Ingredient-Guided Cascaded Multi-Attention Network for Food Recognition

Weiqing Min[1], Linhu Liu[1,2], Zhengdong Luo[1,2],Shuqiang Jiang[1,2]
[1] Key Lab of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
{minweiqing,sqjiang,luozhengdong}@ict.ac.cn;linhu.liu@vipl.ict.ac.cn

ABSTRACT

Recently, food recognition is gaining more attention in the multimedia community due to its various applications, e.g., multimodal foodlog and personalized healthcare. Most of existing methods directly extract visual features of the whole image using popular deep networks for food recognition without considering its own characteristics. Compared with other types of object images, food images generally do not exhibit distinctive spatial arrangement and common semantic patterns, and thus are very hard to capture discriminative information. In this work, we achieve food recognition by developing an Ingredient-Guided Cascaded Multi-Attention Network (IG-CMAN), which is capable of sequentially localizing multiple informative image regions with multi-scale from category-level to ingredient-level guidance in a coarse-to-fine manner. At the first level, IG-CMAN generates the initial attentional region from the category-supervised network with Spatial Transformer (ST). Taking this localized attentional region as the reference, IG-CMAN combined ST with LSTM to sequentially discover diverse attentional regions with fine-grained scales from ingredient-guided sub-network in the following levels. Furthermore, we introduce a new dataset WikiFood-200 with 200 food categories from the list in the Wikipedia, about 200,000 food images and 319 ingredients. We conduct extensive experiment on two popular food datasets and newly proposed WikiFood-200, demonstrating that our method achieves the state-of-the-art performance in Top-1 accuracy. Qualitative results along with visualization further show that IG-CMAN can introduce the explainability for localized regions, and is able to learn relevant regions for ingredients.

KEYWORDS

Food Recognition, Multi-Attention Network, Ingredients

1 INTRODUCTION

Food recognition has attracted more and more attention in computer vision and multimedia [3], [19], [10], [11], [14] in recent years. It is an important and basic step for food image analysis, leading to deep understanding of food from different perspectives, such as health and culture. Automatically recognizing food can also enable various applications. Once we recognize the category of the meal, we can further conduct calorie estimation [19], health-aware recommendation [26], dietary tracking [24] and eating habit analysis [20, 27]. It is particularly helpful for many commercial scenarios, such as fast food restaurants, smart restaurants, grocery stores and supermarkets. For example, in the self-service restaurants, food recognition can help bill the grabbed meal by customers via recognizing food and monitoring food consumption. In addition, food recognition can help restaurant review platforms like Yelp and Foursquare to categorize user-shared content automatically for food photo organization and management.

Figure 1: Some food samples with rich ingredients

Image recognition has undergone a fundamental paradigm shift towards using deep learning as a general-purpose solution for its powerful capability of discriminative feature learning, and food recognition is no exception. To our knowledge, Kagaya *et al.* [14] applied a Convolutional Neural Network (CNN) to the task of food recognition in the multimedia community for the first time. Later, different neural networks are directly used for food recognition, such as GoogLeNet [19], Network-In-Network [29], Inception V3 [8] and ResNet [24]. Recently, Martinel *et al.* [18] combined the wide-slice residual network and its variant for food recognition. However, most of these works simply adopt CNN to extract visual features for food recognition without considering its special features, and thus probably lead to suboptimal performance.

Food recognition belongs to fine-grained recognition, which refers to the task of distinguishing sub-ordinate categories, such as birds and cars. A key to address this problem is to localize discriminative parts for feature extraction. Existing works focused on multiple semantic region localization by category-supervised

MM'19, October 21-25, 2019, Nice, France.

DOI: https://doi.org/10.1145/3123266.3123272

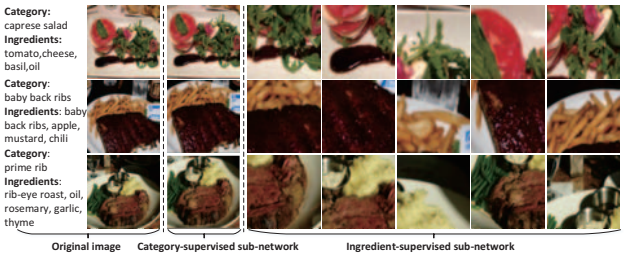

Figure 2: Localized regions of some food images from our proposed method under both category-level and ingredient-level guidance, where five discriminative regions are localized based on multi-ingredient supervision. All these localized regions are resized into the same fixed size.

CNN [36]. However, image-level category labels only provide weak supervised information. Therefore, CNNs trained with category labels can miss fine-grained food regions, which could provide complementary information, and are probably not optimal to guide multiple region localization. Furthermore, existing fine-grained categories have fixed semantic parts, and distinctive relationships between semantic parts and the whole. Under this assumption, these methods mainly localized fixed semantic regions, and meanwhile utilized such relation constraints to remove unreasonable regions for feature extraction. However, compared with these objects, many types of food are non-rigid, and do not exhibit distinctive spatial configuration and fixed semantic patterns. Therefore, it is hard to capture discriminative semantic information from food images via existing fine-grained methods.

On the other hand, the profusion of online recipe-sharing websites with user-uploaded food photos provides additional ingredient information (as shown in Fig. 1). Like the importance of objects for the scene, we argue that the use of semantically meaningful ingredients, as basic units of food images, probably offer one promising venue to empower a visual recognizer to arbitrary food images. Compared with existing fine-grained methods, which explore multiple discriminative regions via the deep network with weakly-supervised category guidance, ingredients can be used to explicitly guide the network to discover diverse semantic regions over fine-grained image scales. These regional features generated from the network under the supervision from different granularity are very complementary. Therefore, integrating diverse regional features are not only based on the global shape or appearance variation but also local parts or patterns, leading to more comprehensive and discriminative representation. Furthermore, associating ingredients with regions can introduce the explainability for localized regions.

To this end, we propose an Ingredient-Guided Cascaded Multi-Attention Network (IG-CMAN) for food recognition, which is capable of sequentially localizing diverse attentional regions over different image scales from category-level to ingredient-level guidance. IG-CMAN first generates the initial attentional region from first-level category-supervised sub-network. Taking this initial attentional region as the reference, at the following levels, IG-CMAN iteratively discovers diverse attentional regions from ingredient-supervised sub-network. In this way, our approach enables to learn contextualized and interpretable multi-scale regions corresponding to ingredients while improving the discriminability for food recognition. Particularly, IG-CMAN mainly consists of two components: 1) a Spatial Transformer (ST) layer to locate attentional regions under different types of supervised signals and 2) a Long-Short Term Memory (LSTM) to sequentially predict ingredient scores on the located regions from ST. Combing LSTM and ST can iteratively localize diverse regions, and associate ingredients with attentional regions. We finally fuse features from discovered multi-scale attentional regions into final feature representation for food classification. Once IG-CMAN has been trained, we can obtain multi-scale representations for food recognition from full-size images to multiple coarse-to-fine attentional regions. Three examples generated by IG-CMAN are illustrated in Fig. 2.

Furthermore, we propose a new food dataset WikiFood-200 for real-world food recognition and analysis with 200 food categories from the list in the Wikipedia, 200,000 images and 319 ingredients. Compared with existing datasets for food recognition, such as FoodCam-256 [15], WikiFood-200 provides additional ingredient information. In addition, it shared less common categories with existing datasets with ingredients, such as VireoFood-172 [4]. Therefore, WikiFood-200 is very complementary to these released datasets for food recognition in the food domain, and is very helpful for prompting the development of food computing in the multimedia.

Our main contributions can be summarized as follows: (1) We develop an Ingredient-Guided Cascaded Multi-Attention Network (IG-CMAN) to sequentially localize diverse multi-scale image regions for discriminative feature extraction from category-level to ingredient-level guidance in a coarse-to-fine manner. (2) We introduce a new food dataset WikiFood-200 for food recognition with 200 food categories, 200,000 food images and 319 ingredients[1]. (3) We conduct extensive experiment and evaluations on three large-scale benchmark datasets including ETH Food-101 [3], VireoFood-172 and newly proposed WikiFood-200, and achieve the state-of-the-art performance on all these datasets in Top-1 accuracy.

2 RELATED WORK

Food Recognition. Recently, Min *et al.* [21] provided a comprehensive survey on food recognition and other food-related works. In the earlier years, various hand-crafted features are extracted from food images for recognition [3, 35]. For example, Yang *et al.* [35] adopted one statistical method to exploit spatial relationships between ingredients for food recognition. Lukas *et al.* [3] utilized random forests to mine discriminative patches of food images as visual representation. Deep learning's revolutionary advances in image recognition have gained significant attention because of its powerful expressive capacity. As a result, more works on food recognition resort to deep learning for feature extraction [14, 18]. Some works further utilized additional context information, such as GPS [10, 32] and restaurant information [30] to improve the recognition performance. Zhou *et al.* [37] exploited rich relationships among ingredients, food category and restaurant information through the bi-partite graph for food classification. Recently, Horiguchi *et al.* [11] further proposed

[1] http://isia.ict.ac.cn/dataset/WikiFood-200.html

a sequential personalized classifier to learn new food classes incrementally for personalized food recognition, where visual features are extracted from GoogLeNet.

In addition, our work is also relevant to multimodal recipe analysis [4, 5, 23, 28]. Different types of food labels, such as food types, ingredients and other attributes can be learned simultaneously via multi-task learning. For example, some works adopted different multi-task deep learning architectures, such as CNN [4] and Deep Boltzmann Machine [22] for food recognition, where ingredients are fully exploited as supervised information for fine-tuning the network. Our work also adopts a multi-task learning framework. However, different from existing works, which simply learn two types of labels by connecting the label layer to the full-connected layer, we sequentially localize diverse multi-scale attentional regions over different image scales from category-level to ingredient-level guidance, and then fuse features from these multi-scale regions for food recognition.

Visual Attention. Attention models have been recently applied to various tasks, such as image classification [25], visual place recognition [38] and person re-identification [34]. Earlier works use the Recurrent Neural Network (RNN) for sequential attentions, and optimize their models with reinforcement learning. For example, Mnih *et al.* [25] present a RNN for object detection by adaptively selecting a sequence of attentional regions and extracting appearance representations in these regions. Jaderberg *et al.* [13] proposed the Spatial Transformer Network (STN), which can provide the spatial transformation capability to extract attentional regions. This makes networks not only select regions of an image that are most relevant, but also to transform those regions to enhance the recognition performance. Our work also utilizes ST for attentional region localization. However, we further combine ST with LSTM in a cascaded way to sequentially localize diverse attentional regions with different scales for food categories and ingredients.

In addition, our work is relevant to fine-grained image recognition [6, 36]. For example, Fu *et al.* [6] proposed a recurrent attention deep network to recursively learn discriminative regions. Different from [6], which tries to focus on only one sub-region, and then delves into more details of this sub-region, our work uses rich ingredient information to explicitly guide the network to localize diverse sub-regions. Our method is also inspired by work [31], but with two important differences: (1) [31] focuses on multi-label classification, while we explore attentional regions under multiple ingredient based guidance for food category recognition; (2) [31] directly localizes the image regions without such hierarchical coarse-to-fine structure modeling. In contrast, we design a cascaded multi-attention network, which is capable of sequentially localizing diverse multi-scale image regions from category-level to ingredient-level guidance in a coarse-to-fine manner.

3 METHOD

In this section, we will introduce the proposed Ingredient-Guided Cascaded Multi-Attention Network (IG-CMAN) for food recognition. Fig. 3 illustrates the architecture of IG-CMAN, which is decomposed into two main components, namely Category-supervised Attention Sub-Network (CASN) and Ingredient-supervised Attention Sub-Network (IASN). CASN is used to localize coarse attentional region of the full food image while IASN is used to capture multiple fine-grained attentional regions with smaller scales based on the localized coarse region from CASN. To implement this, IG-CMAN cascades several CNNs in a hierarchical way, where each sub-network at each level includes ST and LSTM. ST for each level is used to localize the attentional region, while LSTM from different levels is stacked together to model global sequential dependencies of these localized regions, and meanwhile generates transformation parameters for ST in the next level. To this end, IG-CMAN adopts a multi-task learning formulation with both category loss and ingredient loss, and is trained in an end-to-end fashion.

3.1 CASN

As shown in the top of Fig. 3, CASN is a category-supervised STN with an additional LSTM. Compared with traditional CNN, one ST layer is embedded into CNN to form STN, which can spatially transform its input maps to output maps with a given size, which correspond to a subregion of input maps. For ST in CASN, a transformation matrix M_0 is first estimated by a localization network. After that, the corresponding coordinate grid in f_0 is obtained, based on coordinates of f_1, where f_0 is feature maps from one CNN, such as VGG-16. Then the sampled feature maps f_1 that correspond to the attentional region are generated by bilinear interpolation. That is $f_1 = \mathrm{ST}(f_0, M_0)$, where ST is the spatial transformation function. M_0 involves cropping, translation and scaling, and is expressed as

$$\begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix}$$

where s_x, s_y, t_x, t_y are scaling and translation parameters.

Different from traditional STN, one LSTM is introduced in CASN, which is used to combine with the following LSTMs to construct stacked LSTMs for sequential dependency modeling of localized regions.

After STN, we obtain the transformed input x_1 for LSTM.

$$f_1 = \mathrm{ST}(f_0, M_0) \qquad x_1 = \mathrm{relu}(W_{fx} f_1 + b_x) \tag{1}$$

where relu() is the rectified linear function. W_{fx} and b_x are transformation parameters.

We can obtain the hidden representation h_1 and cell state c_1 via LSTM. Based on the output h_1, LSTM can not only predict the category labeling score based on s_1, but also output parameters M_1 for the following ST:

$$\begin{aligned} z_1 &= \mathrm{relu}(W_{hz} h_1 + b_z) \\ s_1 &= W_{zs} z_1 + b_s \\ M_1 &= W_{zm} z_1 + b_m \end{aligned} \tag{2}$$

where W_{hz},W_{zs},W_{zm},b_s,b_z and b_m are transformation parameters.

3.2 IASN

Based on localized regions f_1 and transformation matrix M_1 from CASN, in IASN, the stacked LSTM and ST work collaboratively in an iterative manner: LSTM predicts ingredient scores regarding this localized region from ST and simultaneously updates transformation parameters of ST for the next attentional region localization.

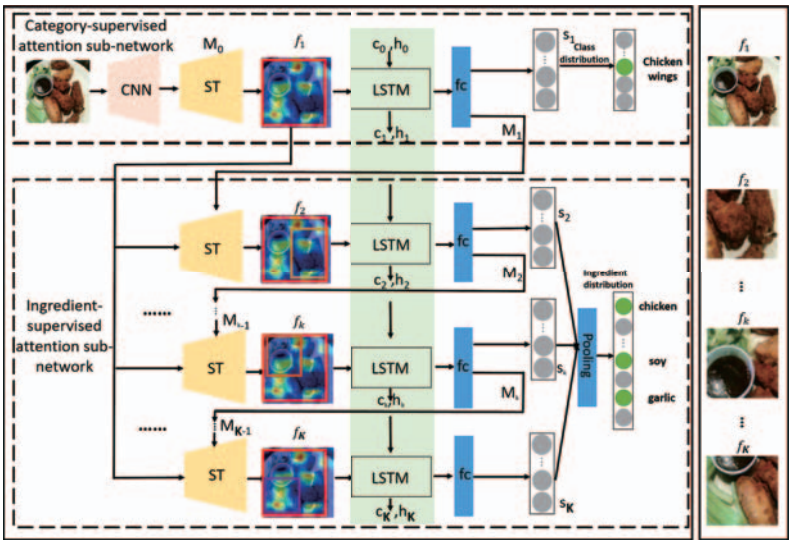

Figure 3: Overview of proposed Ingredient-Guided Cascaded Multi-Attention Network (IG-CMAN) for food recognition, where we also show localized regions corresponding to feature maps on the right of the framework.

For each sub-network in IASN, it all takes localized coarse region f_1 as the reference and used updated parameters M_{k-1} to calculate fine-grained localized region: $f_k = \text{ST}(f_1, M_{k-1})$. LSTM takes the sampled feature map f_k as input to compute memory cells and hidden states:

$$\begin{aligned} x_k &= \text{relu}(W_{fx} f_k + b_x) \\ f_k &= \sigma(W_{xf} x_k + W_{hf} h_{k-1} + b_f) \\ i_k &= \sigma(W_{xi} x_k + W_{hi} h_{k-1} + b_i) \\ g_k &= \tanh(W_{xg} x_k + W_{hg} h_{k-1} + b_g) \\ c_k &= f_k \odot c_{k-1} + i_k \odot g_k \\ o_k &= \sigma(W_{xo} x_k + W_{ho} h_{k-1} + b_o) \\ h_k &= o_k \odot c_k \end{aligned} \tag{3}$$

where $\sigma()$ is the sigmoid function, tanh() is the hyperbolic tangent function, $\odot$ is pointwise multiplication. h_{k-1} and c_{k-1} are the hidden state and memory cell of previous iteration; i_k, f_k, o_k and g_k are outputs of the input gate, forget gate, output gate, and input modulation gate, respectively at the k-level sub-network.

Given the hidden state h_k, we update M_k as follows:

$$z_k = \text{relu}(W_{hz} h_k + b_z) \qquad M_k = W_{zm} z_k + b_m \tag{4}$$

where M_k is the transformation matrix for the $k+1$ level.

3.3 Cascaded Multi-Attention Network

The collaboration between CASN and IASN leads to cascaded multi-attention network. At the first level $k = 1$, CASN localizes the coarse region f_1 from the original input feature map f_0. The following levels of IASN take the localized region f_1 as the input for fine-grained region-localization f_k. They are expressed as follows at the k-th level:

$$\begin{aligned} f_1 &= \text{ST}(f_0, M_0) \qquad k = 1 \\ f_k &= \text{ST}(f_1, M_{k-1}) \qquad k > 1 \end{aligned} \tag{5}$$

The LSTM takes sampled feature map f_k as input to compute the memory cell and hidden state, and it is first changed into the input for LSTM. Then the following computation process from LSTM can be conducted using Eqn. 3 to obtain h_k at the k-th level. Given the hidden state h_k, we update M_k using Eqn. 4. Note that at the first level, we directly estimate M_0 via CASN. Based on M_0, we update M_k in IASN.

3.4 Multi-Task Learning

We finally model IG-CMAN in a multi-task formulation, which is optimized mainly by two types of losses, i.e., category-level classification loss L_{cls} and ingredient-level attribute learning loss L_{ing}, for generating large-scale coarse image region and multiple fine-grained smaller image regions, respectively. In addition, we utilize another type of loss L_{loc} for attentional region localization constraints from ST to guarantee the localization accuracy of attentional regions, leading to the following loss function:

$$L = L_{cls} + \gamma_1 L_{ing} + \gamma_2 L_{loc} \tag{6}$$

where γ_1 and γ_2 are balance parameters.

Category-level Classification Loss. CASN adopts the food category as supervised information to guide STN to localize the attentional region. Through LSTM in CASN, the input is finally changed into s_1 using Eqn. 2. The cross-entropy classification loss function is adopted as follows:

$$L_{cls} = -\frac{1}{N} \sum_i \log(P(y_i | s_{1i})) \tag{7}$$

where N is the number of training samples. s_{1i} is the feature representation for the i-th sample and y_i is the corresponding food category.

Ingredient-level Attribute Learning Loss. In IASN, we obtain its final feature representation s_k for each level as follows:

$$z_k = \text{relu}(W_{hz}h_k + b_z) \quad s_k = W_{zs}z_k + b_s \tag{8}$$

where s_k is the predicted ingredient score distribution of the k-th level sub-network. The following $K-1$ level sub-network in IASN results in K-1 score vectors $\{s_2, ..., s_k, ...s_K\}$, where $s_k = \{s_k^1, ..., s_k^v, ..., s_k^V\}$ denotes scores over V ingredient labels. The ingredient-wise max-pooling is then used to fuse scores into the final result $s = \{s^1, ..., s^v, ..., s^V\}$, and s^v for each ingredient is calculated as follows:

$$s^v = \max(s_2^v, ..., s_K^v), v = 1, 2, 3, ..., V. \tag{9}$$

We then obtain the predicted probability vector p_i

$$p_i^v = \frac{\exp(s_i^v)}{\sum_{m=1}^{V}\exp(s_i^m)} \qquad v = 1, 2, ..., V, \tag{10}$$

This loss function is finally expressed as

$$L_{ing} = \frac{1}{N}\sum_i\sum_v(p_i^v - \hat{p}_i^v)^2 \tag{11}$$

where $\hat{p}_i = q_i/\|q_i\|_1$ is the ground-truth probability vector of the i-th sample, and its ingredient vector $q_i = \{q_i^1, q_i^2, ..., q_i^V\}$, q_i^v is a $0-1$ indicator vector.

Attentional Region Localization Loss Similar to [31], to make STN successfully localize diverse multi-scale image regions, we also adopt the following three types of losses, including anchor constraint,scale constraint and positive constraint.

For anchor constraint, this constraint makes attentional regions scatter over different semantic regions in the food image. It is formulated as

$$\Gamma_A = \frac{1}{2}\{\left(t_x^k - c_x^k\right)^2 + \left(t_y^k - c_y^k\right)^2\} \tag{12}$$

where $\left(c_x^k, c_y^k\right)$ is the location of the k-th anchor point, and t_x^k, t_y^k are horizontal and vertical translation, respectively for the k-level sub-network ($k \geq 2$).

For scale constraint, this constraint is used to push the located attentional region in a certain range, and can be formulated as

$$\Gamma_S = \left(\max\left(|s_x| - \alpha, 0\right)\right)^2 + \left(\max\left(|s_y| - \alpha, 0\right)\right)^2 \tag{13}$$

where α is a threshold value. Not that in different sub-networks, α is different. For example, α should be large in CASN since CASN is used to localize coarse image region. In contrast, α should be small in IASN. This is because IASN is used to localize fine-grained image regions with smaller scales.

For positive constraint, this constraint is used to make attentional regions not be mirrored, and can be formulated as

$$\Gamma_P = \max\left(0, \beta - s_x\right) + \max\left(0, \beta - s_y\right) \tag{14}$$

where β is a threshold value.

Finally, these constraints on the parameters of the transformation matrix are combined to the localization loss:

$$L_{loc} = \Gamma_S + \lambda_1\Gamma_A + \lambda_2\Gamma_P \tag{15}$$

where λ_1 and λ_2 are weighted parameters.

Figure 4: Some food examples from WikiFood-200

Table 1: The statistics of three different datasets.

Dataset	♯categories	♯images	♯ingredients
ETH Food-101 [3]	101	101,000	174
VireoFood-172 [4]	172	110,241	353
WikiFood-200	200	197,323	319

3.5 Multi-scale Joint Representation

Once IG-CMAN has been trained, we can obtain multiple coarse-to-fine attentional regions for each food image. Particularly, there are three types of regions, the full image, the coarse region from CASN and several fine-grained regions from IASN. We train one CNN model for each type of regions. Based on these trained CNN models, we extract three types of features from the full image, coarse region and fine-grained regions: $\{F_0, F_1, ..., F_K\}$, where F_0 denotes visual features from the full image and K is the total number of regions. We normalize each descriptor independently, and then concatenate them as the final feature representation.

4 EXPERIMENT

4.1 Dataset

ETH Food-101 is a dataset with 101 food categories and 101,000 images. There are 1,000 images including 750 training images and 250 test images for each category. Bolaños *et al.* [2] further provided corresponding ingredient list. Our method should utilize ingredients to localize attentional image regions. Therefore, we remove ingredients, which are non-visible to food images. The size of final visible ingredient list is 174.

VireoFood-172 contains 110,241 food images from 172 categories and the size of visible ingredient list is 353. Similar to [4], in each food category, 60%,10%, 30% of images are randomly selected for training, validation and testing, respectively.

WikiFood-200. In order to prove the advantage of our method, we further propose a new dataset with ingredients. To construct this dataset, we first build the vocabulary of food categories according

to "Lists of foods by ingredient" from Wikipedia[2]. We then use the food name as the query to crawl candidate food images from different image search engines, such as Google and Bing for better visual diversity. We finally removed irrelevant and noisy food images through manual annotation. Our resulting food dataset contains 197,323 images with 200 categories and 319 visible ingredients. There are at least 500 images for each category. We coin this dataset WikiFood-200. Similar to [4], the dataset is split into 60%, 10% and 30% images for training, validation and testing, respectively. Fig. 4 shows some examples.

Table 1 provides the statistics of three food datasets. From Table 1, we can see that WikiFood-200 is larger than both ETH Food-101 and VireoFood-172 in the number of food classes and images. Further observation shows that there are very few shared categories between WikiFood-200 and other two datasets (only 15 categories with ETH Food-101 and 2 categories with VireoFood-172). Therefore, WikiFood-200 is complementary to these two datasets, and we expect WikiFood-200 can further promote the development of food community.

4.2 Experimental Setup

For IG-CMAN, the input image I is fed into a VGG-16 network. Here, we use the conv feature maps from the last conv layer as the input of ST. All the training images are resized to 224×224. The models are optimized using Adam with a batch size of 16, momentum of 0.9 and 0.999. The learning rate is set to 10^{-5} initially and divided by 10 after 30 epochs. We select the model with the lowest validation loss for testing. For Food-101, there is no validation dataset, and we thus select the model when the training loss no longer changes.

For multi-task learning in IG-CMAN, we use the standard back-propagation for optimization. The classification loss and ingredient-level attribute loss are set as the same weight without any prior. Therefore, $\gamma_1 = 1.0$. We empirically set γ_2 as 0.1, 0.5 and 0.5 for ETH Food-101, VireoFood-172 and WikiFood-200, respectively. In Eqn 12, we set the number of localized fine-grained image regions as 5 in IASN. Therefore, besides the center $(0, 0)$, four anchor points are empirically set as $(0.4, 0.4)$, $(0.4, -0.4)$, $(-0.4, 0.4)$ and $(-0.4, -0.4)$, respectively. In Eqn 13, α is set as 0.9 and 0.5 for CASN and IASN, respectively. This is because α should be large in CASN since CASN is used to localize the coarse image region. In contrast, α should be small in IASN since IASN is used to localize fine-grained image regions. In Eqn 14, β is empirically set as 0.6 and 0.1 in CASN and IASN, respectively. In Eqn 15, λ_1 and λ_2 are empirically set as 0.01 and 0.5, respectively for food-101, 1 and 1 for VireoFood-172, 0.2 and 0.2 for WikiFood-200.

Once IG-CMAN has been trained, we can obtain multiple coarse-to-fine attentional regions for each food image. We fine-tune DenseNet-161 [12] for the full image and regions from CASN and IASN, respectively. For the fusion strategy, we simply concatenate them as fused representation, and adopt the softmax classifier for fair comparison. Without loss of generality, we all adopt the same feature concatenation method for fusion in the following experiment. Top-1 and Top-5 accuracy is used as evaluation metrics [4, 18].

[2]https://en.wikipedia.org/wiki/Category:Lists_of_foods

Table 2: Performance comparison on feature fusion from different regions in IASN on ETH Food-101 (%).

Method	Top-1	Top-5
IASN(Region1-1)	83.53	96.03
IASN(Region1-2)	86.50	97.13
IASN(Region1-3)	87.17	97.35
IASN(Region1-4)	88.27	97.71
IASN(Region1-5)	88.94	97.87

Table 3: Performance comparison for different components of IG-CMAN on ETH Food-101 (%).

Method	Top-1	Top-5
CASN	85.41	96.67
IASN	88.94	97.87
CASN+IASN	89.89	98.21
IG-CMAN	90.37	98.42

4.3 Experiments on ETH Food-101

Performance Comparison on Region Feature Fusion in IASN. In our experiment, IASN localizes several fine-grained image regions. We compare the fusion results for different number of regions, and show the results in Table 2. For continuously localized fine-grained regions in IASN, we denote Region1-C as top C continuously localized regions. For example, Region1-1 is the first fine-grained region while Region1-5 means all the localized regions. We can see that (1) When more and more localized regions are added into the feature fusion, the recognition performance is improved incrementally. (2) The performance on feature fusion from all the regions is the best, and such best result benefits from the complementary advantage from multiple fine-grained image regions.

Performance Comparison on Different Components of IG-CMAN. Table 3 shows the experimental results from different components of IG-CMAN. We can see that the feature fusion from coarse regions in CASN and fine-grained regions in IASN can further improve the performance compared with single CASN or IASN. After fusing features from the full image, IG-CMAN achieves the best performance in both Top-1 and Top-5 accuracy. We can conclude that the regions from the full image, CASN and IASN are complementary, and fused features from different types are more comprehensive and discriminative.

Comparison with State-of-the-Arts. We list recent state-of-the-art methods on ETH Food-101 in Table 4. The performance on different neural networks such as AlexNet [3], Inception V3 [8], ResNet [9], DenseNet and WRN [18] is provided. From Table 4, we can see that (1) The performance of WRN is better than other single networks. (2) WISeR improves WRN by adding the other slice branch network with slice convolutional layers, which is used to capture specific vertical food layers. (3) Our method achieves the state-of-the-art performance in Top-1 accuracy, and can improve the Top-1 performance of WISeR specifically designed for food recognition by 0.1%. Although marginal performance improvement, our method did not use additional data augmentation strategy like

Table 4: Comparison of our model and state-of-the-art methods on ETH Food-101 (%).

Method	Top-1	Top-5
AlexNet-CNN [3]	56.40	-
DCNN-FOOD [33]	70.41	-
DeepFood [16]	77.40	93.70
FCAN [17]	86.50	-
CurriculumNet [7]	87.30	-
Inception V3 [8]	88.28	96.88
ResNet-200 [9]	88.38	97.85
DenseNet-161 [12]	86.94	97.03
WRN [18]	88.72	97.92
WISeR [18]	90.27	**98.71**
IG-CMAN	**90.37**	98.42

Table 5: Performance comparison on feature fusion from different regions in IASN on VireoFood-172 (%).

Method	Top-1	Top-5
IASN(Region1-1)	82.35	95.35
IASN(Region1-2)	85.96	96.92
IASN(Region1-3)	87.47	97.46
IASN(Region1-4)	88.83	97.91
IASN(Region1-5)	89.43	98.06

Table 6: Performance comparison for different components of IG-CMAN on VireoFood-172 (%).

Method	Top-1	Top-5
CASN	87.39	97.15
IASN	89.43	98.06
CASN+IASN	90.34	98.31
IG-CMAN	90.63	98.40

WISeR, which additionally applied various photometric distortions and AlexNet-style color augmentation.

4.4 Experiments on VireoFood-172

We first compare experimental results on feature fusion from different regions in IASN. As shown in Table 5, we can see that the feature fusion from all the localized regions in IASN achieves the best performance. Table 6 further shows experimental results from different components of our method on VireoFood-172. By combining features from three types of regions, we achieve the best 90.63% in Top-1 accuracy and 98.40% in Top-5 accuracy. The classification accuracy from different methods are summarized in Table 7. We can see that our method achieves the state-of-the-art performance in both Top-1 accuracy and Top-5 accuracy. Compared with existing multi-task approaches [4] with the same basic network, there are performance improvement of about 3.4% and 1.1% in Top-1 and Top-5 accuracy, respectively. This improvement mainly derives from both semantic attentional region localization and multiple attentional region fusion.

Table 7: Comparison of our model and state-of-the-art methods on VireoFood-172 (%).

Method	Top-1	Top-5
AlexNet	64.91	85.32
VGG-16	80.41	94.59
DenseNet-161	86.93	97.17
MultiTaskDCNN(VGG-16)[4]	82.06	95.88
MultiTaskDCNN(DenseNet-161)[4]	87.21	97.29
IG-CMAN	**90.63**	**98.40**

Table 8: Performance comparison on feature fusion from different regions in IASN on WikiFood-200 (%).

Method	Top-1	Top-5
IASN(Region1-1)	58.88	86.18
IASN(Region1-2)	62.09	88.36
IASN(Region1-3)	63.29	89.33
IASN(Region1-4)	64.39	89.92
IASN(Region1-5)	65.59	90.70

Table 9: Performance Comparison for different components of IG-CMAN on WikiFood-200 (%).

Method	Top-1	Top-5
CASN	61.13	87.66
IASN	65.59	90.70
CASN+IASN	66.71	91.45
IG-CMAN	67.47	91.75

Table 10: Comparison of our model and baselines on WikiFood-200 (%).

Method	Top-1	Top-5
AlexNet	49.34	79.30
VGG-16	59.05	86.53
ResNet-152	61.07	87.87
DenseNet-161	62.62	88.28
IG-CMAN	**67.47**	**91.75**

4.5 Experiments on WikiFood-200

As shown in Table 8, we can see that the feature fusion from all the localized regions in IASN achieves the best performance. Table 9 further shows experimental results from different components of our method on WikiFood-200. Our method similarly achieves the best 67.47% in Top-1 accuracy and 91.75% in Top-5 accuracy. The classification accuracy from different methods are summarized in Table 10. Because WikiFood-200 is the proposed new dataset, we conduct different benchmark baselines. The experimental results from different neural networks including AlexNet, VGG, ResNet and DenseNet are listed. From Table 10, we can see that our method achieves the state-of-the-art performance in both Top-1 accuracy and Top-5 accuracy. This again verified the effectiveness of our proposed method.

Figure 5: Localized image regions with probability distribution on top-3 ingredients from some food examples in IASN(Best viewed in magnification).

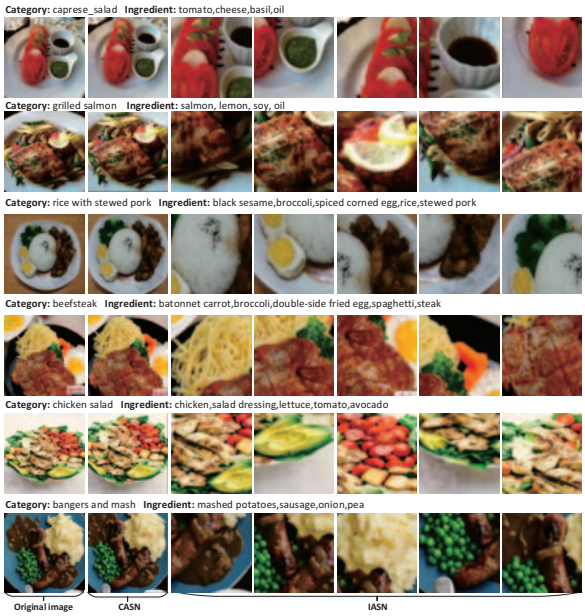

Figure 6: Localized regions from some samples in IG-CMAN.

4.6 Qualitative Analysis and Visualization

To have a richer grasp on this outcome, we conducted qualitative analysis of attentional regions from IG-CMAN. We first showed fine-grained localized regions in IASN. Fig. 5 shows some examples, where at the bottom of each row, we show top-3 ingredient distribution based on softmax transformation from S at each level in IASN. We can observe that these localized regions are discriminative to corresponding food categories. In addition, many localized regions correspond to semantic ingredients and are more interpretable. For example, in the first row, we can observe that many localized regions correspond to their ingredients with the highest score, respectively. The first localized region is chicken and the predicted chicken ingredient has the highest probability. The third localized region is soy and the predicted soy ingredient has the highest probability. Fig 6 further showed coarse and fine-grained localized regions of more image examples.

4.7 Discussions

In our method, we utilize rich ingredients to explicitly guide the network to discover diverse fine-grained attentional regions. In addition, many localized regions correspond to semantic ingredients. Therefore, our proposed method can introduce the interpretability for localized regions. However, it is not always true in the food domain. The reasons are various, such as mixed ingredients without clear division, too small ingredients and the change of spatial structure of ingredients (as shown in Fig. 4). In these cases, our method fail to localize semantic regions for ingredients. However, even these localized regions are not interpretable, they can still also provide complementary visual information, and thus is still helpful in food recognition.

Another aspect is that we pre-define five regions in IASN and IASN can thus only localize five fine-grained image regions. As shown in Table 2, Table 5 and Table 8, we can see that with the increase of localized fine-grained regions, the recognition performance has consistent improvement. We deduced that with the increase of localized regions, there will probably be performance increase. However, with the increase of cascaded sub-networks, the network training needs more time cost and GPU resources. Therefore, how many regions are fixed before or even learned automatically to enable the best recognition performance needs further study. How to balance the network complexity and the number of localized regions is worth exploration in the future.

5 CONCLUSIONS

In this paper, we have proposed an Ingredient-Guided Cascaded Multi-Attention Network (IG-CMAN) for food recognition. It is capable of sequentially localizing diverse multi-scale image regions via combining STN with LSTM under both category-level and ingredient-level guidance. As a result, the fused features from coarse and fine-grained regions are complementary, comprehensive and more discriminative. Furthermore, we present a new dataset WikiFood-200, which is very complementary to existing datasets for food recognition with ingredients, such as ETH Food-101 and VireoFood-172. Comprehensive experimental results on two popular datasets and WikiFood-200 have demonstrated that our method achieves the state-of-the-art recognition performance. Such improvement benefits from both semantic attentional region localization and multiple attentional region fusion. Future directions include introducing more context information, such as cuisine and course [22] for improving the performance and applying our food recognition method into various applications, such as foodlog [1] and health-aware recommendations [26].

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (No. 61602437 and 61532018), in part by the Lenovo Outstanding Young Scientists Program, in part by the National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals.

REFERENCES

[1] Kiyoharu Aizawa. 2013. Multimedia foodLog: diverse applications from self-monitoring to social contributions. *ITE Transactions on Media Technology and Applications* 1, 3 (2013), 214–219.

[2] Marc Bolaños, Aina Ferrà, and Petia Radeva. 2017. Food Ingredients Recognition Through Multi-label Learning. In *New Trends in Image Analysis and Processing – ICIAP 2017*, Sebastiano Battiato, Giovanni Maria Farinella, Marco Leo, and Giovanni Gallo (Eds.). Springer International Publishing, 394–402.

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

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

[5] Jingjing Chen, Chong-Wah Ngo, and Tat-Seng Chua. 2017. Cross-modal recipe retrieval with rich food attributes. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 1771–1779.

[6] Jianlong Fu, Heliang Zheng, and Tao Mei. 2017. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4476–4484.

[7] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang. 2018. CurriculumNet: Weakly Supervised Learning from Large-Scale Web Images. (2018).

[8] Hamid Hassannejad, Guido Matrella, Paolo Ciampolini, Ilaria De Munari, Monica Mordonini, and Stefano Cagnoni. 2016. Food Image Recognition Using Very Deep Convolutional Networks. In *International Workshop on Multimedia Assisted Dietary Management*. 41–49.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Computer Vision and Pattern Recognition*. 770–778.

[10] Luis Herranz, Shuqiang Jiang, and Ruihan Xu. 2017. Modeling restaurant context for food recognition. *IEEE Transactions on Multimedia* 19, 2 (2017), 430 – 440.

[11] S. Horiguchi, S. Amano, M. Ogawa, and K. Aizawa. 2018. Personalized classifier for food image recognition. *IEEE Transactions on Multimedia* 20, 10 (2018), 2836–2848.

[12] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2261–2269.

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. 2015. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems 28*. 2017–2025.

[14] Hokuto Kagaya, Kiyoharu Aizawa, and Makoto Ogawa. 2014. Food detection and recognition using convolutional neural network. In *Proceedings of the ACM International Conference on Multimedia*. 1085–1088.

[15] Yoshiyuki Kawano and Keiji Yanai. 2014. Foodcam-256: a large-scale real-time mobile food recognition system employing high-dimensional features and compression of classifier weights. In *Proceedings of the ACM International Conference on Multimedia*. 761–762.

[16] Chang Liu, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, and Yunsheng Ma. 2016. DeepFood: Deep Learning-Based Food Image Recognition for Computer-Aided Dietary Assessment. (2016), 37–48.

[17] Xiao Liu, Tian Xia, Jiang Wang, Yi Yang, Feng Zhou, and Yuanqing Lin. 2016. Fully Convolutional Attention Networks for Fine-Grained Recognition. *arxiv:1603.06765* (2016).

[18] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. 2018. Wide-slice residual networks for food recognition. In *IEEE Winter Conference on Applications of Computer Vision*. 567–576.

[19] Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P Murphy. 2015. Im2Calories: Towards an automated mobile vision food diary. In *Proceedings of the IEEE International Conference on Computer Vision*. 1233–1241.

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

[21] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A Survey on Food Computing. *ACM Computing Surveys (Accepted)* (2019).

[22] Weiqing Min, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz. 2017. Being a supercook: Joint food attributes and multi-modal content modeling for recipe retrieval and exploration. *IEEE Transactions on Multimedia* 19, 5 (2017), 1100 – 1113.

[23] Weiqing Min, Shuqiang Jiang, Shuhui Wang, Jitao Sang, and Shuhuan Mei. 2017. A Delicious Recipe Analysis Framework for Exploring Multi-Modal Recipes with Various Attributes. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*. 402–410.

[24] Zhao Yan Ming, Jingjing Chen, Yu Cao, Ciaran Forde, Chong Wah Ngo, and Tat Seng Chua. 2018. Food photo recognition for dietary tracking; System and experiment. In *International Conference on Multi Media Modelling*. 129–141.

[25] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent Models of Visual Attention. *Advances in Neural Information Processing Systems* 3 (2014).

[26] Nitish Nag, Vaibhav Pandey, and Ramesh Jain. 2017. Health multimedia: Lifestyle recommendations based on diverse observations. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. 99–106.

[27] Sina Sajadmanesh, Sina Jafarzadeh, Seyed Ali Ossia, Hamid R Rabiee, Hamed Haddadi, Yelena Mejova, Mirco Musolesi, Emiliano De Cristofaro, and Gianluca Stringhini. 2017. Kissing Cuisines: Exploring worldwide culinary habits on the web. In *Proceedings of the International World Wide Web Conference*. ACM, 1013–1021.

[28] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *Computer Vision and Pattern Recognition*. 3020–3028.

[29] Ryosuke Tanno, Koichi Okamoto, and Keiji Yanai. 2016. DeepFoodCam: A DCNN-based real-time mobile food recognition system. In *International Workshop on Multimedia Assisted Dietary Management*. 89–89.

[30] Huayang Wang, Weiqing Min, Xiangyang Li, and Shuqiang Jiang. 2016. Where and what to eat: Simultaneous restaurant and dish recognition from food image. In *Pacific Rim Conference on Multimedia*. 520–528.

[31] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. 2017. Multi-label Image Recognition by Recurrently Discovering Attentional Regions. (2017), 464–472.

[32] Ruihan Xu, Luis Herranz, Shuqiang Jiang, Shuang Wang, Xinhang Song, and Ramesh Jain. 2015. Geolocalized modeling for dish recognition. *IEEE Transactions on Multimedia* 17, 8 (2015), 1187–1199.

[33] Keiji Yanai and Yoshiyuki Kawano. 2015. Food image recognition using deep convolutional network with pre-training and fine-tuning. In *Multimedia & Expo Workshops, 2015 IEEE International Conference on*. IEEE, 1–6.

[34] Jiwei Yang, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 2018. Local Convolutional Neural Networks for Person Re-Identification. In *Proceedings of the 26th ACM International Conference on Multimedia*. 1074–1082.

[35] Shulin Yang, Mei Chen, Dean Pomerleau, and Rahul Sukthankar. 2010. Food recognition using statistics of pairwise local features. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2249–2256.

[36] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. 2017. Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. In *IEEE International Conference on Computer Vision*. 5219–5227.

[37] Feng Zhou and Yuanqing Lin. 2016. Fine-Grained Image Classification by Exploring Bipartite-Graph Labels. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1124–1133.

[38] Yingying Zhu, Jiong Wang, Lingxi Xie, and Liang Zheng. 2018. Attention-based Pyramid Aggregation Network for Visual Place Recognition. In *Proceedings of the 26th ACM International Conference on Multimedia*. 99–107.