Ingredient-Guided Region Discovery and Relationship Modeling for Food Category-Ingredient Prediction

Zhiling Wang, Weiqing Min, *Senior Member, IEEE*, Zhuo Li, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang, *Senior Member, IEEE*

***Abstract*— Recognizing the category and its ingredient composition from food images facilitates automatic nutrition estimation, which is crucial to various health relevant applications, such as nutrition intake management and healthy diet recommendation. Since food is composed of ingredients, discovering ingredient-relevant visual regions can help identify its corresponding category and ingredients. Furthermore, various ingredient relationships like co-occurrence and exclusion are also critical for this task. For that, we propose an ingredient-oriented multi-task food category-ingredient joint learning framework for simultaneous food recognition and ingredient prediction. This framework mainly involves learning an ingredient dictionary for ingredient-relevant visual region discovery and building an ingredient-based semantic-visual graph for ingredient relationship modeling. To obtain ingredient-relevant visual regions, we build an ingredient dictionary to capture multiple ingredient regions and obtain the corresponding assignment map, and then pool the region features belonging to the same ingredient to identify the ingredients more accurately and meanwhile improve the classification performance. For ingredient-relationship modeling, we utilize the visual ingredient representations as nodes and the semantic similarity between ingredient embeddings as edges to construct an ingredient graph, and then learn their relationships via the graph convolutional network to make label embeddings and visual features interact with each other to improve the performance. Finally, fused features from both ingredient-oriented region features and ingredient-relationship features are used in the following multi-task category-ingredient joint learning. Extensive evaluation on three popular benchmark datasets (ETH Food-101, Vireo Food-172 and ISIA Food-200) demonstrates the effectiveness of our method. Further visualization of ingredient assignment maps and attention maps also shows the superiority of our method.**

***Index Terms*— Image classification, object recognition, deep learning.**

Manuscript received 10 August 2021; revised 28 April 2022; accepted 10 July 2022. Date of publication 1 August 2022; date of current version 4 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61972378, Grant U1936203, and Grant U19B2040; and in part by Meituan Group. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kui Jia. *(Corresponding author: Weiqing Min.)*

Zhiling Wang, Weiqing Min, Zhuo Li, and Shuqiang Jiang are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: wangzhiling02@meituan.com; minweiqing@ict.ac.cn; zhuo.li@vipl.ict.ac.cn; sqjiang@ict.ac.cn).

Liping Kang, Xiaoming Wei, and Xiaolin Wei are with Meituan Group, Beijing 100102, China (e-mail: kangliping@meituan.com; weixiaoming@meituan.com; weixiaolin02@meituan.com).

Digital Object Identifier 10.1109/TIP.2022.3193763

I. Introduction

FOOD image analysis has captured numerous attention [1], [2] in the image processing community for supporting many food-relevant applications [3], e.g., nutrition estimation [4], food choice [5], food diary [6], health-aware recommendation [7], [8] and self-service restaurants [9]. Single-label food category recognition and multi-label ingredient prediction are two basic tasks in food image analysis. Therefore, the research on food category recognition and ingredient prediction has great application potentials.

Food recognition can be seen as one fine-grained recognition task, and thus it is important to find the subtle discriminative regions, mainly ingredient regions. Therefore, mining ingredient regions is of great significance for food image analysis. Moreover, multi-label ingredient prediction is generally a harder problem than food recognition. The visual patterns of ingredients change greatly and they are often mixed with each other. Besides, mutual relationships among them highlight their challenges. To fully utilize ingredient information for food recognition and ingredient prediction, there exists three factors for consideration:

1) A dish contains various ingredients, and these ingredients appear in various scales and positions. Specifically, the size, shape, and color of an ingredient can exhibit large visual differences because of diverse ways of cooking and cutting, in addition to changes in viewpoints and lighting conditions. As shown in Fig. 1, all these food contain several different ingredients, and some food like "Shrimp and grits" and "Lobster bisque" have more than ten ingredients. For the ingredient "egg" in "Beef tartare", "Eggs benedict" and "Lobster bisque", there exist large visual differences among them, and its visual pattern changes greatly, which increases the difficulty for extracting ingredient-relevant region features. Therefore, we need to extract various and comprehensive ingredient region features. Moreover, not all the ingredients of a certain food will appear in all the corresponding images, and usually we can only see part of them. Therefore, we should explore some novel regularizations for discovering ingredient regions. In this way, the feature expression ability of the network for ingredients can be further improved.

2) The complicated ingredient composition in the same food category makes food category and ingredient prediction more challenging. Most dishes are often composed of

Fig. 1. Some samples from the experimental datasets, where food category and the ingredients are listed. The detailed ingredient regions of Beef tartare are also marked.

a variety of ingredients being fuzzily mixed, rather than separated clearly or non-overlapping food items. Moreover, some groups of ingredients co-occur more often, and some ingredients are exclusive. For example, "butter" and "milk" always appear together, like "beignets", "bread pudding", "carrot cake", "clam chowder", and "waffles". This motivates us to exploit the mutual relationships among ingredients for better performance.

3) To increase the robustness of recognition, multi-task learning is often employed for the classification of food categories and ingredients [2], [10]–[12]. This is because multi-task learning can leverage food-related information as supplementary supervised information. For example, ingredient prediction can help to obtain the detailed ingredient composition of the food category, which can further promote the performance of food category recognition [12]. Meanwhile, food category recognition can guide the network to predict the corresponding ingredient composition. Therefore, multi-task learning can help exploit the mutual relationship between the food category and ingredients, and we need to optimize the loss function of these two tasks simultaneously to achieve joint learning.

Taking these factors into consideration, we propose a multi-task learning framework for simultaneous food category and ingredient prediction. This framework mainly consists of two components, namely ingredient-oriented visual region discovery and graph relationship modeling. For the ingredient-relevant region discovery, we propose to group the feature map into different ingredient regions and build the corresponding ingredient dictionary to discover regions. Then we can obtain the ingredient assignment map, and further utilize the attention mechanism to enhance these ingredient region features. During the ingredient region discovery, we utilize a regularization term of ingredient occurrence to facilitate ingredient-relevant region discovery, where we enforce the prior U-shaped distribution for the occurrence of each ingredient. Once we can find the corresponding ingredient regions, the ingredient composition of this food thus can be obtained, and we will further update the value vectors of the ingredient dictionary iteratively. For ingredient relationship modeling, we construct an ingredient-oriented semantic-visual graph to explore complex ingredient relationships, where we use the visual representation of ingredients as nodes and semantic similarity between ingredient word embeddings as edges. Then we utilize a graph convolutional network to fuse semantic and visual features simultaneously for better representations learning, resulting in better performance for food category and ingredient prediction.

To evaluate our method, we conduct extensive experiments on three popular food datasets. On western food dataset ETH Food-101 [13], Chinese food dataset Vireo Food-172 [10] and mixed food dataset ISIA Food-200 [14], our method all achieves the performance gain. Moreover, the visualization of ingredient assignment and attention maps demonstrates the superiority of our method, and the detailed ingredient regions prove that our method can discover various meaningful regions. The comparison of the feature map visualization shows that our method can discover multiple and expanded ingredient regions by ingredient relationship modeling.

The contributions of our paper can be summarized as follows:

- We propose a multi-task learning framework for simultaneous food category and ingredient prediction, where we learn an ingredient dictionary for ingredient-relevant region discovery and build an ingredient-oriented semantic-visual graph convolutional network for ingredient relationship learning.
- Our proposed method consists of two branches. The first branch learns an ingredient dictionary and leverages the U-shaped prior of ingredient occurrences to facilitate ingredient-relevant region discovery. The second branch builds an ingredient-oriented semantic-visual graph, and then uses the graph convolutional network to make label embeddings and visual features interact with each other for ingredient relationship learning.
- We conduct extensive evaluation on three benchmark datasets to verify the effectiveness of our method. Further visualization of assignment maps and attention maps demonstrates the advantage of our method.

II. Related Work

A. Food Recognition

In the earlier years, the mainstream recognition methods utilized hand-crafted features [13], [15], such as SIFT and HOG. For instance, Lukas *et al.* [13] adopted random forests to mine discriminative patches of food images as visual representation for food recognition. In the deep learning era, because of its powerful capacity of feature representation, more and more works resort to different deep networks for food recognition [16]–[18]. For example, Qiu *et al.* [19] proposed a PAR-Net to mine discriminative food regions for accurate food recognition. Min *et al.* [20] proposed a stacked global-local attention network to jointly learn global

and local features for food recognition. Due to few samples for some food categories, Zhao *et al.* [21] exploited a fusion learning framework to unify many-shot and few-shot ways for food recognition. In order to further improve the recognition performance, context information and external knowledge, such as ingredients, cuisine and location [10], [14], [22] are leveraged. For example, Zhou and Lin [22] made full use of the relationships among ingredients and restaurant information via the bi-partite graph for food recognition. Min *et al.* [14] utilized ingredients as the additional supervised signal to localize multiple informative regions and fused these regional features as the final representation for recognition.

Similar to [14], we also utilize ingredient information for food recognition. Different from these works [14], which sequentially localizes multiple informative image regions from category level to ingredient level guidance, we not only mine various ingredient-relevant regions, but also utilize the ingredient embeddings and visual features to model the complicated ingredient relationships via the graph convolutional network for better food classification and ingredient prediction.

B. Ingredient Prediction

Compared with food category recognition, ingredient prediction is much more challenging as ingredients are small in size and exhibit larger variance in the appearance. Bolanos *et al.* [23] explored the problem of ingredient prediction from a multi-label perspective and solved it by Convolutional Neural Network (CNN). In [24], the multimodal deep Boltzmann machine is applied for ingredient recognition. Liu *et al.* [12] proposed an attention fusion network and food-ingredient joint learning module for fine-grained food and ingredient recognition. Recently, Chen *et al.* [5], [10] focused on zero-shot ingredient identification by building a multi-relational knowledge graph to model ingredient relationships. Their recent work [5] built a multi-relational Graph Convolutional Network (GCN) to integrate ingredient hierarchy, attribute and co-occurrence for zero-shot ingredient recognition. Different from it, we target the problem of ingredient recognition with sufficient training samples, and evaluate the proposed method using standard multi-label image classification metrics. Moreover, we build an ingredient-oriented semantic-visual GCN to model the complicated ingredient relationships for ingredient prediction. We also propose to localize and fuse the detailed ingredient-relevant visual regions for better prediction performance.

Chen *et al.* [2] provides an insightful analysis of three compelling issues in ingredient recognition, including learning in either single or multi-task manner. Our method is different from it in the following two aspects: (1) Motivation. The paper [2] aims to solve the problem of limited datasets available with ingredient labels, and it proposes Vireo Food-251 and an insightful analysis of three compelling issues for ingredient recognition. In contrast, our proposed method is designed for simultaneous food category recognition and ingredient prediction. (2) Methodology. The paper [2] presents two methods for ingredient recognition. The first method utilizes the global image features for ingredient recognition, and the second one predicts ingredient labels at local image regions. Different from it, our method proposes to discover and extract ingredient region features and model their relationships, where we learn an ingredient dictionary for ingredient-relevant visual region discovery and build an ingredient-based semantic-visual graph for ingredient relationship modeling.

C. Multi-Task Food Attribute Learning

Multi-task learning [25] simultaneously solves multiple tasks at once for enhancing performance and improving generalization. This strategy has been widely used in food analysis [10], [26], [27]. For example, Zhang *et al.* [11] incorporated the cooking attribute recognition into multi-task learning. Ege *et al.* [26] has proved that simultaneously learning food categories, ingredients and calories can boost the performance of all tasks than single-task. Min *et al.* [24] proposed a multimodal multi-task deep belief network to learn joint image-ingredient representation regularized by different attributes. Recently, Liang *et al.* [27] proposed a novel multi-view attention network within the multi-task learning framework to incorporate multiple semantic features into the food recognition task for both ingredient recognition and recipe modeling.

Our method also utilizes a joint learning framework for simultaneous food category and ingredient prediction like [12], [27]. However, their methods ignore the relationships among ingredients, which has been utilized by introducing an ingredient-oriented semantic-visual graph for modeling their relationships in our method. Moreover, we also propose to discover various ingredient-relevant region features, and introduce one U-shaped prior of ingredient occurrence that facilitates ingredient discovery during learning.

III. Our Method

As shown in Fig. 2, in this section, we introduce the proposed ingredient-oriented multi-task food-ingredient joint learning framework, which mainly consists of two components, namely Ingredient-oriented Visual Region Discovery (IVRD) and Ingredient-oriented Graph Relationship Learning (IGRL). In IVRD, when one food image is fed into the proposed network, we first extract the feature map from the last convolution layer and learn a dictionary of ingredient regions by grouping 2D feature maps into detailed ingredient regions. During this process, we also apply one U-shaped prior for the occurrence of ingredients to facilitate ingredient discovery during learning. Thereafter, we pool these ingredient features, followed by an attention mechanism to select a subset of ingredient regions for classification. In IGRL, we construct an ingredient-oriented semantic-visual graph to explore the relationships of various ingredients, where we utilize all the visual ingredient representations as nodes and the similarity between all ingredient semantic embeddings as edges. Then we introduce a graph convolutional network to learn various relations among ingredients. Finally, for the output of two branches IVRD and IGRL, we fuse them together and feed it to the classifiers, and optimize the model in a multi-task

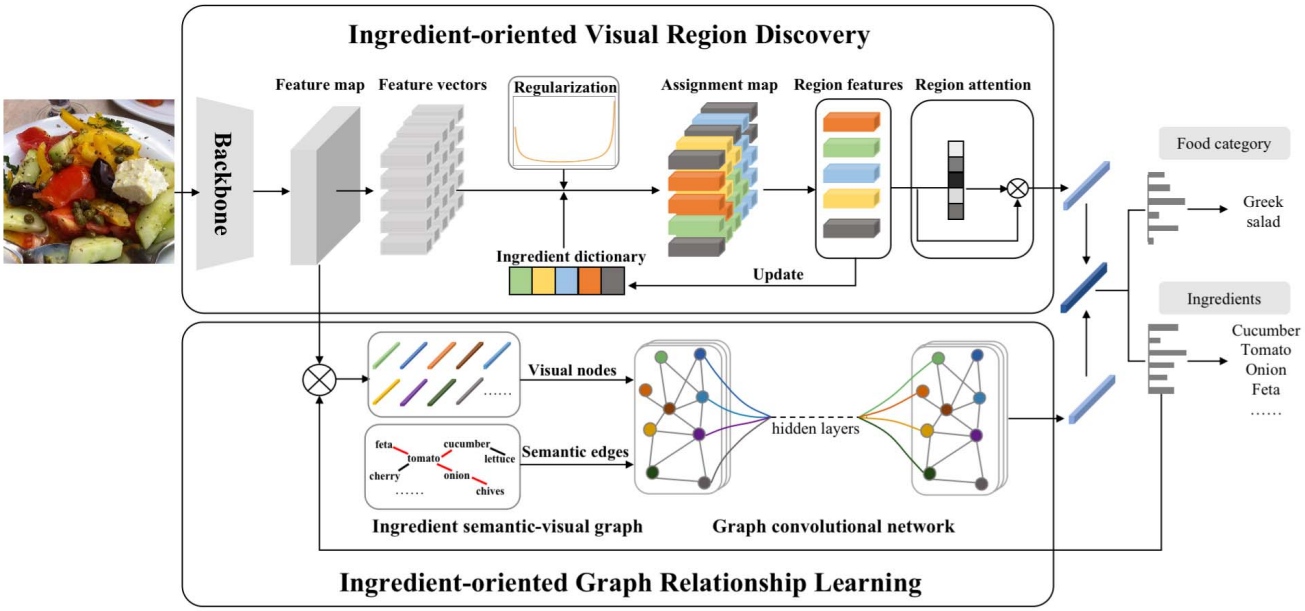


Fig. 2. The framework of the proposed method.

learning way for simultaneous food category and ingredient prediction.

A. Ingredient-Oriented Visual Region Discovery

Most foods have multiple ingredients, and the ingredients in most food images have various scales and views. It is difficult to assign these visual regions into different ingredients. To solve it, we propose to extract region-based ingredient features. Specifically, we design one ingredient dictionary $D = [d_1, d_2, \ldots, d_k, \ldots, d_K]$ for each category, where each vector d_k denotes the unique ingredient concept from the corresponding food category, and K is the total number of ingredients of this category. We build D according to the statistics of the ingredients and initialize it randomly. Then we take the feature map $X \in R^{C\times H\times W}$ from the last convolution layer and the ingredient dictionary D together to learn their detailed region features, where C, H and W are the channel number, height, and width of the feature map. In details, we compare the feature map X and the ingredient dictionary D to generate a soft ingredient assignment map $S = [s_{ij}^k] \in R^{K\times H\times W}$, where s_{ij}^k means the probability of feature vector x_{ij} in X being assigned to the k^{th} ingredient d_k. Besides, we employ the U-shaped distribution as prior to control the ingredient occurrence probability, which can help to better discover various ingredient regions during ingredient assignment. Thereafter, we pool X to obtain the ingredient region features $M \in R^{C\times K}$ based on S and D. Then we utilize M to update the value vectors of D iteratively for generating more precise assignment maps. Finally, we re-weight these region features by a region attention α.

1) Ingredient Region Assignment: For better discovering various ingredient regions in the food images, we utilize a similar projection unit [28] for ingredient assignment. For a feature vector $x_{ij} \in R^C$ at position (i, j) on X, we can obtain its corresponding value s_{ij}^k in the assignment matrix S, where k indexes the ingredient. s_{ij}^k is calculated as:

$$s_{ij}^k = \frac{exp(- \parallel (x_{ij} - d_k)/\beta_k \parallel_2^2 /2)}{\sum_k exp(- \parallel (x_{ij} - d_k)/\beta_k \parallel_2^2 /2)} \tag{1}$$

where $\beta_k \in (0, 1)$ is a learnable factor for d_k. Then we can assemble all the $s_{ij}^k \in R^K$ to generate the ingredient assignment S. Because of the softmax normalization, $s_{ij}^k > 0$ and $\sum_k s_{ij}^k = 1$.

2) Ingredient Occurrence Regularization: During ingredient region assignment, in order to better discover various and comprehensive ingredient regions in images, we enforce a prior U-shaped distribution for the occurrence of each ingredient d_k in a set of image features $X_{1:N}$ to regularize the learning.

Specifically, after obtaining the ingredient assignment S, we need to detect the occurrence of each ingredient d_k. We utilize a Gaussian kernel and a max-pooling operation as ingredient detectors $t_k = max_{ij}\Omega * S^k$, where Ω is a Gaussian kernel and $*$ is the convolution operation. $t_k \in (0, 1)$. Thereafter, we utilize this ingredient detector t_k over the k^{th} ingredient assignment S^k to determine the occurrence of each ingredient. Finally, we concatenate all the outputs of k ingredient detectors into an occurrence vector $\gamma = [t_1, t_2, \ldots, t_k]^T$.

For regularizing the occurrence of each ingredient, we align the empirical distribution of ingredient occurrence with the prior U-shaped distribution. In details, we let $p(d_k|X_{1:N})$ the conditioned probability of ingredient d_k occurrence in $X_{1:N}$, and we can calculate this empirical distribution $p(d_k|X_{1:N})$ by concatenating all occurrence vectors $\gamma_n, n = 1, 2, \ldots, N$ into a matrix $\hat{\mathbf{U}} = [\gamma_1, \gamma_2, \ldots, \gamma_N]$. Meanwhile, we assume a prior known distribution $\hat{p}(d_k|X_{1:N})$, which is the U-shaped distribution in our method. Then we use the Earth Mover

Distance [29] to align $p(d_k|X_{1:N})$ with $\hat{p}(d_k|X_{1:N})$ as:

$$EMD(p(d_k|X_{1:N}), \hat{p}(d_k|X_{1:N})) = \int_0^1 \mid F^{-1}(z) - \hat{F}^{-1}(z) \mid dz \quad (2)$$

where $F(\cdot)$ and $\hat{F}(\cdot)$ are the Cumulative Distribution Function (CDFs) for the empirical and prior distribution and $z \in [0, 1]$.

By applying this regularization, our method can capture more reasonable and comprehensive ingredient regions. Take "Greek salad" for example, some indispensable and common ingredients are always presented in most of the images, like tomato, onion, and cucumber, such that the switch is always on. However, for some preferred ingredients like thyme and chives, the switch will be activated only for some images. Therefore, we regularize the learning to make their occurrence probability close to this prior and discover the correct ingredient regions as possible.

3) Ingredient Dictionary: After obtaining the ingredient assignment S, we pool it to obtain the ingredient region features. The ingredient region feature set $M = [m_1, m_2, \ldots, m_k] \in R^{C\times K}$ from input feature maps can be obtained. We then utilize these to update the value vectors of the ingredient dictionary D, which can help the model to improve the ability of discovering more reasonable ingredient regions. For the ingredient dictionary D, we build the corresponding one for each food category in the dataset, and its length is determined by the number of the ingredients per category. The dictionary is randomly initialized and then it will be updated by the learned region features.

4) Ingredient Attention: In order to highlight the pivotal regions for classification, we need to attach a higher attention vector to it. We transform the feature set of ingredient regions M into $f_m(M)$, where f_m contains multiple 1×1 convolutions with batch norm and ReLU. Next, we utilize the attention mechanism to predict the importance for each ingredient region in M:

$$\alpha = softmax(f_m(M)) \quad (3)$$

Therefore, we can obtain the attention vector $\alpha \in R^K$ for each ingredient region.

Finally, we re-weight $f_m(M)$ by the attention vector α, and get the final ingredient region features f_{region}:

$$f_{region} = f_m(M)\alpha \quad (4)$$

B. Ingredient-Oriented Graph Relationship Learning

There exist various ingredients and discovering the corresponding regions can improve the performance of food category recognition and ingredient prediction. In addition, exploring the relationship between ingredients can further bring more performance gain, like co-occurrence and exclusion. For example, some ingredients are correlated as they share the same cutting or cooking methods (e.g., diced tomato and diced red bell peppers). Other ingredients may be associated because they often co-occur with each other in a dish (e.g., shallots and garlic). Therefore, modeling the relationships among ingredients is significant, and we utilize GCN to explore their interactions and further learn representations of ingredient graphs.

1) Ingredient Semantic-Visual Graph: Particularly, we build up an ingredient-oriented semantic-visual graph for each image, where the nodes represent the visual representations of different ingredients and the edges indicate semantic relations between ingredients.

For the visual nodes in the ingredient graph, we use visual representations of all the ingredients. Specifically, we use the weights of the fully connected layer in the ingredient classifier as the channel attention for different ingredients. Then we multiply these attentions with the last feature map $X \in R^{C\times H\times W}$, and pool the generated visual representations into the visual embeddings for each ingredient. Therefore, these embeddings can serve as the nodes in the ingredient graph, and are defined as:

$$Q_n = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} w_n X(i, j) \quad (5)$$

where $X(i, j)$ is the value of position (i, j) in the feature map X, w_n is the weight in the fully connected layer for the n^{th} ingredient. Therefore, the embedding of the ingredient $Q_n \in N \times N$ can be obtained, and N is the total number of all ingredients in the dataset.Then it can serve as the node in the ingredient graph.

For the semantic edge in the ingredient graph, we use Word2vec [30] for each ingredient word to obtain the corresponding semantic embedding v. Then we can compute the cosine similarity between two ingredient embeddings as:

$$a_{i,j} = cos < v_i, v_j >= \frac{v_i \cdot v_j}{|v_i||v_j|} \quad (6)$$

where $cos < . >$ means the cosine similarity between two ingredient embeddings. Then we utilize these similarities $a_{i,j}$ to form the graph adjacency matrix $A = [a_{i,j}]$, and further employ these similarities as the edges of the constructed graph.

2) Graph Convolutional Network: Then we can explore the ingredient graph for their relationship learning. We utilize GCN to learn various relationships among ingredients from the ingredient graph. The convolutional operation in GCN follows the layer-wise propagation:

$$H^{(l+1)} = \sigma(\hat{O}^{-\frac{1}{2}}\hat{A}\hat{O}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \quad (7)$$

where $H(l) = (Q_1^{(l)}, Q_2^{(l)}, \ldots, Q_n^{(l)})^T$ denotes the feature matrix of nodes in the l^{th} GCN layer, and Q_n is computed using the Equation 5. $A \in R^{N\times N}$ denotes the graph adjacency matrix and N is the total number of all ingredients in the dataset, and it is obtained by Equation 6. $\hat{O} = \sum_j A_{ij}$ denotes the sum of elements in the adjacency matrix. $\hat{A}$ is a normalized version of the graph adjacency matrix A. $W^{(l)}$ is a parameter matrix and σ is a non-linear operation like ReLU.

Finally, we can obtain the output G from GCN, which is the same size as A. Thus we employ the global average pooling f_{GAP} to get the final output features f_{graph} as:

$$f_{graph} = f_{GAP}(G) \quad (8)$$

During ingredient graph learning, we consider both visual and semantic embeddings, and utilize GCN to make label

embeddings and visual features interact with each other. Therefore, the learned representations are more comprehensive.

C. Multi-Task Learning

In this paper, we propose to couple food categorization problem, which is a single-label problem, together with ingredient recognition, which is a multi-label problem, for simultaneous learning. In this paper, we directly add two classifiers to the current network, and train the whole network end-to-end.

After obtaining the ingredient features f_{region} and f_{graph}, we fuse them together and feed them into two classifiers. For food category recognition, we use the cross-entropy loss function L_c during model optimization:

$$L_c = -\sum_{i=1}^{n_1} y_i log(\hat{y}_i) \tag{9}$$

where n_1 means the total number of food categories.

For ingredient prediction, we choose the binary cross-entropy function L_b:

$$L_b = -\sum_{j=1}^{m}\sum_{i=1}^{n_2}\{y_{ij} log(\widehat{y_{ij}}) + (1 - y_{ij}) log(1 - \widehat{y_{ij}})\} \tag{10}$$

where m is the total number of images in a batch and n_2 is the total number of ingredients. y_i, y_{ij} are the ground truth label of the input and $\hat{y}_i$, $\widehat{y_{ij}}$ are the corresponding predicted probability vectors.

Moreover, our model is also trained by minimizing the Earth Mover Distance in Equation 2 for ingredient occurrence regularization. Therefore, the total loss function can be written as follows, and λ is the balance ratio.

$$L = L_c + L_b + \lambda EMD \tag{11}$$

D. Inference

During inference, our model utilizes the corresponding learned dictionary to assign the feature map and obtain ingredient region features, then re-weight these using the attention vector. Furthermore, our model uses the trained weights of fully connected layers in the ingredient classifier (the total ingredient weights are fixed according to the training set and N is the same value) multiplying with the feature map from the last convolutional layer and ingredient semantic embeddings to build the ingredient graph for each image, and feed it to the graph convolutional network. Finally, we concatenate the outputs from two branches for food category and ingredient prediction. Specifically, our model can learn different ingredient dictionary D for each food category and a decision function $y = \varphi(X_i, D; \theta)$ is obtained after end -to-end training, where $\varphi(\cdot)$ takes both the feature maps X_n and the corresponding ingredient dictionary D to predict the food categories and its ingredients, and θ are the parameters.

IV. Experiment

A. Datasets

To further verify the effectiveness of our method, we mainly conduct experiments on three typical food datasets, which contain both food category and ingredient composition annotations, and they are very suitable for our task. The detailed introduction of three datasets is as follows: **ETH Food-101 [13]** is one typical western food dataset, and contains 101,000 images from 101 food categories. There are 1,000 images including 750 training images and 250 test images for each category. Following [14], we use the same ingredient list, and its total size is 174. **Vireo Food-172 [10]** is a Chinese food dataset. It consists of 110,241 food images from 172 categories and the size of the ingredient list is 353. Similar to [10], 60%, 10%, 30% of images are randomly selected for training, validation and testing, respectively. **ISIA Food-200 [14]** is a mixed food dataset. It contains 197,323 images with 200 categories and 399 ingredients. 60%, 10%, 30% of the total images are selected for training, validation and testing, respectively.

Notice that we only use the number of visible ingredients for the construction of ingredient dictionary in IRA and extracts their features. There are two main reasons. First, these visible ingredients are the main components of the food, the corresponding features also become the critical features for recognition. Second, visible ingredient regions can be discovered and obtained via the ingredient assignment map, and then the relevant ingredient features can be obtained.

B. Experimental Setup

For our method, we utilize ResNet-101 pre-trained on ImageNet as our backbone. All the parameters are jointly learned on the target dataset. The ingredient dictionary is randomly initialized following [31]. We adopt the similar parameters setting for the U-shaped prior like [32]. For the balance raito λ in the Equation 11, we set this as 0.1 for three datasets following [32], and this parameter is uniform for all datasets. For the ingredient word embeddings in the ingredient graph, we use Word2vec [30] trained on the cooking instructions of Recipe1M [33]. For ingredient prediction, we set the threshold value of 0.5 for the activation function sigmoid. We use standard stochastic gradient descent with a batch size of 80 and momentum of 0.9 for all datasets. The learning rate is set to 10^{-2} initially and divided by 10 after 60 epochs. The input images are resized to 448×448. Data augmentation including random crop, random horizontal flip and color jittering are applied. We use Pytorch to implement our algorithm. For performance evaluation, Top-1 accuracy (Top-1 acc.) and Top-5 accuracy (Top-5 acc.) are employed for single-label food category recognition. As ingredient recognition is a multi-label problem, we utilize Micro-F1 and Macro-F1 as evaluation metrics, which can take both precision and recall of ingredient prediction into account. The Micro-F1 can be expressed as follows:

$$\overline{TP} = \frac{1}{N}\sum_{i=1}^{N} TP_i$$

TABLE I
ABLATION STUDIES FOR FOOD RECOGNITION ON THREE DATASETS (%)

Method	ETH Food-101		Vireo Food-172		ISIA Food-200	
	Top-1 acc.	Top-5 acc.	Top-1 acc.	Top-5 acc.	Top-1 acc.	Top-5 acc.
FCR	87.65	97.17	88.24	97.35	59.03	86.53
JL	88.21	97.63	90.31	98.33	63.92	84.91
JL+IRA	90.44	97.98	90.92	97.98	65.32	87.21
JL+IRA+IA	91.34	98.54	91.34	98.54	68.98	89.72
JL+IGRL	91.21	98.47	92.83	98.92	69.12	91.64
Our method	**92.36**	**98.68**	**93.33**	**99.15**	**69.47**	**92.98**

TABLE II
ABLATION STUDIES FOR INGREDIENT PREDICTION ON THREE DATASETS (%)

Method	ETH Food-101		Vireo Food-172		ISIA Food-200	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
IP	84.37	82.81	70.32	48.32	52.89	46.12
JL	84.84	83.32	71.58	52.57	53.23	47.36
JL+IRA	87.92	86.76	72.45	54.76	59.92	55.29
JL+IRA+IA	89.42	88.73	73.41	57.23	62.43	58.07
JL+IGRL	90.94	90.53	73.94	57.53	63.18	59.27
Our method	**91.51**	**90.82**	**74.34**	**59.56**	**64.74**	**62.61**

$$\begin{aligned}
\overline{FP} &= \frac{1}{N}\sum_{i=1}^{N} FP_i \\
\overline{FN} &= \frac{1}{N}\sum_{i=1}^{N} FN_i \\
micro\text{-}P &= \frac{\overline{TP}}{\overline{TP}+\overline{FP}}, micro\text{-}R = \frac{\overline{TP}}{\overline{TP}+\overline{FN}} \\
Micro\text{-}F1 &= \frac{2 \times micro\text{-}P \times micro\text{-}R}{micro\text{-}P + micro\text{-}R}
\end{aligned} \tag{12}$$

where N is the number of ingredients, TP_i, FP_i and FN_i are the true positive, false positive and false negative sample quantity of category i ingredients respectively. $micro\text{-}P$ and $micro\text{-}R$ represent the accuracy and recall rates of all category samples, and Micro-F1 is the harmonic average of them.

The Macro-F1 can be expressed as follows:

$$\begin{aligned}
P_i &= \frac{TP_i}{TP_i + FP_i} \\
R_i &= \frac{TP_i}{TP_i + FN_i} \\
macro\text{-}P &= \frac{1}{N}\sum_{i=1}^{N} P_i, macro\text{-}R = \frac{1}{N}\sum_{i=1}^{N} R_i \\
Macro\text{-}F1 &= \frac{2 \times macro\text{-}P \times macro\text{-}R}{macro\text{-}P + macro\text{-}R}
\end{aligned} \tag{13}$$

where P_i and R_i respectively represent the accuracy and recall rates of category i ingredients samples, $macro\text{-}P$ and $macro\text{-}R$ respectively represent the average accuracy and recall rates of all category, and Macro-F1 is the harmonic average of them.

C. Experiment on ETH Food-101

1) Ablation Studies on Different Components of Our Method: In our experiment, we first verify the effect of joint learning framework and then we conduct ablation studies on the food category recognition and ingredient prediction. We just utilize the backbone for food category recognition (FCR) and ingredient prediction (IP), then we conduct the same experiments in a joint learning way (JL). As shown in Table I and II, we can see that the performance of JL exceeds the single task, which means that these two tasks can boost each other. Furthermore, we compare the effect of different components in our method, when we add ingredient region assignment (IRA) into the JL, the Top-1 accuracy and Macro-F1 increase 2.23% and 3.08% respectively. Then when we combine the ingredient attention (IA) to enhance it, we can obtain further improvement. When we add IGRL into JL, the performance of multi-label ingredient prediction is improved incrementally, which means that exploring the ingredient relationships is crucial in our method. Note that for ETH Food-101, the performance of JL+IGRL is slightly lower than JL+IRA+IR, the probable reason is that the food in this dataset is western food, and it is easy to obtain obvious and discriminative ingredient region features. However, for the remaining two food datasets, their ingredients are mixed and exploring their relationships is more significant and thus can obtain better performance.

2) Comparison With the State-of-the-Arts: We compare against the state-of-the-art methods in Table III and Table IV on two tasks. For food category recognition, we can see that our method surpasses all other methods, and can obtain 1.77% performance improvement than the current best method MSMVFA [1]. This indicates the superiority of exploring the ingredient composition and modeling their relationships.

TABLE III
THE PERFORMANCE COMPARISON OF FOOD CATEGORY RECOGNITION ON ETH FOOD-101 (%)

Method	Top-1 acc.	Top-5 acc.
AlexNet-CNN [13]	56.40	-
SELC [34]	55.89	-
ResNet-152+SVM-RBF [35]	64.98	-
DCNN-FOOD [36]	70.41	-
LMBM [37]	72.11	-
Ensemble Net [38]	72.12	91.61
GoogLeNet [39]	78.11	-
DeepFOOD [40]	77.40	93.70
ILSVRC [41]	79.20	94.11
WARN [42]	85.50	-
CNNs Fusion(I_2) [43]	86.71	-
Inception V3 [44]	88.28	96.88
SENet-154 [45]	88.62	97.57
WRN [17]	88.72	97.92
SOTA[46]	90.00	-
DLA[47]	90.00	-
WISeR [17]	90.27	**98.71**
IG-CMAN [14]	90.37	98.42
PAR-Net [19]	89.30	-
Inception-Resnet-v2 SE [48]	90.40	-
MSMVFA [1]	90.59	98.25
SGLANet [20]	89.69	98.01
ReXNet [49]	88.40	-
Our method(ResNet-101)	**92.36**	98.68

TABLE IV
THE PERFORMANCE COMPARISON OF INGREDIENT PREDICTION ON ETH FOOD-101 (%)

Method	Micro-F1	Macro-F1
InceptionV3 [23]	80.06	-
ResNet-50 [23]	80.11	-
ResNet-101 [50]	81.23	80.71
SENet-154 [45]	82.43	81.89
ML-GCN(ResNet-101) [51]	85.49	83.90
SGTN(ResNet-101) [52]	88.89	87.90
DSDL(ResNet-101) [53]	88.50	88.01
ASL(ResNet-101) [54]	82.12	80.27
Our method(ResNet-101)	**91.51**	**90.82**

Moreover, PAR-Net [19] also proposes to mine the discriminative food regions, but ignores the complicated relationships among regions. Our method outperforms PAR-Net by 3.06%, which proves that employing GCN to fully explore the relationships between ingredients has brought larger performance gains. Note that our method has slightly lower performance for Top-5 accuracy compared with WISeR [17], the probable reason is that WISeR is specifically designed to identify western food, especially traditional western food categories in ETH Food-101. For ingredient prediction, since there exist few methods for evaluation on this dataset, we utilize some basic networks like ResNet-101 and some recently proposed multi-label image classification methods like SGTN [52] and DSDL [53] for performance comparison. From Table IV, we can see that our method outperforms all compared state-of-the-arts with significant F1 advantages, and it exceeds

TABLE V
THE PERFORMANCE COMPARISON OF FOOD CATEGORY RECOGNITION ON VIREO FOOD-172 (%)

Method	Top-1 acc.	Top-5 acc.
AlexNet [13]	64.91	85.32
B-CNN [55]	66.07	85.67
DCL [56]	77.06	94.35
VGG-16 [57]	80.41	94.59
VGG16-M [10]	80.73	94.85
DenseNet-161 [58]	86.93	97.17
MTDCNN(VGG-16) [10]	82.06	95.88
MTDCNN(DenseNet-16) [10]	87.21	97.29
SENet-154 [45]	88.71	97.74
PAR-Net [19]	90.20	-
IG-CMAN [14]	90.63	98.40
MSMVFA [1]	90.61	98.31
SGLANet [20]	89.88	97.83
MVANET161 [27]	90.66	98.47
MVANET264 [27]	91.08	98.86
AFN+BFL [12]	89.54	98.05
Our method(ResNet-101)	**93.33**	**99.15**

TABLE VI
THE PERFORMANCE COMPARISON OF INGREDIENT PREDICTION ON VIREO FOOD-172 (%)

Method	Micro-F1	Macro-F1
Arch-A1 [2]	55.17	43.75
Arch-A2 [2]	59.69	43.48
Arch-B [2]	66.32	44.85
Arch-C [2]	63.44	44.26
Arch-D [2]	67.17	47.18
VGG16-M [10]	55.40	34.93
VGG-16 [57]	58.02	33.84
AFN+BFL(VGG-16) [12]	68.65	52.92
AFN+BFL(ResNet-50) [12]	71.38	56.27
AFN+BFL(ResNet-101) [12]	73.47	58.53
ML-GCN(ResNet-101) [51]	70.54	55.32
SGTN(ResNet-101) [52]	73.61	53.97
DSDL(ResNet-101) [53]	66.85	49.44
ASL(ResNet-101) [54]	74.00	58.57
Our method(ResNet-101)	**74.34**	**59.56**

SENet154 [45] by nearly 10% and DSDL by nearly 3%, which suggests that discovering specific ingredient regions can help us better identify the corresponding ingredients, and further verifies the effectiveness of our method.

D. Experiment on Vireo Food-172

Similar to the experiments on ETH Food-101, we first compare various components on Vireo Food-172. As shown in Table I and II, we can see that our method achieves the best performance, which further proves the effectiveness of these components. Table V shows experimental results of Vireo Food-172 about food category recognition. We can see that the performance of our method is better than other methods

TABLE VII

THE PERFORMANCE COMPARISON OF FOOD CATEGORY RECOGNITION ON ISIA FOOD-200 (%)

Method	Top-1 acc.	Top-5 acc.
AlexNet [13]	49.34	79.30
VGG-16 [57]	59.05	86.53
ResNet-152 [50]	61.07	87.87
DenseNet-161 [58]	62.62	88.28
IG-CMAN [14]	67.47	91.75
Our method(ResNet-101)	**69.47**	**92.98**

TABLE VIII

THE PERFORMANCE COMPARISON OF INGREDIENT PREDICTION ON ISIA FOOD-200 (%)

Method	Micro-F1	Macro-F1
VGG-16 [57]	45.66	40.93
ResNet-50 [50]	52.55	45.71
ResNet-101 [50]	52.89	46.12
SENet-154 [45]	54.72	47.38
ML-GCN(ResNet-101) [51]	55.98	48.64
SGTN(ResNet-101) [52]	58.97	55.21
MGTN(ResNet-101) [59]	61.87	60.28
DSDL(ResNet-101) [53]	63.93	62.46
ASL(ResNet-101) [54]	62.74	**64.20**
Our method(ResNet-101)	**64.74**	62.61

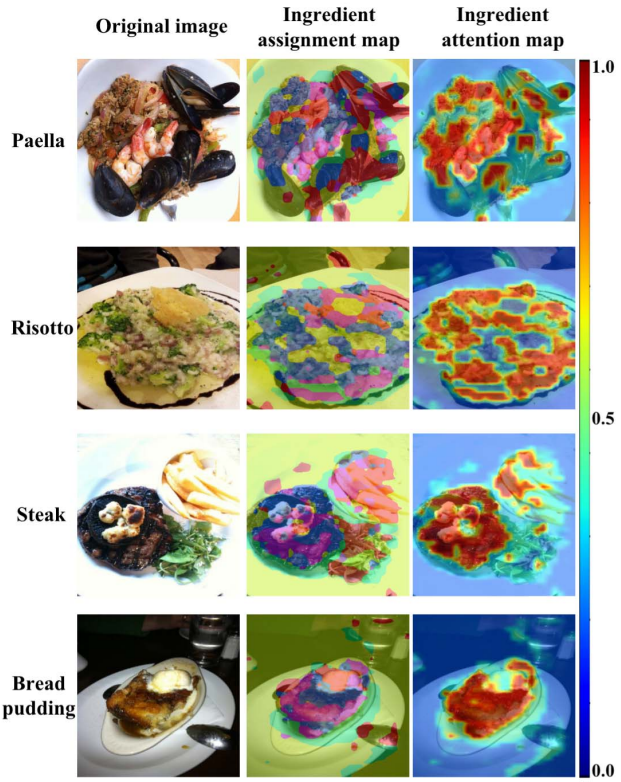


Fig. 3. The assignment maps and attention maps of some samples from our method. For the ingredient assignment maps, different colors mean different ingredient regions, and thus the regions belonging to the same ingredient have been painted with the same color.

for both Top-1 accuracy and Top-5 accuracy. Table VI shows the performance of ingredient prediction on Vireo Food-172. The result of our method is better than all the other methods. For fair comparison, we conduct the experiments under the same backbone as AFN+BFL [12], which is the highest performance for ingredient prediction methods currently. We can see that our method surpasses it by nearly 1.5%. We also utilize the same multi-task strategy in [2] and compare the results with ours, and these results further prove the superiority of our framework. Some recently published methods are also employed for comparison, like DSDL [53] and ASL [54], and we can see that our method also surpasses them. Moreover, the improvement of Vireo Food-172 is less than the one on ETH Food-101. The probable reason is that Vireo Food-172 is a Chinese food dataset, and most ingredients in Chinese food are mixed, which is more difficult to discover them and model their relationships.

E. Experiment on ISIA Food-200

As shown in Table I and II, for the experimental results from various components of our method on ISIA Food-200, we can see that the full model achieves the best 69.47% in Top-1 accuracy and 92.98% in Top-5 accuracy, 64.74% in Micro-F1 and 62.61% in Macro-F1. The food category recognition and ingredient prediction performance from different methods are summarized in Table VII and Table VIII. Because ISIA Food-200 is a newly published dataset and some typical ingredient prediction methods have not experimented on this dataset, we conduct different baselines and some general multi-label image classification methods are used for the ingredient prediction task. The experimental results including VGG, ResNet and DSDL are listed. From Table VII we can see that our method achieves the state-of-the-art performance. This again verifies the effectiveness of the proposed method. For the experimental results for ingredient prediction in Table VIII, we can see that our method surpasses all other methods for Micro-F1, even recently proposed method like DSDL(ResNet-101) [53] and SGTN(ResNet-101) [52]. Notice that our method is weaker than ASL(ResNet-101) [54] for Macro-F1, the probable reason is ASL adopts the asymmetric loss to solve the ground-truth mislabeling problem and high negative-positive imbalance, which makes it discard mislabeled samples and emphasize features learning from both positive and negative samples. Therefore, it can cope well the imbalance distribution of the datasets, and the average precision and the average precision and recall of all categories can be improved, resulting in the improved Macro-F1 of ASL.

F. Qualitative Analysis and Visualization

Our method achieves the state-of-the-art performance on food category recognition and ingredient prediction. We further visualize the assignment maps and attention maps from our method in Fig. 3. We can see that our method can discover various coherent ingredient regions and mark them with different colors, and the discriminative region can be

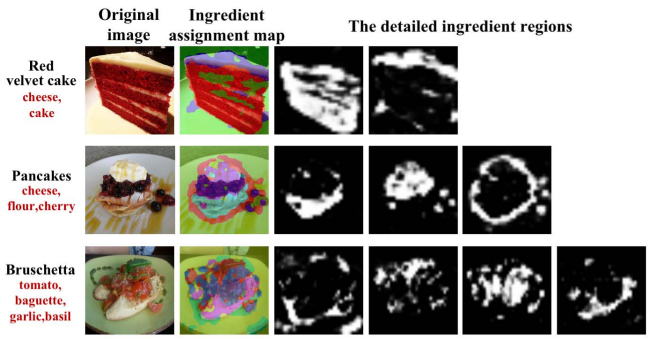


Fig. 4. The detailed ingredient regions from the assignment map. The food category name (in black) and ingredients (in red) are listed in the left.

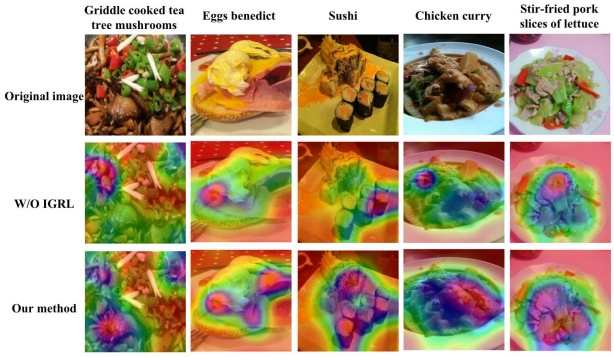


Fig. 5. Visualization results of proposed method and the one without IGRL on some samples.

further attended in the attention map. Take "Bread pudding" for example, some ingredients like bread and egg can be localized,and the regions belonging to the same ingredients are painted with the same color. Besides, we also notice that ingredients cannot be localized very precise. In some cases, the large scale regions are often attended, while small-size ingredients are overlooked. Moreover, there exists an invisible relationship between these two maps. The most discriminative region bread has been discovered in the attention map, and the attended region is exactly the area where the bread is localized in the assignment map.

Moreover, we also display the detailed ingredient regions from our method in Fig. 4. Because different food categories contain different quantities of ingredients, they can be grouped into the corresponding number of ingredient regions. For "Bruschetta" in Fig. 4, we can see that our method can group it into four regions, which are tomato, baguette, garlic and basil. These qualitative results show that our method is able to discover meaningful ingredient regions and extract those regions are discriminative for final recognition.

In order to further reveal the effectiveness of modeling ingredient relationship, we further utilized Grad-CAM to visualize the feature maps of our method for some samples, and compared it with the one w/o IGRL. The experimental results are shown in Figure 1, we can see that for the same food, the complete method can discover expanded ingredient regions, like "Chicken curry". Furthermore, it can extract multiple ingredient regions and obtain a more comprehensive visual representation. For example, the method without IGRL only can find one ingredient region for "Eggs benedict" in Fig. 5, while the complete method can obtain three detailed ingredient regions. The probable reason is that IGRL can model the ingredient relationship by GCN and it can further find those co-occurrence ingredients.

In addition, Fig. 6 shows some experimental results of some samples from our method. The true positive, false positive and false negative ingredients are attached with different colors, respectively. Note that only test results are reported because we only intend to show its generalization capabilities on new data. However, it is not always true for ingredient prediction in Fig. 6. The probable reasons include mixed ingredients

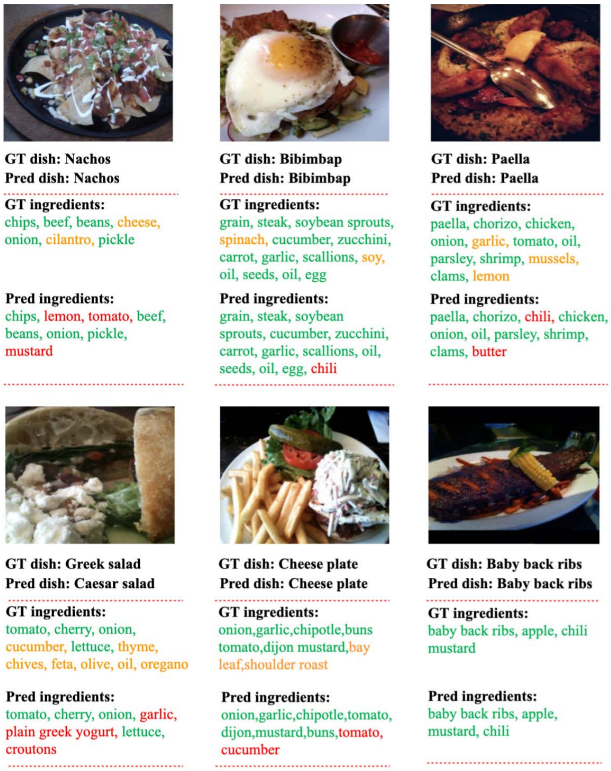


Fig. 6. The experimental results of some samples from our method. True positive ingredients in green, false positive ingredients in red and false negative ingredients in orange. GT means the ground truth.

without clear division, too small ingredients and the change of spatial structure of ingredients, etc. In these cases, our method probably fails to recognize them correctly, just like "Greek salad" and "Caesar salad" in Fig. 6. From Fig. 6, we can see that our model makes a wrong prediction for the "Greek salad". The probable reason is that the visual patterns for these two foods are very similar, and they have some common ingredients like lettuce and garlic.

V. Conclusion

In this paper, we propose a multi-task learning framework for simultaneous food category and ingredient prediction,

where we learn an ingredient dictionary and leverage one U-shaped prior for region-based ingredient discovery, and we also propose to utilize an ingredient-oriented semantic-visual graph convolutional network for ingredient relationship modeling. For the ingredient region discovery, we build the corresponding ingredient dictionary and employ it to group the feature maps into various ingredient regions and further re-weight them with attention. For ingredient relationship modeling, we explore both semantic and visual information for ingredient graph construction, and utilize GCN for better representation learning. These two branches can promote each other via multi-task learning. Comprehensive experimental results on three popular datasets have demonstrated the effectiveness of our method. Further visualization of ingredient assignment maps and attention maps show the superiority of our method. Such improvement benefits from both region-based ingredient discovery and the ingredient-oriented semantic-visual graph convolutional network.

Future work includes: (1) The ingredient distribution among most food datasets is imbalanced, which may seriously influence performance, and thus we need to explore re-balanced samplings [60] or balance loss [12], [61], [62] to solve this problem. (2) We plan to explore transformers for food recognition, which have made a tremendous impact on image recognition [63]–[65], and the performance is higher than CNNs. Therefore, transformer-based recognition methods can be explored to further improve the performance of food category recognition and ingredient prediction in the future. (3) Note that we only use visible ingredients for the region-based ingredient discovery in IAR, and thus IAR can only discover visible ingredients. However, many ingredients cannot be seen and need to be reasoned, like sugar and salt, which play an important role in the food nutritional evaluation. Therefore, exploring the invisible ingredients needs further study, and we plan to create a hierarchical structure relationship [5], [37] of ingredients and extend the model to reason invisible ingredients.

REFERENCES

[1] S. Jiang, W. Min, L. Liu, and Z. Luo, "Multi-scale multi-view deep feature aggregation for food recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 265–276, 2020.

[2] J. Chen, B. Zhu, C.-W. Ngo, T.-S. Chua, and Y.-G. Jiang, "A study of multi-task and region-wise deep learning for food ingredient recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 1514–1526, 2021.

[3] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," *ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–36, Sep. 2020.

[4] Q. Thames *et al.*, "Nutrition5k: Towards automatic nutritional understanding of generic food," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8903–8911.

[5] J. Chen, L. Pan, Z. Wei, X. Wang, C. Ngo, and T. Chua, "Zero-shot ingredient recognition by multi-relational graph convolutional network," in *Proc. Assoc. Advancement Artif. Intell.*, vol. 34, no. 7, 2020, pp. 10542–10550.

[6] A. Myers *et al.*, "Im2Calories: Towards an automated mobile vision food diary," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1233–1241.

[7] N. Nag, V. Pandey, and R. Jain, "Health multimedia: Lifestyle recommendations based on diverse observations," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2017, pp. 99–106.

[8] W. Min, S. Jiang, and R. Jain, "Food recommendation: Framework, existing solutions, and challenges," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2659–2671, Oct. 2020.

[9] E. Aguilar, B. Remeseiro, M. Bolanos, and P. Radeva, "Grab, pay, and eat: Semantic food detection for smart restaurants," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3266–3275, Dec. 2018.

[10] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 32–41.

[11] X.-J. Zhang, Y.-F. Lu, and S.-H. Zhang, "Multi-task learning for food identification and analysis with deep convolutional neural networks," *J. Comput. Sci. Technol.*, vol. 31, no. 3, pp. 489–500, May 2016.

[12] C. Liu, Y. Liang, Y. Xue, X. Qian, and J. Fu, "Food and ingredient joint learning for fine-grained recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2480–2493, Aug. 2021.

[13] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 446–461.

[14] W. Min, L. Liu, Z. Luo, and S. Jiang, "Ingredient-guided cascaded multi-attention network for food recognition," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1331–1339.

[15] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2249–2256.

[16] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1085–1088.

[17] N. Martinel, G. L. Foresti, and C. Micheloni, "Wide-slice residual networks for food recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 567–576.

[18] S. Horiguchi, S. Amano, M. Ogawa, and K. Aizawa, "Personalized classifier for food image recognition," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2836–2848, Oct. 2018.

[19] J. Qiu, F. P. W. Lo, Y. Sun, S. Wang, and B. Lo, "Mining discriminative food regions for accurate food recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2019. [Online]. Available: https://bmvc2019.org/wp-content/uploads/papers/0839-paper.pdf

[20] W. Min *et al.*, "ISIA food-500: A dataset for large-scale food recognition via stacked global-local attention network," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 393–401.

[21] H. Zhao, K.-H. Yap, and A. C. Kot, "Fusion learning using semantics and graph convolutional network for visual food recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1711–1720.

[22] F. Zhou and Y. Lin, "Fine-grained image classification by exploring bipartite-graph labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1124–1133.

[23] M. Bolaños, A. Ferrá, and P. Radeva, "Food ingredients recognition through multi-label learning," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2017, pp. 394–402.

[24] W. Min, S. Jiang, J. Sang, H. Wang, X. Liu, and L. Herranz, "Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1100–1113, May 2017.

[25] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, Jul. 1997.

[26] T. Ege and K. Yanai, "Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions," in *Proc. Thematic Workshops ACM Multimedia Thematic Workshops*, 2017, pp. 367–375.

[27] H. Liang, G. Wen, Y. Hu, M. Luo, P. Yang, and Y. Xu, "MVANet: Multi-task guided multi-view attention network for Chinese food recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 3551–3561, 2021.

[28] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 433–442.

[29] A. Andoni, P. Indyk, and R. Krauthgamer, "Earth mover distance over high-dimensional spaces," in *Proc. SODA*, vol. 8, 2008, pp. 343–352.

[30] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[32] Z. Huang and Y. Li, "Interpretable and accurate fine-grained recognition via region grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8662–8672.

[33] A. Salvador *et al.*, "Learning cross-modal embeddings for cooking recipes and food images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3020–3028.

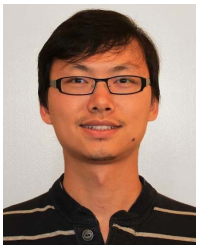
[34] N. Martinel, C. Piciarelli, and C. Micheloni, "A supervised extreme learning committee for food recognition," in *Comput. Vis. Image Understand.*, vol. 148, 2016, pp. 67–86.
[35] P. McAllister, H. Zheng, R. Bond, and A. Moorhead, "Combining deep residual neural network features with supervised machine learning algorithms to classify diverse food image datasets," *Comput. Biol. Med.*, vol. 95, pp. 217–233, Apr. 2018.
[36] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW)*, Jun. 2015, pp. 1–6.
[37] H. Wu, M. Merler, R. Uceda-Sosa, and J. R. Smith, "Learning to make better mistakes: Semantics-aware visual food recognition," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 172–176.
[38] P. Pandey, A. Deepthi, B. Mandal, and N. B. Puhan, "FoodNet: Recognizing foods using ensemble of deep networks," *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1758–1762, Dec. 2017.
[39] S. Ao and C. X. Ling, "Adapting new categories for food recognition with deep representation," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 1196–1203.
[40] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma, "Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment," in *Proc. Int. Conf. Smart Homes Health Telematics*, 2016, pp. 37–48.
[41] M. Bolanos and P. Radeva, "Simultaneous food localization and recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3140–3145.
[42] P. Rodriguez, D. Velazquez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. Gonzalez, "Pay attention to the activations: A modular attention mechanism for fine-grained image recognition," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 502–514, Feb. 2020.
[43] E. Aguilar, M. Bolaños, and P. Radeva, "Food recognition using fusion of classifiers based on CNNs," in *Proc. Int. Conf. Image Anal. Process.*, 2017, pp. 213–224.
[44] H. Hassannejad, G. Matrella, P. Ciampolini, I. D. Munari, M. Mordonini, and S. Cagnoni, "Food image recognition using very deep convolutional networks," in *Proc. 2nd Int. Workshop Multimedia Assist. Dietary Manage.*, Oct. 2016, pp. 41–49.
[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[46] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2661–2671.
[47] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.
[48] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4109–4118.
[49] D. Han, S. Yun, B. Heo, and Y. Yoo, "Rethinking channel dimensions for efficient model design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 732–741.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[51] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5177–5186.
[52] X.-S. Vu, D.-T. Le, C. Edlund, L. Jiang, and H. D. Nguyen, "Privacy-preserving visual content tagging using graph transformer networks," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2299–2307.
[53] F. Zhou, S. Huang, and Y. Xing, "Deep semantic dictionary learning for multi-label image classification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1–10.
[54] E. Ben-Baruch *et al.*, "Asymmetric loss for multi-label classification," 2020, *arXiv:2009.14119*.
[55] T. Y. Lin, A. Roychowdhury, and S. Maji, "Bilinear convolutional neural networks for fine-grained visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1309–1322, Jun. 2017.
[56] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 5157–5166.
[57] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[58] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[59] H. D. Nguyen, X.-S. Vu, and D.-T. Le, "Modular graph transformer networks for multi-label image classification," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1–9.
[60] H. Guo and S. Wang, "Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15089–15098.
[61] M. Hayat, S. Khan, S. W. Zamir, J. Shen, and L. Shao, "Gaussian affinity for max-margin class imbalanced learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6469–6479.
[62] S. Khan, M. Hayat, S. W. Zamir, J. Shen, and L. Shao, "Striking the right balance with uncertainty," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 103–112.
[63] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16478–16488.
[64] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[65] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

Zhiling Wang received the B.E. degree from the School of Geodesy and Geomatics, Wuhan University, Wuhan, China, in 2019, and the master's degree in computer science from the University of Chinese Academy of Sciences, Beijing, China, in 2022. He will work in the Meituan Vision AI Department as an Algorithm Engineer, Beijing. His research interests include food computing and fine-grained image recognition and retrieval.

Weiqing Min (Senior Member, IEEE) is currently an Associate Professor with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences. He has authored or coauthored more than 50 peer-reviewed papers in relevant journals and conferences, including *Patterns* (Cell Press), *ACM Computing Surveys*, *Trends in Food Science and Technology*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, ACM MM, AAAI, and IJCAI. His research interests include multimedia content analysis and food computing. He was a Senior Member of CCF. He was a recipient of the 2016 *ACM TOMM* Nicolas D. Georganas Best Paper Award and the 2017 *IEEE Multimedia Magazine* Best Paper Award. He was the Area Chair of ICME2022/ACM MM2021. He organized several special issues on international journals, such as IEEE MULTIMEDIA and *Neurocomputing* as a Guest Editor.

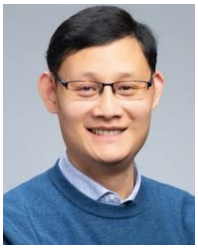
Zhuo Li received the B.E. degree from the School of Computer Science, Northwestern Polytechnical University, Xi'an, China, in 2019, and the master's degree in computer science from the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, 2022. He will work in Tencent as one Researcher, Beijing. His research interests include computer vision, machine learning, and image retrieval.

Liping Kang received the degree from Xi'an Jiaotong University in July 2013 and the master's degree from the Chinese Academy of Sciences University in July 2016. She is currently an Algorithm Expert with Meituan Vision AI Department, Beijing, China. She has applied for 17 patents as the first inventor, and ten have been authorized. Her current research interests include deep learning, computer vision, fine-grained image recognition and retrieval, and their applications.



Xiaoming Wei is currently the Leader of the Vision Understanding Group, Meituan Vision AI Department. He has published over ten papers in CVPR, ECCV, IJCAI, ACM MM, and AAAI. His research interests focus on fine-grained image recognition and retrieval. He has led the team and got top rankings in several fine-grained matches such as Herbarium 2022 FGVC9 (the 1st place) and iMaterialist Challenge on Product Recognition in CVPR2019 (the 2nd place).

Xiaolin Wei received the Ph.D. degree in computer science from Texas A&M University. He worked as a Research Engineer at Google, the CEO of Virtroid, and the Principal Engineer of Magic Leap. He is currently leading the Computer Vision Division, Meituan. He has been granted over 40 patents and published over 30 papers in SIGGRAPH, ICCV, ECCV, ACM MM, and IJCAI. His research interests include computer vision, machine learning, computer graphics, 3D vision, and augmented reality.



Shuqiang Jiang (Senior Member, IEEE) is a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS); and a Professor with the University of CAS. He is also with the Key Laboratory of Intelligent Information Processing, CAS. He was supported by the National Science Fund for Distinguished Young Scholars in 2021. His research interests include multimedia analysis and multimodal intelligence. He has authored or coauthored more than 200 papers on the related research topics. He won the CAS International Cooperation Award for Young Scientists, the CCF Award of Science and Technology, the Wu Wenjun Natural Science Award for Artificial Intelligence, the CSIG Natural Science Award, and the Beijing Science and Technology Progress Award. He is the Vice Chair of the IEEE CASS Beijing Chapter and the ACM SIGMM China Chapter and an Associate Editor of *ACM TOMM*.