2-2024

From canteen food to daily meals: generalizing food recognition to more practical scenarios

Guoshan LIU

Yang JIAO

Jingjing CHEN

Bin ZHU
Singapore Management University, binzhu@smu.edu.sg

Yu-Gang JIANG

Citation

From Canteen Food to Daily Meals: Generalizing Food Recognition to More Practical Scenarios

Guoshan Liu, Yang Jiao, Jingjing Chen, *Member, IEEE*, Bin Zhu and Yu-Gang Jiang *Fellow, IEEE*

Abstract—The precise recognition of food categories plays a pivotal role for intelligent health management, attracting significant research attention in recent years. Prominent benchmarks, such as Food-101 and VIREO Food-172, provide abundant food image resources that catalyze the prosperity of research in this field. Nevertheless, these datasets are well-curated from canteen scenarios and thus deviate from food appearances in daily life. This discrepancy poses great challenges in effectively transferring classifiers trained on these canteen datasets to broader daily-life scenarios encountered by humans. Toward this end, we present two new benchmarks, namely DailyFood-172 and DailyFood-16, specifically designed to curate food images from everyday meals. These two datasets are used to evaluate the transferability of approaches from the well-curated food image domain to the everyday-life food image domain. In addition, we also propose a simple yet effective baseline method named Multi-Cluster Reference Learning (MCRL) to tackle the aforementioned domain gap. MCRL is motivated by the observation that food images in daily-life scenarios exhibit greater intra-class appearance variance compared with those in well-curated benchmarks. Notably, MCRL can be seamlessly coupled with existing approaches, yielding non-trivial performance enhancements. We hope our new benchmarks can inspire the community to explore the transferability of food recognition models trained on well-curated datasets toward practical real-life applications.

Index Terms—Food datasets, Food recognition, Unsupervised Domain Adaptation

I. INTRODUCTION

FOOD computing has developed remarkably with the rise of deep learning [1], [2], [3], which facilitates people to monitor and regulate their nutritional intake and caloric consumption. Several pioneering benchmarks [1], [4], [5], [6] offer abundant food images resources, where Food-101 [5], Food2K [1] are mainly consist of western food, while Vireo Food-172 [4] and Vireo Food-251 [6] are composed of Chinese food. To ensure aesthetics and discernibility, these datasets collect food images from dishes cooked by professionals in canteen scenarios. This data collection behavior introduces inherent bias into the curated datasets. Since culinary experts master extensive recipe knowledge and superb cooking skills, they tend to follow similar cooking procedures and use similar ingredients when preparing the same dishes. Consequently,

Guoshan Liu, Yang Jiao, Jingjing Chen and Yu-Gang Jiang are with Fudan Vision and Learning Laboratory (FVL), Fudan University, Shanghai, China (e-mail: {22210240029, yangjiao21}@m.fudan.edu.cn, {chenjingjing, ygj}@fudan.edu.cn). Bin Zhu is with Singapore Management University, Singapore(e-mail: binzhu@smu.edu.sg).

(a) Images selected from VIREO Food-172.

(b) Images captured in daily life.

Fig. 1: Variations in visual appearance of "Braised Tofu" from VIREO Food-172 and daily meals. The first row shows three examples of dishes in VIREO Food-172, followed by examples from daily meals in the second row.

this bias leads to a reduced variance among samples within the same type of dishes. Nevertheless, dishes in daily meals generally exhibit higher variance due to the non-standard cooking procedures and the randomness of shooting angles. As shown in Fig. 1, we compare the images of the same dish "Braised Tofu" from Vireo Food-172 and daily meals in the first and second rows, respectively. First, the comparison of examples within each row reveals that the same type of dishes from Vireo Food-172 exhibit high appearance consistency, while those from daily life vary dramatically. Besides, dishes from Vireo Food-172 are more exquisite than their daily-life counterparts. Therefore, classifiers pre-trained on these canteen food datasets struggle to effectively generalize to daily meals, thereby limiting their broader applicability.

To the best of our knowledge, there is no available dataset that encompasses food images captured in daily meals at present, which hinders the development of practical food recognition. To mitigate this gap, we curate two new benchmarks, namely DailyFood-172 and DailyFood-16, which include 172 and 16 categories of daily-life food images respectively. The DailyFood-172 is collected from the "Go Cooking" [1] website, where culinary enthusiasts share their everyday dishes. On the other hand, DailyFood-16 gathers images from "Eggplant Health"[2], a platform for weight loss purposes. Due to the cumbersome and resource-intensive nature of collecting

[1]https://www.xiachufang.com/category/
[2]https://www.qiezilife.com/

and annotating a vast number of daily-life food images, merely utilizing the proposed datasets as extensional training resources still limits the scalability of models. Therefore, in this paper, we employ our datasets as the benchmark for evaluating the transferability of existing approaches. In this scenario, deep models have access to labeled food images from existing benchmarks such as Food-101 and Vireo Food-172, while our DailyFood-172 and DailyFood-16 datasets only provide unlabeled images during training.

With the above restrictions, our benchmarks can be formulated as an Unsupervised Domain Adaptation (UDA) problem, where deep models are trained with both label-rich source dataset and label-scarce target dataset. By mitigating the distribution discrepancy between source and target domains, pioneering domain-level UDA approaches [7], [8], [9], [10], [11] have demonstrated improved generalization abilities. Another line of works [12], [13], [14], [15] explores learning the pseudo labels for target data to address the category-level shift. In this paradigm, the pseudo-label prediction has been extensively studied [16], [17] as it determines the domain shift aiming to learn within each category. Approaches leveraging fine-grained alignment for category-level domain shift learning often demonstrate superior generalization compared to those focused solely on domain-level shifts

Despite their effectiveness, existing approaches fall short of achieving satisfactory generalization results due to the inherent variability present in the target domain of daily-food images. Compared to existing food datasets, our DailyFood series exhibit a notable characteristic of significant intra-class variance, leading to the challenge of "category ambiguity". As illustrated in the last image of the second row in 1, the assigned category label "Braised Tofu" can be ambiguous to distinguish from another class "Shredded cabbage" due to some shared ingredients. Previous methods [13], [14], [17] have primarily focused on learning single-category cross-domain alignment, which is considered a sub-optimal solution for our datasets. In light of this, we propose a simple yet effective baseline method called Multi-Cluster Reference Learning (MCRL). Our MCRL relaxes the restriction that image features from the target domain should be pulled closer to one specific source cluster while being distant from other clusters in the source domain. Specifically, MCRL first predicts pseudo-labels for each sample in the target domain dataset and groups them into distinct semantic clusters following previous methods [17], [18], [19]. Subsequently, MCRL encourages samples within these clusters to simultaneously learn distribution shifts towards multiple source cluster features dynamically. This dynamic learning enables the model to adaptively weigh the relevance of each target sample to multiple source clusters during training. Through this learning approach, MCRL not only addresses the aforementioned "category ambiguity" problem but also mitigates the adverse effects of imperfect pseudo-label predictions. Furthermore, it is noteworthy that MCRL can be seamlessly integrated with existing UDA methods, resulting in significant performance improvements.

We summarize our contributions as following three folds:

1) We introduce two new high-quality daily-life food recognition benchmarks DailyFood-172 and DailyFood-16 to unlock the potential of transferring food recognition models trained on well-curated canteen food datasets to the daily life scenario.

2) We propose a simple yet effective baseline method named Multi-Cluster Reference Learning (MCRL) for our DailyFood-172 and DailyFood-16 datasets. Within our MCRL, multiple source clusters are dynamically referred to comprehensively learn the domain gap.

3) We extensively evaluate the transferability of existing approaches on our proposed DailyFood-172 and DailyFood-16. Besides, we couple our proposed MCRL with several state-of-the-art UDA approaches, and comprehensive performance enhancements prove the effectiveness of our proposed method.

II. RELATED WORK

Dataset for Food Recognition. Over the years, the research community has developed several high-quality food datasets. For instance, the ETH Food-101 dataset compiled by [5] encompasses a collection of 101,000 images distributed across 101 Western food categories. VIREO Food-172 [4] consists of 110,241 images from 172 Chinese food categories. FoodX-251 [20] has also been frequently utilized for classification tasks. In addition, there are other food-relevant large recipe datasets, such as Food2K [1] and Recipe1M [21]. Food2K [1] and Recipe1M [21] are representative of large-scale food-related datasets but with two important differences. First, the primary purpose of Recipe1M is to facilitate cross-modal embedding and retrieval between recipes and images, whereas Food2K is tailored towards propelling scalable food visual feature learning. Secondly, Recipe1M predominantly encompasses over 1 million structured cooking recipes, each linked to a series of food images. Conversely, Food2K is an aggregation of more than 1 million images, falling under 2,000 food categories. Based on these datasets, many works related to food have emerged as a topic of significant interest. These efforts include, but are not limited to, food quantity estimation based on depth images [22], image segmentation for volume estimation [23], multi-food recognition [24], and food image recognition [25], [26]. Nonetheless, the majority of existing datasets for food recognition are either composed of restaurant food photos or collected from cooking enthusiast websites. Most of these images depict food that has undergone relatively professional cooking, plating, and photography, which significantly differs from the appearance of food images captured in daily life primarily for documenting meals. To evaluate the performance of food recognition models in real-life scenarios, we constructed two food datasets, namely DailyFood-172 and DailyFood-16, which closely resemble everyday usage scenarios.

Unsupervised Domain Adaptation. While food recognition has recently captured numerous research attention, cross-domain food recognition has been less studied. To the best of our knowledge, there is only one work that investigates the problem of cross-domain food computing in the literature [27]. This work proposes a cross-domain cross-modal food transfer network that employs adversarial learning to reduce the gap between source and target domains for recipe retrieval tasks.
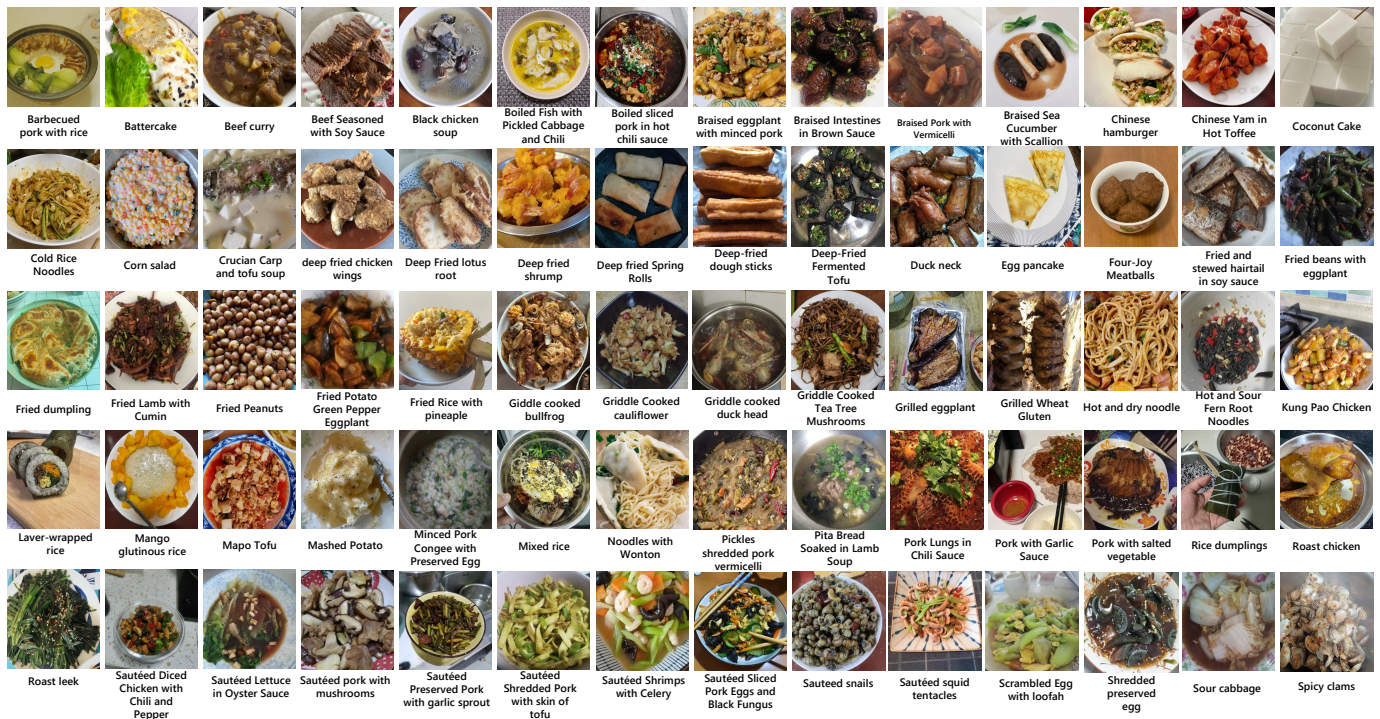
Fig. 2: Examples of food categories in DailyFood-172.

As our work focuses on unsupervised cross-domain adaptation for food recognition, therefore, we review recent work on unsupervised cross-domain adaptation.

Current unsupervised domain adaption (UDA) methods basically make alignment between source and target domain either at domain-level [7], [28], [10], [29] or in category level. Domain-level UDA minimizes the distribution divergence between the source and target domain by pulling them into the same distribution at different scale levels. The majority of the approaches rely on discrepancy minimization and adversarial training. The former minimizes the discrepancy between domains using a statistical distance function. With the introduction of MMD (Maximum Mean Discrepancy) [30], [8], [31] used MMD to learn more transferable features for domain adaptation. In the context of adversarial training, supervision is provided by a learned domain discriminator in a GAN [32] framework to encourage domain-invariant features [33], [34], [35], [28]. In the past few years, several works focus on the fine-grained category-level label distribution alignment [16], [36], [15], [37]. Compared to coarse-grained alignment at the domain level, such kind of methods attempt to align each category distribution between the source and target domain, by using the pseudo label [38] as a guide, to improve the classification accuracy. Generally, previous methods for category-level alignment involve aligning the conditional distribution [15], [39], [40]. To increase the robustness and accuracy of pseudo labels, [41] utilizes the prediction variance of two classifiers as an estimation of pseudo label uncertainty, prioritizing samples with low prediction variances for increased reliability. [16] uses a spanning tree to delete low-confidence samples to train the model. However, the aforementioned methods do not adequately adapt classification models trained on the fine-dining food dataset to everyday food images. Therefore, we propose a baseline method MCRL to address this issue.

III. DATASET CONSTRUCTION

We construct two curated datasets DailyFood-172 and DailyFood-16[3], which are made publicly available, both of which consist of food images captured from daily life dishes.

A. DailyFood-172

Since the UDA problem requires the homogeneity of categories of the source and the target domain, our newly constructed dataset, DailyFood-172, has categories that perfectly align with the VIREO Food-172 dataset, which consists of the 172 most common Chinese dishes. The images in the DailyFood-172 dataset are crawled from the "Go Cooking" [4] website. On this website, each dish category has several corresponding recipes, each with a unique ingredient list and cooking procedure. Users have uploaded numerous images for each recipe. Considering that variations in ingredient composition and cooking procedures can lead to differences in food appearance, we carefully select images from every recipe for each dish category to ensure the diversity of our curated dataset. Subsequently, we resize each image to a resolution of 280×280 pixels. Following this selection process, our DailyFood-172 dataset now contains 42,312 images, with minimum 47 and maximum 329 samples per category. Fig.2 showcases some examples of food categories included in DailyFood-172. For a detailed distribution of the sample count per category, please refer to the DailyFood-172 GitHub repository[3]. On average, there are 246 samples per category.

---

[3] https://github.com/SLKAlgs/DailyFood-172
[4] https://www.xiachufang.com/category/

B. DailyFood-16

Different from DailyFood-172, DailyFood-16 is collected from the diet recording platform "eggplant health" which is specialized for users who want to lose weight. On this platform, users are required to upload pictures of the food they consume each meal. These images are used for analysis by nutritionists to provide advice. Therefore, the photos on this platform serve as a means of fulfilling the task rather than being intended for sharing purposes. As a result, these pictures often do not prioritize aesthetic appeal. We construct DailyFood-16 following the subsequent steps. First, since the uploaded images only have ingredient descriptions rather than category labels, we manually label a subset of images with the predefined labels in VIREO Food-172 [4] according to ingredient overlappings. For instance, if the user's descriptions include "spinach", we will annotate it as "spinach" related dishes, e.g., "stir-fried spinach". However, such an annotating manner will lead to the one-to-many problem, i.e., a certain ingredient can correspond to multiple dishes. Toward this end, we require multiple annotators to manually assign the most likely category label from multiple possible choices for each image according to their consensus and select the label obtaining the most votes. Lastly, considering that the uploaded images often include cluttered backgrounds, which will disturb the model's attention, we crop the food-relevant part to guarantee the quality of our dataset. With the above steps, our DailyFood-16 includes 1,695 images in total with minimum 20 and maximum 413 samples per category and covers 16 food categories in VIREO Food-172. In DailyFood-16, each food category approximately contains 106 samples on average. Due to the irregularity of image cropping, the sizes of images in the DailyFood-16 dataset are different. Fig.3 shows some examples of food categories in DailyFood-16 and Fig.4 presents the distribution of sample numbers in each category.

It is worth mentioning that although the ingredients can provide more fine-grained cues for food recognition and such resources are also available in both VIREO Food-172 and our DailyFood-16, we do not formulate our DailyFood-16 as an ingredient-level UDA benchmark because the ingredients from DailyFood-16 exhibit high variety and randomness, exceeding the ranges of those in the VIREO Food-172, which deviates the traditional UDA problem. Therefore, we only use the category labels as ground truths in our DailyFood-16 benchmark.

IV. METHOD

In this section, we will first preliminarily introduce the problem definition of Unsupervised Domain Adaptation (UDA) and the basic formulation of the majority of the existing solutions. Then, we will elaborate on the motivation as well as detailed designs of our proposed Multi-Cluster Reference Learning (MCRL) method.

A. Preliminary

In Unsupervised Domain Adaptation (UDA), a source domain $\mathcal{D}_s = \{x_i^s, y_i^s\}_{i=1}^{n_s}$ of n_s labeled examples and a target domain $\mathcal{D}_t = \{x_j^t\}_{j=1}^{n_t}$ of n_t unlabeled examples are generally defined. y^s are labels of samples in the source domain. The
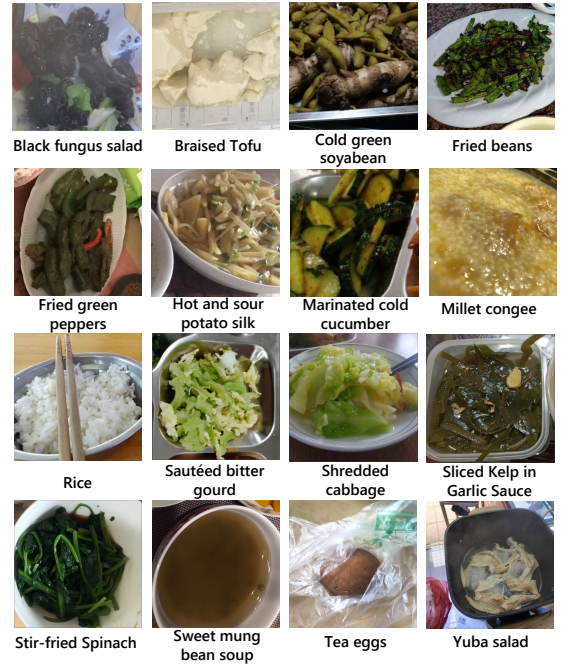

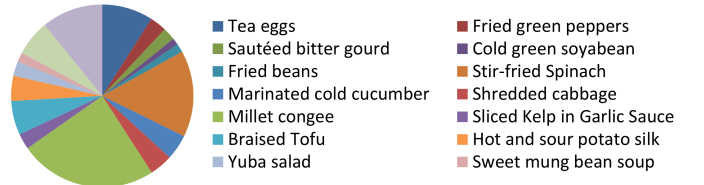Fig. 3: Examples of food categories in DailyFood-16.

Fig. 4: Pie distribution chart of categories in DailyFood-16.

purpose of UDA is to transfer the model trained on source domain $\mathcal{D}_s$ to the target domain $\mathcal{D}_t$ without catastrophic performance loss by mitigating the domain shift between source and target domains. As summarized in Section II, existing leading methods, i.e., category-based UDA solutions, focus on obtaining a good alignment of the source and target domains in a class-wise manner. Specifically, a feature extractor g_θ and a classifier f_θ can be trained on the source domain using labeled data as:

$$f_i = g_\theta(x_i)$$
$$L_y = -\frac{1}{n_s} \sum_{x_i \in \mathcal{D}_s} \sum_{c=1}^{C} P_{x_i \to c} \log f_\theta(f_i), \tag{1}$$

where C is the number of classes, $P_{x_i \to c}$ is the probability of x_i belonging to class c. Afterward, f_θ is then used to generate pseudo labels $\hat{y}^t$ for target domain samples $\mathcal{D}_t = [x_1^t, ..., x_{n_t}^t]$ as:

$$\hat{y}^t = \underset{c}{argmax}(\text{Softmax}(f_\theta(f_i))), \tag{2}$$

we define the features of source samples $X_s = \{x_i^s\}_{i=1}^{n_s}$ as $F_s = \{f_i^s\}_{i=1}^{n_s}$, Subsequently, the labels $Y_s = \{y_i^s\}_{i=1}^{n_s}$ are employed to partition the feature f_i^S, yielding class-specific clusters:

$$C_i^{(c)} = \text{MERGE}(F_s^{(c)}), \tag{3}$$

and for each sample x_i^s with the corresponding ground truth y_i^s, the index operator $[y_i^s]$ can be used to extract the feature
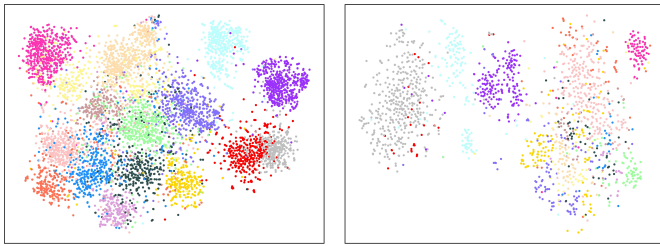
(a) Feature distributions of images in VIREO Food-172.   (b) Feature distributions of images in DailyFood-16.

Fig. 5: (Best viewed in color) The t-SNE visualization of features in different datasets. (a) and (b) respectively represent the feature distributions extracted from the VIREO Food-172 [4] and DailyFood-16 dataset, using the ResNet-50 [?] as backbone. (Colors identical are utilized for the same categories.)

of the corresponding category. Then, the features of the same category c are merged to obtain the cluster. The cluster for each category is represented as $F_s^{(c)}$:

$$F_s^{(c)} = \{(f_i^s, y_i^s), y_i^s = c\}, \qquad (4)$$

here, f_i^s indicates the feature of the ith sample in its corresponding category c, while $F_s^{(c)}$ denotes the features corresponding to category c. Then, the shift between the target domain and its corresponding category within the source domain is minimized by learning the domain-invariant features between the target samples and the source domain clusters. This process aims to achieve maximal alignment between the source and target domains within the feature space. Numerous methods exist for the computation of distances between two distributions, including but not limited to Cosine Distance, Kullback-Leibler Divergence, and Maximum Mean Discrepancy(MMD) [42]. Formally, MMD defines the following difference measure based on category:

$$d_{\mathcal{H}}(p, q) \triangleq \mathbf{E}_c \left\| \mathbf{E}_{p^{(c)}} \phi(x^s) - \mathbf{E}_{q^{(c)}} \phi(x^t) \right\|_{\mathcal{H}}^2, \qquad (5)$$

where x^s and x^t are the instances in $\mathcal{D}_s$ and $\mathcal{D}_t$, and $p^{(c)}$ and $q^{(c)}$ are the distributions of $\mathcal{D}_s^{(c)}$ and $\mathcal{D}_t^{(c)}$ respectively. By minimizing Equation (4) in deep networks, the distributions of the same category are drawn close.

B. Multi-Cluster Reference Learning

As illustrated in Fig 5, taking the t-SNE visualizations of VIREO Food-172 (Fig. 5(a)) with the same categories in DailyFood-16 (Fig. 5(b)) dataset as examples, compared to the standardized features in VIREO Food-172 [4], the data in the DailyFood-16 dataset shows a state of large intra-class feature variations and unclear inter-class feature differences, which leads to the problem of "category ambiguity" mentioned in Section I. To solve this issue, we propose the Multi-Cluster Reference Learning (MCRL) framework, as shown in Fig. 6. It simultaneously learns the domain gap between a certain target sample and multiple source clusters.

In the proposed MCRL, the features obtained from the source data through g_θ are used for clustering, where samples of the same category are gathered around the same clusters. According to Equation (2), the category corresponding to

the highest probability following the Softmax operation of the classifier f_θ has traditionally been used as the pseudo label for the sample. However, due to the ambiguity and diversity of ingredients among food images, we choose K categories with the highest probabilities as the pseudo labels for each target sample. Consequently, each target sample can correspond to multiple category clusters for reference. Finally, the distribution shifts between the target domain samples and several source domain clusters are minimized during training. We employed two strategies when deciding the target clusters, i.e., MCRL with hard and soft selection, which will be elaborated on in the following parts.

1) MCRL with Hard Selection: Compared to Equation (2), we now aim to obtain the K categories with the highest probability after the Softmax function. First, we calculate the probability of each target sample $P = \{p_1, ..., p_C\}$ as:

$$P = \text{Softmax}(f_\theta(f_i^T)), \qquad (6)$$

where f_i^T denotes the features obtained from the target domain samples through the feature extractor g_θ. Then we select the K highest probability of P and then obtain their corresponding pseudo labels. To align the distribution of the target samples with cluster features $F_s^{(K)}$ of these K categories in the source domain, we compute the distribution discrepancy between the target domain samples and the source domain clusters as learning targets:

$$d_{\mathcal{H}}(p, q) \triangleq \frac{1}{K} \sum_{k=1}^{K} \left\| F_s^{(k)} - \frac{1}{|\mathcal{D}_t|} \sum_{x_j^t \in \mathcal{D}_t} \phi(x_j^t) \right\|_{\mathcal{H}}^2, \qquad (7)$$

where $F_s^{(K)}$ are the feature clusters of the $\text{Top} - K$ categories corresponding to each target sample in the source domain and $\phi(x_j^t)$ is the distribution of target sample x_i^t. In this way, for each target domain sample, we can simultaneously reduce its domain shifts from the multiple source clusters that correspond to its categories. For each $\text{Top} - K$ category, we utilize identical weights to reduce the discrepancy between domains to learn distribution shifts.

2) MCRL with Soft Selection: Considering the varying resemblance between individual target samples and source categories, it is reasonable to encourage the model to adaptively control the magnitude of domain shift learning with different source category clusters. Therefore, we also devise a dynamic scoring mechanism that enables the model to learn weights for controlling the domain shift learning process. To avoid the adverse effect of introducing too much noise when setting the learning target, we posit that a single image most likely correlates with the features of a maximum of K categories. Consequently, we initially procure the probability $\hat{Y}^t$ of the $\text{Top} - K$ categories by leveraging equations (6). Inspired by the fact that the sigmoid function is more suitable for multi-label classification tasks [43], we generate K controlling weights with sigmoid functions as:

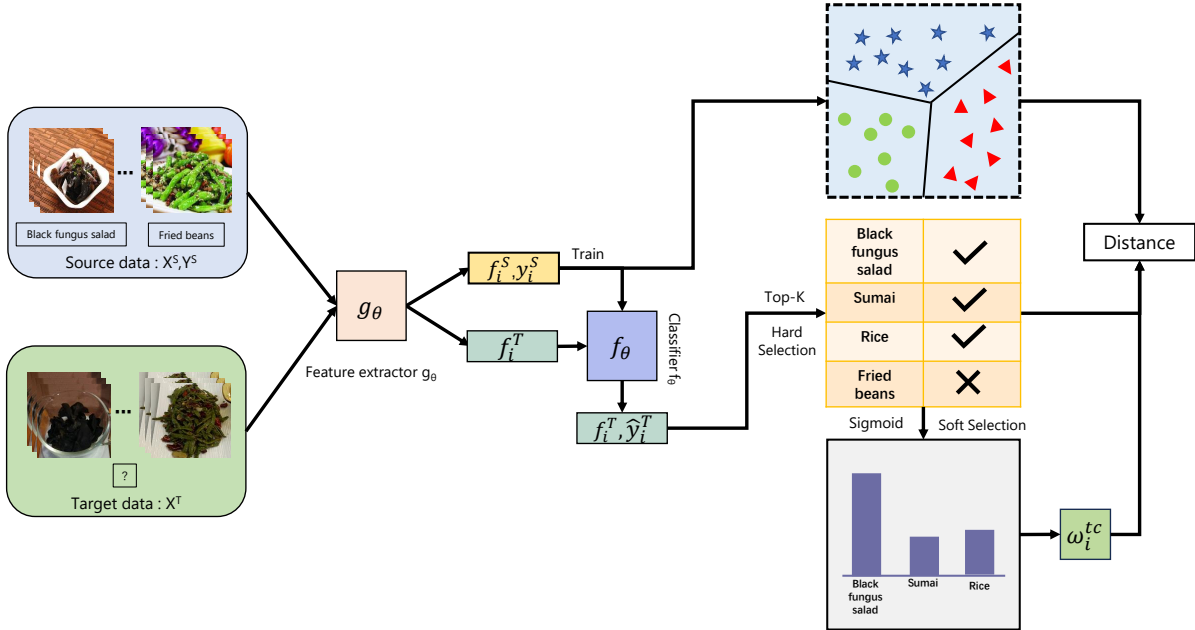$$w_j^{tc} = \text{Sigmoid}(f_\theta(f_i^T)), w_j^c = w_j^1, ..., w_j^K, \qquad (8)$$

Fig. 6: **The architecture of the proposed Multi-Cluster Reference Learning.** It consists of a feature extractor g_θ, a classifier f_θ, as well as two multi-cluster reference learning objectives. Within the two learning objectives, the distribution gap between target samples and multiple categories of source clusters is narrowed for pursuing better generalization ability.

where w_j^{tc} denotes the individual probabilities of the target sample x_j^t among c categories, and $c \leq K$. Accordingly, equation (7) is modified as follows:

$$d_{\mathcal{H}}(p,q) \triangleq \frac{1}{K} \sum_{k=1}^{K} \left\| F_s^{(k)} - \frac{1}{|\mathcal{D}_t|} \sum_{x_j^t \in \mathcal{D}_t} w_j^{tc} \phi(x_j^t) \right\|_{\mathcal{H}}^2, \quad (9)$$

where w_i^{sc} denotes the weights of each sample within the K selected categories corresponding to the target samples in the source domain, and w_j^{tc} denotes the weight of x_j^t belonging to class c. Hard selection, assigning equal weights to each sample, is a specific case of soft selection, which accounts for the uncertainty of target samples.

In summary, by incorporating the controlling scores of the K selected categories as weights into the original category-based UDA method, our MCRL method can dynamically learn and enable the model to adaptively balance the relevance of each target sample to multiple source clusters during the training process. With this learning approach, MCRL not only solves the previously mentioned "category ambiguity" problem but also alleviates the negative impacts brought by imprecise pseudo-label prediction.

V. EXPERIMENT

A. Implementation Details

Unlike typical UDA datasets like Office-Home [44] and VisDA [45] with multiple domains, we use the VIREO Food-172 training set [4] as the source domain and evaluate on two target datasets DailyFood-172 and DailyFood-16.

We implement all deep models using the PyTorch framework and utilize ImageNet [48] pre-trained ResNet-50,

TABLE I: Performance comparison. "Source-only" denotes that the classifiers are pre-trained on VIREO Food-172 without using any UDA approaches.

Backbone	Methods	DailyFood-172(%)	DailyFood-16(%)
ResNet50	Source-only	66.62	74.34
	DSAN[13]	66.96	82.24
	DAAN[14]	68.26	79.06
	DANN[28]	67.41	79.47
	DeepCoral[9]	66.36	75.04
	CoVi[46]	73.82	85.25
ResNet101	Source-only	67.48	78.41
	DSAN[13]	68.04	81.53
	DAAN[14]	68.48	80.00
	DANN[28]	67.49	84.37
	DeepCoral[9]	67.14	78.11
ResNet152	Source-only	67.95	78.88
	DSAN[13]	68.25	80.94
	DAAN[14]	69.22	81.24
	DANN[28]	67.67	79.82
	DeepCoral[9]	67.89	74.16
ViT	CDTrans[47]	81.93	90.86
	CDTrans+MCRL (Ours)	**82.10**	**93.63**

ResNet-101, and ResNet-152 models [?] as backbone networks. Note that the labels for the target domain are solely used for evaluation purposes. To validate the effectiveness of our method, we integrate it into three algorithms: Deep Subdomain Adaptation Network (DSAN) [13], Dynamic Adversarial Adaptation Network (DAAN) [14], and Cross-Domain Transformer (CDTrans) [47]. For DSAN and DAAN, we employ mini-batch stochastic gradient descent (SGD) with a momentum of 0.9 and a learning rate of 0.01. Throughout the experiments, we fix the epoch at 20 and the batch size at 32. As for CDTrans, we utilize the Adam optimizer [49] with a momentum of 0.9 and a learning rate of 0.0003. Similarly, we keep the epoch at 40 and the batch size at 32. CDTrans specifically employ the Vision Transformer (ViT) [50] as its backbone. For initializing our model, we used backbones trained from Data-efficient Image Transformers (DeiT) [51],

namely DeiT-Small (DeiT-S) and DeiT-Base (DeiT-B). To ensure a fair comparison, we set the parameters of the extended methods to be the same as those of the original methods. The evaluation metric used is the classification accuracy in the target domain.

B. Performance of Multi-Cluster Reference Learning

To examine the effectiveness of our proposed datasets DailyFood-172 and DailyFood-16 when used as target domains, for image classification UDA, we combine our method with the state-of-the-art method CDtrans. To further illustrate the effectiveness of MCRL, we compared CDTrans+MCRL which used DeiT-B as the backbone with several conventional UDA methods, which are DSAN [13], DAAN [14], DANN [28], DeepCoral [9], CoVi [46], and CDTrans [47], respectively. To ensure a fair comparison, the results for the comparative methods were sourced from their public codes or directly cited from the original papers. The TABLE I lists the comparison results on the DailyFood-172 and DailyFood-16 datasets. It is not hard to observe that CDTrans+MCRL achieves the best performance in all tasks. This indicates that, as it currently stands, our method exhibits the best performance in terms of domain adaptation for food image recognition.

MCRL can be easily combined with many UDA algorithms. Taking into consideration that accuracy alone cannot provide a comprehensive evaluation of the model's performance, the macro-F1 score is employed to assess the overall balance between classification precision and recall of the model. TABLE II lists the comparative performances of the Top-1, Top-3 accuracies and Macro-F1 scores of DSAN [13], DAAN [14], CDTrans [47] (DeiT-S), and CDTrans (DeiT-B) before and after the extended of our method, with the DailyFood-172 and DailyFood-16 datasets serving as the target domains respectively, while VIREO Food-172 [4] is always the source domain. From the table, we can see that, after incorporating MCRL, the performance of DSAN, DAAN, and CDTrans has been significantly improved. To be specific, in terms of Top-1 accuracy, the majority of the baselines exhibited varying degrees of improvement after employing the MCRL. Specifically, the improvement rates for CDTrans using DeiT-S backbone, CDTrans with DeiT-B backbone, and DSAN with ResNet50 as the backbone on the DailyFood-16 dataset were 3.11%, 2.77%, and 2.18% respectively. Eventually, in the vast majority of cases, these three metrics witnessed varying degrees of improvement after the application of MCRL. This demonstrates that MCRL can be effectively integrated into existing UDA methods, not only achieving a good cross-domain distribution alignment on food-related cross-domain datasets but also perfectly resolving the issue of "category ambiguity" mentioned before.

C. Ablation Study

In this section, we present a comprehensive ablation study to evaluate the contribution of MCRL and investigate its influence on overall performance. By systematically analyzing and comparing the results, we aim to gain deeper insights into the key factors that contribute to the success of our proposed

approach. We compared three distinct implementations of multi-label methods on the DailyFood-16 dataset. All these methods can be seen as special cases of our most efficacious Multi-Cluster Reference Learning via soft-label strategy.

1) Ratio Adjustment Method(RAM): We extract the logits of the two highest-confidence categories for each target sample after f_θ and calculate their $ratio$. If the $ratio$ exceeds a predefined threshold, it signifies that the feature of the target sample is closer to the feature of the source domain from the same category with the highest confidence. In such scenarios, we only use the feature of the highest confidence category of the source domain, minimizing the distance between it and the feature distribution of the target sample. Conversely, if the $ratio$ is below the specified threshold, it implies that the target sample's feature closely resembles the features of the two categories with the highest confidence. In such cases, we utilize the features of the top two categories, reducing the distance between their feature distributions and the feature distribution of the target sample. This design rationale acknowledges the fact that a single image of food may contain multiple potential primary ingredients, which could originate from 2 dishes. Consequently, we assign up to 2 possible categories to each food image to adaptively align with the source domain.

2) Hard-Selection Method(HM): Considering that a single food image may closely align with the features of more than two categories of food, for instance, a dish containing cucumber, tofu skin, and black fungus, it is challenging to determine whether it belongs to "Marinated cold cucumber", "Yuba salad", or "Black fungus salad". In such cases, we aim to draw the feature distribution of this image closer to the feature distributions of all three categories in the source domain. Therefore, we set different values for n, where n represents how many classes of source domain features each target domain sample needs to align in the feature space. In this case, we give the same weight to n classes.

3) Soft-Selection Method(SM): We adopt SM, which uses MCRL to align the source and target domains, wherein each class of samples is assigned varying weights based on classification confidence to approximate the source domain.

The results on each dataset in TABLE III indicate that it is not enough to align them by using PAM, or aligning them with equal weights. Therefore, the proposed MCRL is able to perform dynamic distribution alignment between domains and achieve better performance on food recognition. This property of MCRL is extremely important in real applications since given an unknown target domain, we can never know the specific number of classes in the source domain whose features more closely resemble the target sample. MCRL makes it possible to easily, dynamically, and quantitatively evaluate their relative importance in everyday food recognition.

D. Qualitative Results

Food classification is a complex task, particularly for data wherein the distribution of source and target domain samples diverges, as it relies heavily on the model's capability to extract the correct features corresponding to food categories. Fig. 7 presents examples of correct and incorrect classifications by our model when the target domain is the DailyFood-16

TABLE II: Performance comparison on target datasets DailyFood-172 and DailyFood-16.

Backbone	Methods	DailyFood-172			DailyFood-16		
		Top-1(%)	Top-3(%)	Macro-F1(%)	Top-1(%)	Top-3(%)	Macro-F1(%)
ResNet50	DSAN[13]	66.96	83.43	65.35	82.24	95.75	71.90
	+MCRL(Ours)	**68.68**	**84.42**	**67.48**	**84.42**	**96.28**	**74.49**
	DAAN[14]	68.26	84.21	67.48	79.06	94.57	65.63
	+MCRL(Ours)	**68.84**	**84.83**	**68.03**	**80.65**	**95.46**	**67.06**
ResNet101	DSAN[13]	68.04	84.51	66.38	81.53	95.99	72.55
	+MCRL(Ours)	**69.31**	**84.84**	**68.36**	81.47	95.69	72.21
	DAAN[14]	68.48	84.54	67.51	80.00	95.75	72.05
	+MCRL(Ours)	**68.66**	**84.79**	**67.82**	**81.88**	**96.76**	**73.64**
ResNet152	DSAN[13]	68.25	84.48	66.34	80.94	94.40	72.35
	+MCRL(Ours)	**69.64**	**85.47**	**68.69**	**82.30**	**95.51**	**72.72**
	DAAN[14]	69.22	84.88	68.40	81.24	94.93	70.20
	+MCRL(Ours)	**69.44**	**85.01**	**68.55**	81.17	**95.10**	**71.18**
ViT	CDTrans(DeiT-S)[47]	79.00	90.52	78.31	90.09	96.34	82.97
	+MCRL(Ours)	79.00	**91.00**	77.95	**93.20**	**97.94**	**85.38**
	CDTrans(DeiT-B)[47]	81.93	92.55	81.43	90.86	97.35	83.35
	+MCRL(Ours)	**82.10**	**93.03**	**81.46**	**93.63**	**98.47**	**88.38**

TABLE III: Ablation study of multi-label on DailyFood-16.

Methods		DSAN[13]	CDTrans[47]
Baselines		82.24	90.86
RAM	$ratio = 1.1$	83.01	93.22
	$ratio = 1.2$	82.48	92.92
	$ratio = 1.5$	82.83	93.2
HM	$k = 2$	83.95	93.3
	$k = 3$	83.89	93.6
	$k = 4$	82.77	90.9
SM	$k = 2$	83.07	93.27
	$k = 3$	**84.42**	**93.63**
	$k = 4$	83.01	93.51

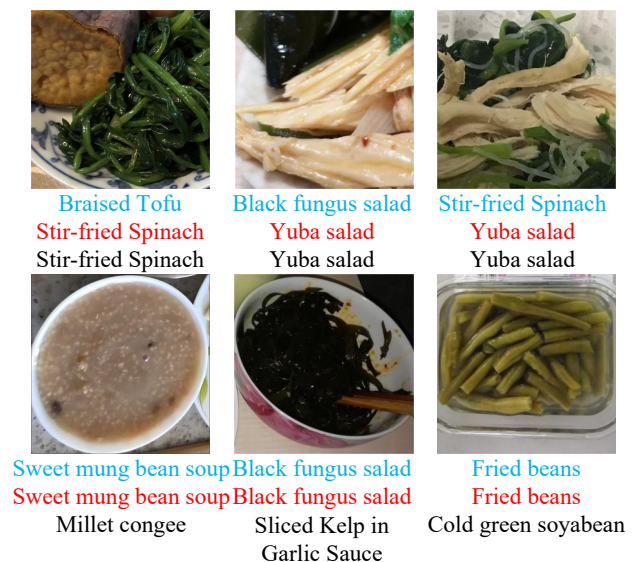|		
Braised Tofu	Black fungus salad	Stir-fried Spinach
Stir-fried Spinach	Yuba salad	Yuba salad
Stir-fried Spinach	Yuba salad	Yuba salad
Sweet mung bean soup	Black fungus salad	Fried beans
Sweet mung bean soup	Black fungus salad	Fried beans
Millet congee	Sliced Kelp in Garlic Sauce	Cold green soyabean

Fig. 7: Qualitative examples on DailyFood-16 dataset. For each image, the three lines of captions below indicate the predictions of CDTrans (blue), the predictions of MCRL (red), and the ground truth (black), respectively.

dataset. For the successful samples, we show SOTA method CDTrans [47] fail but our MCRL could make it. As shown in Fig. 7, images in the top row are misclassified by the SOTA method CDTrans but correctly classified by our method MCRL. In the second row of the Fig. 7, we present the samples where MCRL made incorrect predictions. These images often pose challenges in accurate classification due to issues such as dim lighting, cluttered ingredients, and subtle appearance differences, making them difficult to be clearly distinguished even by human vision. In summary, our model may misclassify images where the inherent features are not clear, or images that contain distinctive characteristics of other food categories. However, these errors usually result from the nature of the images themselves, as even human vision would find it challenging to accurately distinguish these pictures.

E. Discussion

While previous UDA methods adopt an end-to-end training strategy that simultaneously train the classifier and perform domain adaptation, we use a two-stage process here. Specifically, we first train the classifier and subsequently conduct domain adaptation, which we refer to as the "2-stage" approach. The results are shown in Tab. IV. It indicates that the pseudo-labels assigned to target domain samples, acquired post-training the model on the source domain, are more accurate. Hence, category-level Unsupervised Domain Adaptation (UDA) methods can extract more robust features, enhancing the classification of food images.

In addition, considering the similarity among images from the same class in our three datasets: VIREO Food-172, DailyFood-172, and DailyFood-16, we leverage an alternate

target domain as an intermediate domain for training on DailyFood-172 and DailyFood-16 datasets to enlarge the dataset in order to enhance the generalization capability of the model. Specifically, VIREO Food-172 consistently serves as the source domain. Initially, we train using DailyFood-16 as the target domain and save the model weights, then switch the target domain to DailyFood-172, and re-train using the pre-trained model to obtain the final results. The experiment results are presented in the "intermediate domain" section of TABLE IV. The results indicate that due to the similarities between images of the same categories of food, the neural network was able to learn more features, demonstrating the effective mutual assistance between the two target domains. It's worth noting that by employing the MCRL method, we have surpassed the baseline in all three settings, thus positioning ourselves as the leading-edge approach. To detail further, ResNet50 was employed as the backbone in all the aforementioned experiments and Deit-S is used in CDTrans.

TABLE IV: Experimental results of the 2-stage and Enlarged Dataset settings.

Methods	DailyFood-172(%)
DSAN[13]	66.96
DSAN-2-stage[13]	68.42
DSAN-Enlarged Dataset[13]	67.26
DSAN+MCRL[13]	68.80
DSAN-2-stage+MCRL[13]	**69.51**
DSAN-Enlarged Dataset+MCRL[13]	68.57
DAAN[14]	68.26
DAAN-2-stage[14]	68.78
DAAN-Enlarged Dataset[14]	68.29
DAAN+MCRL[14]	68.84
DAAN-2-stage+MCRL[14]	**69.31**
DAAN-Enlarged Dataset+MCRL[14]	68.70
CDTrans[47]	78.98
CDTrans-Enlarged Dataset[47]	78.94
CDTrans+MCRL[47]	**79.00**

VI. CONCLUSIONS

In this paper, we presented two novel food datasets DailyFood-172 and DailyFood-16, along with a proposed baseline method called Multi-Cluster Reference Learning (MCRL). The DailyFood-172 dataset is collected from regular individuals who uploaded their home-cooked dishes to a website following recipes, whereas the DailyFood-16 dataset comes from daily meals uploaded to a website by people on a weight loss regime, acquired through searches and cropping of the images. And the proposed MCRL learns the domain shift between a certain target sample and multiple source clusters. Experimental results on extensive visual cross-domain tasks have shown that applying our method to conventional UDA methods can improve the classification accuracy of the target domain and further lead to more discriminative and domain invariant features than the conventional UDA baselines.

REFERENCES

[1] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, and S. Jiang, "Large scale visual food recognition," *CoRR*, vol. abs/2103.16107, 2021.

[2] Q. Thames, A. Karpur, W. Norris, F. Xia, L. Panait, T. Weyand, and J. Sim, "Nutrition5k: Towards automatic nutritional understanding of generic food," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8903–8911.

[3] Y. Liang, J. Li, Q. Zhao, W. Rao, C. Zhang, and C. Wang, "Image segmentation and recognition for multi-class chinese food," in *International Conference on Image Processing*, 2022, pp. 3938–3942.

[4] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *ACM Multimedia*, 2016, pp. 32–41.

[5] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *European Conference on Computer Vision*. Springer, 2014, pp. 446–461.

[6] J. Chen, B. Zhu, C.-W. Ngo, T.-S. Chua, and Y.-G. Jiang, "A study of multi-task and region-wise deep learning for food ingredient recognition," *IEEE Trans. on Image Processing*, vol. 30, pp. 1514–1526, 2020.

[7] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[8] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning*. PMLR, 2015, pp. 97–105.

[9] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *ECCV Workshop*. Springer, 2016, pp. 443–450.

[10] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," *Advances in neural information processing systems*, vol. 29, 2016.

[11] M. Peng, Z. Li, and X. Juan, "Similarity-based domain adaptation network," *Neurocomputing*, vol. 493, pp. 462–473, 2022.

[12] J. Huang, D. Guan, A. Xiao, S. Lu, and L. Shao, "Category contrast for unsupervised domain adaptation in visual tasks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1203–1214.

[13] Y. Zhu, F. Zhuang, J. Wang, G. Ke, J. Chen, J. Bian, H. Xiong, and Q. He, "Deep subdomain adaptation network for image classification," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 32, no. 4, pp. 1713–1722, 2020.

[14] C. Yu, J. Wang, Y. Chen, and M. Huang, "Transfer learning with dynamic adversarial adaptation network," in *IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 778–786.

[15] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4893–4902.

[16] J. Wang and X.-L. Zhang, "Improving pseudo labels with intra-class similarity for unsupervised domain adaptation," *Pattern Recognition*, vol. 138, p. 109379, 2023.

[17] Y. Du, Z. Tan, Q. Chen, X. Zhang, Y. Yao, and C. Wang, "Dual adversarial domain adaptation," *arXiv preprint arXiv:2001.00153*, 2020.

[18] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu, "Visual domain adaptation with manifold embedded distribution alignment," in *ACM Multimedia*, 2018, pp. 402–410.

[19] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *IEEE International Conference on Computer Vision*, 2013, pp. 2200–2207.

[20] P. Kaur, K. Sikka, W. Wang, S. Belongie, and A. Divakaran, "Foodx-251: a dataset for fine-grained food classification," *arXiv preprint arXiv:1907.06167*, 2019.

[21] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, "Learning cross-modal embeddings for cooking recipes and food images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3020–3028.

[22] M.-Y. Chen, Y.-H. Yang, C.-J. Ho, S.-H. Wang, S.-M. Liu, E. Chang, C.-H. Yeh, and M. Ouhyoung, "Automatic chinese food identification and quantity estimation," in *SIGGRAPH Asia 2012 Technical Briefs*, 2012, pp. 1–4.

[23] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney, "Recognition and volume estimation of food intake using a mobile device," in *Workshop on Applications of Computer Vision*, 2009, pp. 1–8.

[24] Y. Matsuda and K. Yanai, "Multiple-food recognition considering co-occurrence employing manifold ranking," in *International Conference on Pattern Recognition*. IEEE, 2012, pp. 2017–2020.

[25] A. Şengür, Y. Akbulut, and Ü. Budak, "Food image classification with deep features," in *International Artificial Intelligence and Data Processing Symposium (IDAP)*. Ieee, 2019, pp. 1–6.

[26] B. Arslan, S. Memiş, E. B. Sönmez, and O. Z. Batur, "Fine-grained food classification methods on the uec food-100 database," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 2, pp. 238–243, 2021.

[27] B. Zhu, C.-W. Ngo, and J.-j. Chen, "Cross-domain cross-modal food transfer," in *ACM Multimedia*, 2020, pp. 3762–3770.

[28] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[29] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[30] M. Ghifary, W. B. Kleijn, and M. Zhang, "Domain adaptive neural networks for object recognition," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2014, pp. 898–904.

[31] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2208–2217.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[33] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1180–1189.

[34] R. Gong, W. Li, Y. Chen, and L. V. Gool, "Dlow: Domain flow for adaptation and generalization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2477–2486.

[35] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International Conference on Machine Learning*. Pmlr, 2018, pp. 1989–1998.

[36] X. Zhang, F. X. Yu, S.-F. Chang, and S. Wang, "Deep transfer network: Unsupervised domain adaptation," *arXiv preprint arXiv:1503.00591*, 2015.

[37] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[38] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.

[39] X. Gu, J. Sun, and Z. Xu, "Spherical space domain adaptation with robust pseudo-label loss," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9101–9110.

[40] Y. Zhang, Y. Zhang, Y. Wei, K. Bai, Y. Song, and Q. Yang, "Fisher deep domain adaptation," in *International Conference on Data Mining*. SIAM, 2020, pp. 469–477.

[41] Z. Zheng and Y. Yang, "Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1106–1120, 2021.

[42] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[43] Z. Qin, D. Kim, and T. Gedeon, "Rethinking softmax with cross-entropy: Neural network classifier as mutual information estimator," *arXiv preprint arXiv:1911.10688*, 2019.

[44] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5018–5027.

[45] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko, "Visda: The visual domain adaptation challenge," *arXiv preprint arXiv:1710.06924*, 2017.

[46] J. Na, D. Han, H. J. Chang, and W. Hwang, "Contrastive vicinal space for unsupervised domain adaptation," in *European Conference on Computer Vision*. Springer, 2022, pp. 92–110.

[47] T. Xu, W. Chen, P. Wang, F. Wang, H. Li, and R. Jin, "Cdtrans: Cross-domain transformer for unsupervised domain adaptation," *arXiv preprint arXiv:2109.06165*, 2021.

[48] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[50] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[51] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

Yang Jiao received the B.E. degree from University of Electronic Science and Technology of China, Chengdu, China, in 2021. He is currently pursuing his Ph.D. degree in Computer Science at Fudan University. His research interests include multi-media analysis and 3D vision.

Jingjing Chen is now an Associate Professor at the School of Computer Science, Fudan University. Before joining Fudan, she was a postdoc research fellow at the School of Computing in the National University of Singapore. She received her Ph.D. degree in Computer Science from the City University of Hong Kong in 2018. Her research interest lies in diet tracking and nutrition estimation based on multi-modal processing of food images, including food recognition, cross-modal recipe retrieval.

Bin Zhu received the B.E. degree from Southeast University, Nanjing, China, the M.E. degree from Zhejiang University, Hangzhou, China, and Ph.D. degree from City University of Hong Kong, Hong Kong, China. He is currently an Assistant Professor of Computer Science in the School of Computing and Information Systems , Singapore Management University (SMU). Prior to SMU, He was a postdoctoral researcher in department of computer science, University of Bristol, United Kingdom. His research interests mainly lie in cross-modal recipe retrieval, domain adaptation and egocentric video understanding. Dr. Zhu is the co-organizer of EPIC KITCHENS100 CHALLENGES and co-organizer of Special Session on Multimedia on Cooking and Eating Activities hosted in ACM Multimedia Asia 2023.

Guoshan Liu received the B.E. degree in Computer Science from Jilin University, Changchun, China, in 2022. She is currently pursuing her M.S. degree in Computer Science at Fudan University. Her research interests include food recognition and domain adaptation.

Yu-Gang Jiang received the Ph.D. degree in Computer Science from City University of Hong Kong in 2009 and worked as a Postdoctoral Research Scientist at Columbia University, New York, during 2009-2011. He is currently a Professor of Computer Science at Fudan University, Shanghai, China. He has been elected as an IEEE Fellow in 2023. His research lies in the areas of multimedia, computer vision, and robust and trustworthy AI. His work has led to many awards, including the inaugural ACM China Rising Star Award, the 2015 ACM SIGMM Rising Star Award, the Research Award for Excellent Young Scholars from NSF China, and the Chang Jiang Distinguished Professorship appointed by Ministry of Education of China.