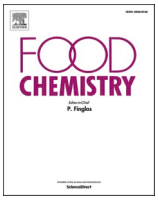
Food Chemistry 424 (2023) 136309

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem



Vision-based food nutrition estimation via RGB-D fusion network

Wenjing Shao [a], Weiqing Min [b,c,*], Sujuan Hou [a,*], Mengjiang Luo [b,c], Tianhao Li [b,c], Yuanjie Zheng [a], Shuqiang Jiang [b,c]

[a] School of Information Science and Engineering, Shandong Normal University, Shandong 250358, China
[b] The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Keywords:
Food nutrient
Nutrition estimation
Food composition
Deep learning
RGB-D fusion

ABSTRACT

With the development of deep learning technology, vision-based food nutrition estimation is gradually entering the public view for its advantage in accuracy and efficiency. In this paper, we designed one RGB-D fusion network, which integrated multimodal feature fusion (MMFF) and multi-scale fusion for visioin-based nutrition assessment. MMFF performed effective feature fusion by a balanced feature pyramid and convolutional block attention module. Multi-scale fusion fused different resolution features through feature pyramid network. Both enhanced feature representation to improve the performance of the model. Compared with state-of-the-art methods, the mean value of the percentage mean absolute error (PMAE) for our method reached 18.5%. The PMAE of calories and mass reached 15.0% and 10.8% via the RGB-D fusion network, improved by 3.8% and 8.1%, respectively. Furthermore, this study visualized the estimation results of four nutrients and verified the validity of the method. This research contributed to the development of automated food nutrient analysis (Code and models can be found at http://123.57.42.89/codes/RGB-DNet/nutrition.html).

1. Introduction

Over the past few decades, the dietary structure has been investigated in 195 countries and regions around the world, and a strong correlation between diet structure and disease was found (The GBD 2015 Obesity Collaborators, 2017). One-fifth of all deaths worldwide were related to poor diet. Understanding and estimating food nutrition played an important role in improving the poor dietary structure and guiding people to make scientific food choices. Accurate estimation of food nutrient composition promoted further development of rapid analysis of food composition and precision nutrition (Kirk, Catal, & Tekinerdogan, 2021). Therefore, the estimation of food nutritional content had become a hot topic of research in various fields such as food science (Ma, Lau, Yu, Li, & Sheng, 2022), computer vision (Thames et al., 2021; Min, Jiang, Liu, Rui, & Jain, 2020) and nutritional health (Lu et al., 2021). Professional food nutrition evaluators had slightly lower accuracy in estimating the food nutrient content, and professionals were not able to meet the public demand for daily nutrition evaluation. To alleviate the shortage of nutritionists and improve the accuracy and efficiency of nutrition estimation, dietary estimation applications for smartphone and food nutrition estimation systems (Myers et al., 2015) emerged in large numbers, which provide the possibility of food nutrition estimation in daily life, but these applications and systems were still lacking in accuracy and efficiency. For example, some applications required users to individually record the ingredients of each food as well as the portion size, and then convert these values to nutrient content (Shim, Oh, & Kim, 2014). It was tedious and laborious. As a result, developing automated and effective food nutrition content assessment methods have become particularly important.

In recent years, a variety of food nutrition assessment methods (Wang et al., 2022) have been proposed. Some traditional dietary assessment methods (Shim et al., 2014) have contributed to the study of nutrition estimation. However, these methods were time-consuming, laborious, and inaccurate in nutrition assessment. With the development of artificial intelligence and computer vision, vision-based food recognition (Min et al., 2021; Wang et al., 2022) and nutrition estimation methods (Foster & Bradley, 2018; Thames et al., 2021) were gradually emerged. These methods monitored the public's dietary structure and predicted the nutrient content from food, promoting the development of precision nutrition (Juan, Benoit, Jean-Pierre, & Marie-

* Corresponding authors at: The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (W. Min); School of Information Science and Engineering, Shandong Normal University, Shandong 250358, China (S. Hou).
E-mail addresses: minweiqing@ict.ac.cn (W. Min), sujuanhou@sdnu.edu.cn (S. Hou).

https://doi.org/10.1016/j.foodchem.2023.136309
Received 6 July 2022; Received in revised form 23 February 2023; Accepted 2 May 2023
Available online 17 May 2023

Claude, 2017). Ege and Yanai (2017) used convolutional neural networks to predict the calories, categories, and ingredients to estimate food calories from food images. They demonstrated that multi-task convolutional neural networks were more effective than single-task ones for food category prediction and food calorie estimation. Liu, Pu, and Sun (2021) combined non-destructive detection techniques and computer vision systems based on convolutional neural networks to detect and analyse complex food matrices. Ruede et al. (2021) proposed an end-to-end nutrition estimation method based on a single image and built a large-scale food database. For the nutritional intake of hospital patients, Lu et al. (2021) took pre and post-meal RGB-D image pairs as network inputs and employed a multi-task contextual network, a food classifier based on few-shot learning, and a volume estimation method to accurately estimate nutrition intake. Existing methods (Thames et al., 2021; Lu et al., 2021) performed nutrient estimation based on a single image or simply added RGB and depth images into the network for nutrient estimation. For example, Thames et al. (2021) took the depth of the input image as the fourth channel, which was sampled as a 3 channel tensor and input to the model. They didn't consider the characteristics of RGB and depth images. Although these methods demonstrated the feasibility of nutrition assessment based on food images, they failed to fully exploit the image features used for nutrition prediction. One possible reason was that using only RGB information or simple addition of RGB and depth images for prediction did not take full advantage of the complementary information between RGB and depth images.

In this paper, we proposed an RGB-D fusion-based approach for food nutrition assessment. Unlike existing nutritional estimation networks (Myers et al., 2015; Thames et al., 2021; Lu et al., 2021), we exploited the feature information of RGB and depth images and explored the fusion of complementary information in both modalities to improve the performance of model nutritional prediction. The RGB-D feature fusion network was constructed by the multimodal feature fusion module and the multi-scale fusion module to effectively represent the food image features. In order to fully exploit the features of RGB and depth images, the multimodal feature fusion module fused the image features of both modalities by the Balanced Feature Pyramid (BFP) and the Convolutional Block Attention Module (CBAM). The multi-scale fusion module enhanced the feature representation of the fused images by fusing the features of different resolutions through the Feature Pyramid Network (FPN), so as to construct the fusion network model to realize multimodal representation learning and vision-based food nutrition assessment. Furthermore, we improved a loss function to balance the loss values of the five subtasks, which solved the problem of difficult subtask optimization and improved the performance of the model. In our paper, we utilized 2,960 pairs of RGB and depth images from the Nutrition5k dataset (Thames et al., 2021) as the input of our network. The experimental results proved the effectiveness of our proposed method. Compared to the experimental results from Thames et al. (2021), the mean of our experimental results reached 18.5%, which improved by 1.6%. The PMAE of calories and mass reached 15.0% and 10.8%, improved by 3.8% and 8.1%, respectively.

Our main contributions were summarized as follows: (1) An RGB-D feature fusion network for food nutrition assessment was proposed, which combined a multimodal feature fusion module and a multi-scale fusion module to fully fuse the complementary information of different modalities and improve the accuracy of nutrition prediction. (2) A novel multimodal feature fusion network was presented, which incorporated the BFP and CBAM modules as a way to construct effective fusion feature information of food images and improve the performance of the model. (3) Experimental results on Nutrition5k demonstrated that our approach for food nutrition assessment was effective while achieving competitive performance.

2. Materials and methods

2.1. Nutrition5k dataset

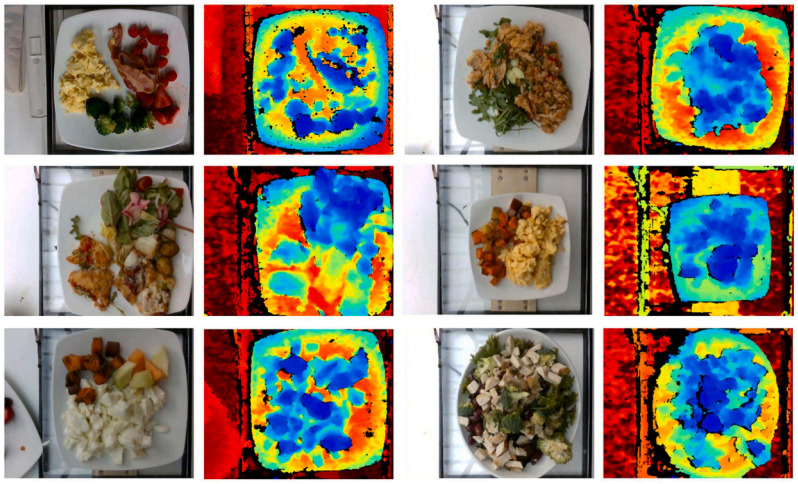
In this paper, we used 2,960 pairs RGB-D images from the Nutrition5k dataset (Thames et al., 2021), as shown in Fig. 1 (a). The RGB-D set consisted of RGB and depth images that were captured by an Intel RealSense D435. The camera consisted of four main sensors, including two infrared sensors, an infrared laser emitter, and a color sensor. The first two were responsible for forming depth images, while the latter was responsible for generating RGB images. The Intel RealSense D435 captured a color depth image that was a readable depth image, where objects closer to the camera were shown in blue and objects farther away were shown in red. All depth information was less than or equal to 0.4 meters. According to the USDA Food and Nutrient Database (Montville et al., 2013), the RGB-D set was constructed following the principle of incrementality, i.e., adding one at a time to the plates and scanning after each food item was added. The dataset consisted of more than 250 food categories and was annotated with nutritional information, as shown in Fig. 1 (b). Each dish had a complete food nutrition label, which ranged from a single ingredient to as many as 35 food ingredients. We perform experiments on the RGB-D set. The RGB-D set was divided into training set and test set with a ratio of 5:1. Meanwhile, the splitting of the dataset followed the incremental nature of the scan, with all dishes of a single incremental scan present in both the training and test sets. We used RGB-D image pairs as the input of the network to explore the effectiveness of the RGB-D feature fusion network.

2.2. Methods

2.2.1. Overall network architecture

The overall structure of our network was shown in Fig. 2. Our method supported arbitrary backbone networks. Without loss of generality, we selected ResNet-101 (He, Zhang, Ren, & Sun, 2016) as our backbone network. We performed the initial feature extraction for RGB and depth images, respectively, and used the extracted RGB and depth features of the four layers as the input of each layer of the FPN (Lin et al., 2017). We concatenated the feature maps of each level obtained from two branches after the FPN module and used the feature maps of each level after concatenation its as the input of the MMFF module. The MMFF contained the BFP and CBAM modules, where the BFP performed feature refinement and enhancement, and the CBAM combined attention mechanisms for adaptive feature optimization. The feature maps of each layer output from BFP were input to CBAM module for feature enhancement. The MMFF module fused the features of RGB and depth modalities and generated the final feature maps that contain more details and semantic information. The final feature maps went through the global average pooling layer and the fully connected layer to output the final prediction results for calories, mass, and macronutrients (fat, carbohydrate, and protein). The network not only fully exploited the features of RGB and depth images but also fused the features of different modalities to improve the accuracy of food nutrition estimation prediction.

In general, our network consisted of the backbone network, the multi-scale fusion module, and the MMFF module, as shown in Fig. 2. The backbone network was ResNet-101, which extracted the unique feature information of RGB and depth images. The multi-scale fusion module obtained feature maps at multiple scales with the help of the FPN and fused the low-level feature maps with local information and the high-level feature maps with rich semantic information to construct features at different levels. The MMFF consisted of the BFP and CBAM modules, which fully fused the unique feature information of RGB and depth images obtained by the multi-scale fusion module. After the multi-scale fusion module and MMFF module, rich complementary features of RGB and depth images were obtained to improve the accuracy of nutrition prediction.



(a)

Dish_1560368019					
Ingredient	Cal(kcal)	Mass(g)	Fat(g)	Carb(g)	Protein(g)
pizza	332.5	125	12.125	41.25	13.75
salt	0	0.59	0	0	0
green onions	0.17	0.54	0.001	0.038	0.0097
onions	0.86	2.16	0.002	0.19	0.024
olive oil	13.62	1.54	1.54	0	0
country rice	16.42	10.80	0.73	3.41	0.42
eggplant	11.43	46.08	0.09	2.72	0.46
spinach (raw)	6.21	27	0.11	0.97	0.78
fish	48.61	54.01	7.02	0	9.94
asparagus	6.73	33	0.03	1.29	0.73
garlic	0.40	0.27	0.001	0.09	0.017
Sum	436.97	301	21.66	49.96	26.13

(b)

Fig. 1. (a) RGB and depth images examples from Nutrition5k. (b) Nutritional content annotation information.

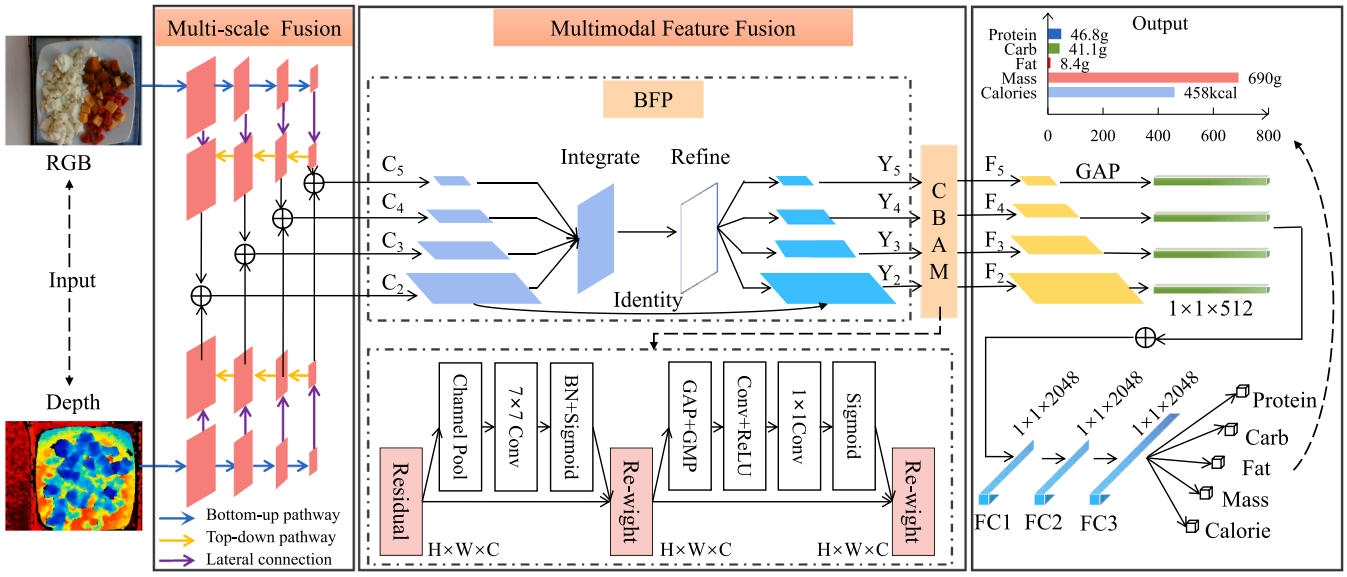


Fig. 2. The overall framework of our proposed method. We adopted a multi-scale fusion method, which fused different resolution features through FPN. We introduced a multimodal feature fusion module to integrate RGB and depth image features. ⊕ meaned element-wise addition. "GAP" and "GMP" indicated global average pooling and global max pooling.

2.2.2. *Multi-scale fusion*

Multi-scale fusion integrated features of different resolutions to enhance feature representation. As shown in Fig. 2, we extracted the features of RGB and depth images by two ResNet-101 branches respectively. We took the extracted features as the input of FPN, which consisted of three modules: bottom-up pathway, top-down pathway, and lateral connection. Because the convolutional module had the hierarchical semantic features, FPN used this property to build the feature pyramid network. The feature maps at different scales were obtained by convolution and pooling operations on the original images. Since feature maps at different scales had different feature information, we fused the feature maps at different scales to generate a feature map with rich semantic information. We performed a bottom-up pathway to generate features of different resolutions, and the features of the last layer in each convolutional block were selected as the features of the corresponding level of FPN. For each level of feature maps, top-down pathway and lateral connections were used to add the semantic features of the lower-level layers. The top-down pathway was to enlarge the feature maps of the upper layers to the same size as the adjacent feature maps by upsampling. Each lateral connection adopted a 1×1 convolution to adjust the number of feature map channels and fused the features of the upper layer after upsampling with the features of the current layer. The method is formulated as Eq. (1):

$$C_i = R_i \oplus D_i, (i = 2, \cdots 5) \tag{1}$$

where R_i denotes the i-th feature map obtained by the multi-scale fusion module of the RGB image, D_i represents the i-th feature map obtained by the multi-scale fusion module of the depth image, C_i denotes the i-th feature map output by the multi-scale fusion module. Both branches of the feature map performed the element-wise addition to obtain the final output of the multi-scale fusion module. In this way, we made use of

both the semantic information from the upper layers and fused the location information from the lower layers. Multi-scale fusion laid the foundation for the later nutrition assessment.

2.2.3. Multimodal feature fusion

To obtain more robust image features, multimodal feature fusion were adopted besides the ordinary feature extraction. The essence of multimodal fusion was to fuse two and more modal information by rational processing to obtain richer feature information. The multimodal fusion methods were applied in various tasks, such as recipe recognition (Wang, Kumar, Thome, Cord, & Precioso, 2015), food intake detection (Bahador, Ferreira, Tamminen, & Kortelainen, 2021), image classification (Kim, Son, & Kim, 2021), and scene parsing (Zhou, Liu, Lei, Yu, & Hwang, 2021; Zhou, Guo, Lei, Yu, & Hwang, 2022; Zhou, Lin, Lei, Yu, & Hwang, 2021). Most of the existing multimodal fusion methods (Zhou, Guo, Lei, Yu, & Hwang, 2021; Liao, Gao, Jiang, Wang, & Li, 2020; Zhou, Yang, Lei, & Yu, 2022; Zhou, Yang, Lei, Wan, & Yu, 2022) mainly focused on processing the data from different tasks and exploring better fusion methods to obtain great feature presentation.

Multimodal fusion was aimed at integrating extracted information from different modalities into a stable multimodal representation. Zhou et al. (2021) designed a dual-stream interactive recursive feature reshaping network to detect salient objects, which utilized a gated attention fusion mechanism to obtain fused complementary features of RGB and depth information. Zhou et al. (2021) adopted two multimodal fusion modules (namely shallow feature fusion module and deep feature fusion module) to solve the problem of urban scene semantic segmentation. The multimodal fusion strategy was designed to improve the segmentation accuracy according to the respective characteristics of shallow features and deep features. Liao et al. (2020) designed a Multi-stage Multiscale fusion network (MMNet), which includes cross-modal multi-stage fusion (CMFM) and a bi-directional decoder (BMD). The CMFM module was aimed at mining the important feature representations in the feature response stage, and the BMD module learned combinations of cross-modal features from multiple levels to capture local and global information. Our multimodal fusion network was designed to take advantage of the characteristics of RGB images and depth images and explore the fusion of complementary information between the two.

Considering the complementarity between RGB and depth images, we extracted features of each modality and fused the complementary information between two different modalities through the multimodal feature fusion module. This module first concatenated the output of RGB and depth branches from the multi-scale fusion module, and then refined the features by Balanced Feature Pyramid (BFP) (Pang et al., 2019). The BFP module utilized the integrated semantic information to reinforce the multi-level information and alleviated the semantic dilution problem in the top-down feature fusion process performed by FPN. The module first concatenated the features of two branches, and the concatenated features were represented by $\{C_2, C_3, C_4, C_5\}$, which were four levels with decreasing resolution step by step. Considering that lower-level features had more detailed information and higher-level features had more semantic information, we adjusted the feature maps of each level by max pooling to the same size as C_4, then summed and averaged to get the balanced semantic features, as shown in Eq. (2):

$$C = \frac{1}{N}\sum_{i=i_{\min}}^{i=i_{\max}} C_i \tag{2}$$

where $i = \{2, 3, \cdots\} \in \mathbb{R}$ and N is the number of feature maps involved in the computation. We used the embedded Gaussian non-local attention for feature enhancement to obtain higher resolution features (Wang, Girshick, Gupta, & He, 2018). The method is formulated as Eq. (3) and (4):

$$Y_i = \frac{1}{C(x)}\sum_{\forall j} f(x_i, x_j) g(x_j) \tag{3}$$

$$f(x_i, x_j) = e^{\theta(x_i)^T \phi(x_j)}, \quad C(x) = \sum_{\forall j} f(x_i, x_j) \tag{4}$$

where x denotes the input, Y_i denotes the output, i and j denote certain spatial location of the input, respectively. $f(x_i, x_j)$ is a function that calculates the similarity between two points. $g(x_j)$ is a mapping function that maps a point into a vector. We set it to a 1×1 convolution. Similarly, θ and $\phi$are both convolution operations. The BFP module utilized the integrated semantic information to reinforce the multi-level information and alleviated the semantic dilution problem in the top-down feature fusion process performed by FPN. In this way, the BFP enhanced the representation of feature map at each layer.

After the BFP, four levels of feature maps were obtained, and each level of feature maps then passed through the CBAM (Woo, Park, Lee, & Kweon, 2018), which applied channel attention and spatial attention to the feature maps to obtain the attention maps, and then multiplied the attention maps with the original feature maps for adaptive feature optimization as follows:

$$M_C(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{5}$$

$$M_S(F) = \sigma(f^{(7\times7)}([AvgPool(F), MaxPool(F)])) \tag{6}$$

where F is the feature maps. σ is sigmoid function, which is often used as an activation function for neural networks, mapping variables to between (0, 1). *MLP* denotes multilayer perceptron, also called artificial neural network, and the method defines a two-layer perceptron. 7×7 represents the convolution kernel as a 7×7 convolution, and the parameters and computation are greatly reduced by this convolution operation, which is conducive to the establishment of high-dimensional spatial feature correlation. *AvgPool* indicates that the values of several pixel blocks are averaged, and *MaxPool* indicates that the maximum value is taken. The two pooling methods are mainly used for feature dimensionality reduction. M_C denotes the attention map which is obtained after the channel attention module, and M_S is the attention map which is obtained after applying the spatial attention module. After the CBAM module, attention weights of channels and spatial dimensions can be obtained in the new feature map, which improves the correlation of a single feature in channels and spatial dimensions. Therefore, it can enrich the representation of fine-grained image features and improve the performance of the network.

The final output feature representation of the MMFF module was obtained after the BFP and CBAM modules, and was formulated as Eq. (7):

$$F_i = M_C(Y_i) + M_S(Y_i) \tag{7}$$

where F_i denotes the i-th layer of feature maps output by MMFF and Y_i denotes the i-th feature map output by BFP module. After the MMFF, the output feature maps had extremely rich semantic and detailed information, which provided a stronger impetus for the following nutrition prediction.

2.2.4. Loss function

We used loss function to measure the degree of difference between the predicted and groundtruth values of the model, and the model reduced the loss between the predicted and groundtruth values by back-propagating to update the parameters of the network. Our model predicted nutrition content for five subtasks (calories, mass, fat, carbohydrate and protein). To balance the loss values of the five subtasks, we improved a loss function as follows:

$$L = l_{\text{cal}} + l_{\text{mass}} + l_{\text{carb}} + l_{\text{fat}} + l_{\text{protein}} \tag{8}$$

$$l_{cal} = \frac{\sum_{i=1}^{N} |\hat{y}_i^{\text{cal}} - y_i^{\text{cal}}|}{\sum_{i=1}^{N} y_i^{\text{cal}}} \tag{9}$$

The total loss is the sum of five sub-task loss functions: calorie regression loss l_{cal}, mass regression loss l_{mass}, and regression losses of three macronutrients l_{carb}, l_{fat}, and l_{protein}. We express carbohydrate as carb. The loss function of each subtask is used as the regression loss as the ratio of the MAE to the mean of all groundtruth values for that task. $\hat{y}_i^{\text{cal}}, y_i^{\text{cal}}$ denotes the predicted and groundtruth values of the calorie regression sub-task, respectively. The other four sub-tasks are represented in the same way. Thames et al. (2021) directly added up the loss values of the five subtasks without considering the variability between the loss values of different tasks. Unlike the loss function used by Thames et al. (2021), we normalized the loss function. We found that the values of mass and calorie loss were on a different order of magnitude from those of other three subtasks (protein, fat, and carbohydrate). If the loss values of five subtasks were added directly, it would lead to difficulties in optimizing three subtasks of protein, fat, and carbohydrate. Therefore, we normalized the loss values of each nutrient to balance the loss values of five subtasks.

2.3. Evaluation

2.3.1. Evaluation metrics

We adopted the evaluation metric of the percentage of mean absolute error (PMAE) (Thames et al., 2021). Based on the mean absolute error (MAE) (De Myttenaere, Golden, Le Grand, & Rossi, 2016), we used the PMAE to measure the accuracy of regressions for calories, mass, and macronutrients, as shown in Eq. (10) and (11).

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{10}$$

$$\text{PMAE} = \frac{\text{MAE}}{\frac{1}{N} \sum_{i=1}^{N} y_i} \tag{11}$$

where $\hat{y}_i$ is the predicted value of i-th for a given test image and y_i is the groundtruth value of the i-th image. Calorie values are expressed in kilocalories, and mass and macronutrient values are expressed in grams. The PMAE indicates the percentage of the mean absolute error relative to the average of all the groundtruth values for that field. The lower value of the PMAE for each evaluation metric, the more accurate the nutrient content assessment is represented. In all the experiments, we used PMAE as our evaluation metric.

2.3.2. Visualization

To investigate the effectiveness of the multimodal feature fusion module, we applied Gradient-weighted Class Activation Map (Grad-CAM) (Selvaraju et al., 2020) operation for the obtained feature maps after this module, which presented the degree of similarity between each location and class in the original image in the form of a heat map, and then judged the effectiveness of the module. The image was input to the network and the feature maps were obtained as A, the weights of each channel in the feature maps were first calculated using the back-propagation gradient, and then the channels of the feature maps were weighted and added together with the original image to obtain the heat map. The method is formulated as Eq. (12) and (13).

$$\alpha_{\text{k}}^{c} = \frac{1}{Z} \sum_{i} \sum_{j} \frac{\partial y^{c}}{\partial A_{ij}^{k}} \tag{12}$$

$$L_{\text{Grad-CAM}}^{c} = ReLU\left(\sum_{k} \alpha_{\text{k}}^{c} \text{A}^{\text{k}}\right) \tag{13}$$

where c denotes the category, A denotes the feature mapping output from the convolutional layer, k is the feature mapping channel, i and j are the horizontal and vertical coordinates in the feature maps, respectively. Z are the size of the feature maps, and y^c represents the score predicted by the network for category c. The *ReLU* is used to retain the feature regions that have a positive effect for category c.

3. Results

In this section, we first presented our experimental setup in Section 3.1. Then, we introduced the experimental results from our method in Section 3.2, and compared the backbone networks and loss functions in Sections 3.2.1 and 3.2.2, respectively. We compared our method with existing multimodal fusion methods in Section 3.2.3. In Section 3.3, performance analysis of each module of our proposed nutrition assessment model was performed to verify the validity of each module. Finally, we visualized our results by Grad-CAM method.

3.1. Experimental setup

The PyTorch toolbox with an Nvidia GTX 3090 GPU was used to complete all our experiments. We initialized our backbone network with the weights pre-trained on the Food2k dataset. Pre-training was the process of pre-training the model, which provided better initialization parameters for the model. We used the Food2k dataset for pre-training to improve the performance of nutrition estimation model. In addition, Food2k dataset was a large food dataset in the food computing field, in terms of food categories and number of images (Min et al., 2021). We performed preprocessing on the training data, which mainly included resizing, random horizontal flipping, and center cropped. In the training stage, a multi-scale training approach was adopted to randomly selects one size from [(256, 352), (288, 384), (320, 448), (352, 480), (384, 512)] as the input size of the current batch per 10 iterations. The batch size was 4, the number of training epochs was 150, and the initial learning rate was 5×10^{-5}. During training, the learning rate was dynamically updated using an exponential decay, where the decay rate was 0.99.

3.2. Experimental results

3.2.1. Comparison of backbone networks

Comparing six basic backbone networks, the experimental results were shown in Table 1a(a). For fair comparison, the same experimental setup was adopted for all networks. All networks adopted a two-branch network structure, and the two branches were used to extract the features of RGB and depth images respectively. The multimodal fusion of all the backbone networks adopted simple feature vector concatenation. Among them, AlexNet (Krizhevsky, Sutskever, & Hinton, 2017), VGG-16 (Simonyan & Zisserman, 2014), Inception V3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), and ResNet (He et al., 2016) networks were the mainstream backbone networks in convolutional neural networks, and CoTNet (Li, Yao, Pan, & Mei, 2022) was a variant of ResNet, which was a new kind of attention network. We found that ResNet-101 achieved better prediction performance. The mean of PMAE values reached 23.4%. The ResNet-101 effectively avoided the gradient disappearance and accelerated the convergence of the network. Therefore, we used ResNet-101 as our backbone network for all the experiments, if not otherwise specified.

3.2.2. Different loss functions

We evaluated the performance of different loss functions. Specifically, our loss function was compared with the loss function (L_{multi}) proposed by Thames et al. (2021). The results were shown in Table 1a (a). As we can see, the experiments achieved optimal results by adopting our proposed loss function. The PMAE values for calories, mass, fat,

Table 1a
Comparing the performance of different methods. The best results were shown in bold blue.

Method types	Methods	Calories PMAE (%)	Mass PMAE (%)	Fat PMAE(%)	Carb. PMAE (%)	Protein PMAE (%)	Mean PMAE (%)
Backbone networks	AlexNet (Krizhevsky et al., 2017)	19.0	18.2	30.1	29.2	29.0	25.1
	VGG-16 (Simonyan & Zisserman, 2014)	19.2	17.9	29.6	29.0	28.4	24.8
	Inception V3 (Szegedy et al., 2016)	19.5	16.8	29.0	28.7	28.5	24.5
	ResNet-50 (He et al., 2016)	19.0	16.2	28.9	28.1	28.3	24.1
	ResNet-101 (He et al., 2016)	18.3	14.9	28.4	27.9	27.6	23.4
	CoTNet (Li et al., 2022)	18.4	15.1	29.1	28.0	27.9	23.7
	ACNet (Ding et al., 2019)	18.1	17.2	30.4	28.1	27.3	24.2
	RDFNet (Lee et al., 2017)	18.2	15.7	31.7	29.0	27.6	24.4
	RDFNet + CBAM	17.9	14.9	28.2	26.5	25.3	22.6
	Google-Nutrition (Thames et al., 2021)	18.8	18.9	**18.1**	23.8	**20.9**	20.1
Loss functions	L_{multi} (Thames et al., 2021)	15.3	11.7	24.4	24.2	22.6	19.6
	Ours	**15.0**	**10.8**	23.5	**22.4**	21.0	**18.5**

carbohydrate, and protein reached 15.0%, 10.8%, 23.5%, 22.4%, and 21.0%, respectively, and the mean value of PMAE decreased by 1.1% compared to the use of L_{multi}. We normalized the loss values of each nutrient to balance the loss values of five subtasks. Our loss function made the loss values of the five subtasks in the same order of magnitude, which solved the problem of difficult subtask optimization and also improved the performance of the model.

3.2.3. Comparison with the state-of-the-art methods

Our model was compared with different multimodal fusion methods, and the comparison results were shown in Table 1a(a). To ensure the fairness of the experimental results, all experimental settings and the division of the dataset were the same for all methods. We improved multimodal fusion methods by referring to the methods in several vision tasks, such as food recognition and object detection. We called the method proposed by Thames et al. (2021) as Google-Nutrition. We compared different methods including ACNet (Ding, Guo, Ding, & Han, 2019), RDFNet (Lee, Park, & Hong, 2017), and Google-Nutrition.

We referred to the methods in the field of computer vision and improved RDFNet to make it applicable to the food nutrition assessment. The RDFNet contained Multimodel Feature Fusion Network (MMFNet) and RefineNet. MMFNet mainly fused the features obtained from the feature maps of the two branches after feature extraction, and RefineNet took fused features from MMFNet and the previously refined features as input to obtain more details and semantic information. We took advantage of RDFNet in RGB-D feature fusion and applied the improved RDFNet to the food nutrition estimation field. Similarly, we improved ACNet, which used an attention complementary module to collect feature information from two modalities. The network adopted two independent branches of ResNet to extract the features of RGB and depth images respectively, and a third ResNet-based branch was used to process the merged features. The approach was designed with an attention complementary module that integrates information from different channels. The ACNet predicted a calorie PMAE of 18.1%, a mass PMAE of 17.2%, and a mean PMAE of 24.2%. For Google-Nutrition, the method used Inception V2 as the backbone network to predict the content of five nutrients and took RGB and depth images as input (depth image as the 4th channel). The method predicted a caloric PMAE of 18.8%, a mass PMAE of 18.9%, and a protein PMAE of 20.9%, as shown in Table 1a(a).

We presented the experimental results of different multimodal fusion methods on the Nutrition5k dataset in Table 1a(a). As we can see, our method achieved the best results compared to other methods, with the PMAE values of 15.0%, 10.8%, 22.4%, and 21% for calories, mass, carbohydrate, and protein, respectively. Compared with the experimental results of Thames et al. (2021), all values surpassed it except for fat. The mean value reached 18.5% and exceeded it by 1.6%. The experimental results showed that the proposed multimodal fusion method achieved the optimal performance. For the high PMAE value of fats, we guessed that it might be because fat was mostly found in edible oils such as olive oil and fish oil. However, these cooking oils were attached to the surface of food or plates, and it was difficult to detect the oil ingredients by image-based detection methods. In addition, our network was compared with other RGB-D fusion networks in terms of the model size (Params), complexity (FLOPs), and speed (FPS) of the model. The comparison results were shown in Table 1a(b), where Google-Nutrition's data was not included because its model was not released. It can be seen that our method outperformed other comparative fusion networks on all three metrics. The reason for the optimal results of our network was that the network adopted a two-branch structure and did not introduce a third branch in the multimodal fusion phase, skillfully combining BFP and CBAM to obtain the complementary features of the two modalities.

To show our prediction results more visually, we compared the predicted values (y-axis) of the model output with the groundtruth values (x-axis), and the results were shown in Fig. 3. We compared the experimental results from applying our method with those from other methods. The original method denoted the removal of the MMFF module and the new loss function on the base of our method. The loss function of the original method was the same as that of the Google-Nutrition. The green line in the figure represented the predicted values were the same as the groundtruth values. The more concentrated the discrete points in each graph represented a lower PMAE value, the better the nutrient content estimation. In Fig. 3, we found that the predicted values from our method are less discrete from the groundtruth values, which proved the validity of our proposed method. In addition, we found that the predicted results for the sub-task of the mass had the smallest deviation from the groundtruth values and the lowest dispersion. This was consistent with the experimental results in Table 1a(a).

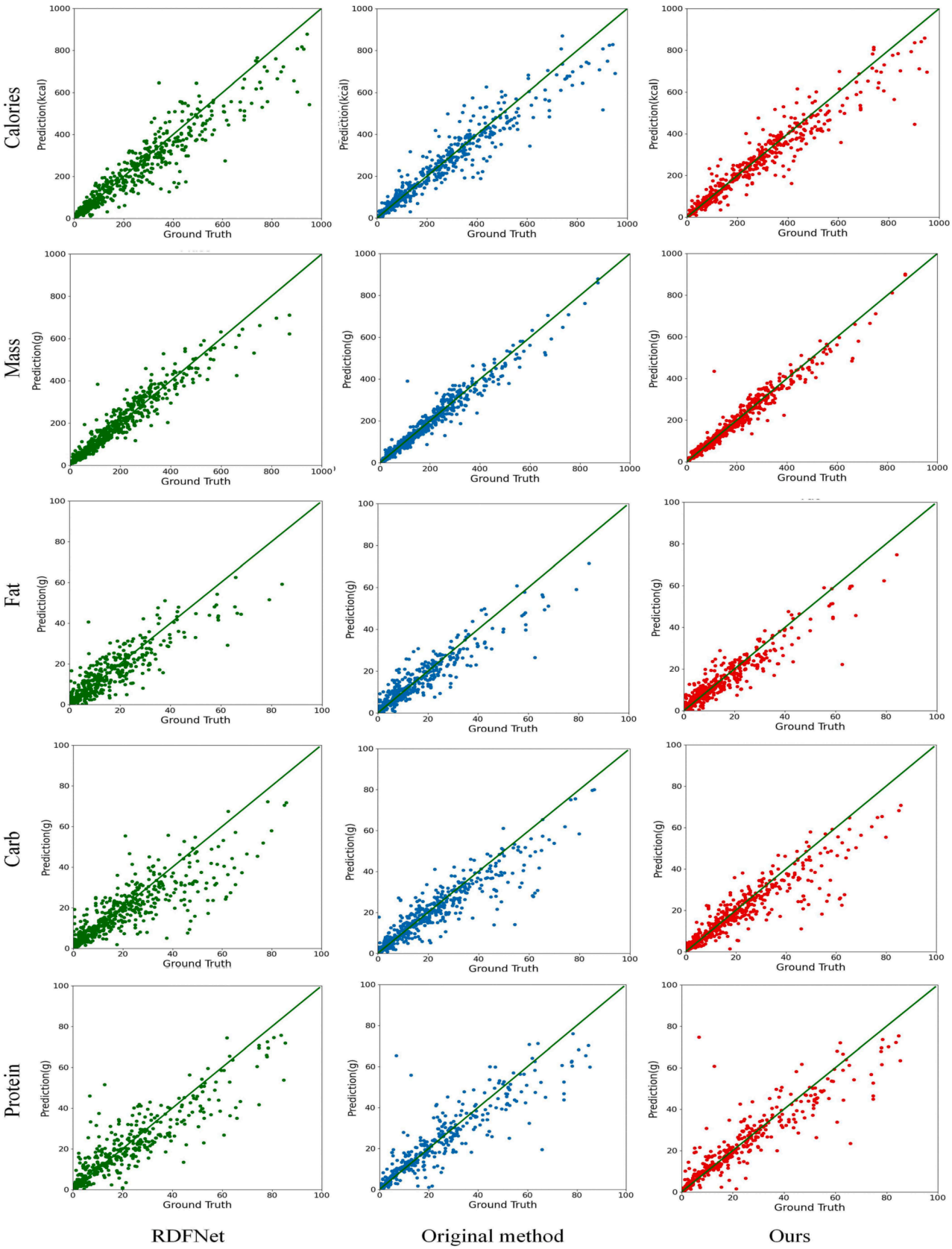


Fig. 3. Relationship between the prediction values and the groundtruth values from our proposed method and other methods. The original method removed the MMFF module from our method and used the loss function L_{multi}.

Most of the predicted results for the calorie and carbohydrate subtasks deviated less from the groundtruth values. However, there was a small minority of values for the predicted output of the fat subtask that deviated more from the groundtruth values. This was the same as the higher PMAE values for the fat in Table 1a(a).

3.3. Performance analysis of nutrition assessment methods

To investigate the impact of different components of our model, we analyzed the performance of each module from method. The baseline was our architecture removing five modules and using only ResNet-101 and a simple multi-stage fusion approach in the fusion stage. The effectiveness of the five modules in our approach was compared.

First, we analysed the FPN module, which performed multi-scale fusion by the top-down pathway. Considering that the size of ingredients varies greatly in images, we used the FPN module to fuse feature maps of different scales and generate more expressive feature maps. Second, we performed performance analysis on the pre-training module. The experiment was performed by pre-training our model using the Food2k dataset, and the best weight values were obtained at the end of the training to be used as the initialized weight values for our model training. Thames et al. (2021) used JFT-300 M for pre-training. Considering that the JFT-300 M dataset was not publicly available and the Nutrition5K dataset was relatively small, pre-training model became important to improve the performance of our model. In addition, the Food2k dataset was a massive food recognition dataset that contained 2,000 categories and over one million images. Third, we introduced multi-scale training. We pre-defined several fixed scales during training and randomly selected one scale for training per 10 iterations. By this data enhancement method, the robustness of the model was to some extent improved. To investigate how multimodal fusion module plays a role in food nutrition assessment, we added the MMFF module in the previous base for the analysis study. In addition, we added a CBAM module to introduce spatial attention and channel attention for adaptive feature optimization.

As shown in Table 1b, these five modules had different degrees of impact on improving the performance of the food nutrition evaluation model. We introduced the multiscale fusion module, and the PMAE values for calories and mass reached 17.8% and 14.0%, respectively. This module had high robustness to detect the different sizes of ingredients because it used the FPN to construct multi-scale feature information. Therefore, it can fuse low-level feature maps and high-level feature maps with the help of top-down pathway and lateral connections when building the features at different levels. When we pre-trained with the Food2k dataset, the mean value reached 20.3%. The PMAE values for fat and carbohydrate were reduced by 2.1% and 2.5% with this module, respectively. Meanwhile, the PMAE value of the carbohydrate reached the lowest value, 21.8%. One possible reason it worked so well was that it was pre-trained on the Food2K dataset and has optimized initial parameters. The performance of our model was improved by adding the multi-scale training method, and the PMAE values of both calories and protein were reduced by 1.4% and 0.8%. The possible reason why the multi-scale training module can play a positive role was that the model was trained on images of different scales, which improved the robustness of the model to detect objects of different sizes. The MMFF module played an important role in the food nutrition content assessment by fusing the features of the two modalities, which provided a robust feature representation for the nutrition assessment model. Among them, the BFP utilized feature information from multiple layers to enhance the representation of feature map at each layer. Because BFP adopted the balanced semantic information to enhance the features of different layers, the dilution of semantic information of non-adjacent layer features can be alleviated in the fusion process. The CBAM used spatial attention and channel attention mechanisms to obtain attention weights for both dimensions in a new feature map, which improved the correlation of a single feature in both channel and spatial dimensions. This module achieved a 1.1% gain of the mean value. In the end, the mean value of the PMAE of our model reached 18.5%. Compared to the Google-Nutrition, our model obtained a 1.6% gain of the mean value. The PMAE values for calories and mass were reduced by 3.8% and 8.1%, respectively (See Table 2).

Table 1b
Comparison of the proposed method with others in terms of parameters(Params), complexity(FLOPs), and speed(FPS).

Model	Params(M)	FLOPs(G)	FPS
ACNet	129.61	72.10	11.9
RDFNet	140.56	78.83	13.3
RDFNet + CBAM	143.98	78.87	13.1
Ours	110.87	68.96	15.2

To confirm the effectiveness of our proposed method and present the results of the nutrition assessment more intuitively, we applied the Grad-CAM method on the feature maps obtained from the last convolution layer to generate the heat maps, as shown in Fig. 4 (a). Fig. 4 (a) presented us with the original images of the three dishes and the results of the four subtasks output by the model, which were marked in red text for original images. The composition of these dishes and the amount of each nutrient contained in each ingredient were represented in Fig. 4 (b). Given the concern about the response area of calories and macronutrients in the food images, the results of the mass subtask visualization were not presented. For example, the total calorie content of dish_1558114609 was 540.9 kcal, of which the calorie content of sausage reached 258.45 kcal. The calorie content of sausage accounted for 47.8% of the total calorie content. Therefore, in Fig. 4 (a), the response area for the sub-task of the calories in this dish was mainly at the location of the sausages. However, the carbohydrate response area was mainly concentrated in the location of grapes and nuts, since these two ingredients accounted for 75.8% of the total carbohydrate content. The total protein content of the dish was 29.14 g, and the sausage and bacon had 24.84 g of protein, which accounted for 85.2% of the total protein content in Fig. 4 (b). Consequently, the response area of the protein in the picture was the location of the sausage and bacon. Likewise, the other two dishes were observed in the same way. In Fig. 4 (b), we found that olive oil had the highest fat content in dish15664602160, accounting for 85.6% of the total fat content. However, in Fig. 4 (a), the visualization of fat in this dish was poor and failed to respond correctly to where the olive oil was located. We found it difficult to observe the olive oil in the image with our eyes. In addition, we found that the ingredients with high fat content in the ingredient database are mainly edible oils, such as olive oil, vegetable oil, and fish oil. These edible oils were difficult to detect in the images, which resulted in higher PMAE values of fats. Therefore, the prediction of fat remained more challenging. For the prediction of other nutrients, the visualization results in Fig. 4 (a) showed that the performance of our model proved to be reliable. For food nutritional assessment, we acquired more accurate assessment results using multimodal fusion methods.

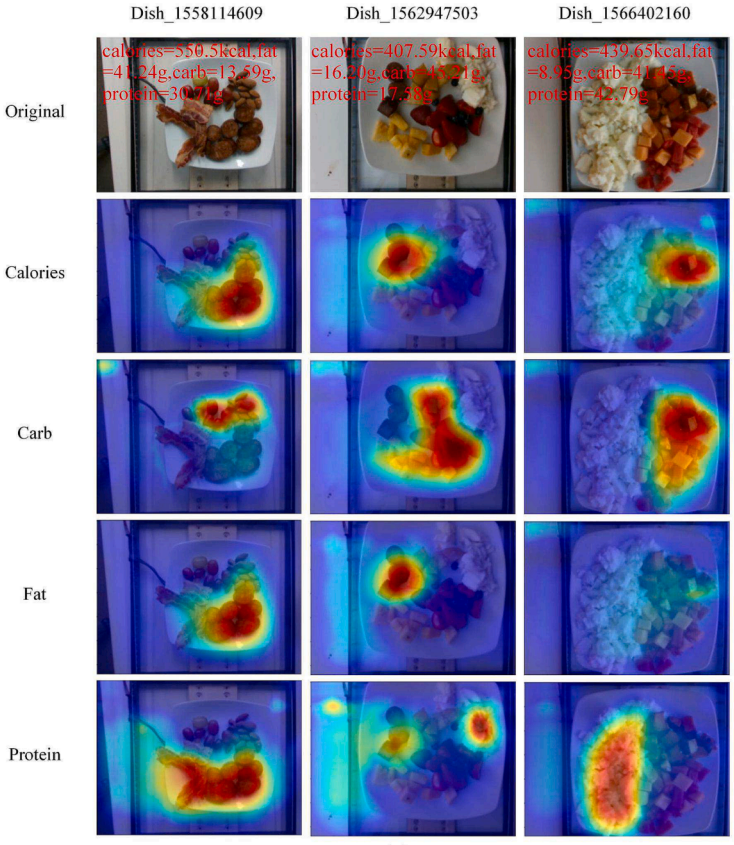
4. Discussion

Some limitations existed in the annotation information of the Nutrition5k dataset. We manually screened RGB-D set. We compared the ingredients in the images with the annotated nutrition information one by one and found a few images that had serious discrepancies with the real label information. For example, the ingredients contained in the dish (peppers, onions, quinoa, mixed greens, and olive oil) did not match the annotated information (olive oil and quinoa) in dish_1566414225. We removed the images with the same problem as dish_1566414225 and performed the experiment. The performance of our method was improved. The PMAE value of calories, mass, fat, carbohydrate, and protein reached 14.1%, 10.5%, 22.1%, 21.8%, and 19.6%, respectively.

There were some limitations in our nutrition estimation method.

Table 2
Performance comparison results for each module on the Nutrition5k dataset. The best results were shown in boldface.

Method					Nutrition5k					
FPN	Pretraining	Multi-scale training	MMFF		Calories PMAE (%)	Mass PMAE (%)	Fat PMAE(%)	Carb. PMAE (%)	Protein PMAE (%)	Mean PMAE (%)
			BFP	CBAM						
✓					17.8	14.0	27.2	24.3	25.1	21.7
✓	✓				17.5	12.9	25.1	21.8	24.0	20.3
✓	✓	✓			16.1	12.9	24.1	**21.8**	23.2	19.6
✓	✓	✓	✓		15.8	12.4	23.6	22.9	21.7	19.3
✓	✓	✓	✓	✓	**15.0**	**10.8**	**23.5**	22.4	**21.0**	**18.5**



(a)

Dish_1558114609				
Ingredient	Cal(kcal)	Fat(g)	Carb(g)	Protein(g)
grapes	26.91	0.08	7.02	0.27
sausage	258.45	20.48	3.15	14.48
almonds	104.06	8.91	4.09	4.03
bacon	151.48	11.76	0.39	10.36
Sum	540.90	41.22	14.65	29.14
Dish_1562947503				
Ingredient	Cal(kcal)	Fat(g)	Carb(g)	Protein(g)
sausage	179.19	14.20	2.18	10.04
roasted potatoes	109.98	5.30	14.04	1.64
egg whites	45.76	0.18	0.62	8.80
berries	66.69	0.35	16.38	0.82
pineapple	36.50	0.07	9.49	0.37
Sum	438.12	20.10	42.71	21.66
Dish_1566402160				
Ingredient	Cal(kcal)	Fat(g)	Carb(g)	Protein(g)
egg whites	231.92	0.89	3.12	44.60
watermelon	24.00	0.16	6.08	0.48
yam	76.68	0.12	18.17	0.99
cantaloupe	13.60	0.08	3.20	0.32
sweet potato	46.96	0	11.12	0.62
olive oil	65.55	7.42	0	0
Sum	458.72	8.67	41.69	47.01

(b)

Fig. 4. (a) Visualization of the prediction results for our method using Grad-CAM. The prediction results were shown in red font for the original images. (b) Nutritional content annotations associated with (a).

Firstly, when many foods were stacked together, the prediction of food nutrient content model could be biased, which was a challenge in food nutrition assessment. Specifically, the stacking of many foods resulted in some of the foods being obscured and covered, leading the model to get incorrect estimation results. For a plate without a lot of food stacks, our model was able to successfully estimate the calories, fat, carbohydrates, and protein content and correctly highlighted the region where each nutrient was in the original image. Secondly, when ingredients that are difficult to observe with the naked eye are present in images, such as cooking oils and smaller particle ingredients, the prediction of fat remains a challenge. Thirdly, there were still relatively few food nutrition datasets, and the construction of the datasets was labor-intensive. Finally, the wide variety of foods and the interference of other environmental factors in real life make nutritional assessment difficult. These will be the urgent problem to solve in the future food nutrition assessment.

In a nutshell, we explored a food nutrient content estimation approach from RGB-D fusion perspective, and an end-to-end food nutrient estimation network was proposed. However, there are still some challenges to overcome before automated food nutrient content estimation can be applied to daily life, such as realistic problems of food obscuration and coverage. In the future, we will continue to explore more effective food nutrition estimation methods to meet the requirements of the public for nutrition estimation in real-world scenarios.

5. Conclusions

In this paper, we proposed an RGB-D feature fusion-based network for food nutrition estimation, which aimed to improve the accuracy of nutrition assessment. Specifically, the network extracted features from RGB and depth images, and effectively fused the features of both modalities through multimodal fusion and multi-scale fusion to improve the performance of the nutrition assessment model. We improved a loss function to balance loss values of five subtasks and enhance the robustness of the model. We demonstrated the effectiveness of our method on the Nutrition5k dataset. We hope that our method can

provide a novel perspective for further development of food nutrition estimation and expect that automated food nutrition assessment will be widely applied in people's daily life shortly to provide a scientific and healthy life for people and meet the increasing demand of diet monitoring.

CRediT authorship contribution statement

Wenjing Shao: Writing – original draft, Conceptualization, Methodology, Software. **Weiqing Min:** Conceptualization, Methodology, Writing – review & editing, Project administration. **Sujuan Hou:** Conceptualization, Supervision. **Mengjiang Luo:** Data curation, Software, Validation, Writing – review & editing. **Tianhao Li:** Data curation, Investigation, Writing – review & editing. **Yuanjie Zheng:** Writing – review & editing, Supervision. **Shuqiang Jiang:** Writing – review & editing, Supervision.

Declaration of Competing 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.

Acknowledgment

This work was supported in part by the National Nature Science Foundation of China [Grant No. 62072289, 61972378, U19B2040, 62125207] and CAAI-Huawei MindSpore Open Fund.

References

Bahador, N., Ferreira, D., Tamminen, S., & Kortelainen, J. (2021). Deep learning–based multimodal data fusion: Case study in food intake episodes detection using wearable sensors. *JMIR mHealth and uHealth, 9*(1), Article e21926.

De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing, 192*, 38–48.

Ding, X., Guo, Y., Ding, G., & Han, J. (2019). ACNet: strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1911–1920).

Ege, T. and Yanai, K. (2017). Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions. In Proceedings of the on Thematic Workshops of ACM Multimedia 2017 - Thematic Workshops '17, pages 367–375.

Foster, E., & Bradley, J. (2018). Methodological considerations and future insights for 24-hour dietary recall assessment in children. *Nutrition Research, 51*, 1–11.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Juan, T. M., Benoit, A., Jean-Pierre, D., & Marie-Claude, V. (2017). Precision nutrition: A review of personalized nutritional approaches for the prevention and management of metabolic syndrome. *Nutrients, 9*(8).

Kim, W., Son, B., & Kim, I. (2021). ViLT: vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 5583–5594).

Kirk, D., Catal, C., & Tekinerdogan, B. (2021). Precision nutrition: A systematic literature review. *Computers in Biology and Medicine, 133*, Article 104365.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84–90.

Lee, S., Park, S.-J., & Hong, K.-S. (2017). RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4990–4999).

Li, Y., Yao, T., Pan, Y., & Mei, T. (2022). Contextual transformer networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liao, G., Gao, W., Jiang, Q., Wang, R., & Li, G. (2020). MMNet: multi-stage and multi-scale fusion network for RGB-D salient object detection. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 2436–2444).

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 936–944).

Liu, Y., Pu, H., & Sun, D.-W. (2021). Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices. *Trends in Food Science & Technology, 113*, 193–204.

Lu, Y., Stathopoulou, T., Vasiloglou, M. F., Christodoulidis, S., Stanga, Z., & Mougiakakou, S. (2021). An artificial intelligence-based system to assess nutrient intake for hospitalised patients. *IEEE Transactions on Multimedia, 23*, 1136–1147.

Ma, P., Lau, C. P., Yu, N., Li, A., & Sheng, J. (2022). Application of deep learning for image-based Chinese market food nutrients estimation. *Food Chemistry, 373*, Article 130994.

Min, W., Jiang, S., Liu, L., Rui, Y., & Jain, R. (2020). A survey on food computing. *ACM Computing Surveys, 52*(5), 1–36.

Min, W., Wang, Z., Liu, Y., Luo, M., Kang, L., Wei, X., et al. (2021). Large scale visual food recognition. arXiv preprint arXiv:2103.16107.

Montville, J. B., Ahuja, J. K., Martin, C. L., Heendeniya, K. Y., Omolewa-Tomobi, G., Steinfeldt, L. C., et al. (2013). USDA food and nutrient database for dietary studies (FNDDS), 5.0. *Procedia Food Science, 2*, 99–112.

Myers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., et al. (2015). Im2Calories: towards an automated mobile vision food diary. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1233–1241).

Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., & Lin, D. (2019). Libra R-CNN: towards balanced learning for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 821–830).

Ruede, R., Heusser, V., Frank, L., Roitberg, A., Haurilet, M., & Stiefelhagen, R. (2021). Multi-task learning for calorie prediction on a novel large-scale recipe dataset enriched with nutritional information. *International Conference on Pattern Recognition*, 4001–4008.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision, 128*(2), 336–359.

Shim, J.-S., Oh, K., & Kim, H. C. (2014). Dietary assessment methods in epidemiologic studies. Epidemiology and health. *Epidemiology and health, 36*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818–2826).

Thames, Q., Karpur, A., Norris, W., Xia, F., Panait, L., Weyand, T., et al. (2021). Nutrition5k: towards automatic nutritional understanding of generic food. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8899–8907).

The GBD 2015 Obesity Collaborators. (2017). Health effects of overweight and obesity in 195 countries over 25 years. *New England Journal of Medicine, 377*(1), 13–27.

Wang, W., Min, W., Li, T., Dong, X., Li, H., & Jiang, S. (2022). A review on vision-based analysis for automatic dietary assessment. *Trends in Food Science & Technology, 122*, 223–237.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7794–7803).

Wang, X., Kumar, D., Thome, N., Cord, M., & Precioso, F. (2015). Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops* (pp. 1–6).

Wang, Z., Min, W., Li, Z., Kang, L., Wei, X., Wei, X., et al. (2022). Ingredient-guided region discovery and relationship modeling for food category-ingredient prediction. *IEEE Transactions on Image Processing, 31*, 5214–5226.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: convolutional block attention module. In *Proceedings of the European Conference on Computer Vision* (pp. 3–19).

Zhou, W., Guo, Q., Lei, J., Yu, L., & Hwang, J.-N. (2021). IRFR-Net: Interactive recursive feature-reshaping network for detecting salient objects in rgb-d images. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

Zhou, W., Guo, Q., Lei, J., Yu, L., & Hwang, J.-N. (2022). ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(3), 1224–1235.

Zhou, W., Lin, X., Lei, J., Yu, L., & Hwang, J.-N. (2021). MFFENet: Multiscale feature fusion and enhancement network for rgb–thermal urban road scene parsing. *IEEE Transactions on Multimedia, 24*, 2526–2538.

Zhou, W., Liu, J., Lei, J., Yu, L., & Hwang, J.-N. (2021). GMNet: Graded-feature multilabel-learning network for RGB-Thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing, 30*, 7790–7802.

Zhou, W., Yang, E., Lei, J., Wan, J., & Yu, L. (2022). PGDENet: Progressive guided fusion and depth enhancement network for rgb-d indoor scene parsing. *IEEE Transactions on Multimedia*.

Zhou, W., Yang, E., Lei, J., & Yu, L. (2022). FRNet: Feature reconstruction network for RGB-D indoor scene parsing. *IEEE Journal of Selected Topics in Signal Processing, 16*(4), 677–687.