OmniFood8K: Single-Image Nutrition Estimation via Hierarchical Frequency-Aligned Fusion

Dongjian Yu[1] Weiqing Min[2] Qian Jiang[1] Xing Lin[1] Xin Jin[1*] Shuqiang Jiang[2]

[1]School of Software, Yunnan University, China
[2]State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, China

Abstract

Accurate estimation of food nutrition plays a vital role in promoting healthy dietary habits and personalized diet management. Most existing food datasets focus on Western cuisines, with limited coverage of Chinese dishes, leading to limitations in accurate nutritional estimation for Chinese meals. Moreover, many state-of-the-art nutrition prediction methods rely on depth sensors, restricting their applicability in daily scenarios. To address these limitations, we introduce OmniFood8K, a comprehensive multimodal dataset comprising 8,036 food scenes with detailed nutritional annotations and multi-view images for each scene. In addition, to enhance models' capability in nutritional prediction, we construct NutritionSynth-115K, a large-scale synthetic dataset that introduces compositional variations while preserving precise nutritional labels. Moreover, we propose an end-to-end framework to predict nutritional information from a single RGB image. We first predict a depth map from a single RGB image, then refine it using our Scale-Shift Residual Adapter (SSRA), which enforces global scale consistency and preserves local structural details. Second, the Frequency-Aligned Fusion Module (FAFM) hierarchically fuses RGB and adapted depth features, aligning multi-modal representations in the frequency domain across layers. Third, the Mask-based Prediction Head (MPH) emphasizes key ingredient regions via dynamic channel selection, improving prediction accuracy. Extensive experiments on multiple datasets demonstrate that our method outperforms existing approaches, providing a practical solution for daily dietary assessment. Project homepage: https://yudongjian.github.io/OmniFood8K-food/

1. Introduction

With the improvement of living standards and the diversification of dietary habits, nutritional assessment of food has become a key research focus in both public health and personal health management [20, 40]. Accurate estimation of the energy and nutrient content of food plays a vital role in preventing chronic diseases such as obesity, diabetes, and cardiovascular disorders [14, 32]. However, traditional nutritional assessment methods [26, 33, 42] often rely on manual recording or expert estimation, which are not only time-consuming and labor-intensive but also prone to subjective bias. This makes accurate and consistent evaluation difficult to achieve, particularly in daily dietary applications [7, 27].

	Mass	Calories	Protein	Fat	Carb.
Total	206.0	2729.5	19.0	55.9	19.7
Garlic scape	102.9	308.7	2.0	0.1	15.8
Pork belly	127.8	2088.2	16.8	47.2	3.1
Vegetable oil	8.5	314.1	0	8.5	0
...	...	...	...	...	...

	Mass	Calories	Protein	Fat	Carb.
Total	300.0	1652.4	48.9	17.2	11.2
Red pepper	113.3	201.8	1.3	0.1	10.0
Chicken Breast	193.4	955.3	47.5	3.6	1.1
...	...	...	...	...	...

	Mass	Calories	Protein	Fat	Carb.
Total	255.0	1577.8	7.3	19.3	43.5
Rice	113.0	560.4	2.9	0.3	29.2
Pumpkin	69.5	568.8	0.6	14.4	1.4
Prawns	32.0	60.7	1.5	0.2	1.5
...	...	...	...	...	...

Figure 1. Representative examples from the OmniFood8K dataset.

Recent advances in computer vision and multimodal learning have significantly advanced the development of automated food nutrition assessment [29–31]. Deep learning techniques have shown great potential in automatically recognizing, segmenting, and analyzing food images to estimate food mass, volume, and nutritional composition [9, 41]. Moreover, leveraging multimodal features such as textual descriptions and sensor data can further enhance estimation accuracy and robustness, facilitating personalized nutrition monitoring and healthcare systems [37].

Despite these advances, current methods still encounter two fundamental limitations that constrain their practical usability. **(1) Data Limitation.** Existing datasets are heavily biased toward Western cuisines, with limited representation of Chinese food. The inherent diversity, complex ingredient composition, and non-standardized preparation of Chinese food make reliable annotation and quantitative nutritional analysis particularly challenging. **(2) Algorithmic Limitation.** Many advanced approaches depend on depth information for accurate nutrition estimation. Nevertheless, in most real-world contexts, food photographs are taken with conventional RGB cameras rather than depth sensors, thereby restricting the practical adoption of these methods.

To address these two limitations, this work makes two key contributions. First, we introduce OmniFood8K, a comprehensive multimodal food nutrition dataset designed for real-world scenarios, as shown in Figure 1. The dataset covers the entire food preparation process, including raw ingredient images with mass information, structured recipe descriptions, full cooking videos, and multi-view images of the finished dishes, all paired with detailed nutritional annotations, as illustrated in Figure 2. Unlike existing datasets, OmniFood8K employs a unified, process-oriented design that preserves causal links from ingredient preparation to final presentation and provides comprehensive nutritional annotations. In addition, to further improve the model's nutritional prediction capability, we construct NutritionSynth-115K, a large-scale synthetic dataset with compositional diversity and precise nutritional annotations.

Second, we propose an end-to-end framework that predicts nutritional information directly from a single RGB image. Compared with depth-sensor-based approaches, our method is more generalizable and scalable for real-world dietary assessment. Specifically, we first employ a pretrained depth estimation model [45] to predict the depth information of food. To correct scale bias and local structural errors, we design a Scale-Shift Residual Adapter (SSRA) for consistent global and local calibration. We then employ a Frequency-Aligned Fusion Module (FAFM) to hierarchically fuse RGB and adapted depth features, aligning multi-modal representations in the frequency domain for enhanced cross-modal learning. Finally, a Mask-based Prediction Head (MPH) leverages dynamic channel selection and region-aware attention to emphasize key ingredient regions, improving nutrition prediction accuracy.

Overall, our contributions are summarized as follows:

- We introduce OmniFood8K, a comprehensive multimodal food dataset with detailed nutritional annotations, covering ingredients, recipes, cooking videos, and multi-view food images. Additionally, we construct NutritionSynth-115K, a large-scale synthetic dataset featuring diverse food compositions and precise nutritional annotations.
- We propose an end-to-end framework for predicting food nutritional information from a single RGB image. It incorporates a Scale-Shift Residual Adapter (SSRA) to refine depth estimation and achieve consistent calibration of both global scale and local geometry.
- We further design a Frequency-Aligned Fusion Module (FAFM) to hierarchically fuse RGB and depth features in the frequency domain, and a Mask-based Prediction Head (MPH) that dynamically selects informative channels, thereby enhancing the accuracy of nutritional prediction.
- Extensive experiments on multiple datasets demonstrate that our method outperforms state-of-the-art approaches, validating its effectiveness and scalability for food nutrition assessment.

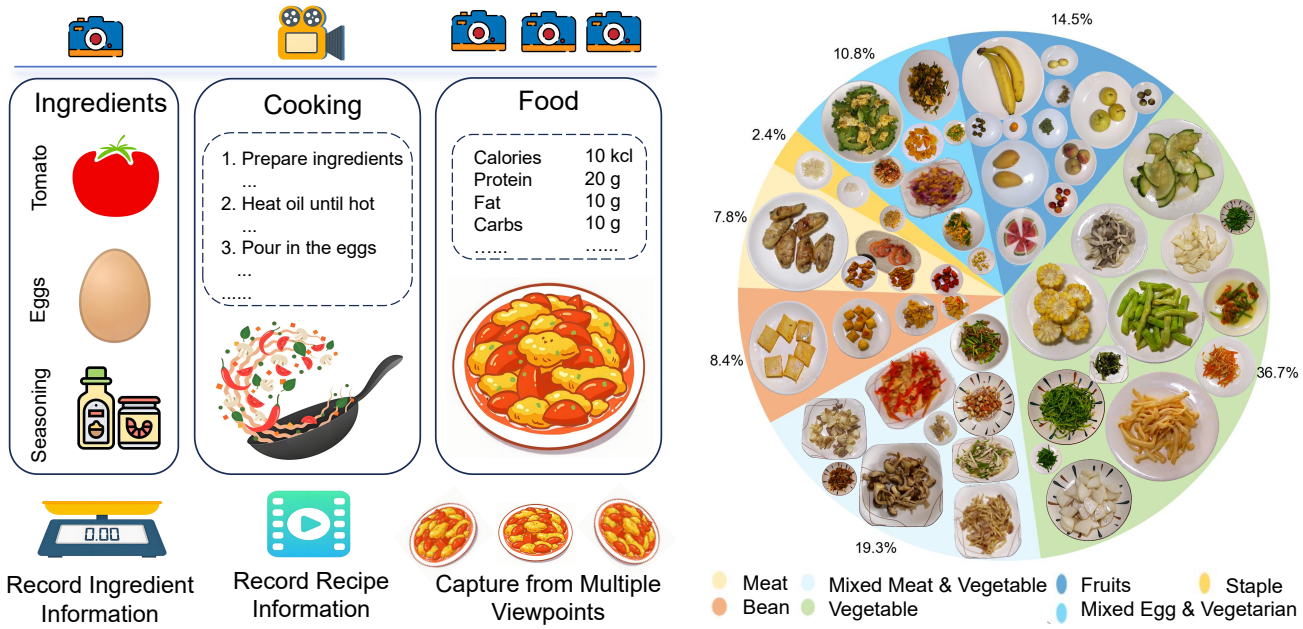
2. Related work

2.1. Food datasets

Recent advances in food computing have led to the release of numerous publicly available datasets [8], which have substantially promoted research across multiple domains, such as food recognition, cross-modal analysis, and nutritional estimation. Representative datasets such as Food-101 [3], VIREO Food-172 [4], ISIA Food-500 [22], and Food2K [23] cover diverse food categories and have been widely adopted for recognition and classification tasks. Additionally, Recipe1M [28] aligns food images with textual recipes, serving as a fundamental resource for cross-modal retrieval and recipe generation tasks. Nutrition5k [37] includes detailed nutritional annotations and has been instrumental in advancing research on food nutrition assessment. However, its primary focus on Western dishes, with limited coverage of complex cuisines such as Chinese food, constrains its general applicability. FoodSeg103 [44] is a dataset designed for food image segmentation, aiming to facilitate fine-grained understanding of food components. MetaFood3D [5] is a multimodal dataset comprising 637 3D food objects across 108 categories, accompanied by detailed nutritional information and diverse visual modalities. The FastFood dataset [27] collects food images and their corresponding nutritional information from official fast-food brand websites, providing valuable data for nutrition analysis and related studies.

2.2. Food nutrition assessment methods

Traditional nutrition assessment methods are time-consuming, labor-intensive, and often limited in predictive accuracy. With advances in computer vision, AI-based approaches for nutrition prediction have increasingly emerged [9, 13, 29, 37, 39]. NR *et al.* [25] employed CNN to automatically recognize food and predict its nutritional information. Swin-Nutrition [29] introduced an efficient,



(a) Food Data Collection Process **(b) Distribution of Food Categories**

Figure 2. Overview of the OmniFood8K dataset: data collection process and category distribution.

non-destructive AI-based framework for accurately predicting multiple food nutrient components using Swin Transformer features. Vinod *et al.* [38] integrated energy density maps with depth information to effectively improve calorie estimation accuracy. Shao *et al.* [31] proposed estimating food energy through 3D shape reconstruction from a single-view image, achieving promising results. Feng *et al.* [7] proposed a method that incorporates ingredient information to complement visual features, thereby improving the accuracy of nutrition prediction and achieving promising results. RoDE [13] introduced a multi-expert framework to improve the accuracy and efficiency of large-scale food multi-modal models for tasks such as nutrition estimation. Boyuan Ma [18] *et al.* proposed FBFPN, a network that fuses RGB and depth images via a bidirectional feature pyramid to improve nutrition estimation accuracy.

3. OmniFood8K dataset construction

3.1. Motivation

Existing nutrition datasets primarily focus on Western foods and lack sufficient representation of complex cuisines such as Chinese dishes, limiting their applicability to accurate nutritional estimation across diverse dietary contexts. To address these limitations, we introduce OmniFood8K, a comprehensive multimodal dataset for practical nutrition assessment, supporting diverse tasks such as nutrition estimation, food image generation, ingredient recognition, and recipe generation.

3.2. Data collection

The dataset comprises 8,036 food scenes, representing a wide range of commonly consumed Chinese dishes. Data were collected from real-world cooking scenarios, including raw ingredients, structured recipes, cooking videos, and final dish presentations. The data acquisition process, illustrated in Figure 2, follows a structured and systematic workflow. Raw ingredients were first weighed and photographed, followed by video recording of the cooking process and textual documentation of the recipes. Finally, the prepared food was captured simultaneously from six viewpoints at a horizontal height of 50 cm, approximating the typical cubit height of an adult for practical daily use [34]. Two cameras were positioned directly above the food, while the remaining four were placed at the front, back, left, and right sides.

3.3. Annotation

For each food scene, ingredients were first photographed and their corresponding weights recorded to enable precise computation of nutritional information. Based on these measurements, detailed nutritional annotations were generated for each scene, including calorie, protein, fat, and carbohydrate contents, which were subsequently validated using standardized nutritional databases. Recipes were doc-

Table 1. Comparison between the proposed dataset and existing food datasets.

Dataset	Categories	Size	Ingredient Image	Food Image	Multi-view Food Image	Recipe	Cooking Video	Nutritional Info
Food101 [3]	101	101,000	-	✓	-	-	-	-
VIREO Food-172 [4]	172	110,241	-	✓	-	✓	-	-
Recipe1M [28]	1480	1,047,000	-	✓	-	✓	-	-
Yummly-66K [19]	10	66,615	-	✓	-	✓	-	-
ISIA Food-200 [21]	200	197,323	-	✓	-	-	-	-
ISIA Food-500 [22]	500	399,726	-	✓	-	-	-	-
Food2K [23]	2000	1,036,564	-	✓	-	-	-	-
YouCook2 [50]	89	2000	✓	✓	-	✓	✓	-
Menu-Match [1]	41	646	-	✓	-	-	-	✓
Fang *et al.* [6]	3	45	-	✓	-	-	-	✓
Nutrition5k [37]	-	5066	-	✓	✓	-	-	✓
MetaFood3D [5]	108	637	-	✓	✓	-	-	✓
OmniFood8K (**Ours**)	165	8,036	✓	✓	✓	✓	✓	✓
NutritionSynth-115K (**Ours**)	165	115,000	✓	✓	-	✓	✓	✓

Table 2. Data types and corresponding applications in the OmniFood8K dataset.

Data Type	**Applications**
Ingredient images	Ingredient recognition, detection, and quantity estimation.
Recipe descriptions	Recipe generation, cross-modal retrieval, and cooking reasoning.
Cooking process videos	Action recognition, cooking stage understanding, and multimodal alignment with recipes.
Multi-view images of food	Food recognition, detection, image synthesis, and 3D-aware food understanding.
Food nutritional data	Nutrition prediction, dietary assessment, and health-oriented analysis.

umented using structured textual descriptions, while the entire cooking process was comprehensively recorded on video. Additionally, key steps in each recipe were captured as corresponding images to support multimodal analysis.

3.4. Dataset characteristics

OmniFood8K provides nutritional annotations for 8,036 food scenes, accompanied by diverse multimodal data, whose types and corresponding applications are summarized in Table 2.

By spanning the entire food preparation pipeline, OmniFood8K serves as a comprehensive resource for research in food computing and nutritional analysis. Table 1 presents a comparison of OmniFood8K with existing food datasets across multiple characteristics. Notably, OmniFood8K is the first dataset to cover the entire food preparation process, from raw ingredients to finished food. Its main features are summarized as follows:

- **Multimodal**: OmniFood8K includes images, structured recipe descriptions, cooking videos, and detailed nutritional annotations.
- **Full pipeline coverage**: OmniFood8K spans the complete cooking process, from annotated ingredient images and weights to multi-view final food images, accompanied by recipe descriptions and cooking videos.
- **Causal relationship**: OmniFood8K preserves the causal relationship from raw ingredients to finished dishes through recipe descriptions and cooking videos.
- **Accuracy:** Unlike other datasets [13, 27] that rely on web-scraped labels or LLM-generated annotations, OmniFood8K provides labels based on real-world weight measurements and standardized nutritional database calculations, ensuring higher accuracy.

3.5. NutritionSynth-115K

We performed foreground segmentation on OmniFood8K food images to isolate individual items. These segmented items were subsequently cropped or recombined according to their original proportions, allowing the synthesized scenes to faithfully reflect the nutritional composition of each meal. Through this process, we constructed a large-scale and controllable synthetic dataset, named NutritionSynth-115K, which contains 115,000 images with precise nutritional annotations.

4. Method

The overall architecture of the proposed method is illustrated in Figure 3. A single RGB image is first fed into a depth estimation model to produce an initial depth map, which is subsequently refined using the proposed Scale-

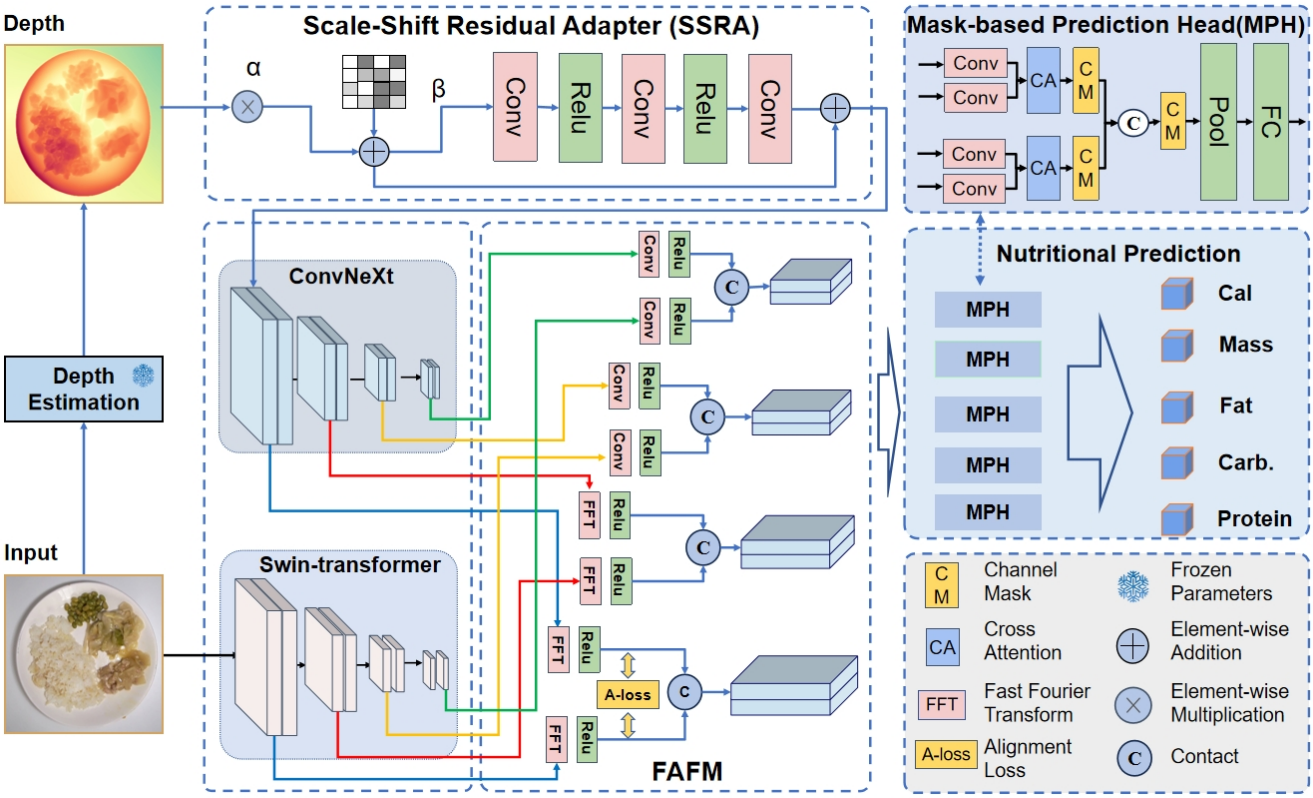


Figure 3. Overview of the proposed method. The figure illustrates the overall pipeline of our method, consisting of three proposed modules: the Scale-Shift Residual Adapter (SSRA), Frequency-Aligned Fusion Module (FAFM), and Mask-based Prediction Head (MPH).

Shift Residual Adapter (SSRA) module for both global and local corrections. The RGB image and the refined depth map are then used to extract multi-scale features, which are fused by the Frequency-Aligned Fusion Module (FAFM) through frequency-domain alignment of multimodal representations. Finally, the Mask-based Prediction Head (MPH) predicts the nutritional information, forming an end-to-end framework for nutrition estimation.

4.1. Scale-Shift Residual Adapter (SSRA)

To ensure practical applicability in daily scenarios, we focus on nutrition estimation from a single RGB image. Since RGB images inherently lack explicit depth cues, we employ a monocular depth estimation model [45] to infer depth information. To further mitigate estimation errors, we design a lightweight Scale-Shift Residual Adapter (SSRA) that efficiently refines both global and local structures.

Specifically, at the global level, we introduce a learnable scaling factor α and shift parameter β to align the overall distribution of the predicted depth map d_{mono}:

$$d_{global} = \alpha \cdot d_{mono} + \beta. \tag{1}$$

This step effectively mitigates the common issues of scale ambiguity and offset errors in monocular depth prediction. Building upon this, a shallow convolutional residual refinement module $f_\theta(\cdot)$ is employed to enhance local structural details and correct geometric distortions. The refinement process can be expressed as:

$$d_{res} = f_\theta(d_{global}). \tag{2}$$

The final adapted depth is then obtained as:

$$d_{out} = d_{global} + d_{res}. \tag{3}$$

By combining global alignment with local refinement, the proposed SSRA ensures consistency in overall depth estimation while improving local precision, thus enhancing the accuracy of nutrition estimation tasks.

4.2. Frequency-Aligned Fusion Module (FAFM)

To effectively integrate complementary RGB and depth information while preserving feature consistency, we propose a Frequency-Aligned Fusion Module (FAFM). This module fuses multi-frequency features and enforces cross-modal consistency via an inter-modal alignment loss.

Given RGB and depth feature maps $r, d \in \mathbb{R}^{C\times H\times W}$, the module first applies a 2D Fast Fourier Transform (FFT) to obtain the frequency representations $R_f = \mathcal{F}(r)$ and

$D_f = \mathcal{F}(d)$. Based on a predefined frequency threshold τ, a low-frequency mask M_L is constructed to separate low-frequency (global structure) and high-frequency (local detail) components:

$$R_L = \mathcal{F}^{-1}(R_f \odot M_L), \quad R_H = \mathcal{F}^{-1}(R_f \odot (1 - M_L)), \tag{4}$$

$$D_L = \mathcal{F}^{-1}(D_f \odot M_L), \quad D_H = \mathcal{F}^{-1}(D_f \odot (1 - M_L)). \tag{5}$$

The high-frequency and low-frequency components from both modalities are then fused via element-wise addition:

$$F_H = R_H + D_H, \quad F_L = R_L + D_L. \tag{6}$$

The fused high-frequency and low-frequency features are concatenated along the channel dimension and passed through a learnable convolutional layer to adaptively compress and integrate them, producing the final cross-modal feature representation:

$$O = \text{Conv}([F_H, F_L]). \tag{7}$$

In this design, high-frequency components enhance texture and local details, while low-frequency components preserve global structure, achieving Multi-Frequency Fusion.

To enhance semantic consistency between RGB and depth features, we introduce an inter-modal alignment loss $\mathcal{L}_{align}$, defined as:

$$\mathcal{L}_{align} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{f}_i^r, \mathbf{f}_i^d)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{f}_i^r, \mathbf{f}_j^d)/\tau)}, \tag{8}$$

where $\mathbf{f}_i^r$ and $\mathbf{f}_i^d$ are the global feature vectors of the i-th sample in RGB and depth modalities, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, N is the batch size, and τ is a temperature parameter. This loss encourages features from the same sample across modalities to be close, while separating features from different samples, thus improving cross-modal representation for subsequent fusion.

4.3. Mask-based Prediction Head (MPH)

To fully leverage multi-scale features and emphasize key ingredient regions, we propose the Mask-based Prediction Head (MPH). This module takes feature maps from different sources as input, including multi-scale RGB features and semantic features. First, MPH applies convolutional operations and adaptive average pooling to each feature map to extract high-dimensional representations and unify the spatial dimensions to a fixed size, ensuring the controllability and stability of subsequent fusion operations. Next, MPH employs a Cross-Attention mechanism to enable information interaction between RGB and semantic features, allowing the model to capture correlations across different modalities. This is followed by Gated Fusion, which further reweights the fused features to highlight the most critical channels. To further enhance the focus on key ingredient regions, MPH introduces a dynamic channel mask (Channel Mask), which automatically selects the most informative channels, thereby preserving the features most relevant to the target nutritional information during fusion. Finally, the RGB and semantic features are integrated through a global fusion module and processed by global pooling and a fully connected layer to produce the predicted nutritional values.

4.4. Loss function

In the nutrient prediction task, we adopt a Dynamic Task-weighted Loss [46] for nutrient estimation, which dynamically adjusts task weights according to the current prediction difficulty. The nutrient prediction loss is formulated as:

$$\mathcal{L}_{\text{nutri}} = \sum_{i=1}^{n} w_i \cdot \text{PMAE}(\text{nutri}[i]), \tag{9}$$

where n is the number of nutrient components, w_i is the dynamically updated weight for the i-th nutrient task, and PMAE denotes the Percentage Mean Absolute Error. Specifically, each task i corresponds to a nutrient component such as Calorie, Mass, Fat, Carbohydrate, or Protein. The dynamic weights are updated according to the task difficulty via a KPI metric:

$$\begin{aligned} \text{KPI}_i &= \frac{1}{1 - \text{PMAE}(\text{nutri}[i])}, \\ w_i^{(t)} &= \alpha \cdot \text{KPI}_i^{(t)} + (1 - \alpha) \cdot w_i^{(t-1)}, \end{aligned} \tag{10}$$

where t denotes the training iteration, and α is a smoothing factor controlling the update momentum, which is empirically set to 0.3 in our experiments.

To enhance cross-modal representation consistency, we further introduce an inter-modal alignment loss $\mathcal{L}_{\text{align}}$ (Eq. 8), applied to fused RGB–Depth features. Finally, the total loss for training the nutrient prediction model is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{nutri}} + \lambda \, \mathcal{L}_{\text{align}}, \tag{11}$$

where λ balances the contribution of the inter-modal alignment loss. This formulation enables joint optimization of nutrient prediction and cross-modal feature alignment, thereby improving the model's predictive performance in multimodal fusion.

5. Experiments

5.1. Experimental setup

1) Evaluation metrics. The percentage of mean absolute error (PMAE) is adopted to evaluate the prediction performance of our method. This metric measures the relative

Table 3. Comparison of different backbones on the OmniFood8K dataset.

Input	Method	Calories	Mass	Fat	Carb.	Protein	Mean
RGB images	Inception V3 [35]	17.2	10.0	17.8	13.8	20.5	16.0
	ResNet-50 [10]	10.2	5.1	11.0	11.1	14.8	10.4
	ResNet-101 [10]	10.9	6.0	12.2	9.1	15.2	10.6
	DenseNet-121 [12]	12.1	7.4	13.7	10.5	18.2	12.4
	DenseNet-161 [12]	12.2	7.6	12.7	12.6	16.3	12.3
	EfficientNet [36]	27.1	10.4	28.3	38.9	40.0	28.9
	MobileNetV3 [11]	17.1	8.5	18.5	21.2	27.1	18.5
	Swin-Transformer [17]	27.7	15.1	29.6	42.6	41.0	31.2
	ConvNeXt [16]	10.1	5.6	10.8	9.5	13.3	9.8
	ConvNeXt V2 [43]	15.5	6.4	11.3	9.3	**12.6**	11.2
	Ours	**8.8**	**4.5**	**9.5**	**8.0**	**12.6**	**8.6**

Table 4. Comparison of our method and baselines on the Nutrition5k dataset.

Input	Method	Calories	Mass	Fat	Carb.	Protein	Mean
RGB images	Google-Nutrition-rgb [37]	26.1	18.8	34.2	31.9	29.5	29.1
	Coarse-to-Fine Nutrition [39]	24.1	19.4	36.0	32.1	33.5	29.0
	Swin-Nutrition [29]	16.2	13.7	24.9	21.8	25.4	20.4
	Portion-Nutrition [31]	15.8	-	-	-	-	-
	RoDE [13]	52.4	38.4	67.1	47.8	53.9	51.9
	DPF-Nutrition [9]	14.7	10.6	22.6	20.7	20.2	17.8
RGB-D images	CMX [48]	21.8	20.7	34.8	37.0	33.2	29.5
	HINet [2]	24.5	25.2	43.4	39.9	38.8	34.3
	CDINet [47]	21.1	20.4	37.1	37.1	32.8	29.7
	DEFNet [51]	32.7	34.2	48.9	40.3	43.8	39.9
	TriTransNet [15]	22.1	20.1	37.5	34.8	38.0	30.5
	Deliver [49]	29.5	25.9	48.3	47.7	46.1	39.5
	Google-Nutrition-rgbd [37]	18.8	18.9	**18.1**	23.8	20.9	20.1
	Domain Adaptation-Nutrition [38]	16.8	-	-	-	-	-
	RGB-D Net [30]	15.0	10.8	23.5	22.4	21.0	18.5
	IMIR-Net [24]	14.5	10.4	21.8	20.4	20.0	17.4
	FBFPN [18]	**14.0**	10.3	22.6	19.5	20.2	17.3
RGB images	**Ours**	14.1	**10.2**	21.0	**18.9**	**18.4**	**16.5**

deviation between predicted and true values, with smaller values indicating more accurate predictions. PMAE is defined as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_t - y_p|, \tag{12}$$

$$\text{PMAE} = \frac{\text{MAE}}{\frac{1}{N}\sum_{i=1}^{N} y_t} \times 100\%, \tag{13}$$

where y_t denotes the true value, y_p denotes the predicted value, and N denotes the total number of samples. A smaller value indicates higher prediction accuracy and better model performance.

2) Baseline methods. To evaluate the effectiveness of our proposed method, we conduct a comparison with recent state-of-the-art approaches. On the OmniFood8K dataset, we select representative models including [10–12, 17, 35, 36]. For the Nutrition5k dataset, we compared our method with recent approaches, including Google-Nutrition [37], Portion-Nutrition [31], Coarse-to-Fine Nutrition [39], Swin-Nutrition [29], DPF-Nutrition [9], RGB-D Net [30], Domain Adaptation-Nutrition [38], IMIR-Net [24], FBFPN [18]. We also compare our method with prior image fusion approaches [2, 15, 47–49, 51] .

Table 5. Ablation study of the proposed modules on the Nutrition5k dataset.

Baseline	FAFM	SSRA	MPH	Calories	Mass	Fat	Carb.	Protein	Mean
✓				14.8	11.3	23.6	20.1	19.7	17.9
✓	✓			14.4	**10.2**	21.7	19.7	19.2	17.0
✓	✓	✓		14.2	10.4	21.9	19.2	18.6	16.8
✓	✓	✓	✓	**14.1**	**10.2**	**21.0**	**18.9**	**18.4**	**16.5**

3) Implementation details. All models were implemented in PyTorch and trained on an NVIDIA A100 GPU. The Adam optimizer was used with a weight decay of 1e-5 to prevent overfitting. A cosine annealing scheduler adjusted the learning rate during training. Models were trained for 150 epochs with a batch size of 8. The OmniFood8K and Nutrition5k datasets are split into training and testing sets at ratios of 7:3 and 5:1 [9, 30], respectively.

5.2. Experimental results

To validate the effectiveness of the proposed method, we conduct experiments on multiple food nutrition estimation datasets. Table 3 presents the comparison with different baseline methods on our proposed OmniFood8K dataset. It can be observed that our method achieves the best performance across multiple metrics, with the lowest PMAE, indicating superior prediction accuracy. Table 4 reports the comparison results on the Nutrition5k dataset. Our method outperforms existing state-of-the-art approaches on the PMAE metric across multiple nutrient components, achieving the best overall performance. Notably, even when compared with a number of state-of-the-art methods that directly leverage RGB-D inputs, our method still achieves superior performance, further demonstrating its effectiveness and robustness.

As shown in Tables 3 and 4, our method achieves a lower PMAE on OmniFood8K than on Nutrition5k, demonstrating superior predictive accuracy. This improvement can be attributed to the larger scale of OmniFood8K, which contains 8,036 images (5,600 used for training) compared to Nutrition5k's 3,500 images (2,800 for training) [7, 30]. The roughly twofold increase in training samples exposes the model to more diverse food appearances and nutritional compositions, enhancing feature learning and prediction performance.

5.3. Ablation study

To evaluate the contribution of each component in our method, we conduct an ablation study on the Nutrition5k dataset. We gradually add the proposed components and observe the resulting performance changes. As shown in Table 5, the model's performance steadily improves with the inclusion of each component, demonstrating the effectiveness of our method.

Table 6. Evaluation of our method pre-trained on NutritionSynth-115K and fine-tuned on OmniFood8K.

Method	Calories	Mass	Fat	Carb.	Protein	Mean
Ours	8.8	**4.5**	9.5	8.0	12.6	8.6
+ Pre-trained	**8.0**	**4.5**	**9.0**	**7.4**	**10.0**	**7.8**

5.4. Pretraining on NutritionSynth-115K

To enhance the performance of our method on the OmniFood8K dataset, it is first pretrained on NutritionSynth-115K. This pretraining enables the model to capture rich food features and multimodal representations, providing a robust initialization for subsequent fine-tuning on OmniFood8K. As shown in Table 6, pretraining significantly improves predictive performance, demonstrating its effectiveness in enhancing generalization.

6. Conclusion

To enable practical and accessible nutritional estimation, we presented an end-to-end framework capable of predicting nutritional information directly from a single RGB image. In addition, we introduced OmniFood8K, a comprehensive dataset containing 8,036 real-world food scenes with multi-view images, ingredient weights, recipes, cooking videos, and detailed nutritional annotations. To further improve nutritional prediction, we also introduced NutritionSynth-115K, a large-scale synthetic dataset that preserves accurate nutritional labels. Our framework first predicts monocular depth, which is globally and locally refined by the Scale-Shift Residual Adapter (SSRA) to ensure scale consistency and preserve structural accuracy. The Frequency-Aligned Fusion Module (FAFM) hierarchically fuses RGB and depth features in the frequency domain, capturing both spatial and cross-modal semantic information. Finally, the Mask-based Prediction Head (MPH) dynamically selects informative channels to improve prediction accuracy. Extensive experiments across multiple datasets demonstrate that our approach consistently outperforms existing methods in nutritional prediction, validating its effectiveness and generalizability.

References

[1] Oscar Beijbom, Neel Joshi, Dan Morris, Scott Saponas, and Siddharth Khullar. Menu-match: Restaurant-specific food logging from images. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 844–851, 2015. 4

[2] Hongbo Bi, Ranwan Wu, Ziqi Liu, Huihui Zhu, Cong Zhang, and Tian-Zhu Xiang. Cross-modal hierarchical interaction network for rgb-d salient object detection. *Pattern Recognition*, 136:109194, 2023. 7

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer, 2014. 2, 4

[4] Jingjing Chen and Chong-Wah Ngo. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 32–41, 2016. 2, 4

[5] Yuhao Chen et al. Metafood3d: Large 3d food object dataset with nutrition values. *arXiv e-prints*, pages arXiv–2409, 2024. 2, 4

[6] Shaobo Fang, Chang Liu, Fengqing Zhu, Edward J Delp, and Carol J Boushey. Single-view food portion estimation based on geometric models. In *2015 IEEE International Symposium on Multimedia (ISM)*, pages 385–390, 2015. 4

[7] Zhihui Feng et al. Ingredient-guided rgb-d fusion network for nutritional assessment. *IEEE Transactions on AgriFood Electronics*, 2024. 1, 3, 8

[8] Yinxuan Gui, Bin Zhu, Jingjing Chen, Chong Wah Ngo, and Yu-Gang Jiang. Navigating weight prediction with diet diary. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 127–136, 2024. 2

[9] Yuzhe Han, Qimin Cheng, Wenjin Wu, and Ziyang Huang. Dpf-nutrition: Food nutrition estimation via depth prediction and fusion. *Foods*, 12(23), 2023. 1, 2, 7, 8

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 7

[11] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019. 7

[12] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 7

[13] Pengkun Jiao, Xinlan Wu, Bin Zhu, Jingjing Chen, Chong-Wah Ngo, and Yugang Jiang. Rode: Linear rectified mixture of diverse experts for food large multi-modal models. *arXiv preprint arXiv:2407.12730*, 2024. 2, 3, 4, 7

[14] Fotios S Konstantakopoulos, Eleni I Georga, and Dimitrios I Fotiadis. A review of image-based food recognition and volume estimation artificial intelligence systems. *IEEE Reviews in Biomedical Engineering*, 17:136–152, 2023. 1

[15] Zhengyi Liu, Yuan Wang, Zhengzheng Tu, Yun Xiao, and Bin Tang. Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4481–4490, 2021. 7

[16] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 7

[17] Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 7

[18] Boyuan Ma, Donglin Zhang, and Xiao-Jun Wu. Food nutrition estimation with rgb-d fusion module and bidirectional feature pyramid network. *Multimedia Systems*, 31(2):1–11, 2025. 3, 7

[19] Weiqing Min, Bing-Kun Bao, Shuhuan Mei, Yaohui Zhu, Yong Rui, and Shuqiang Jiang. You are what you eat: Exploring rich recipe information for cross-region food analysis. *IEEE Transactions on Multimedia*, 20(4):950–964, 2018. 4

[20] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A survey on food computing. *ACM Computing Surveys*, 52(5):1–36, 2019. 1

[21] Weiqing Min, Linhu Liu, Zhengdong Luo, and Shuqiang Jiang. Ingredient-guided cascaded multi-attention network for food recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1331–1339, 2019. 4

[22] Weiqing Min et al. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 393–401, 2020. 2, 4

[23] Weiqing Min et al. Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9932–9949, 2023. 2, 4

[24] Fudong Nian, Yujie Hu, Yanhong Gu, Zhize Wu, Shimeng Yang, and Jianhua Shu. Ingredient-guided multi-modal interaction and refinement network for rgb-d food nutrition assessment. *Digital Signal Processing*, 153:104664, 2024. 7

[25] Deepak NR et al. A framework for food recognition and predicting its nutritional value through convolution neural network. In *Proceedings of the International Conference on Innovative Computing & Communication*, page 6, 2022. 2

[26] Cathal O'Hara and Eileen R Gibney. Dietary intake assessment using a novel, generic meal–based recall and a 24-hour recall: Comparison study. *Journal of Medical Internet Research*, 26:e48817, 2024. 1

[27] Huiyan Qi, Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Ee-Peng Lim. Advancing food nutrition estimation via visual-ingredient feature fusion. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 1091–1099, 2025. 1, 2, 4

[28] Amaia Salvador et al. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3020–3028, 2017. 2, 4

[29] Wenjing Shao, Sujuan Hou, Weikuan Jia, and Yuanjie Zheng. Rapid non-destructive analysis of food nutrient content using swin-nutrition. *Foods*, 11(21), 2022. 1, 2, 7

[30] Wenjing Shao, Weiqing Min, Sujuan Hou, Mengjiang Luo, Tianhao Li, Yuanjie Zheng, and Shuqiang Jiang. Vision-based food nutrition estimation via rgb-d fusion network. *Food Chemistry*, 424:136309, 2023. 7, 8

[31] Zeman Shao, Gautham Vinod, Jiangpeng He, and Fengqing Zhu. An end-to-end food portion estimation framework based on shape reconstruction from monocular image. In *2023 IEEE ICME*, pages 942–947, 2023. 1, 3, 7

[32] Zhidong Shen, Adnan Shehzad, Si Chen, Hui Sun, and Jin Liu. Machine learning based approach on food recognition and nutrition estimation. *Procedia Computer Science*, 174: 448–453, 2020. 1

[33] Rohan Singh, Mathieu Théo Eric Verest, and Marcel Salathé. Minimum days estimation for reliable dietary intake information: findings from a digital cohort. *European Journal of Clinical Nutrition*, pages 1–11, 2025. 1

[34] Mark H. Stone. The cubit: A history and measurement commentary. *Journal of Anthropology*, 2014(1):489757, 2014. 3

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 7

[36] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 7

[37] Quin Thames et al. Nutrition5k: Towards automatic nutritional understanding of generic food. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8903–8911, 2021. 1, 2, 4, 7

[38] Gautham Vinod, Zeman Shao, and Fengqing Zhu. Image based food energy estimation with depth domain adaptation. In *2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval*, pages 262–267, 2022. 3, 7

[39] Binglu Wang, Tianci Bu, Zaiyi Hu, Le Yang, Yongqiang Zhao, and Xuelong Li. Coarse-to-fine nutrition prediction. *IEEE Transactions on Multimedia*, 26:3651–3662, 2023. 2, 7

[40] Huimin Wang, Yong Zhang, Xiaoping Liang, and Yingying Zhang. Smart fibers and textiles for personal health management. *ACS nano*, 15(8):12497–12508, 2021. 1

[41] Wei Wang, Weiqing Min, Tianhao Li, Xiaoxiao Dong, Haisheng Li, and Shuqiang Jiang. A review on vision-based analysis for automatic dietary assessment. *Trends in Food Science & Technology*, 122:223–237, 2022. 1

[42] Clare Whitton et al. Accuracy of energy and nutrient intake estimation versus observed intake using 4 technology-assisted dietary assessment methods: a randomized crossover feeding study. *The American journal of clinical nutrition*, 120(1):196–210, 2024. 1

[43] Sanghyun Woo et al. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 7

[44] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven CH Hoi, and Qianru Sun. A large-scale benchmark for food image segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 506–515, 2021. 2

[45] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 2, 5

[46] Dongjian Yu, Weiqing Min, Xin Jin, Qian Jiang, and Shuqiang Jiang. Spatial-aware multi-modal information fusion for food nutrition estimation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 8863–8871, 2025. 6

[47] Chen Zhang et al. Cross-modality discrepant interaction network for rgb-d salient object detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2094–2102, 2021. 7

[48] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24(12): 14679–14694, 2023. 7

[49] Jiaming Zhang et al. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023. 7

[50] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 4

[51] Wujie Zhou, Yi Pan, Jingsheng Lei, Lv Ye, and Lu Yu. Defnet: Dual-branch enhanced feature fusion network for rgb-t crowd counting. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):24540–24549, 2022. 7