. <https://doi.org/10.1016/j.foodres.2021.110437>
- 6 C.-H. Hsia, S.-C. Yen, and J.-H. Jang: *Sens. Mater.* **31** (2019) 1803. <https://doi.org/10.18494/SAM.2019.2351>
- 7 C.-H. Hsia, T.-Y. Lin, J.-L. Lin, H. Prasetyo, S.-L. Chen, and H.-W. Tseng: *Sens. Mater* **32** (2020) 3235. <https://doi.org/10.18494/SAM.2020.2862>
- 8 J. Chen and C.-W. Ngo: *Proc. 24th ACM Int. Conf. Multimedia* (2016) 32. <https://doi.org/10.1145/2964284.2964315>
- 9 W. Min, L. Liu, Z. Luo, and S. Jiang: *Proceedings of the 27th ACM Int. Conf. Multimedia* (2019) 1331. <https://doi.org/10.1145/3343031.3350948>
- 10 M. Luo, W. Min, Z. Wang, J. Song, and S. Jiang: *IEEE Trans. Image Process.* **32** (2023) 5509. <https://doi.org/10.1109/TIP.2023.3318958>
- 11 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby: *arXiv preprint arXiv:2010.11929* (2020). <https://doi.org/10.48550/arXiv.2010.11929>
- 12 Y. Liu, W. Min, S. Jiang, and Y. Rui: *IEEE Trans. Image Process.* **33** (2024) 2572. <https://doi.org/10.1109/TIP.2024.3374211>
- 13 Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu: *Adv. Neural Inf. Process. Syst.* **37** (2024) 103031. <https://doi.org/10.52202/079017-3273>
- 14 Y.-T. Chen, J. Liao, L.-Y. Ke, and C.-H. Hsia: *2025 Int. Symp. Intelligent Signal Processing and Communication Systems (ISPACS)* (2025) 1. <https://doi.org/10.1109/ISPACS68724.2025.11383414>
- 15 L.-Y. Ke, K. Shen, and C.-H. Hsia: *2025 Int. Symp. Intelligent Signal Processing and Communication Systems (ISPACS)* (2025) 1. <https://doi.org/10.1109/ISPACS68724.2025.11382925>
- 16 C.-A. Chen, S. K. A. Bosshart, L.-Y. Ke, and C.-H. Hsia: *2025 Second Int. Conf. Artificial Intelligence for Medicine, Health and Care (AIXMHC)* (2025) 262. <https://doi.org/10.1109/AIXMHC65380.2025.00056>
- 17 Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo: *Proc. IEEE/CVF Int. Conf. Computer Vision* (2021) 10012. <https://doi.org/10.1109/ICCV48922.2021.00986>
- 18 A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam: *Proc. IEEE/CVF Int. Conf. Computer Vision* (2019) 1314. <https://doi.org/10.1109/ICCV.2019.00140>
- 19 Y. Tang, K. Han, J. Guo, C. Xu, C. Xu, and Y. Wang: *Adv. Neural Inf. Process. Syst.* **35** (2022) 9969. <https://doi.org/10.52202/068431-0724>
- 20 M. Tan and Q. Le: *arXiv preprint arXiv:1905.11946* (2019). <https://doi.org/10.48550/arXiv.1905.11946>
- 21 Z. Xu, R. Liu, S. Yang, Z. Chai, and C. Yuan: *2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)* (2023) 15793. <https://doi.org/10.1109/CVPR52729.2023.01516>
- 22 T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor: *Proc. IEEE/CVF Int. Conf. Computer Vision* (2021) 82. <https://doi.org/10.1109/ICCV48922.2021.00015>
- 23 F. Zhou, S. Huang, and Y. Xing: *Proc. AAAI Conf. Artificial Intelligence* **35** (2021) 3572. <https://doi.org/10.1609/aaai.v35i4.16472>
- 24 H. D. Nguyen, X.-S. Vu, and D.-T. Le: *Proc. AAAI Conf. Artificial Intelligence* **35** (2021) 9092. <https://doi.org/10.1609/aaai.v35i10.17098>
- 25 X.-S. Vu, D.-T. Le, C. Edlund, L. Jiang, and H. D. Nguyen: *Proc. 28th ACM Int. Conf. Multimedia* (2020) 2299. <https://doi.org/10.1145/3394171.3414047>

- 26 K. Simonyan and A. Zisserman: arXiv preprint arXiv:1409.1556 (2014). <https://doi.org/10.48550/arXiv.1409.1556>
- 27 G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2017) 4700. <https://doi.org/10.1109/CVPR.2017.243>.
- 28 L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan: IEEE Trans. Pattern Analysis and Machine Intelligence **45** (2022) 6575. <https://doi.org/10.1109/TPAMI.2022.3206108>
- 29 A. Kendall, Y. Gal, and R. Cipolla: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2018) 7482. <https://doi.org/10.1109/CVPR.2018.00781>
- 30 Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich: Proc. 35th Int. Conf. Machine Learning (2018) 794. <https://doi.org/10.48550/arXiv.1711.02257>