

# Multi-Scale Multi-View Deep Feature Aggregation for Food Recognition

Shuqiang Jiang<sup>1</sup>, Senior Member, IEEE, Weiqing Min, Member, IEEE, Linhu Liu, and Zhengdong Luo

**Abstract**—Recently, food recognition has received more and more attention in image processing and computer vision for its great potential applications in human health. Most of the existing methods directly extracted deep visual features via convolutional neural networks (CNNs) for food recognition. Such methods ignore the characteristics of food images and are, thus, hard to achieve optimal recognition performance. In contrast to general object recognition, food images typically do not exhibit distinctive spatial arrangement and common semantic patterns. In this paper, we propose a multi-scale multi-view feature aggregation (MSMVFA) scheme for food recognition. MSMVFA can aggregate high-level semantic features, mid-level attribute features, and deep visual features into a unified representation. These three types of features describe the food image from different granularity. Therefore, the aggregated features can capture the semantics of food images with the greatest probability. For that solution, we utilize additional ingredient knowledge to obtain mid-level attribute representation via ingredient-supervised CNNs. High-level semantic features and deep visual features are extracted from class-supervised CNNs. Considering food images do not exhibit distinctive spatial layout in many cases, MSMVFA fuses multi-scale CNN activations for each type of features to make aggregated features more discriminative and invariable to geometrical deformation. Finally, the aggregated features are more robust, comprehensive, and discriminative via two-level fusion, namely multi-scale fusion for each type of features and multi-view aggregation for different types of features. In addition, MSMVFA is general and different deep networks can be easily applied into this scheme. Extensive experiments and evaluations demonstrate that our method achieves state-of-the-art recognition performance on three popular large-scale food benchmark datasets in Top-1 recognition accuracy. Furthermore, we expect this paper will further the agenda of food recognition in the community of image processing and computer vision.

**Index Terms**—Food recognition, ingredient knowledge, feature aggregation, convolutional neural networks.

Manuscript received December 23, 2018; revised June 2, 2019; accepted July 11, 2019. Date of publication July 29, 2019; date of current version September 12, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61532018 and Grant 61602437, in part by the Beijing Natural Science Foundation under Grant L182054, and in part by the National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shuicheng Yan. (*Corresponding author: Shuqiang Jiang.*)

The authors 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: sqjiang@ict.ac.cn; minweiqing@ict.ac.cn; linhu.liu@vip1.ict.ac.cn; luozhengdong@ict.ac.cn).

Digital Object Identifier 10.1109/TIP.2019.2929447

## I. INTRODUCTION

HUMANS have historically faced the task of identifying food to further gather food for their survival. In recent years, food recognition has attracted increasing attention in image processing and computer vision [1]–[6]. It is of paramount importance for analyzing and understanding food images from different perspectives, such as health, culture and marketing. Automatically recognizing food can also enable various applications, such as mobile visual food diary [4] and self-service restaurants [7].

Like generic object recognition, the key of food recognition is to extract discriminative visual features. Early researches on food recognition mainly extracted hand-crafted features [1], [3], [8]. For example, Yang *et al.* [1] first used a semantic texton forest to compute the distribution over ingredients for each pixel in the image, and then constructed multi-dimensional histogram features as the visual representation. Bettadapura *et al.* [8] combined different types of feature descriptors, such as original SIFT [9] and their variants into fused features for food recognition. Recently, there have been more attempts to utilize deep learning in food recognition. For example, Meyers *et al.* [4] used the GoogLeNet network to train a multi-label classifier to predict the type of food present in the meal. Martinel *et al.* [10] proposed a wide-slice residual network to capture the vertical structure from food images. Deep learning based approaches generally obtain better performance than hand-crafted features because of their advantage in representation learning.

However, as a special object recognition task, food recognition has not been fully addressed due to the following reasons. First, different from general object recognition, many types of food do not exhibit distinctive spatial layout and configuration. They are typically non-rigid, and the structure information can not be easily exploited. Therefore, standard object recognition approaches probably perform poorly on this task. Existing food recognition methods such as [1], [10] are only limited to food types with certain visually distinctive spatial arrangement, such as vertical structures (e.g., hamburgers). Second, food recognition can be considered as fine-grained recognition [11]. The first step of fine-grained object recognition is generally to discover fixed semantic parts of certain object, such as birds and cars. However, common semantic parts do not exist in many types of food images. Therefore, it is hard to capture semantic information from food images via existing fine-grained methods. Third, similar to object recognition,



Fig. 1. Some food images with different geometrical deformations.



Fig. 2. Some food examples from VireoFood-172 [15].

food images have also various geometrical variants, such as different viewpoints, rotation and scales. Fig. 1 shows some food images from some existing datasets. It requires that food recognition methods should have the geometrical invariance to robustly recognize food images. Existing food recognition methods [4], [12] generally used the CNNs to directly extract visual features from the whole food image, and may fail when geometric variants are larger. This is because the CNNs can only process images with small-scale deformations through max-pooling. Fourth, in spite of recent development in food recognition, it does not receive enough attention for researchers like object recognition and scene recognition. There are no trained models available like ImageNet [13] and PlacesNet [14] to help advance its development in the computer vision community.

On the bright side, there are additional ingredient information available associated with food images from the web. Like the importance of objects for the scene, ingredients within food images are also very important for food recognition as mid-level attributes. Fig. 2 shows some food examples with ingredients from VireoFood-172 [15]. Although Fig. 2(a) and Fig. 2(b) belong to the same class, their visual appearance has larger difference. However, they have many common ingredients. Such mid-level ingredient attributes can help recognize them. Similarly, the visual appearance of Fig. 2(a) is similar to Fig. 2(c). However, they do not belong to the same food type.

We can distinguish them via their ingredients. For example, the scrambled egg is one representative ingredient of Yangzhou fried rice. Hence, mid-level ingredient attribute learning can provide another clue, which is helpful for food recognition. Besides mid-level ingredient representation, high-level food semantic distribution and deep visual features from CNNs can also provide complementarity information from different perspectives and granularity. If we aggregate these three types of features together, we can capture semantic information from food images with the greatest probability.

Furthermore, although food typically does not exhibit distinctive spatial arrangement, we can explore image patches from different scales and then fuse them into multi-scale representation. Such representation can fuse patch features from the coarse scale to the fine scale, and thus their features contain information from discriminative image regions. In addition, multi-scale fusion can be more robust to the geometrical deformation. Some works such as [16], [17] have verified the effectiveness of multi-scale features in classification and retrieval tasks.

Taking these factors into consideration, in this paper, we proposed a Multi-Scale Multi-View Feature Aggregation (MSMVFA) scheme for food recognition, where multi-view means different types of feature sets. Different types of features with different granularity are jointly utilized in MSMVFA. Particularly, MSMVFA consists of two-level fusion, namely multi-scale fusion for each type of features and multi-view aggregation for different types of features. Considering food typically does not exhibit distinctive spatial arrangement, we utilize multi-scale fusion methods for each type of features. The coarsest scale is the whole image, so the global spatial layout is preserved, and the finer scales allow us to capture more local, fine-grained details of the food image. Therefore, such fused features are more robust and invariable to the geometrical deformation. Based on multi-scale representation for each type of features, MSMVFA can further aggregate high-level semantic features, mid-level attribute features and deep visual features into a unified representation. These three types of features describe food images from different granularity. Therefore, the aggregated features can capture semantic information with the greatest probability. For that solution, we utilize additional ingredient information to fine-tune the deep network to extract mid-level attribute features. The high-level semantic features and deep visual features are extracted from class-supervised deep neural network. The resulting representation is more robust, comprehensive and discriminative as generic features for food recognition.

MSMVFA is general and different types of CNNs can be applied into this framework. We benchmark several deep learning networks including VGG-16 [18], ResNet-152 [19] and Densenet-161 [20] in MSMVFA on three large-scale food datasets and the released trained model<sup>1</sup> for furthering the agenda of food recognition and other food-related study.

The contributions of our paper can be summarized as follows:

<sup>1</sup><http://isia.ict.ac.cn/dataset/MSMVFA-Models.html>

- We propose a Multi-Scale Multi-View Feature Aggregation (MSMVFA) scheme for food recognition, which can conduct two-level fusion, namely multi-scale fusion for each type of features and multi-view aggregation for various types of features with different granularity to produce more robust, discriminative and comprehensive fine-grained representation.
- We conduct comprehensive experimental evaluation on three popular food benchmark datasets, and the experimental results demonstrate our approach achieves the state-of-the-art performance in food recognition on all these three food benchmark datasets for the Top-1 accuracy.
- We benchmark several deep learning networks in our proposed framework on three different food datasets. It could further the agenda of food-related study in the community of image processing and meanwhile forms a contribution to other food-related fields, such as computational gastronomy and food science.

The rest of this paper is organized as follows. Section II reviews related work. section III elaborates the proposed food recognition framework. Experimental results and analysis are reported in Section IV. Finally, we conclude the paper and give future work in Section V.

## II. RELATED WORK

Our work is closely related to two research fields: (1) food recognition and (2) multi-scale visual recognition.

### A. Food Recognition

Food recognition is one of the promising applications in visual recognition. After we recognize the category of the meal, we can further conduct various health-related applications, e.g., mobile food diary [4] and self-restaurant service [7]. For these reasons, we have seen an explosion of food recognition algorithms in image processing and computer vision [1]–[5]. A more comprehensive survey of food recognition and food-related works is provided in [21]. In the earlier years, they extracted various hand-crafted features from food images for recognition [1], [3], [22]. For example, Yang *et al.* [1] calculated pairwise feature statistics between local features computed over a soft pixel-level image segmentation to exploit spatial relationships between ingredients for food recognition. Bossard *et al.* [3] adopted random forests to mine discriminative patches of food images as the visual representation. Some works such as [23] combined different types of hand-crafted features to represent the food image.

In contrast, deep features learned by deep networks have been confirmed to be far more effective than hand-crafted features in food recognition because of its powerful expressive capacity. For example, Kagaya *et al.* [12] adopted the AlexNet network [13] to extract deep visual features for food detection and recognition. There are also some works [4], [24], [25], which combined deep visual features and other context information, such as GPS and restaurant information to improve the performance of food recognition. For example, Xu *et al.* [24] leveraged the geolocation and external information about

restaurants for geolocalized modeling. Myers *et al.* [4] took the GoogLeNet deep network to extract deep visual features to recognize the content of the meal from one image. The GPS information is further introduced for predicting their nutritional content. Zhou and Lin [5] exploited rich relationships among ingredients, food category and restaurant information through the bi-partite graph for food image classification. Different from existing works, we aggregate high-level semantic features, mid-level ingredient features and deep visual features together for food recognition. Such feature aggregation is capable of capturing the semantic information with the greatest probability. In addition, we further consider multi-scale feature fusion on each type of features to obtain more robust and discriminative feature representation to achieve the state-of-the-art performance.

In addition, our work is also relevant to recipe analysis [6], [26]–[29]. For example, some works [15], [28] adopted a multi-task deep learning architecture such as CNNs or Deep Boltzmann Machines [30] for simultaneous ingredient and food recognition, where the ingredients are fully exploited as supervised information for fine-tuning the network. Min *et al.* [29] combined topic model and deep learning methods to discover topics of ingredient bases and visualize them to conduct cross-region recipe analysis. Salvador *et al.* [6] released a new large-scale dataset with over 1 million cooking recipes and 800K food images for cross-modal image-recipe retrieval.

### B. Multi-Scale Visual Recognition

It has been shown that in addition to the entire image, it is consistently better to extract CNN features from multi-scale local patches arranged in regular grids to improve the performance of image recognition [16], [17], [31]. For example, Wu *et al.* [17] proposed an architecture in which dense sampling of patches is replaced by region proposals and discriminative patch mining. Herranz *et al.* [32] analyzed multi-scale CNN architectures and showed that careful multi-scale combinations of ImageNet-CNNs and Places-CNNs can improve the performance of scene recognition. In our work, we also utilize multi-scale feature fusion for the task of food recognition to handle the geometrical deformation.

In addition, our work is also relevant to fine-grained classification [11], [33], [34], which aims to distinguish among different breeds or species. They generally first discover fixed semantic parts, and then fuse features extracted from these semantic parts as the final representation. For example, in the bird classification, some semantic parts, such as head and breast should be first localized. However, they are not directly used into food recognition. The reason is that the concepts of common semantic parts do not exist in many types of food images, and there are no fixed semantic parts or patterns for many types of food images.

## III. OUR APPROACH

Food images have more sophisticated visual complexities than general object images, including vague spatial arrangement, larger food deformations, severe intra-class variations

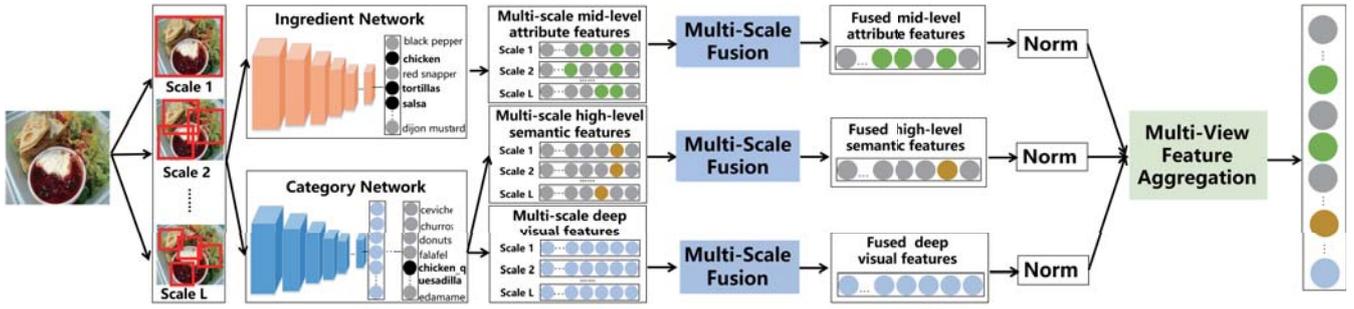


Fig. 3. Our proposed Multi-Scale Multi-View Feature Aggregation (MSMVFA) framework.

ETH Food-101			VireoFood-172			ChineseFoodNet		
<b>Category:</b> Chicken_wings	<b>Category:</b> French_fries	<b>Category:</b> Risotto	<b>Category:</b> Hot and sour	<b>Category:</b> Deep-Fried	<b>Category:</b> Shredded cabbage	<b>Category:</b> Pig ears	<b>Category:</b> Fried Tofu	<b>Category:</b> Shredded cabbage
<b>Ingredient:</b> Chicken,Soy, Garlic,Sugar	<b>Ingredient:</b> Potato,Flour, Salt,Black pepper,Cayen ne pepper,Oil	<b>Ingredient:</b> Oil,Arborio rice,Risotto,Broth, Butter,Cheese	<b>Ingredient:</b> Shredded pepper,Crushed hot and dry chili, Shredded potato	<b>Ingredient:</b> Crushed pepper,Chinese Parsley,coriander, Stinky tofu	<b>Ingredient:</b> Crushed hot and dry chili,Cabbage			

Fig. 4. Some food examples from three datasets.

and smaller inter-class variations, and thus require more elaborate solutions. In order to solve this, we propose a Multi-Scale Multi-View Feature Aggregation (MSMVFA) framework for food recognition. Two key technologies are exploited in MSMVFA. First, we aggregate high-level semantic features, mid-level attribute features and deep visual features into unified representation. Different types of features describe food images from different granularity. Therefore, the aggregated features can capture semantic information of food images with the greatest probability. Second, unlike general objects, food typically does not exhibit distinctive spatial patterns. In order to solve this, for each type of features, multi-scale patches based feature fusion is utilized to obtain more robust and discriminative representation. Such multi-scale feature representation not only contains ones from discriminative image regions, but also is insensitive to geometrical deformation.

As shown in Fig. 3, given an input image, MSMVFA is capable of extracting and aggregating three types of features with various scales and different granularity. For that solution, two types of deep neural networks are introduced in MSMVFA, namely ingredient network and category network. We can use any one of existing popular neural networks, such as VGG [18], ResNet [19] and DenseNet [20] as one basic network for these two types of networks. Through the category network, we can extract category-oriented semantic distribution and more abstract deep visual features with multi-scales. In order to obtain mid-level attribute features, we utilized additional ingredient information and also designed the ingredient network to extract mid-level attribute features with multi-scales. Compared with food category, ingredients from each food category can describe food images at fine-grained and local level. Therefore, ingredient network can extract region-oriented mid-level attribute features compared with

global semantic distribution. For each type, the features from different scales are then fused via multi-scale fusion. The fused features from three different types are further normalized and aggregated into the final representation via multi-view feature aggregation. The classifier based on aggregated features is finally trained for food recognition. In the following sections, we introduce main components of MSMVFA in details.

#### A. Multi-View Feature Aggregation

1) *Mid-Level Attribute Representation*: Food classification belongs to fine-grained classification and is very challenging because of its visual complexities. Exploring only food class information is probably not enough for food recognition. Fortunately, there is rich ingredient information associated with food images. As shown in Fig. 4, many ingredients can describe visual attributes of food images at local level. Therefore, ingredient-based representation provides more fine-grained feature representation for food images.

To obtain such mid-level representation, we should design an ingredient network to extract ingredient-level representation. Different types of deep attribute networks can be adopted, such as PANDA [35] and cascading CNN [36]. PANDA combines part-based models and CNNs for human attribute prediction while cascading CNN first localizes face regions and then conducts face attribute prediction based on localized regions. Actually, among these attribute networks, one simple method for the ingredient network is to directly fine-tune the deep network using multi-label ingredient information as supervised information for each scale, and then fuse ingredient-level attribute features from all the scales into unified representation. Therefore, our method is similar to PANDA, but the difference is that PANDA first fuses the

features from all the scales and the fused features are then fed into the classifier for attribute learning.

Particularly, we consider one scale, namely the whole image as one example. Through the ingredient network, we can obtain the attribute representation for each image  $x$  as one vector.  $\hat{\mathbf{a}} = (\hat{\mathbf{a}}^1, \dots, \hat{\mathbf{a}}^m, \dots, \hat{\mathbf{a}}^M)$ , where  $M$  is the size of ingredient vocabulary.  $\hat{\mathbf{a}}^m \in [0, 1]$  is predicted probability for attribute  $m$ , and is calculated via a sigmoid activation function as follows:

$$\hat{\mathbf{a}}^m = P(\mathbf{a}^m | x) = \frac{1}{1 + \exp(-f(x))} \quad (1)$$

where  $f(\cdot)$  denotes the final features from the ingredient network we adopted and the groundtruth ingredient labels are binary representation  $\mathbf{a} = (\mathbf{a}^1, \dots, \mathbf{a}^m, \dots, \mathbf{a}^M)$ : whether this image has this ingredient  $m$  or not.

We consider multi-label ingredient attribute learning with  $M$  ingredients as  $M$  binary attribute classification tasks. Correspondingly, the loss function with the cross-entropy function  $L$  is used as follows for single image  $x$ .

$$L_I = - \sum_{m=1}^M (\mathbf{a}^m \log(\hat{\mathbf{a}}^m) + (1 - \mathbf{a}^m) \log(1 - \hat{\mathbf{a}}^m)) \quad (2)$$

which allows to have multiple highly activated outputs. For other types of ingredient networks, some additional terms probably need to be added.

Through the ingredient network, we can obtain predicted ingredient distribution  $\hat{\mathbf{a}} = (\hat{\mathbf{a}}^1, \dots, \hat{\mathbf{a}}^m, \dots, \hat{\mathbf{a}}^M)$  as mid-level representation for each image.

2) *High-Level Semantic Representation*: The output layer of a category network, namely the prediction layer is a semantic probability distribution, and it generally denotes high-level semantic information. High-level semantic representation is very useful and has been verified in some tasks, such as [33]. In order to obtain high-level semantic representation, we fine-tune one deep network, such as VGG using the food class label as supervised information. In this category network, CNNs typically use the softmax activation in the last layer. The softmax function allows to obtain a semantic probability distribution for the input sample  $x$  over all possible output classes, and thus predicts the most probable outcome,  $\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} P(\mathbf{y} | x)$ . The softmax activation is usually combined with the categorical cross-entropy loss function  $L_c$  during model optimization, which penalizes the model when the optimal output value is far away from 1:

$$L_c = - \sum_x \log(P(\mathbf{y} | x)) \quad (3)$$

After fine-tuning the category network, we can predict class-probability distribution  $\hat{\mathbf{y}} = (\hat{\mathbf{y}}^1, \dots, \hat{\mathbf{y}}^c, \dots, \hat{\mathbf{y}}^C)$  as high-level semantic features for each test image  $x$ , where  $C$  is the number of categories.

3) *Deep Visual Feature Representation*: The layers of a deep neural network close to the output layer also contain independent class-relevant information that is not contained in the output layer [37]. Therefore, besides high-level semantic features, deep visual features should also be extracted. For this type of features, we can directly extract deep visual features

$\hat{\mathbf{h}} = (\hat{\mathbf{h}}^1, \dots, \hat{\mathbf{h}}^d, \dots, \hat{\mathbf{h}}^D)$  from the category network, where  $D$  is the number of feature dimensions. For example, we extract 4096-D features of the FC-layer from the VGG-16 network as the deep visual features.

4) *Multi-View Feature Aggregation*: After obtaining all types of features, we next aggregate them into unified representation. Considering different types of features have different range of values, we first normalize each type of features via  $\text{Norm}(\cdot)$  and then obtain the aggregated representation via the aggregating operator  $\text{Agg}(\cdot)$ :

$$\mathbf{F} = \text{Agg}(\text{Norm}(\hat{\mathbf{a}}), \text{Norm}(\hat{\mathbf{y}}), \text{Norm}(\hat{\mathbf{h}})) \quad (4)$$

where  $\text{Norm}(\cdot)$  can be certain normalization operation, such as  $l_2$  and z-score. Similarly, the aggregation operator  $\text{Agg}(\cdot)$  can be one of many aggregation methods, such as simple concatenation [33] and deep feedforward networks [38]. Without loss of generality, in our experiment, z-score normalization is used and simple concatenation aggregation method is adopted.

## B. Multi-Scale Feature Fusing

Different types of features work better at different scales. For example, we probably extract more discriminative mid-level ingredient features at smaller scales. In addition, many types of food images have no distinctive spatial arrangement. Fusing on various scales for each type of features is also one way to circumvent this problem. Furthermore, various multi-scale fusion methods have been proposed, and have been verified as one effective way for robust feature representation in many tasks, such as scene recognition, image retrieval and image restoration [16], [31], [39].

For each type of features, we can adopt multi-scale CNNs to extract features for each scale, and then fuse features from different types into unified representation. Take mid-level ingredient representation as one example, Let  $l$  denote certain scale.  $l = 1$  means the whole food image while  $l = N$  means the finest scale. For each scale, we take the multi-label ingredient information as supervised information to extract intermediate attribute features. For the first scale, we generally train the ingredient network for the whole image to obtain the feature representation. For the remaining small scales, we train the network using these local patches, and extract the features from each patch of one image. Next, we fuse the activations of these multiple patches to summarize these scales such as mean-pooling, max-pooling or other fusion methods. Finally, we obtain corresponding the representation from different scales  $\{\hat{\mathbf{a}}_l\}_{l=1}^N$ . Similarly, we can obtain multi-scale high-level semantic features and deep visual features from the category network  $\{\hat{\mathbf{y}}_l\}_{l=1}^N$  and  $\{\hat{\mathbf{h}}_l\}_{l=1}^N$ .

After we obtain features at each scale, we next fuse them into multi-scale representation via  $\text{Fus}(\cdot)$ . The fused features for three types are listed as  $\text{Fus}(\{\hat{\mathbf{h}}_l\}_{l=1}^N)$ ,  $\text{Fus}(\{\hat{\mathbf{a}}_l\}_{l=1}^N)$  and  $\text{Fus}(\{\hat{\mathbf{y}}_l\}_{l=1}^N)$ . Before multi-scale fusion, the normalization should also be applied like multi-view aggregation. Similar to  $\text{Agg}(\cdot)$ , the fusion operator  $\text{Fus}(\cdot)$  can be one of many fusion methods, such as simple concatenation and deep feedforward networks.

### C. Multi-Scale Multi-View Feature Aggregation

Our final feature representation is obtained from two-level fusion, namely multi-scale fusion for each type of features and multi-view aggregation for different types of features via

$$\mathbf{F} = \text{Agg}(\text{Norm}(\text{Fus}(\{\widehat{\mathbf{h}}_l\}_{l=1}^N)), \text{Norm}(\text{Fus}(\{\widehat{\mathbf{a}}_l\}_{l=1}^N)), \text{Norm}(\text{Fus}(\{\widehat{\mathbf{y}}_l\}_{l=1}^N))) \quad (5)$$

For the first level fusion, we conduct multi-scale fusion to make fused features contain ones from discriminative regions of food images and insensitive to the geometrical deformation. For the second level fusion, we fuse three different types of features to capture semantic features of food images with the biggest probability. Therefore, our proposed two-level fusion, namely multi-scale multi-view feature aggregation is suitable for food images.

In the test stage, a softmax classifier [40] is first trained based on aggregated representations for each image from the training dataset. Given one test image, we first obtain predicted mid-level attribute representation, predicted high-level semantic representation and deep visual features, and then aggregate these different types of features into the final representation via MSMVFA. Finally the aggregated features are fed them into the classifier to obtain predicted results. Considering excising food recognition methods adopt CNN based methods. They all use an end-to-end CNNs for classification, where the sigmoid layer is generally adopted for classification. For the technical consistency in classification and fair comparison in the following experiment, we thus adopt the softmax classifier to classify final fused representations.

### D. The Analysis of MSMVF

The advantage of MSMVF can be derived from two-fold. First, MSMVF can obtain different types of deep features under different supervised signals. Category-supervised deep network can provide high-level semantic features while ingredient-supervised deep network can provide fine-grained attribute features. They are complementarity from different perspectives and granularity. Second, MSMVF can explore discriminative image regions with different scales. Fusing these regional features from the coarse scale to the fine scale contain discriminative information with the greatest probability. In addition, such fused features can also be more robust to the geometrical deformation. The final fused features from MSMVF are thus comprehensive, complementarity and discriminative. MSMVF for each food image is summarized in Algorithm 1.

## IV. EXPERIMENT

### A. Dataset

To evaluate the performance of MSMVFA, we conduct extensive experiment on three large-scale datasets, including two benchmark datasets, namely ETH Food-101 [3] and VireoFood-172 [15], and recently released large-scale Chinese food dataset ChineseFoodNet [41]. The details of three datasets are described below:

---

### Algorithm 1 Visual Representation via MSMVF

---

**Input:** One food image  $x$

**Output:** The final feature representation  $\mathbf{F}$ .

- 1: Compute mid-level ingredient representation  $\widehat{\mathbf{a}}_l = (\widehat{\mathbf{a}}_l^1, \dots, \widehat{\mathbf{a}}_l^n, \dots, \widehat{\mathbf{a}}_l^M)$  for each scale  $l$  via ingredient network, fine-tuned by optimizing Eqn. 2.
  - 2: Compute high-level semantic representation  $\widehat{\mathbf{y}}_l = (\widehat{\mathbf{y}}_l^1, \dots, \widehat{\mathbf{y}}_l^c, \dots, \widehat{\mathbf{y}}_l^C)$  for each scale  $l$  via the category network, fine-tuned by optimizing Eqn. 3.
  - 3: Compute deep visual feature representation  $\widehat{\mathbf{h}}_l = (\widehat{\mathbf{h}}_l^1, \dots, \widehat{\mathbf{h}}_l^d, \dots, \widehat{\mathbf{h}}_l^D)$  for each scale  $l$  via the category network, fine-tuned by optimizing Eqn. 3.
  - 4: Compute fused ingredient representation  $\widehat{\mathbf{a}}_f = \text{Fus}(\{\widehat{\mathbf{a}}_l\}_{l=1}^N)$  via multi-scale fusion
  - 5: Compute fused high-level semantic representation  $\widehat{\mathbf{y}}_f = \text{Fus}(\{\widehat{\mathbf{y}}_l\}_{l=1}^N)$  via multi-scale fusion
  - 6: Compute fused deep visual representation  $\widehat{\mathbf{h}}_f = \text{Fus}(\{\widehat{\mathbf{h}}_l\}_{l=1}^N)$  via multi-scale fusion
  - 7: Compute final aggregated feature representation  $\mathbf{F}$  via  $\mathbf{F} = \text{Agg}(\text{Norm}(\widehat{\mathbf{h}}_f), \text{Norm}(\widehat{\mathbf{a}}_f), \text{Norm}(\widehat{\mathbf{y}}_f))$
- 

**ETH Food-101.** It consists of 101,000 images with 101 categories. For each class, there are 1,000 images including 750 training images and 250 test images. Furthermore, Ingredients101 [42] provides the ingredient vocabulary with the size of 227 for this dataset<sup>2</sup>.

**VireoFood-172.** It contains 110,241 food images from 172 categories with 353 ingredients. For each food category, 60% of images are randomly selected for training, 10% for validation and the remaining 30% for testing.

**ChineseFoodNet.** It consists of 185,628 images with 208 Chinese food categories. The whole dataset is split into 145,065, 20,253 and 20,310 images for training, validation and testing, respectively. However, the label information for the test set is not provided. Therefore, we divide the validation set into two parts: about 20% (4,050) is used as the validation set and the remaining 80% (16,503) as the test set. This dataset does not provide associated ingredient information. Considering that both ChineseFoodNet and VireoFood-172 belong to Chinese cuisine, we simply use the ingredient list from VireoFood-172 for this dataset.

Fig. 4 shows some food images from different datasets. We can see that there are overlapped food categories for VireoFood-172 and ChineseFoodNet, such as Shredded cabbage. This is because these two datasets both belong to Chinese cuisine.

### B. Implementation Details

VGG [18], ResNet [19] and DenseNet [20] are currently three basic and also popular CNN architectures. In order to validate the effectiveness and robustness of our proposed method,

<sup>2</sup><http://www.ub.edu/cvub/ingredients101/>

TABLE I

THE PERFORMANCE COMPARISON IN % OF DIFFERENT COMBINATIONS OF SCALES FOR DEEP VISUAL FEATURES, MID-LEVEL ATTRIBUTE FEATURES AND HIGH-LEVEL SEMANTIC FEATURES ON THE ETH FOOD-101 USING VGG-16, WHERE L1 CORRESPONDS TO THE GLOBAL CNN REPRESENTATION AND L1 + L2 + L3 CORRESPONDS TO THE COMBINATION AMONG L1, L2 AND L3. FD MEANS THE FEATURE DIMENSION. THE SAME FOR THE FOLLOWING TABLES

(a) Deep visual features				(b) Mid-level attribute features				(c) High-level semantic features			
method	FD	Top-1	Top-5	method	FD	Top-1	Top-5	method	FD	Top-1	Top-5
L1	4,096	78.76	94.19	L1	227	77.67	88.65	L1	101	78.38	94.07
L2	4,096	84.73	96.47	L2	227	76.32	91.32	L2	101	81.08	95.38
L3	4,096	83.03	96.03	L3	227	74.97	<b>92.27</b>	L3	101	76.66	94.18
L1+L2	8,192	83.26	96.02	L1+L2	454	78.50	90.06	L1+L2	202	84.37	96.38
L1+L3	8,192	81.34	95.31	L1+L3	454	76.61	88.42	L1+L3	202	82.54	95.75
L2+L3	8,192	83.95	96.27	L2+L3	454	75.80	90.09	L2+L3	202	81.83	95.56
L1+L2+L3	12,288	<b>85.89</b>	<b>96.98</b>	L1+L2+L3	681	<b>78.60</b>	90.36	L1+L2+L3	303	<b>84.94</b>	<b>96.68</b>

TABLE II

THE PERFORMANCE COMPARISON IN % OF DIFFERENT COMBINATIONS OF SCALES FOR DEEP VISUAL FEATURES, MID-LEVEL ATTRIBUTE FEATURES AND SEMANTIC FEATURES ON THE ETH FOOD-101 USING RESNET-152

(a) Deep visual features				(b) Mid-level attribute features				(c) High-level semantic features			
Method	FD	Top-1	Top-5	Method	FD	Top-1	Top-5	Method	FD	Top-1	Top-5
L1	2,048	83.61	95.72	L1	227	80.42	90.23	L1	101	83.41	95.67
L2	2,048	87.02	97.13	L2	227	83.37	94.76	L2	101	82.86	95.74
L3	2,048	82.71	95.32	L3	227	76.88	92.99	L3	101	76.68	93.17
L1+L2	4,096	88.39	97.70	L1+L2	454	82.71	92.40	L1+L2	202	88.08	97.53
L1+L3	4,096	86.07	96.85	L1+L3	454	82.00	92.05	L1+L3	202	87.36	97.22
L2+L3	4,096	88.15	97.39	L2+L3	454	82.82	94.36	L2+L3	202	85.80	96.56
L1+L2+L3	6,144	<b>89.00</b>	<b>97.85</b>	L1+L2+L3	681	<b>83.81</b>	<b>95.70</b>	L1+L2+L3	303	<b>89.05</b>	<b>97.79</b>

we conduct experiments on all these three basic CNN architectures. Without loss of generality, VGG-16, ResNet-152 and DenseNet-161 are selected in our framework. The learning rate of VGG-16 is set to 0.0001 while the learning rate of both ResNet-152 and DenseNet-161 is set to 0.001 initially. They are divided by 10 after 10 epochs on both category network and ingredient network. The batch size of VGG-16, ResNet-152 and DenseNet-161 is 48, 8 and 8, respectively. The training epochs are 30 for each network. For VireoFood-172 and ChineseFoodNet, we select the model with the highest validation accuracy as the best model for testing. For Food-101, there is no the validation dataset, and we select the model when the training loss no longer changes. All the deep networks are optimized using the stochastic gradient descent with momentum of 0.9 and weight decay of 0.0001. We implemented all the deep networks via the Caffe platform [43] on Nvidia GPUs Titan X. Each model is pre-trained on the ImageNet.

In our experiment, we adopt three different scales [16], corresponding to the global  $256 \times 256$  images(L1),  $128 \times 128$  patches (L2) and  $64 \times 64$  patches (L3), respectively. For the global scale, we directly use the whole image to fine-tune the model. As for the L2 scale, one image is divided into four patches, and they share the same food category or ingredient labels. We use all the patches that are resized to  $256 \times 256$  to fine-tune the model. Similar strategy is adopted for the L3 scale. For multi-scale fusion Fus( $\cdot$ ) and multi-view aggregation Agg( $\cdot$ ), we both adopt simply concatenation operation. In addition, for the feature fusion from different patches at certain scale, we adopt max-pooling. Note that although other choices for types of scales and feature fusion methods

are possible, in this work, we emphasize the contribution of our proposed framework.

For deep visual features, we extract 4096-dimensional activations from the FC7 layer of VGG-16, 2048-dimensional activations from the ResNet-152, 2208-dimensional activations from the DenseNet-161. We extract high-level semantic features at the category prediction layer from the category network. Similarly, we extract mid-level attribute features at the ingredient prediction layer from the ingredient network.

Similar to [10], we adopt both Top-1 and Top-5 accuracy as the evaluation metrics.

### C. Performance Analysis on ETH Food-101

In this subsection, we first conduct the performance comparison on feature fusion from different scales, and then the performance comparison on multi-view feature aggregation. Finally, we give the comparison with the state-of-the-art.

1) *Performance Comparison on Multi-Scale Feature Fusion*: The results on ETH Food-101 from three types of network architectures are displayed in Table I, Table II and Table III, respectively. We can see that (1) For single scale based method, L2 ( $128 \times 128$  patches) works better than L1 and L3 for deep visual features for all three types of networks. The reason is that different types of features works better at different scales. When adopting the deep visual features, L2 may contain more discriminative information compared with L1 and L3. The large scale probably contains much background while L3 scale contains incomplete appearance and patterns for deep visual features. (2) In

TABLE III

COMPARISON OF TOP-1 AND TOP-5 ACCURACY IN % FROM DIFFERENT COMBINATIONS OF SCALES FOR DEEP VISUAL FEATURES, MID-LEVEL ATTRIBUTE FEATURES AND SEMANTIC FEATURES ON THE ETH FOOD-101 USING DENSENET-161

(a) Deep visual features				(b) Mid-level attribute features				(c) High-level semantic features			
Method	FD	Top-1	Top-5	Method	FD	Top-1	Top-5	Method	FD	Top-1	Top-5
L1	2,208	86.94	97.03	L1	227	82.84	93.30	L1	101	86.66	96.89
L2	2,208	89.08	97.91	L2	227	84.50	<b>95.30</b>	L2	101	86.28	97.36
L3	2,208	85.93	96.95	L3	227	78.10	93.98	L3	101	82.16	96.15
L1+L2	4,416	89.57	97.94	L1+L2	454	84.88	94.82	L1+L2	202	88.76	97.79
L1+L3	4,416	88.64	97.69	L1+L3	454	83.57	94.06	L1+L3	202	88.28	97.50
L2+L3	4,416	90.04	98.11	L2+L3	454	83.74	95.04	L2+L3	202	86.10	97.31
L1+L2+L3	6,624	<b>90.14</b>	<b>98.11</b>	L1+L2+L3	681	<b>84.89</b>	94.83	L1+L2+L3	303	<b>89.32</b>	<b>97.94</b>

Table I, the performance of the combination among two scales is higher than single scale based method for all three types of features in many cases. Concatenating all three scale levels gives the best performance over any subset in Top-1 accuracy. We can also see similar trends when adopting different networks in Table II and Table III. Note that the better result benefits from the complementary advantages from three different scales. Multi-scale fusion can improve the recognition performance. (3) For three types of networks, the performance of multi-scale fusion on the DenseNet is best for its unique network architecture and efficient training method.

### 2) Performance Comparison on Multi-View Feature Fusion:

In our experiment, three types of features, namely deep visual features, mid-level attribute features and high-level semantic features are used. Table IV shows the experimental results from different combinations of different types of features, where F1 denotes deep visual features with multi-scale fusion. F2 and F3 denote mid-level attribute features and high-level semantic features with multi-scale fusion, respectively. Considering different types of features are in different range of values, we first normalize each type of features and then concatenate them. In our experiment, each type of features is normalized to the [0, 1] interval, then standardized using z-score method. From Table IV, we can see that (1) The performance of the combination among two or three types of features is generally higher than single type based method for all three networks. Concatenating all three types of features gives the best performance over any subset for all three types of networks in both Top-1 and Top-5 accuracy. We can conclude that the features from different types describe food from different aspects, and are thus very complementary. (2) The performance of final features via two-level fusion is better than ones via multi-scale fusion for all three types of networks. This verifies MSMVFA can obtain more satisfactory results through two-level fusion. (3) For three types of networks, the performance of multi-scale multi-view feature aggregation on the DenseNet-161 is best. Therefore, we use the DenseNet as the basic network for another two datasets VireoFood-172 and ChineseFoodNet in the following experiment.

3) Comparison With the State-of-the-Art: We compare against the state-of-the-art. Table V shows the results achieved by existing methods and our method on the ETH Food-101. The performance on different neural networks including AlexNet, Inception V3, ResNet-200 and WRN is listed. From Table V, we can see that (1) The performance of WRN is better

TABLE IV

COMPARISON OF TOP-1 AND TOP-5 ACCURACY IN % FROM MULTI-VIEW FEATURE AGGREGATION UNDER THREE DIFFERENT NETWORK ARCHITECTURES ON ETH FOOD-101. F1, F2 AND F3 DENOTE DEEP VISUAL FEATURES, MID-LEVEL ATTRIBUTE FEATURES AND HIGH-LEVEL SEMANTIC FEATURES FROM THE COMBINATION OF THREE DIFFERENT SCALES, RESPECTIVELY. F1 + F2 + F3 CORRESPONDS TO THE COMBINATION AMONG F1, F2 AND F3

Network	Method	FD	Top-1	Top-5
VGG-16	F1	12,288	85.89	96.97
	F2	681	78.60	90.36
	F3	303	84.94	96.68
	F1+F2	12,969	87.66	97.43
	F1+F3	12,591	87.41	97.33
	F2+F3	984	84.30	95.88
	F1+F2+F3	13,272	<b>87.68</b>	<b>97.45</b>
ResNet-152	F1	6,144	89.00	97.85
	F2	681	83.81	95.70
	F3	303	89.05	97.79
	F1+F2	6,825	90.36	98.12
	F1+F3	6,447	89.80	98.00
	F2+F3	984	89.15	97.99
	F1+F2+F3	7,128	<b>90.37</b>	<b>98.15</b>
DenseNet-161	F1	6,624	90.14	98.11
	F2	681	84.89	94.83
	F3	303	89.32	97.94
	F1+F2	7,305	90.51	98.17
	F1+F3	6,927	90.57	98.19
	F2+F3	984	89.54	98.02
	F1+F2+F3	7,608	<b>90.59</b>	<b>98.25</b>

TABLE V

COMPARISON OF TOP-1 AND TOP-5 ACCURACY IN % OF OUR MODEL AND OTHER METHODS ON THE ETH FOOD-101. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD TYPE

Method	Top-1	Top-5
AlexNet-CNN [6]	56.40	-
DCNN-FOOD [53]	70.41	-
DeepFood [26]	77.4	93.7
FCAN [27]	86.5	-
CurriculumNet [14]	87.3	-
Inception V3 [15]	88.28	96.88
ResNet-200 [17]	88.38	97.85
WRN [45]	88.72	97.92
WiSeR [30]	90.27	<b>98.71</b>
MSMVFA(ResNet-152)	<b>90.37</b>	98.15
MSMVFA(DenseNet-161)	<b>90.59</b>	98.25

than other single networks. (2) WiSeR [10] improved WRN by adding the other slice branch network with slice convolutional layers, which is used to capture specific vertical food layers.

TABLE VI

COMPARISON OF TOP-1 AND TOP-5 ACCURACY IN % FROM DIFFERENT COMBINATIONS OF SCALES FOR DEEP VISUAL FEATURES, MID-LEVEL ATTRIBUTE FEATURES AND SEMANTIC FEATURES ON THE VIREOFOOD-172 USING DENSENET-161

(a) Deep visual features				(b) Mid-level attribute features				(c) High-level semantic features			
Method	FD	Top-1	Top-5	Method	FD	Top-1	Top-5	Method	FD	Top-1	Top-5
L1	2,208	87.40	97.25	L1	353	82.84	95.55	L1	172	86.93	97.17
L2	2,208	89.70	98.02	L2	353	83.15	96.35	L2	172	87.53	97.53
L3	2,208	83.86	96.02	L3	353	77.86	94.44	L3	172	77.86	94.28
L1+L2	4,416	89.96	98.10	L1+L2	706	84.96	96.75	L1+L2	344	89.34	97.98
L1+L3	4,416	88.63	97.67	L1+L3	706	84.51	96.46	L1+L3	344	88.29	97.54
L2+L3	4,416	90.23	98.14	L2+L3	706	83.85	96.62	L2+L3	344	88.29	97.73
L1+L2+L3	6,624	<b>90.28</b>	<b>98.20</b>	L1+L2+L3	1,059	<b>85.87</b>	<b>97.13</b>	L1+L2+L3	516	<b>89.75</b>	<b>98.08</b>

TABLE VII

COMPARISON OF TOP-1 AND TOP-5 ACCURACY IN % FROM DIFFERENT COMBINATIONS OF SCALES FOR DEEP VISUAL FEATURES, MID-LEVEL ATTRIBUTE FEATURES AND SEMANTIC FEATURES ON CHINESEFOODNET USING DENSENET-161

(a) Deep visual features				(b) Mid-level attribute features				(c) High-level semantic features			
Method	FD	Top-1	Top-5	Method	FD	Top-1	Top-5	Method	FD	Top-1	Top-5
L1	2,208	75.49	94.33	L1	353	63.56	88.44	L1	172	75.22	93.97
L2	2,208	81.11	96.60	L2	353	63.06	88.41	L2	172	78.67	95.75
L3	2,208	79.02	95.72	L3	353	59.00	85.72	L3	172	75.17	96.07
L1+L2	4,416	81.04	96.56	L1+L2	706	66.03	90.36	L1+L2	344	79.07	95.87
L1+L3	4,416	79.35	96.04	L1+L3	706	66.01	90.23	L1+L3	344	77.26	95.31
L2+L3	4,416	81.94	96.82	L2+L3	706	64.09	89.10	L2+L3	344	78.28	95.78
L1+L2+L3	6,624	<b>81.96</b>	<b>96.92</b>	L1+L2+L3	1,059	<b>66.41</b>	<b>90.32</b>	L1+L2+L3	516	<b>79.47</b>	<b>96.25</b>

TABLE VIII

THE RESULTS IN % FROM MULTI-VIEW FEATURE AGGREGATION FROM DENSENET-161 ON VIREOFOOD-172 AND CHINESEFOODNET

Network	Method	FD	Top-1	Top-5
VireoFood-172	F1	6,624	90.28	98.20
	F2	1,059	85.87	97.13
	F3	516	89.75	98.08
	F1+F2	7,683	90.56	98.22
	F1+F3	7,140	90.55	98.22
	F2+F3	1,575	90.06	<b>98.40</b>
	F1+F2+F3	8,199	<b>90.61</b>	98.31
ChineseFoodNet	F1	6,624	81.96	96.92
	F2	1,059	66.41	90.32
	F3	624	79.47	96.26
	F1+F2	7,683	81.91	96.82
	F1+F3	7,248	<b>81.99</b>	96.89
	F2+F3	1,683	81.17	96.86
	F1+F2+F3	8,307	81.94	<b>96.94</b>

The output of two branches is fused via the concatenation and then fed to two fully connected layers for food classification. (3) When using general ResNet, MSMVFA(ResNet) has better performance than WISer in Top-1 accuracy. When adopting the DenseNet network, MSMVFA achieved the best performance in Top-1 accuracy, and can improve the Top-1 performance of WISer specifically designed for food recognition by 0.3%. Although marginal, MSMVFA achieves the state-of-the-art food recognition performance of Top-1 accuracy. This again verifies the effectiveness of MSMVFA.

#### D. Performance Analysis on VireoFood-172 and Chinese FoodNet

The classification accuracy from multi-scale feature fusion on VireoFood-172 is summarized in Table VI. Different

TABLE IX

COMPARISON OF TOP-1 AND TOP-5 ACCURACY IN % OF OUR MODEL AND OTHER METHODS ON THE VIREOFOOD-172 DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD TYPE

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)[8]	82.06	95.88
MultiTaskDCNN(DenseNet-161)[8]	87.21	97.29
MSMVFA(DenseNet-161)	<b>90.61</b>	<b>98.31</b>

TABLE X

COMPARISON OF TOP-1 AND TOP-5 ACCURACY IN % OF OUR MODEL AND OTHER METHODS ON THE CHINESEFOODNET DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD TYPE

Method	Top-1	Top-5
DenseNet-121[10]	78.07	95.42
DenseNet-169[10]	78.87	95.80
DenseNet-201[10]	79.05	95.79
DenseNet Fusion[10]	80.47	96.26
MSMVFA(DenseNet-161)	<b>81.94</b>	<b>96.94</b>

from Food-101, VireoFood-172 belongs to Chinese cuisine. As shown in Table VI, for each type of features, we obtain the highest Top-1 and Top-5 recognition accuracy by leveraging the power of feature ensemble, which integrates three different scales compared with single or two-scale fusion. Table VII shows the classification accuracy from multi-scale feature fusion on ChineseFoodNet. The number of categories and samples from ChineseFoodNet is larger than VireoFood-172. Because the ChineseFoodNet does not provide the ingredients, we directly extracted the ingredient features from ingredient

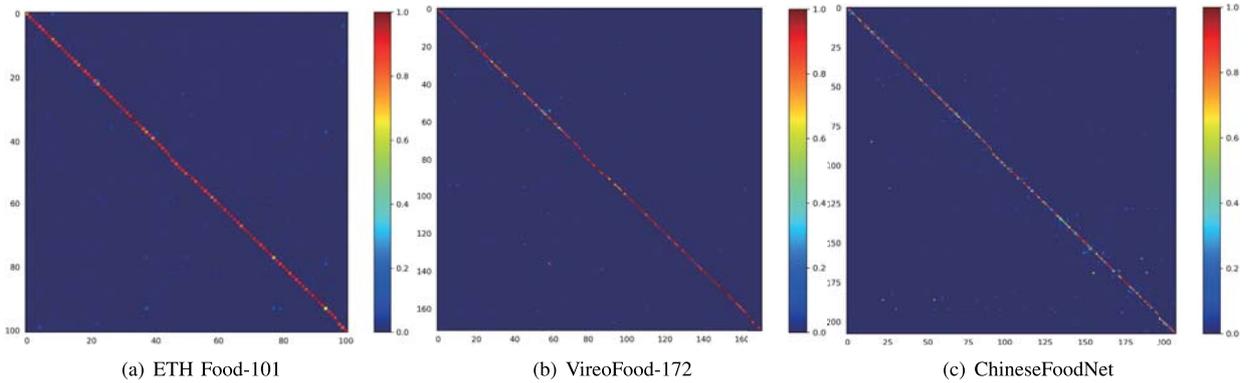


Fig. 5. The detailed comparison over each individual food category for MSMVFA via the confusion matrix. The column denotes the true food category and the row denotes estimated category (best viewed under magnification).



Fig. 6. Some confused food categories from three datasets.

model of VireoFood-172. Again, we can see that concatenating all three scale levels gives the best performance over any subset in Top-1 and Top-5 accuracy.

Table VIII shows experimental results of multi-view fusion on two datasets VireoFood-172 and ChineseFoodNet, respectively. In VireoFood-172, the performance of fusing three types of features achieves the best performance in Top-1 accuracy. In ChineseFoodNet, the performance of fusing three types of features is comparable with the best performance. The possible reason is that we simply adopted the features from the ingredient model of VireoFood-172 and the ingredient features are not best for ChineseFoodNet.

Finally, we compared our method with other baselines on both VireoFood-172 and ChineseFoodNet. The recognition results on VireoFood-172 are summarized in Table IX. MultiTaskDCNN [15] is multi-task method with two types of output layers, one is classification layer and the other is ingredient prediction layer. For fair comparison, we realized the version of MultiTaskDCNN based on the DenseNet-161 network. As Table IX shown, although VireoFood-172 is far different from ETH Food-101, similar results have been observed, our methods also achieve the best performance. Similar results can also be observed on ChineseFoodNet in Table X.

### E. Discussions

To our knowledge, our proposed framework has achieved the state-of-art performance for food recognition on three popular large-scale food datasets. However, there are still some food images, which are hard to recognize. This section lays out additional observations that follow from our results to find the probable reasons.

Fig. 5 showed the confusion matrix of MSMVFA over each individual food category. We can see that our method still does not provide perfect accuracy for some food categories. We further observe some confused food categories based on Fig. 5, and Fig. 6 shows some confused food categories from three datasets. We can see that these food categories are very similar in the visual appearance and texture. Even the humans are not easy to distinguish among these food categories. We probably need to design more fine-grained visual feature learning methods to classify these food categories.

## V. CONCLUSION

In this paper, we propose a Multi-Scale Multi-View Feature Aggregation (MSMVFA) scheme for food recognition. MSMVFA incorporates both food images and ingredient context information to aggregate high-level semantic features,

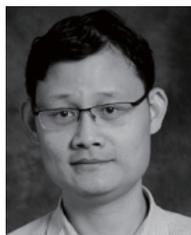
mid-level attribute features and deep visual features into the unified representation to capture the semantics of food images with the greatest probability. Besides, by multi-scale fusion of deep convolutional activation features for each type of features, MSMVFA is able to learn more robust and discriminative features to handle geometric deformation. By aggregating features via two-level fusion, namely multi-scale feature fusion for each type of features and multi-view feature aggregation among three types of features, MSMVFA can generate more robust, discriminative and comprehensive representation to cope with sophisticated visual complexities unique for food images. Extensive experimental results have demonstrated that MSMVFA outperforms all the baseline models on all popular large-scale benchmark food datasets in Top-1 accuracy.

In the future, we plan to conduct the research on four directions: (1) We plan to introduce the attention mechanism [38] into our scheme to localize discriminative regions rather than fixed patch division to improve the performance of food recognition. Also salient features can be learned from different methods, such as compressive sensing [50] and variational auto-encoder [51] for the performance improvement. (2) Existing multi-scale methods should predefine the number of regions, and we can utilize the deep reinforcement learning method [52] to automatically detect discriminative regions for food recognition, such as [53]. (3) We can extend our scheme to the problem of multiple items in one food image. (4) We generalize the proposed scheme to other fields such as social images with rich attributes to enable social image recognition.

## REFERENCES

- [1] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2249–2256.
- [2] G. M. Farinella, M. Moltisanti, and S. Battiato, "Classifying food images represented as bag of textons," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 5212–5216.
- [3] 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.
- [4] A. Meyers *et al.*, "Im2Calories: Towards an automated mobile vision food diary," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1233–1241.
- [5] F. Zhou and Y. Lin, "Fine-grained image classification by exploring bipartite-graph labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1124–1133.
- [6] A. Salvador *et al.*, "Learning cross-modal embeddings for cooking recipes and food images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3020–3028.
- [7] E. Aguilar, B. Remeseiro, M. Bolaños, 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.
- [8] V. Bettadapura, E. Thomaz, A. Parnami, G. D. Abowd, and I. Essa, "Leveraging context to support automated food recognition in restaurants," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 580–587.
- [9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
- [10] 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.
- [11] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1487–1500, Mar. 2018.
- [12] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *Proc. 22nd ACM Int. Conf.*, Nov. 2014, pp. 1085–1088.
- [13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [14] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
- [15] 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.
- [16] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 392–407.
- [17] R. Wu, B. Wang, W. Wang, and Y. Yu, "Harvesting discriminative meta objects with deep cnn features for scene classification," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1287–1295.
- [18] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
- [20] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.
- [21] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," *ACM Comput. Surv.*, to be published.
- [22] N. Martinel, C. Piciarelli, C. Micheloni, and G. L. Foresti, "A structured committee for food recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Dec. 2015, pp. 484–492.
- [23] O. Beijbom, N. Joshi, D. Morris, S. Saponas, and S. Khullar, "Menu-match: Restaurant-specific food logging from images," in *IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 844–851.
- [24] R. Xu, L. Herranz, S. Jiang, S. Wang, X. Song, and R. Jain, "Geolocalized modeling for dish recognition," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1187–1199, Aug. 2015.
- [25] H. Wang, W. Min, X. Li, and S. Jiang, "Where and what to eat: Simultaneous restaurant and dish recognition from food image," in *Proc. Pacific Rim Conf. Multimedia*, Nov. 2016, pp. 520–528.
- [26] J. Chen, L. Pang, and C. W. Ngo, "Cross-modal recipe retrieval: How to cook this dish?" in *Proc. Int. Conf. Multimedia Model.*, Dec. 2017, pp. 588–600.
- [27] W. Min, S. Jiang, S. Wang, J. Sang, and S. Mei, "A delicious recipe analysis framework for exploring multi-modal recipes with various attributes," in *Proc. ACM Multimedia Conf.*, Oct. 2017, pp. 402–410.
- [28] 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.
- [29] W. Min, B.-K. Bao, S. Mei, Y. Zhu, Y. Rui, and S. Jiang, "You are what you eat: Exploring rich recipe information for cross-region food analysis," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 950–964, Apr. 2017.
- [30] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *J. Mach. Learn. Res.*, vol. 15, pp. 2949–2980, Oct. 2014.
- [31] X. Song, S. Jiang, and L. Herranz, "Multi-scale multi-feature context modeling for scene recognition in the semantic manifold," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2721–2735, Jun. 2017.
- [32] L. Herranz, S. Jiang, and X. Li, "Scene recognition with CNNs: Objects, scales and dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 571–579.
- [33] S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, vol. 2014, pp. 1–14.
- [34] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4476–4484.
- [35] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "PANDA: Pose aligned networks for deep attribute modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1637–1644.
- [36] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3730–3738.
- [37] E. Orhan, "A simple cache model for image recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 10128–10137.

- [38] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. Int. Conf. Comput. Vis.*, vol. 6, Oct. 2017, pp. 5219–5227.
- [39] V. Papan and M. Elad, "Multi-scale patch-based image restoration," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 249–261, Jan. 2016.
- [40] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2231–2239.
- [41] X. Chen, H. Zhou, Y. Zhu, and L. Diao, "Chinesefoodnet: A large-scale image dataset for chinese food recognition," 2017, *arXiv:1705.02743*. [Online]. Available: <https://arxiv.org/abs/1705.02743>
- [42] M. Bola nos, A. Ferrà, and P. Radeva, "Food ingredients recognition through multi-label learning," in *Proc. Int. Conf. Image Anal. Process.*, 2017, pp. 394–402.
- [43] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," 2014, *arXiv:1408.5093*. [Online]. Available: <https://arxiv.org/abs/1408.5093>
- [44] 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)*, Jul. 2015, pp. 1–6.
- [45] 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.
- [46] X. Liu, T. Xia, J. Wang, Y. Yang, F. Zhou, and Y. Lin, "Fully convolutional attention networks for fine-grained recognition," 2016, *arXiv:1603.06765*. [Online]. Available: <https://arxiv.org/abs/1603.06765>
- [47] S. Guo *et al.*, "Curriculumnet: Weakly supervised learning from large-scale Web images," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 139–154.
- [48] H. Hassannejad, G. Matrella, P. Ciampolini, I. D. Munari, M. Mordonini, and S. Cagnoni, "Food image recognition using very deep convolutional networks," in *Proc. Int. Workshop Multimedia Assist. Dietary Manage.*, 2016, pp. 41–49.
- [49] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. British Mach. Vis. Conf. (BMVC)*, R. C. Wilson, E. R. Hancock, and W. A. P. Smith, Eds., Sep. 2016, pp. 87.1–87.12. doi: [10.5244/C.30.87](https://doi.org/10.5244/C.30.87).
- [50] L. T. Tan and L. B. Le, "Joint data compression and MAC protocol design for smartgrids with renewable energy," *Wireless Commun. Mobile Comput.*, vol. 16, no. 16, pp. 2590–2604, 2016.
- [51] L. T. T. Abrar Zahin and R. Q. Hu, "Sensor-based human activity recognition for smart healthcare: A semi-supervised machine learning," in *Proc. Int. Conf. Artif. Intell. Commun. Netw.*, 2019, pp. 450–472.
- [52] L. T. Tan and R. Q. Hu, "Mobility-aware edge caching and computing in vehicle networks: A deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10190–10203, 2018.
- [53] X. He, Y. Peng, and J. Zhao, "StackDRL: Stacked deep reinforcement learning for fine-grained visual categorization," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 741–747.



**Shuqiang Jiang** (SM'08) is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, and a Professor with the University of CAS. He is also with the Key Laboratory of Intelligent Information Processing, CAS. He has authored or coauthored over 150 papers on the related research topics. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He was supported by the New-Star program of Science and Technology of Beijing

Metropolis in 2008, NSFC Excellent Young Scientists Fund in 2013, and Young top-notch talent of Ten Thousand Talent Program in 2014. He has served as a TPC member for over 20 well-known conferences, including ACM Multimedia, CVPR, ICCV, IJCAI, AAAI, ICME, ICIP, and PCM. He is a Senior Member of CCF and a member of ACM. He received the Lu Jiaxi Young Talent Award from Chinese Academy of Sciences in 2012 and the CCF Award of Science and Technology in 2012. He is the Vice Chair of the IEEE CASS Beijing Chapter and the Vice Chair of ACM SIGMM China chapter. He is the General Chair of ICIMCS 2015 and the Program Chair of ACM Multimedia Asia2019 and PCM2017. He is an Associate Editor of the IEEE MULTIMEDIA and *Multimedia Tools and Applications*.



**Weiqing Min** received the B.E. degree from Shandong Normal University, Jinan, China, in 2008, the M.E. degree from Wuhan University, Wuhan, China, in 2010, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2015. He 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 over 20 peer-referenced papers in relevant journals and conferences, including *ACM Computing Surveys*, the IEEE TRANSACTIONS ON MULTIMEDIA, *ACM TOMM*, the *IEEE Multimedia Magazine*, *ACM Multimedia*, and *IJCAI*. His current research interests include multimedia content analysis, understanding and applications, food computing, and geo-multimedia computing. 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 is the Reviewer of some international journals, including the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON NEURAL NETWORK AND LEARNING SYSTEM, *ACM TOMM*, and the *IEEE Multimedia Magazine*.



**Linhu Liu** is currently pursuing the M.E. degree with the University of Chinese Academy of Sciences, Beijing, China. His current research interests include multimedia content analysis, understanding, and applications.



**Zhengdong Luo** is currently pursuing the M.E. degree with the University of Chinese Academy of Sciences, Beijing, China. His current research interests include multimedia content analysis, understanding, and applications.