ISIA Food-500: A Dataset for Large-Scale Food Recognition via Stacked Global-Local Attention Network

Weiqing Min[1,2], Linhu Liu[1,2], Zhiling Wang[1,2], Zhengdong Luo[1,2], Xiaoming Wei[3], Xiaolin Wei[3], 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

[3] Meituan-Dianping Group

{minweiqing,sqjiang,luozhengdong}@ict.ac.cn;{linhu.liu,zhiling.wang}@vipl.ict.ac.cn;{weixiaoming,weixiaolin02}@meituan.com

ABSTRACT

Food recognition has received more and more attention in the multimedia community for its various real-world applications, such as diet management and self-service restaurants. A large-scale ontology of food images is urgently needed for developing advanced large-scale food recognition algorithms, as well as for providing the benchmark dataset for such algorithms. To encourage further progress in food recognition, we introduce the dataset ISIA Food-500 with 500 categories from the list in the Wikipedia and 399,726 images, a more comprehensive food dataset that surpasses existing popular benchmark datasets by category coverage and data volume. Furthermore, we propose a stacked global-local attention network, which consists of two sub-networks for food recognition. One subnetwork first utilizes hybrid spatial-channel attention to extract more discriminative features, and then aggregates these multi-scale discriminative features from multiple layers into global-level representation (e.g., texture and shape information about food). The other one generates attentional regions (e.g., ingredient relevant regions) from different regions via cascaded spatial transformers, and further aggregates these multi-scale regional features from different layers into local-level representation. These two types of features are finally fused as comprehensive representation for food recognition. Extensive experiments on ISIA Food-500 and other two popular benchmark datasets demonstrate the effectiveness of our proposed method, and thus can be considered as one strong baseline. The dataset, code and models can be found at http://123.57.42.89/FoodComputing-Dataset/ISIA-Food500.html.

CCS CONCEPTS

• **Computing methodologies → Image representations**; **Object recognition**.

KEYWORDS

Food Recognition, Food Datasets, Benchmark, Deep Learning

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3414031

ACM Reference Format:

Weiqing Min[1,2], Linhu Liu[1,2], Zhiling Wang[1,2], Zhengdong Luo[1,2], Xiaoming Wei[3], Xiaolin Wei[3], Shuqiang Jiang[1,2]. 2020. ISIA Food-500: A Dataset for Large-Scale Food Recognition via Stacked Global-Local Attention Network. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3414031

1 INTRODUCTION

Food computing [38] is emerging as a new field to ameliorate the issues from many food-relevant fields, such as nutrition, agriculture and medicine. As one significant task in food computing, food recognition has received more attention in multimedia and beyond [15, 25, 36, 41] for its various applications, such as visual food diary [36], health-aware recommendation [42] and self-service restaurants [2].

Despite its great potential applications, recognizing food from images is still a challenging task, and the challenge derives from three-fold:

- **There is a lack of large-scale food dataset for food recognition.** Existing works mainly focus on utilizing smaller datasets for food recognition, such as ETH Food-101 [6] and Vireo Food-172 [7]. For example, Bossard *et al.* [6] released one food dataset ETH Food-101 from western cuisines with 101 food categories and 101,000 images. Chen *et al.* [7] introduced the Vireo Food-172 dataset from 172 Chinese food categories. These data-sets is lack of diversity and coverage in food categories and do not include a wide range of food images. Therefore, they are probably not sufficient to construct more complicated deep learning models for food recognition.
- **There are larger intra-class variations in the global appearance, shape and other configurations for food images.** As shown in Fig. 1, there are different shapes for the butter pecan and different textures appear in the mie goreng dish. Although numerous methods have been developed for addressing the problem of food recognition, most of these methods mainly focus on extracting features with certain type or some types while ignoring other aspects. For example, works on [4] mainly extracted color features while Niki *et al.* [32] designed a network to capture certain vertical structure for food recognition.
- **There are subtle discriminative details from food images, which are harder to capture in many cases.** Food recognition belongs to fine-grained recognition. Therefore, discriminative details are too subtle to be well-represented by existing

CNNs in many cases. As shown in Fig. 1, global features are not discriminative enough to distinguish between corn stew and leek soup. Although local regional features are probably more useful, we should carefully design one network to capture and represent such subtle difference. In order to improve the recognition performance, additional context information, such as location and ingredients [4, 41, 51, 59] is utilized. However, when these information is unavailable, these methods probably do not work.

Figure 1: Some samples from ISIA Food-500

In this work, we address data limitations by introducing a new large-scale dataset ISIA Food-500 with 399,726 images and 500 categories. In contrast with existing popular benchmark datasets, it is a more comprehensive food dataset with larger category coverage, larger data volume and higher diversity. To solve another two challenges, we propose a Stacked Global-Local Attention Network (SGLANet) to jointly learn complementary global and local visual features for food recognition. This is achieved by two sub-networks, namely Global Feature Learning Subnetwork (GloFLS) and Local-Feature Learning Subnetwork (LocFLS). GloFLS first utilizes hybrid spatial-channel attention to obtain more discriminative features for each layer, and then aggregates these features from different layers with both coarse and fine-grained levels, such as shape and texture cues about food into global-level features. LocFLS adopts cascaded Spatial Transformers (STs) to localize different attentional regions (e.g., ingredient-relevant regions), and aggregates fused regional features from different layers into local-level representation. In addition, SGLANet is trained with different types of losses in an end-to-end fashion to maximize their complementary effect in terms of discriminative power.

The contributions of our paper can be summarized as follows:

- We introduce a new large-scale and highly diverse food image dataset with 500 categories and about 400,000 images, which will be made publicly available to further the development of scalable food recognition.
- We propose a stacked global-local attention network architecture to jointly learn food-oriented global and local features via combining hybrid spatial-channel attention and multi-scale strategy for food recognition.
- We conduct extensive evaluation on our proposed dataset and other two popular food benchmark datasets to verify the effectiveness of our approach. As one strong baseline, code and models will also be released upon publication to support future research.

Table 1: Summary of available datasets for food recognition.

Dataset	#Images	#Categories	#Coverage
PFID [9]	4,545	101	Japanese
UEC Food100 [34]	14,361	100	Japanese
UEC Food256 [27]	25,088	256	Japanese
ETHZ Food-101 [6]	101,000	101	Western
UPMC Food-101 [48]	90,840	101	Western
UNIMIB2015 [12]	2,000	15	Misc.
UNIMIB2016 [13]	1,027	73	Misc.
ChineseFoodNet [10]	192,000	208	Chinese
Vireo Food-172 [7]	110,241	172	Chinese
KenyanFood13 [23]	8,174	13	Kenyan
Sushi-50 [44]	3,963	50	Japanese
FoodX-251 [26]	158,846	251	Misc.
ISIA Food-200 [41]	197,323	200	Misc.
ISIA Food-500	**399,726**	**500**	**Misc.**

2 RELATED WORK

Food-centric datasets More and more food datasets have been developed [6, 7, 26, 27, 34, 41] in recent years. Table 1 summarizes statistics of publicly available datasets for food recognition. The first benchmark is the PFID dataset [9] with only 4,545 images from 101 fast food categories. ETHZ Food-101 dataset [6] and VIREO Food-172 dataset [7] consist of more food images. However, these datasets failed in term of more comprehensive coverage of food categories, like object-centric ImageNet [14] and place-centric Places [58]. We hence introduce a new large scale food dataset ISIA Food-500 with 399,726 images and 500 food categories, and it aims at advancing multimedia food recognition and promoting the development of food-oriented multimedia intelligence.

There are some recipe-relevant multimodal datasets, such as Yummly28K [39], Yummly66K [37] and Recipe1M [45]. Recipe1M is the most known dataset, which contains about 1 million structured cooking recipes and their images for cross-modal retrieval. In contrast, the goal of our proposed ISIA Food-500 is for advancing multimedia food recognition.

Food Recognition Recently, Min *et al.* [38] gave a survey on food computing including food recognition. In the earlier years, various hand-crafted features are utilized for recognition [6, 53]. For example, Lukas *et al.* [6] utilized random forests to mine discriminative image patches as visual representation. Recent advances in deep learning have gained significant attention due to its impressive performance. As a result, existing methods resort to deep learning for food recognition [18, 25, 32]. There are also literatures, which utilize additional context information, such as ingredients and location [7, 41, 59] to improve the recognition performance. For example, Zhou *et al.* [59] exploited rich relationships among ingredients and restaurant information through the bi-partite graph for food recognition. Different from these works, our work does not introduce additional context information, and design a two-branch network to jointly learn food-oriented global features(e.g., texture and shape) and local features (e.g.,ingredient-relevant regional features) to enable comprehensive and discriminative feature representation for food recognition.

In addition, our work is also very relevant to fine-grained image recognition [49], which aims to classify subordinate categories. Food recognition belongs to fine-grained image recognition. However, compared with other types of fine-grained objects, we should take characteristics of food images into consideration, and design the targeted network for food recognition.

3 ISIA FOOD-500

3.1 Dataset Construction

In order to obtain one high-quality food dataset with broad coverage, high diversity and density of samples, we build ISIA Food-500 from the following four steps:

(1) Constructing the Food Category List. In order to guarantee high-coverage of the categorical space, we resort to Wikipedia to construct the food concept system. Particularly, we built the food list according to “Lists of foods by ingredient" from Wikipedia[1]. The Deep-First-Search algorithm is used to traverse links of the website to find food categories more completely. After that, we obtained the original food list with 4,943 types. We then removed redundant food types and conducted the combination for synonyms. Finally, we obtained 3,309 food categories.

(2) Collecting Food Images. Using a query term from the constructed food category list, we crawled candidate images from various search engines (i.e., Google, Bing and Baidu) for broader coverage and higher diversity of food images compared with other datasets from only one data source. In order to ensure that crawled images are less noisy, we expanded search terms by adding keywords, such as “food" and “dish". In this case, images for each term are retrieved and these images are then combined from different search engines. Because some images crawled from different search engines are repeated, we conducted hash based duplication detection to remove repeated ones.

(3) Cleaning and Pre-processing Food Images. Images are cleaned up through both automatic and manual processing. For automatic data cleaning, we removed candidate images with incomplete RGB channels, and the length or width of an image less than 100 pixels. We next trained a food/non-food binary classifier to further remove non-food images. Particularly, we combined images from the training set of both ETHZ Food-101 (western dishes) and VireoFood-172 (eastern dishes) as positive samples of the training set. We then randomly selected about 400,000 non-food images from both ImageNet and Places365 as negative samples of the training set. All the test samples of both ETHZ Food-101 and VireoFood-172 and the other 100,000 non-food images randomly selected from both ImageNet and Places365 constitute the test set. We trained a deep network (VGG-16 in our work) on the constructed training set and the classification accuracy of the network achieved 99.48% on the test set. The trained model is then used to filter out non-food images from downloaded images. After automatic cleaning, we then conduct manual verification by crowd-sourcing the task to 20 Lab members.

(4) Scaling Up the Dataset. After image collection and annotation, there are still many food categories with few images. To further increase the number of the candidate dataset, we translated the name of these food categories into different languages, such as Chinese and French, and then crawled images from three search engines. We also crawled more food images from other recipe/food shared websites, such as Allrecipes.com and foodgawker.com. We finally selected 500 categories with more than 500 images per category as our resulting dataset.

3.2 Dataset Statistics and Characteristics

ISIA Food-500 consists of 399,726 images with 500 categories. The average number of images per category is about 800. Fig. 2 shows sorted distribution of the number of images from sampled classes while Fig. 3 shows some samples. Note that we represented the food category with more than two words by concatenating them using ‘-’. ISIA Food-500 is a more comprehensive food dataset that surpasses existing popular benchmark datasets, such as ETH Food-101 and Vireo Food-172 from the following three aspects: **(1) Larger data volume**. It has 399,726 images from 500 food categories, which has created a new milestone for the task of complex food recognition. **(2) Larger category coverage**. It consists of 500 categories, which is about $3 \sim 5$ times that of existing datasets, such as Food-101 and Vireo Food-172. **(3) Higher diversity**. Food categories from this dataset covers various countries and regions including both eastern and western cuisines. Fig. 4 provided the comparisons of distributions of food categories on food types, such as ETH Food-101 (western food), Vireo Food-172 (eastern food) and ISIA Food-200 (Misc. food). According to the GSFA standard[2], the food from our dataset and existing typical ones mainly belongs to the following 11 categories: Meat, Cereals, Vegetables, Fish, Fruits, Dairy, Bakery, Fats, Confectionary, Beverages and Eggs. We can see that for most of food types, the number of food categories from ISIA Food-500 is larger than these existing datasets. Furthermore, some food types are covered in ISIA Food-500, but missing in other ones, such as Dairy and Beverages.

4 FRAMEWORK

Fig. 5 illustrates the proposed Stacked Global-Local Attention Network (SGLANet), which can jointly learn complementary global and local features for food recognition. SGLANet mainly consists of two components, namely **Glo**bal **F**eature **L**earning **S**ub-network (GloFLS) and **Loc**al-**F**eature **L**earning **S**ub-network (LocFLS). GloFLS first adopts hybrid Spatial-Channel Attention (SCA) to obtain more discriminative features from each layer of the network, and then aggregates a set of features from these layers to capture different types of global level features, such as shape and texture cues about food. LocFLS adopts cascaded STs to localize different local regions for each layer, and then aggregates fused features with different regions from different layers into final local feature representation. Finally, SGLANet fuses both global and local features

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

[2]http://www.fao.org/gsfaonline/index.html?lang=en

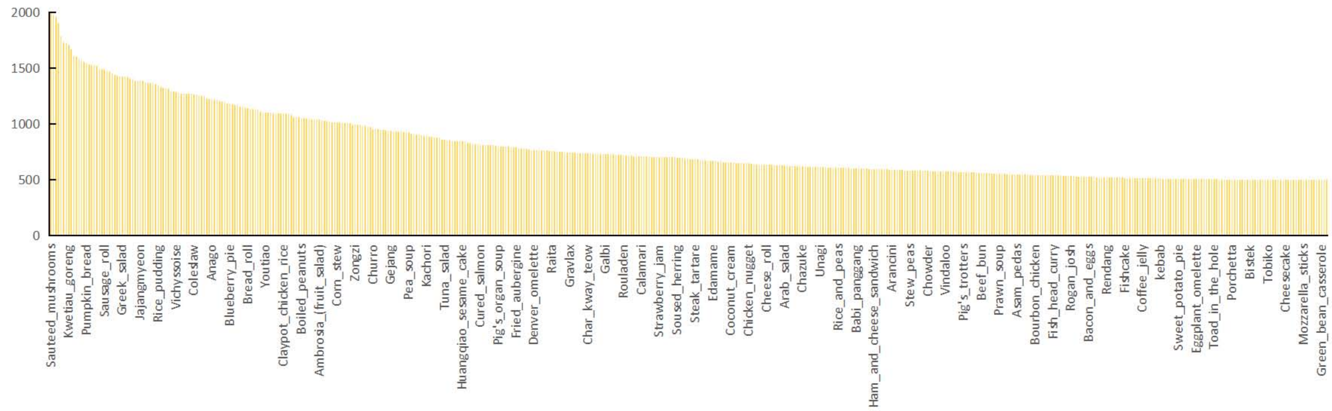


Figure 2: Sorted distribution of the number of images from sampled classes in the ISIA Food-500.

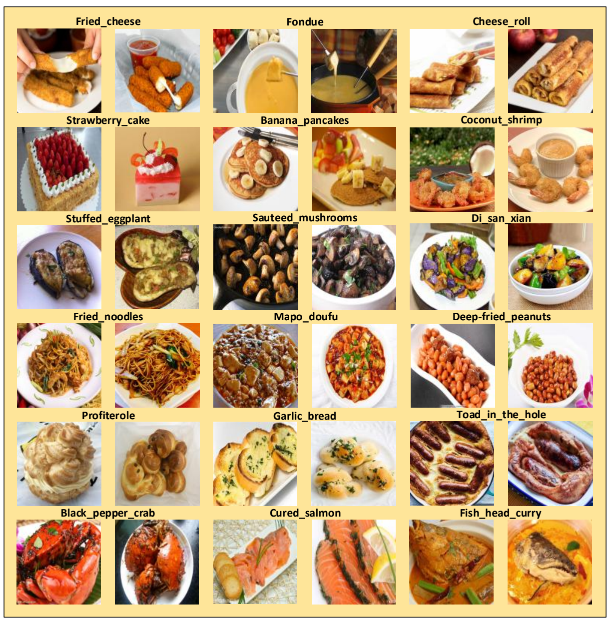


Figure 3: Image samples from the ISIA Food-500 dataset

for food recognition. In addition, SGLANet is trained with different types of losses, including global loss, local loss and joint loss in an end-to-end fashion to maximize their complementary benefit in terms of the discriminative power.

4.1 GloFLS

Given the whole input image, GloFLS first learns more discriminative features via hybrid Spatial-Channel Attention (SCA) for each layer, and then aggregates these discriminative features from different layers into global level representations via multi-layer feature fusing. Considering features extracted from different layers contain low-level, mid and high ones, GloFLS can capture various types of global level features, such as shape, texture and edge cues about food.

Spatial-Channel Attention (SCA) The combination of both spatial and channel attention can capture discriminative features comprehensively from different dimensions, and thus have been successfully applied in many tasks, such as image captioning [8] and person ReID [29]. Different from these works, we apply SCA to the food recognition task by capturing food-oriented discriminative features.

The input to a SCA module is a 3-D tensor $\mathbf{X}^l \in \mathrm{R}^{\,h\times w\times c}$ with width w, height h, channels c and the layer of GloFLS l, respectively. The output of this module is a saliency weight map $\mathbf{A}^l \in \mathrm{R}^{\,h\times w\times c}$ of the same size as $\mathbf{X}$. We calculate $\mathbf{A}^l \in \mathrm{R}$ for SCA learning [29]:

$$\mathbf{A}^l = \mathbf{S}^l \times \mathbf{C}^1 \tag{1}$$

where $\mathbf{S}^l \in \mathrm{R}^{h\times w\times 1}$ and $\mathbf{C}^l \in \mathrm{R}^{1\times 1\times c}$ mean spatial and channel attention maps, respectively.

The Global Averaging Pooling (GAP) is used to calculate the spatial attention as follows:

$$\mathbf{S}^l = \frac{1}{c}\sum_{i=1}^{c} \mathbf{X}^l_{1:h,1:w:i} \tag{2}$$

The channel attention from the squeeze-and-excitation block [19] is computed as follows:

$$\begin{aligned} \mathbf{C}^l_1 &= \frac{1}{h\times w}\sum_{i=1}^{h}\sum_{j=1}^{w} \mathbf{X}^l_{i,j,1:c} \\ \mathbf{C}^l &= ReLU(\mathbf{M}^{ca}_2 \times \mathrm{Relu}(\mathbf{M}^{ca}_1 \mathbf{C}^l_1)) \end{aligned} \tag{3}$$

where $\mathbf{M}^{ca}_1 \in \mathbf{R}^{\frac{c}{r}\times c}$ and $\mathbf{M}^{ca}_2 \in \mathbf{R}^{c\times\frac{c}{r}}$ represent the parameter matrix of 2 conv layers respectively, and r denotes the bottleneck reduction rate.

Multi-Layer Feature Fusing By extracting attentional features from multiple layers, we can obtain low, mid and high-level features, which include various types of global features, such as texture, shape and edge information [54]. Such global features are important cues for food recognition. Therefore, we aggregate discriminative attentional features from different layers into global level feature representation for food recognition via a concatenation layer and a fully connected layer.

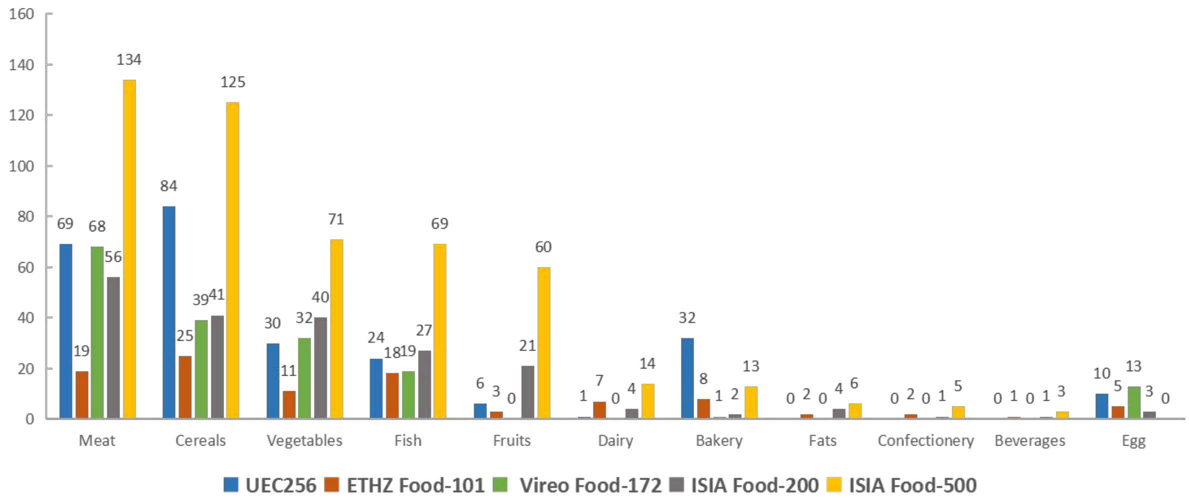


Figure 4: Comparison on distributions of categories on ISIA Food-500 and other existing typical ones.

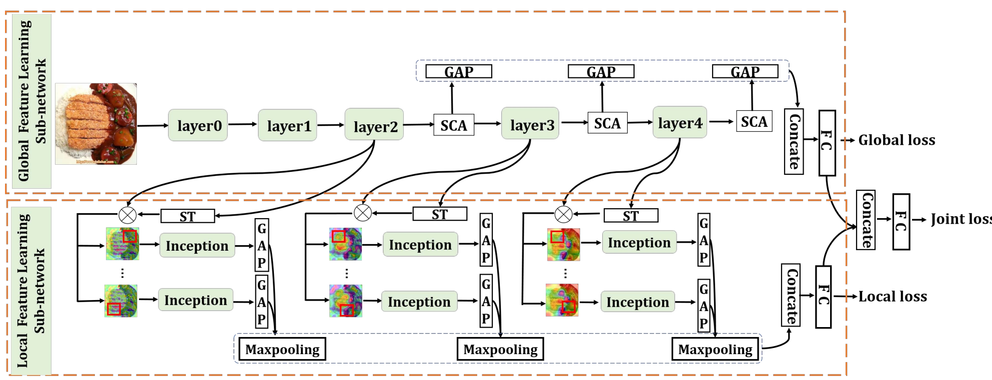


Figure 5: The proposed framework. GAP: Global Average Pooling layer. SCA: Spatial-Channel Attention. ST: Spatial Transformer. FC: Full-Connected layer.

4.2 LocFLS

LocFLS localizes discriminative regions with different positions and scales to capture local features. It uses stacked STs [22] to localize regions from different layers. For each layer, one inception block is introduced to extract regional features, and followed by a global average pooling layer and a max-pooling layer for fusing these regional features. The features from each layer are fused to final local features via a concatenation layer and a fully connected layer.

Spatial Transformer (ST) For each layer, we adopt ST to locate latent T regions, and model this regional attention by a transformation matrix as:

$$\mathbf{A}^l = \begin{bmatrix} s_h & 0 & t_x \\ 0 & s_w & t_y \end{bmatrix} \tag{4}$$

which allows for image cropping, translation, and isotropic scaling operations by varying two scale factors (s_h , s_w) and 2-D spatial position (t_x, t_y).

4.3 Learning with Multiple Losses

SGLANet is jointly optimized by three types of losses, i.e., joint loss L_{Joi}, global loss L_{Glo}, and local loss L_{Loc} respectively, leading to the final loss function:

$$L = L_{Joi} + \gamma_1 L_{Glo} + \gamma_2 L_{Loc} \tag{5}$$

where γ_1 and γ_2 are balance parameters, and the cross-entropy classification loss function is used for all three types of losses.

Such learning with different types of losses can maximize their complementary benefit in terms of the discriminative power.

5 EXPERIMENT

5.1 Experimental Setup

Our model is implemented on the Pytorch platform. The images are resized to 224×224. The models are optimized using stochastic gradient descent with a batch size of 80 and momentum of 0.9. The learning rate is set to 10^{-2} initially and divided by 10 after 30 epochs. For GloFLS, we selected SENet [19] as the backbone, and

Table 2: Evaluating individual modules in GloFLS on ISIA Food-500 (%).

Method	Top-1 acc.	Top-5 acc.
SENet-154	63.83	88.61
SENet-154+SCA	64.42	89.05
SENet-154+Multi-scale	64.60	89.08
GloFLS	**64.63**	**89.14**

Table 3: Ablation experiments on ISIA Food-500 with global & local-level features (%).

Method	Top-1 acc.	Top-5 acc.
GloFLS	64.63	**89.14**
LocFLS	64.10	88.86
SGLANet	**64.74**	89.12

the bottleneck reduction rate $r = 16$. For LocFLS, we selected simple Inception-B unit as basic building block. For each layer, $T = 4$ and the scale of ST is fixed as $s_h = s_w = 0.5$. We set $\gamma_1 = \gamma_2 = 0.5$ in Eq. 5. Top-1 accuracy (Top-1 acc.) and Top-5 accuracy (Top-5 acc.) are used as evaluation metrics.

5.2 Experiment on ISIA Food-500

ISIA Food-500 is divided into 60%, 10% and 30% images for training, validation and testing, respectively. All the experiments adopt a single centered crop (1-crop) at test time in the defaulting setting.

Ablation Study We first evaluated the effect of each individual component in GloFLS in Table 2. It shows that: (1) Any of two components in isolation brings recognition performance gain; (2) The combination of SCA and Multi-scale gives further accuracy boost, which suggests the complementary effect. We then evaluated the effect of joint global and local feature learning by comparing their individual global and local features. Table 3 shows that a performance gain is obtained in Top-1 accuracy by joining two representations, which validates the complementary effect of jointly learning global and local features from GloFLS and LocFLS.

We finally evaluate the effect of different losses as shown in Table 4. The experimental results demonstrate that we obtain the best recognition performance when different losses are utilized. The reason is that different loss functions can regulate the deep network from different aspects and work together to improve the recognition performance. Another observation is that the performance with one additional loss does not improve the performance compared with the baseline without both global and local losses. The probable reason is that the performance improvement needs joint work from two losses.

Comparisons with State-of-the-Art We evaluated SGLANet against different baseline methods on Table 5. These baselines include not only various typical deep networks, such as VGG16 and SENet, but also some recently proposed fine-grained methods, such as NTS-NET [55] and WS-DAN [20]. Note that for these fine-grained methods, we followed the same setting in their mentioned papers. We observed that the performance superiority of SGLANet over all the state-of-the-arts in both Top-1 accuracy and Top-5 accuracy. Compared with best baseline SENet-154, there is the performance improvement of about 0.9 percent in Top-1 accuracy for the test set. These results validate the advantage of joint global and local feature learning of SGLANet.

Table 4: Evaluating individual losses on ISIA Food-500 (%).

Method	Top-1 acc.	Top-5 acc.
$\lambda_1 = \lambda_2 = 0$	64.16	88.94
$\lambda_1 = 0, \lambda_2 = 0.5$	63.95	88.57
$\lambda_1 = 0.5, \lambda_2 = 0$	64.02	88.59
$\lambda_1 = 0.5, \lambda_2 = 0.5$	**64.74**	**89.12**

Table 5: Performance comparison on ISIA Food-500 (%).

Method	Top-1 acc.	Top-5 acc.
VGG-16 [47]	55.22	82.77
GoogLeNet [36]	56.03	83.42
ResNet-152 [17]	57.03	83.80
WRN-50 [46]	60.08	85.98
DenseNet-161 [21]	60.05	86.09
NAS-NET [60]	60.66	86.38
SE-ResNeXt101_32x4d [19]	61.95	87.54
NTS-NET [55]	63.66	88.48
WS-DAN [20]	60.67	86.48
DCL [11]	64.10	88.77
SENet-154 [19]	63.83	88.61
SGLANet	**64.74**	**89.12**

Visualization of GloFLS and LocFLS We visualized both SCA from GloFLS and STs from LocFLS at three different layers of SGLANet. Fig. 6 shows: (1) in GloFLS, SCA captures different global level features at different layers, such as shape information for Boiled_beef and texture information from Pumpkin_bread. Meanwhile, with increased depth of SGLANet, SCA captures more focused and discriminative features (2) in LocFLS, STs capture different local regions with less background at different layers from LocFLS, such as Crudites. This again verified complementary effect of joint global and local feature learning.

Qualitative Analysis We selected 20 classes in the test phase to further evaluate our method. Particularly, we listed the Top-1 accuracy of both 10 best and 10 worst performing classes in Fig. 7. We can observe that some categories can be easily recognized, such as Chakli and Edamame, and their Top-1 accuracy is above 97%. However, there are some categories, which are very hard to recognize, such as Curry_rice and kebab, and their Top-1 accuracy is below 10%. We further demonstrate some challenging recognized examples from the 10 worst performing classes, and Fig. 8 shows that too small inter-class variations is the main reason for bad performance. We have shown that existing methods are far from tackling large-scale recognition task with high accuracy like ImageNet, pointing to exciting future directions.

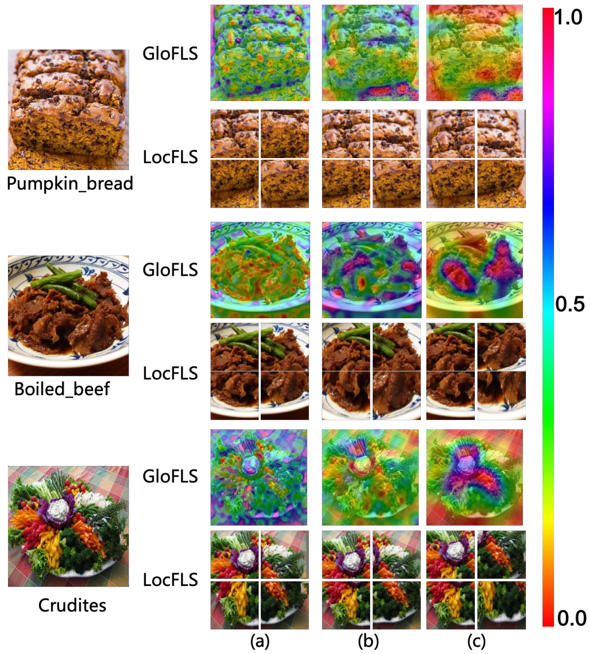
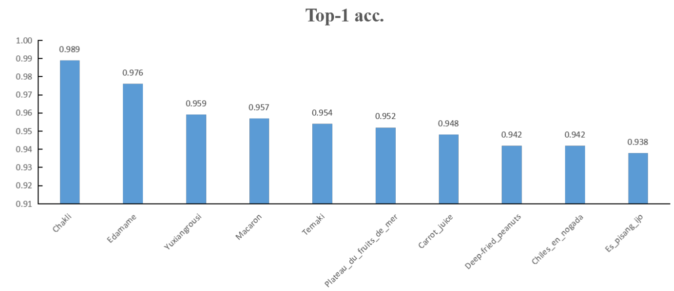


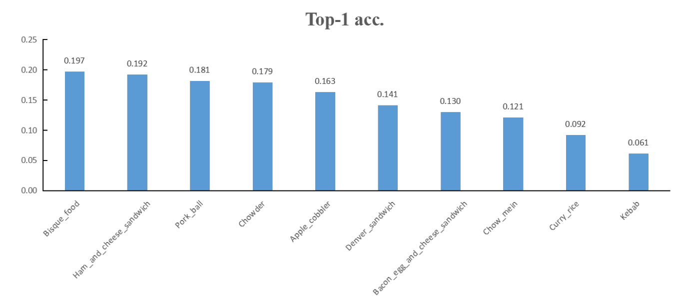
Figure 6: Visualization of SCA in GloFLS and STs in LocFLS from (a) The 2th layer,(b) The 3th layer and (c) The 4th layer.

Table 6: Performance comparison on ETHZ Food-101 (%).

Method	Setting	Top-1 acc.	Top-5 acc.
AlexNet-CNN [6]	1-crop	56.40	-
SELC [33]	1-crop	55.89	-
ResNet-152+SVM-RBF [35]	1-crop	64.98	-
DCNN-FOOD [52]	1-crop	70.41	-
LMBM [50]	1-crop	72.11	-
Ensemble Net [43]	1-crop	72.12	91.61
GoogLeNet [3]	1-crop	78.11	-
DeepFOOD [30]	1-crop	77.40	93.70
ILSVRC [5]	1-crop	79.20	94.11
WARN [31]	1-crop	85.50	-
CNNs Fusion(I_2) [1]	1-crop	86.71	-
Inception V3 [16]	1-crop	88.28	96.88
SENet-154 [19]	1-crop	88.62	97.57
WRN [32]	10-crop	88.72	97.92
SOTA[28]	1-crop	90.00	-
DLA[57]	1-crop	90.00	-
WISeR [32]	10-crop	90.27	**98.71**
IG-CMAN [41]	1-crop	90.37	98.42
PAR-Net [44]	1-crop	89.30	-
PAR-Net [44]	10-crop	90.40	-
Inception-Resnet-v2 SE [56]	1-crop	90.40	-
MSMVFA [24]	1-crop	90.59	98.25
SGLANet	1-crop	89.69	98.01
SGLANet	10-crop	90.33	98.20
SGLANet(Pretrained)	1-crop	90.47	98.21
SGLANet(Pretrained)	10-crop	**90.92**	98.24



(a) The 10 best performing classes



(b) The 10 worst performing classes

Figure 7: Selected categories from (a)The 10 best and (b)The 10 worst performing classes.

Figure 8: Some confused classes, where the first column denotes some classes from the 10 worst performing classes and for each class, 3 more confused classes are listed for each row.

5.3 Experiment on Other Benchmarks

We further conduct extensive evaluation on other two popular food benchmark datasets to verify the effectiveness of our approach, and also assessed the generalizability of our model learned using ISIA Food-500 to the two datasets. Considering some evaluations from

Table 7: Performance comparison on Vireo Food-172 (%).

Method	Setting	Top-1 acc.	Top-5 acc.
AlexNet	1-crop	64.91	85.32
VGG-16 [47]	1-crop	80.41	94.59
DenseNet-161 [21]	1-crop	86.93	97.17
MTDCNN(VGG-16) [7]	1-crop	82.06	95.88
MTDCNN(DenseNet-16) [7]	1-crop	87.21	97.29
SENet-154 [19]	1-crop	88.71	97.74
PAR-Net [44]	1-crop	89.60	-
PAR-Net [44]	10-crop	90.20	-
IG-CMAN [41]	1-crop	90.63	**98.40**
MSMVFA [24]	1-crop	90.61	98.31
SGLANet	1-crop	89.88	97.83
SGLANet	10-crop	90.30	98.03
SGLANet(Pretrained)	1-crop	90.78	98.16
SGLANet(Pretrained)	10-crop	**90.98**	98.35

existing works are conducted in the setting of 10-crop test, we show the experimental results of our method in the setting of both 1-crop and 10-crop at test time.

Experiments on ETHZ Food-101 ETHZ Food-101 contains 101,000 images from 101 food categories. There are 1,000 images including 750 training images and 250 test images for each category [6]. We evaluated SGLANet against existing methods on Food-101. Table 6 shows that our method exceeds many baseline methods except some ones, such as MSMVFA [24], IG-CMAN [41] and Inception-Resnet-v2 SE [56] under the 1-crop test setting. The reason is that MSMVFA and IG-CMAN require multiple stages training for feature extraction and introduced additional ingredient information as the supervision. Inception-Resnet-v2 SE used additional data and adopted transfer learning method. When we used the pretrained model on ISIA Food-500, namely SGLANet(Pretrained), there is the performance improvement of about 0.8 percent and 0.6 percent in Top-1 accuracy on 1-crop and 10-crop test respectively. These results also verify the generalization of models learned using ISIA Food-500.

Experiments on Vireo Food-172 Vireo Food-172 consists of 110,241 food images from 172 categories. In each food category, 60%, 10%, 30% of images are randomly selected for training, validation and testing, respectively [7]. Table 7 shows experimental results on Vireo Food-172. We can see that the performance from SGLANet is better than many baselines, except that few ones, such as IG-CMAN. This is because that these methods, such as IG-CMAN introduced additional ingredient information for food recognition. In addition, these methods generally need multi-stage feature learning. When we fine-tuned SGLANet pre-trained on ISIA Food-500, there is the performance improvement of about 0.9 percent and 0.7 percent in Top-1 accuracy on 1-crop and 10-crop test respectively, and achieved the best performance under the 1-crop setting. These results again demonstrate the generalization of our model learned using ISIA Food-500.

6 CONCLUSIONS

In this paper, we present a new large-scale dataset ISIA Food-500 with larger data volume, larger category coverage, and higher diversity compared with existing typical datasets. We then propose a stacked global-local attention network to jointly exploit complementary global and local features via the designed two subnetworks for food recognition. Extensive evaluation on ISIA Food-500 and another two benchmark datasets have verified its effectiveness, and thus can be considered as one strong baseline.

Future work includes: (1) We are expanding ISIA Food-500 dataset, and aim to complete the construction of about 1.5 million food images spread over about 2,000 food categories. We expect it will serve as a new challenge to train high-capacity models for large-scale food recognition in the multimedia community. (2) We plan to collect rich attribute information, e.g., ingredients, cooking instructions and flavor information [40] to support multimodal food recognition.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant 61972378, 61532018, U1936203, U19B2040. This research was also supported by Meituan-Dianping Group.

REFERENCES

[1] Eduardo Aguilar, Marc Bolaños, and Petia Radeva. 2017. Food recognition using fusion of classifiers based on CNNs. In *International Conference on Image Analysis and Processing*. 213–224.
[2] Eduardo Aguilar, Beatriz Remeseiro, Marc Bolaños, and Petia Radeva. 2018. Grab, Pay and Eat:Semantic Food Detection for Smart Restaurants. In *IEEE Transactions on Multimedia*, Vol. 20. 3266–3275.
[3] Shuang Ao and Charles X. Ling. 2015. Adapting new categories for food recognition with deep representation. In *IEEE International Conference on Data Mining Workshop*. 1196–1203.
[4] Vinay Bettadapura, Edison Thomaz, Aman Parnami, Gregory D Abowd, and Irfan Essa. 2015. Leveraging context to support automated food recognition in restaurants. In *IEEE Winter Conference on Applications of Computer Vision*. 580–587.
[5] Marc Bolanos and Petia Radeva. 2017. Simultaneous food localization and recognition. In *International Conference on Pattern Recognition*. 3140–3145.
[6] 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.
[7] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the ACM on Multimedia Conference*. 32–41.
[8] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua. 2017. SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6298–6306.
[9] Mei Chen, Kapil Dhingra, Wen Wu, Lei Yang, Rahul Sukthankar, and Jie Yang. 2009. PFID: Pittsburgh fast-food image dataset. In *IEEE International Conference on Image Processing*. 289–292.
[10] Xin Chen, Hua Zhou, and Liang Diao. 2017. ChineseFoodNet: A large-scale image dataset for Chinese food recognition. In *CoRR*, Vol. abs/1705.02743.
[11] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. 2019. Destruction and Construction Learning for Fine-Grained Image Recognition. *In: IEEE International Conference on Computer Vision and Pattern Recognition* (2019), 5157–5166.
[12] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. 2015. Food Recognition and Leftover Estimation for Daily Diet Monitoring. In *International Conference on Image Analysis and Processing*. 334–341.
[13] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. 2016. Food recognition: a new dataset, experiments, and results. In *IEEE Journal of Biomedical and Health Informatics*, Vol. 21. 588–598.
[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.
[15] Lixi Deng, Jingjing Chen, Qianru Sun, Xiangnan He, Sheng Tang, Zhaoyan Ming, Yongdong Zhang, and Tat-Seng Chua. 2019. Mixed-dish Recognition with

Contextual Relation Networks. In *ACM Multimedia*. ACM, 112–120.
[16] 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.
[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[18] 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.
[19] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7132–7141.
[20] Tao Hu and Honggang Qi. 2019. See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification. *CoRR* abs/1901.09891.
[21] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. 2017. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2261–2269.
[22] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*. 2017–2025.
[23] Mona Jalal, Kaihong Wang, Sankara Jefferson, Yi Zheng, Elaine O. Nsoesie, and Margrit Betke. 2019. Scraping Social Media Photos Posted in Kenya and Elsewhere to Detect and Analyze Food Types. In *Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management*. 50–59.
[24] Shuqiang Jiang, Weiqing Min, Linhu Liu, and Zhengdong Luo. 2019. Multi-Scale Multi-View Deep Feature Aggregation for Food Recognition. *IEEE Transactions on Image Processing* 29, 1, 265–276.
[25] 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.
[26] Parneet Kaur, Karan Sikka, Weijun Wang, Serge J. Belongie, and Ajay Divakaran. 2019. FoodX-251: A Dataset for Fine-grained Food Classification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.
[27] Yoshiyuki Kawano and Keiji Yanai. 2014. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *European Conference on Computer Vision*. 3–17.
[28] Simon Kornblith, Jonathon Shlens, and Quoc Le. 2019. Do Better ImageNet Models Transfer Better?. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2661–2671.
[29] W. Li, X. Zhu, and S. Gong. 2018. Harmonious Attention Network for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2285–2294.
[30] 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. In *International Conference on Smart Homes and Health Telematics*. 37–48.
[31] Pau RodrÍguez López, Diego Velazquez Dorta, Guillem Cucurull Preixens, Josep M. Gonfaus, and Jordi Gonzàlez Sabaté. 2020. Pay attention to the activations: a modular attention mechanism for fine-grained image recognition. In *IEEE Transactions on Multimedia*, Vol. 22. 502–514.
[32] 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.
[33] Niki Martinel, Claudio Piciarelli, and Christian Micheloni. 2016. A supervised extreme learning committee for food recognition. In *Computer Vision and Image Understanding*, Vol. 148. Elsevier, 67–86.
[34] Yuji Matsuda and Keiji Yanai. 2012. Multiple-food recognition considering co-occurrence employing manifold ranking. In *International Conference on Pattern Recognition*. 2017–2020.
[35] Patrick McAllister, Huiru Zheng, Raymond Bond, and Anne Moorhead. 2018. Combining deep residual neural network features with supervised machine learning algorithms to classify diverse food image datasets. In *Computers in Biology and Medicine*, Vol. 95. Elsevier, 217–233.
[36] 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.
[37] Weiqing Min, Bing-Kun Bao, 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.
[38] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A Survey on Food Computing. In *ACM Computing Surveys*, Vol. 52. 1–36.
[39] Weiqing Min, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz. 2017. Being a Super Cook: Joint Food Attributes and Multi-Modal Content Modeling for Recipe Retrieval and Exploration. *IEEE Transactions on Multimedia* 19, 5 (2017), 1100–1113.
[40] 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 *ACM Multimedia*. ACM, 402–410.
[41] Weiqing Min, Linhu Liu, Zhengdong Luo, and Shuqiang Jiang. 2019. Ingredient-Guided Cascaded Multi-Attention Network for Food Recognition. In *Proceedings of the ACM International Conference on Multimedia*. 1331–1339.
[42] 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.
[43] Paritosh Pandey, Akella Deepthi, Bappaditya Mandal, and N. B. Puhan. 2017. FoodNet: Recognizing foods using ensemble of deep networks. In *IEEE Signal Processing Letters*, Vol. 24. 1758–1762.
[44] Jianing Qiu, Frank P.-W. Lo, Yingnan Sun, Siyao Wang, and Benny Lo. 2019. Mining Discriminative Food Regions for Accurate Food Recognition. In *British Machine Vision Conference (Accepted)*.
[45] 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 *IEEE Conference on Computer Vision and Pattern Recognition*. 3020–3028.
[46] Zagoruyko Sergey and Komodakis Nikos. 2016. Wide Residual Networks. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 87.1–87.12.
[47] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.
[48] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. 2015. Recipe recognition with large multimodal food dataset. In *IEEE International Conference on Multimedia and Expo Workshops*. 1–6.
[49] Xiu-Shen Wei, Jianxin Wu, and Quan Cui. 2019. Deep Learning for Fine-Grained Image Analysis: A Survey. *CoRR* abs/1907.03069 (2019).
[50] Hui Wu, Michele Merler, Rosario Uceda-Sosa, and John R Smith. 2016. Learning to make better mistakes: Semantics-aware visual food recognition. In *ACM Multimedia Conference*. 172–176.
[51] Ruihan Xu, Luis Herranz, Shuqiang Jiang, Shuang Wang, Xinhang Song, and Ramesh Jain. 2015. Geolocalized modeling for dish recognition. In *IEEE Transactions on Multimedia*, Vol. 17. 1187–1199.
[52] Keiji Yanai and Yoshiyuki Kawano. 2015. Food image recognition using deep convolutional network with pre-training and fine-tuning. In *IEEE International Conference on Multimedia and Expo Workshops*. 1–6.
[53] 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.
[54] S. Yang and D. Ramanan. 2015. Multi-scale Recognition with DAG-CNNs. In *IEEE International Conference on Computer Vision*. 1215–1223.
[55] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. 2018. Learning to Navigate for Fine-Grained Classification. In *European Conference on Computer Vision*. 438–454.
[56] Cui Yin, Song Yang, Sun Chen, Howard Andrew, and Belongie Serge. 2018. Large Scale Fine-Grained Categorization and Domain-Specific Transfer Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4109–4118.
[57] F. Yu, D. Wang, E. Shelhamer, and T. Darrell. 2018. Deep Layer Aggregation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2403–2412.
[58] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40. 1452–1464.
[59] 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.
[60] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc Le. 2018. Learning Transferable Architectures for Scalable Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 8697–8710.