Robust Visual Food Recognition for Enriching Nutrition Knowledge Bases

Zhaoyan Ming, Zeyu Xie, Chao Zhang, Kui Su, Changzheng Yuan, and Tat-Seng Chua

***Abstract*—Acquiring nutrition information and health-related knowledge about food is a common need among individuals. However, using conventional food names as search queries often fails to yield accurate matches to entries within food nutrition knowledge bases (FoodnKB), which frequently utilize scientific or product names. In this study, we present a method for enriching FoodnKB entries with imagery and facilitating visual access to food-related knowledge through image recognition. We start with an official food nutrition database and propose a consensus-based approach using Large Language Models to identify visually discernible and directly edible foods, expanding food synonyms and harnessing diverse web-based food images for comprehensive visual representation. To minimize manual annotation of noisy web images, we introduce a cyclic training-based area under the margin metric (cAUM) approach that effectively distinguishes appropriate images, including rare instances, from noisy ones. Additionally, we design a generic accuracy gap (AccGap) algorithm to automatically estimate the noise ratio of the web-harnessed data. Our integrated cAUM and AccGap method demonstrates superior performance in noise detection and enhancement of image recognition accuracy compared to existing noise-robust frameworks. Furthermore, we successfully apply the visually enriched FoodnKB and food recognition capabilities within a smart nutritionist mobile application.**

***Index Terms*—Image classification, knowledge representation, large language models.**

I. Introduction

In recent years, there has been a growing interest in health-conscious dietary habits [1], fueled by increased awareness of the links between nutrition and well-being [2]. Consequently, more and more people are seeking nutritional information and health-related knowledge about the food they consume. Food nutrition knowledge bases (FoodnKB) [3] serve as an essential resource to address this need, providing detailed information about the nutritional content of various food items. However, accessing the desired information from these knowledge bases is often hindered by the use of common food names as search queries, which may not correspond accurately to the scientific or product names included in FoodnKB entries.

One effective way to bridge this gap is to enhance FoodnKB entries by incorporating relevant images, which can provide an intuitive and accessible visual representation of food items [4]. Image recognition technologies can then be employed to facilitate access to food-related knowledge [5], enabling users to visually identify nutritional information associated with a specific food item. This approach can substantially broaden the applicability of FoodnKB in various food-related fields, such as diet planning, grocery shopping, and nutrition education.

Therefore, opting for a visually-driven methodology to access and utilize the food items and associated nutritional knowledge database can be profoundly beneficial. This approach will empower individuals to initiate a search using images of food items, a process that is not only more intuitive but also significantly more efficient than typing out exact food names to align with entries in the FoodnKB. As illustrated in Fig. 1, the integration of Food Images and an Image Classifier serves to enhance the existing FoodnKB by enabling visually-driven access to the enriched knowledge database of specific food items.

Collecting a diverse set of food images for FoodnKBs can be a significant challenge, particularly given the large number of entries typically involved [6]. The standard practice of displaying a single icon image per food item in commercial food nutrition apps is often insufficient for effective visual recognition algorithms, and harnessing the wealth of web images to provide food visual representations can present its own set of challenges. Quality and quantity form a dilemma in the context of web-harnessed food images—with the inclusion of a higher number of images to cater to food visual diversity, the quality may suffer, resulting in noisily labeled images [7]. Considering the vast scale of food items and the lengthy image lists for each, manual inspection for label and image correspondence is not tenable, calling for an automatic label quality evaluation of candidate images for FoodnKBs.

To mitigate these challenges, various techniques and models have been applied, including "Decoupling" [8] and

Received 1 August 2023; revised 22 September 2024 and 22 October 2024; accepted 22 October 2024. Date of publication 17 February 2025; date of current version 27 May 2025. This work was supported in part by the Natural Science Foundation of Zhejiang Province under Grant LQ23F020002, in part by the National Key Technologies Research and Development Program in China under Grant 2023YFF1104202-02, and in part by the advanced computing resources provided by the Supercomputing Center of Hangzhou City University. The guest editor coordinating the review of this article and approving it for publication was Weiqing Min. *(Corresponding authors: Kui Su; Changzheng Yuan.)*

Zhaoyan Ming is with the School of Computer and Computing, Hangzhou City University, Hangzhou 310015, China, and also with the Innovation Center of Yangtze River Delta, Zhejiang University, Hangzhou 310058, China (e-mail: mingzhaoyan@gmail.com).

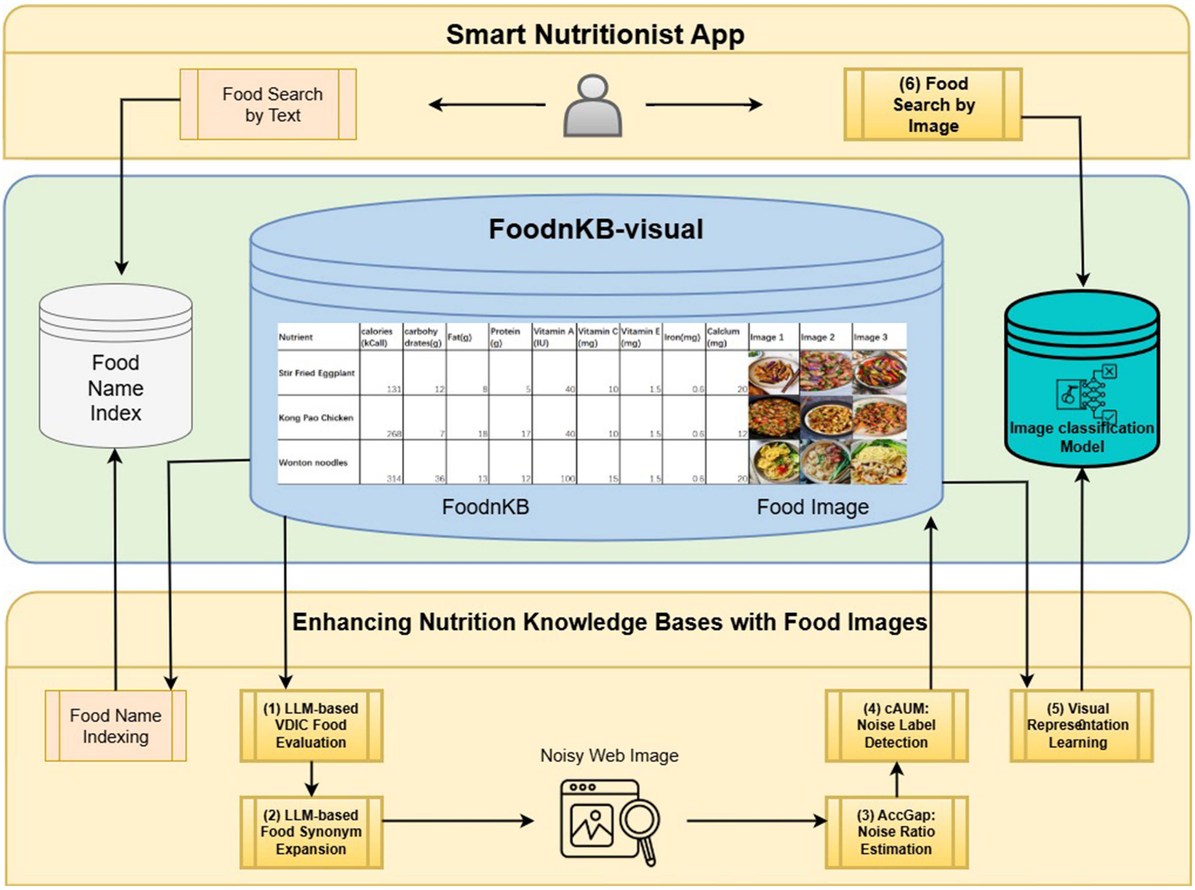
Zeyu Xie and Kui Su are with the School of Computer and Computing, Hangzhou City University, Hangzhou 310015, China (e-mail: xiezy@hzcu.edu.cn; suk@hzcu.edu.cn).

Chao Zhang is with Beijing Zuoyi Technology Company Ltd., Beijing 100099, China (e-mail: 1226.dahai@163.com).

Changzheng Yuan is with the School of Public Health, Zhejiang University School of Medicine, Hangzhou 310058, China (e-mail: chy478@zju.edu.cn).

Tat-Seng Chua is with the School of Computing, National University of Singapore, Singapore 119077 (e-mail: dcscts@nus.edu.sg.).

Digital Object Identifier 10.1109/TMM.2025.3542962


Smart Nutritionist App
Food Search by Text
(6) Food Search by Image
FoodnKB-visual
Food Name Index
Image classification Model
FoodnKB
Food Image
Enhancing Nutrition Knowledge Bases with Food Images
Food Name Indexing
(1) LLM-based VDIC Food Evaluation
(2) LLM-based Food Synonym Expansion
Noisy Web Image
(3) AccGap: Noise Ratio Estimation
(4) cAUM: Noise Label Detection
(5) Visual Representation Learning

Nutrient	calories (kCall)	carbohy drates(g)	Fat(g)	Protein (g)	Vitamin A (IU)	Vitamin C (mg)	Vitamin E (mg)	Iron(mg)	Calcium (mg)	Image 1	Image 2	Image 3
Stir Fried Eggplant	131	12	8	5	40	10	1.5	0.6	20			
Kong Pao Chicken	268	7	18	17	40	10	1.5	0.6	12			
Wonton noodles	314	36	13	12	100	15	1.5	0.6	20			

Fig. 1. Overview of the system architecture designed to enhance nutrition knowledge bases using food images. The framework leverages web data to identify and filter noise, employing image recognition techniques. Our primary contributions include "Enhancing Nutrition Knowledge Bases with Food" and "FoodnKB-visual." These components facilitate the functionality of a smart nutritionist application, enabling users to perform both image and text searches for food items.

"Co-teaching" [9], cyclical learning rates [10] and metrics like 'forgetting event' [11] and AUM [12], which have shown excellence in handling noisy labeled data. However, these approaches usually work on lab-generated systematic noisy data instead of random web noise [13], [14].

In this study, we propose a comprehensive method for enriching FoodnKB entries with visually discernible and instantaneously consumable food images. Our approach makes use of state-of-the-art large language models to identify suitable food items and their synonyms, which are then harnessed to gather diverse web-based images for an extensive visual representation. To address the noisy nature of images obtained from the web, we introduce a novel cyclic-training-based area under the margin metric (cAUM) technique, as well as a generic accuracy gap (AccGap) algorithm, both of which effectively estimate the data's noise ratio and filter out irrelevant images.

Our method demonstrates superior performance in noise detection and image recognition accuracy enhancement when compared to existing noise-robust frameworks. Furthermore, we successfully applied the visually enriched FoodnKB and food recognition within a smart nutritionist mobile application. This innovative approach not only improves access to the food nutrition knowledge base entries but also fosters a more seamless and accessible user experience.

The primary contributions of this work are:

- Proposing a comprehensive method to enrich FoodnKB entries with visually discernible and instantaneously consumable food images, leveraging state-of-the-art large language models for food item identification and synonym expansion.
- Introducing a novel cyclic-training-based area under the margin metric (cAUM) technique and a generic accuracy gap (AccGap) algorithm to handle noise in web-harnessed images, effectively improving image recognition accuracy.
- Demonstrating the practical applicability of the visually enriched FoodnKB through successful integration with a smart nutritionist mobile application, showcasing real-world usability and enhanced user experience in various food-related fields.

In the following sections, we detail our methodology, techniques, and algorithms employed in this research study, focusing on data collection, preprocessing, image acquisition and cleaning, deep learning-based image classification, and FoodnKB query and nutritional information retrieval.

The rest of this paper is structured as follows: Section II introduces the related works. Section III describes the details of our proposed method in curating the images for FoodnKB. The analysis of our results and the application of visual access in a nutritionist app is presented in Section IV. Finally, we conclude our work and discuss the future directions for improvement in Section V.

II. Related Work

In this section, we review the existing research that relates to our study, focusing on three main areas: food nutrition knowledge bases, image recognition in the food domain, and noise reduction techniques for large-scale web-harvested datasets.

A. Food Nutrition Knowledge Bases

Food nutrition knowledge bases (FoodnKBs) are repositories of nutritional information associated with various food items. These knowledge bases aim to facilitate informed dietary decisions by providing users with essential nutritional information. Researchers have explored different aspects of FoodnKBs, such as the development of ontology-based dietary knowledge systems [15], standardization of food descriptions and units [16], and data acquisition and curation methods for FoodnKBs [17]. Our work extends the existing body of research by enriching FoodnKB entries with visually discernible and instantaneously consumable food images to enable more intuitive access to food-related knowledge.

B. Image Recognition in the Food Domain

Deep learning-based image recognition approaches [4], [18] have gained significant attention in the food domain due to their remarkable performance, particularly in food recognition [19], [20], food categorization [21], and food recommendation [22]. Some studies have focused on building fine-grained food classification models [23] and utilizing global features in combination with local features for food image recognition [24]. In contrast, others have explored the use of domain-specific convolutional neural networks (CNNs) for improved food classification performance [25]. Our study builds upon these advances by developing a deep learning-based hierarchical image classification model to access FoodnKBs visually, thus improving the accessibility and usability of the food knowledge base.

C. Noise Reduction Techniques for Labeled Image Datasets

Navigating the noise prevalent in large-scale web-harvested image datasets is critical for accurate image recognition [26], [27]. Various noise reduction techniques have been proposed to address this issue, including label noise-robust deep learning methods [26], [28], sample selection techniques [29], [30], and weak supervision strategies [31], [32].

Techniques such as “Decoupling” [8] and “Co-teaching” [9] provide different models to handle noisy labeled data, often involving small-loss criteria and peer network learning, respectively. Curriculum Learning-based approaches, like “CurriculumNet” [33] and MentorNet [34], also prove valuable in managing noisy and outlier data, with strategies mimicking human learning progression from simple to complex tasks. O2U [10] has applied unique techniques such as cyclical learning rates to improve noisy data detection, managing both the loss metric and noise ratio considerations. Other valuable metrics include ’forgetting event’ [11] and AUM [12], which collect valuable data on sample progress during the training phase. Our study contributes to the field with a novel cyclic training-based area under the margin metric (cAUM) and a generic accuracy gap (AccGap) algorithm. These display superior performance in managing noise in web-harvested food images, thereby enhancing image recognition accuracy.

In summary, our study builds upon and extends the existing body of research in FoodnKBs, deep learning-based food image recognition, and noise reduction techniques. We propose an innovative approach to enriching FoodnKBs and harnessing the power of image recognition technology, thus enabling more intuitive access to food-related knowledge and information.

III. Proposed Method

The methodology for enriching a food nutrition knowledge base with images and enabling access to the knowledge base through image recognition is illustrated in Fig. 1. The process begins by identifying visually discernible and instantaneously consumable (VDIC) food from the food names in the knowledge base. This is achieved by using a sophisticated consensus-based approach, leveraging advanced Large Language Models (LLM), to procure the salient embeddings of the comestible nomenclature and subsequently engender the visually discernible and instantaneously consumable (VDIC) classification. Further, a refined synonym determination process reliant on the capabilities of LLM is put forth, thereby enabling the generation of synonymous food terminology.

Web food images are then harnessed by using the food names and synonyms as queries in web search engines. The irrelevant images in the collected dataset are removed by our proposed novel noise-label detection method, cAUM, the cyclic training-based area under the margin evaluation. The resultant clean images are then used to build an image classification model. Given a user’s query image, the image classification model will output a probability for each image that it is a food image. The images with the highest probabilities can then be linked with the food nutrition information in the knowledge base.

A. Consensus-Based Identification of Visually Discernible and Instantaneously Consumable (VDIC) Food

To enrich FoodnKB entries with visual content, we adopt a consensus-based approach using Large Language Models (LLMs). The motivation for this approach stems from the need to identify foods that possess distinguishable visual features, which are essential for instant recognition and consumption. Many foods lack clear visual characteristics, making them unsuitable for image recognition techniques. Therefore, our focus is solely on those foods that are visually discernible and can be readily identified based on their appearance.

By generating synonyms for these visually identifiable foods, we enhance the database with comprehensive food name variants, facilitating better user interaction and searchability. This approach ensures that we are not burdened with the complexities of modeling non-visually discernible foods, allowing us to concentrate our efforts on those that can be effectively recognized and categorized.

We employ a consensus-based representation using LLMs to harness their synergistic capabilities for the identification and classification of instantaneously consumable food. This section delineates the theoretical underpinnings and formal articulations of the algorithms and functions, detailing the consensus-based approach for VDIC label generation and synonym extraction.

Let $\mathcal{F}$ denote the set of food names within the nutrition knowledge base FoodnKB. For each food term $f_i \in \mathcal{F}$, embeddings are generated using a set of LLMs denoted as $\mathcal{M} = \{M_1, M_2, \ldots, M_n\}$, where n is the number of models in use. The embeddings for each food term can be expressed generically as:

$$\mathbf{v}_{f_i}^{(m)} = M_m(f_i) \quad \text{for } m = 1, 2, \ldots, n \tag{1}$$

where $\mathbf{v}_{f_i}^{(m)}$ represents the embedding generated by the m-th model in the set $\mathcal{M}$. To determine the VDIC label, each model predicts the VDIC status, $\ell_{f_i}^{(m)}$, of each food term $f_i \in \mathcal{F}$:

$$\ell_{f_i}^{(m)} = \text{VDIC}\left(M_m, \mathbf{v}_{f_i}^{(m)}\right) \quad \text{for } m = 1, 2, \ldots, n \tag{2}$$

where $\ell_{f_i}^{(m)}$ can be either 1 (*true*) or 0 (*false*). The final VDIC label L_{f_i} is obtained by comparing the predictions of the models:

$$L_{f_i} = \begin{cases} \ell_{f_i}^{(m)} & \text{if } \ell_{f_i}^{(m)} = \ell_{f_i}^{(k)} \text{ for some } m, k \\ \text{HumanExpert}(f_i) & \text{otherwise} \end{cases} \tag{3}$$

The LLM-based representations enable a thorough and robust disambiguation of the visually discernible and instantaneously consumable food items in the knowledge base, amalgamating the merits of various state-of-the-art LLMs and human expertise to establish an exhaustive and precise classification.

To label these food items based on whether they are visually discernible and instantaneously consumable, we have to consider the characteristics of these items. Table I shows some sample labeling with the rationale for each case. For our implementation, we will specifically utilize DeepSeek [35] and LLaMA [36] as examples of the LLMs in this approach.

B. Leveraging Web Images to Enhance Food Nutrition Knowledge

To obtain our web dataset, FoodnKB-visual, we utilized a text-to-image search based on food names derived from publicly available official food composition tables. This work employs a FoodnKB with 7276 entries, incorporating data from the official Chinese Food Composition Table and various Chinese dishes. Each entry includes a food name, category, and associated nutrients, encompassing three macronutrients and 32 micronutrients.

The catalog of synonyms and the original food names serve as invaluable resources for extracting relevant visual data from the World Wide Web, thereby augmenting our understanding of food nutrition knowledge. This section delineates the systematic process of transforming synonymic information into a diverse array of candidate images, creating a comprehensive representation of the visual attributes of the foods.

TABLE I
Classification of Food Items as True or False Based on the Criteria of Discernible and Instantaneously Consumable (VDIC)

Food Item	Label	Explanation
Salt	False	Although salt is visually recognizable, it is not instantaneously consumable as it's generally consumed as a condiment in small quantities, not as a full food item on its own.
Pepper Powder	False	Similar to salt, pepper powder is visually discernible but not instantaneously consumable, as it's a spice used for seasoning other food items.
Dry Seaweed	True	Dry seaweed is visually discernible and can be eaten directly without any additional preparation as a snack or a part of a meal.
Apple	True	Apples are visually discernible, and they can be eaten directly without any preparation.
Pizza	True	Pizzas are visually discernible, and they are typically consumed directly without any additional preparation needed.
Raw Rice	False	Although raw rice is visually recognizable, it is not instantaneously consumable, as it requires cooking before consumption to be palatable.
Wheat Flour	False	Wheat flour can be recognized visually as a powder; however, it is not instantaneously consumable, as it is typically used as an ingredient in making various baked goods and other recipes.

For synonym generation, we employ multiple Large Language Models (LLMs) to generate lists of synonyms $\mathcal{S}_{f_i}^{(1)}, \mathcal{S}_{f_i}^{(2)}, \ldots, \mathcal{S}_{f_i}^{(m)}$ for each food term $f_i \in \mathcal{F}$:

$$\mathcal{S}_{f_i}^{(j)} = \text{Synonyms\_LLM}_j(\mathbf{v}_{f_i}^{(j)}) \quad \text{for } j = 1, 2, \ldots, m \tag{4}$$

The ultimate list of synonyms, $\mathcal{S}_{f_i}$, for each food term f_i is formulated by taking the union of the respective synonymous terms generated by all employed LLMs:

$$\mathcal{S}_{f_i} = \bigcup_{j=1}^{m} \mathcal{S}_{f_i}^{(j)} \tag{5}$$

This proposed synonym generation harnesses the synergistic power of various LLMs, thereby extending the repertoire of food terminology and optimizing our understanding of the food domain.

Both the food name and its synonyms, $\mathcal{S}_{f_i}$, are utilized to formulate a set of web search engine queries $\mathcal{Q}_{f_i}$ for each food item, $f_i \in \mathcal{F}$, as illustrated in the following equation:

$$\mathcal{Q}_{f_i} = \{f_i\} \cup \mathcal{S}_{f_i} \tag{6}$$

Subsequently, the queries from the compendium $\mathcal{Q}_{f_i}$ are employed to gather a collection of candidate images $\mathcal{I}_{f_i}$, utilizing an ensemble of web search engines to glean a diverse sampling of visual data, represented as:

$$\mathcal{I}_{f_i} = \bigcup_{q \in \mathcal{Q}_{f_i}} \text{ImageSearchEngine}(q) \tag{7}$$

The resulting candidate image compilation $\mathcal{I}_{f_i}$ enables further examination of the food composition properties and visual characteristics. By effectively amalgamating linguistic diversity with computer vision techniques, this approach transcends traditional text-based knowledge bases, enriching food nutrition knowledge and facilitating well-founded assessments of dietary relations. The harvested corpus of candidate images will be further processed to create a high-quality visual complement to the knowledge base.

For our implementation, we will utilize DeepSeek [35] and LLaMA [36] as examples of the LLMs in this approach.

C. cAUM: Cyclical Area Under the Margin for Enhanced Noisy Label Detection

In this section, we introduce a refined technique termed **cAUM** (Cyclical Area Under the Margin), designed to effectively differentiate between "noisy" samples, "hard" examples, and conventional "easy" ones. The cAUM metric builds upon the Area Under the Margin (AUM) metric [12] and integrates concepts from cyclical training [10], addressing the challenge of distinguishing correctly labeled hard samples (rare images) from incorrectly labeled noisy samples.

Let $(x, y) \in D_{\text{train}}$ represent a sample, and $z^{(t)}(x) \in \mathbb{R}^c$ its logits output prior to softmax at epoch t, where c denotes the total number of classes. The margin at epoch t is defined as the difference between the logit of the assigned label and the logit of the highest predicted class excluding the assigned label:

$$M^{(t)}(x, y) = z_y^{(t)}(x) - \max_{i \neq y} z_i^{(t)}(x). \tag{8}$$

A positive margin signifies that the model's prediction aligns with the assigned label, while a negative margin indicates a disagreement. To capture the overall margin statistics, we compute the Area Under the Margin (AUM):

$$AUM(x, y) = \frac{1}{T} \sum_{t=1}^{T} M^{(t)}(x, y), \tag{9}$$

where T represents the total number of training epochs.

To enhance the performance of the AUM metric, we incorporate cyclical training by adapting the learning rates during model training. The cyclical learning rate is defined as follows:

$$\begin{aligned} r(t) &= (1 - s(t)) \times r_1 + s(t) \times r_2, \\ s(t) &= \frac{1 + ((t-1) \mod c)}{c}, \end{aligned} \tag{10}$$

where r_1 and r_2 are the maximum and minimum learning rates, respectively, with $r_1 > r_2$, and c signifies the total number of epochs in each cyclical round.

By averaging the margin statistics captured during cyclical training, we derive the cAUM metric. This approach alleviates the early stopping issue commonly encountered in traditional training methods while effectively distinguishing between noisy samples and hard examples. The training process consists of three stages:

1) *Initial Model Training:* The model is trained on the entire dataset using standard training procedures until convergence.

Algorithm 1: Estimating Noise Ratio from Training Set with Label Noise

```
Data: $D_{\text{train}}$: training set with label noise
Result: Estimated noise ratio
begin
    Divide $D_{\text{train}}$ into $D_r$ (noisy) and $D_l$ (clean);
    Generate $D_r^*$, the correctly labeled version of $D_r$;
    Train a multi-way classification model $\Theta_l$ using
      $D_l$;
    Test the model using $D_r^*$ to obtain accuracy $Acc_r^*$;
    Test the model using $D_r$ to obtain accuracy $Acc_r$;
    Estimate the noise ratio as:
        $\text{Noise Ratio} = \frac{Acc_r^* - Acc_r}{Acc_r^*}$
```

2) *Cyclical Training:* The model undergoes cyclical training, where the learning rates fluctuate to maintain the network's status between overfitting and underfitting. During this phase, we record the cAUM values for all samples.

3) *Refinement:* After ranking samples based on their cAUM values, we filter out the top $k\%$ of samples with the lowest cAUM values, which are likely to be mislabeled. Finally, the model is retrained using only the refined dataset.

The combination of the cAUM metric and cyclical training enhances the quality of the cleansed dataset. Samples that meet both the cAUM threshold and the cyclical loss criteria are considered correctly labeled clean samples, resulting in a more accurate food nutrition knowledge base.

This methodology not only improves the detection of mislabeled data but also maintains the integrity of rare, correctly labeled samples, thereby enhancing the overall performance of neural networks trained on real-world datasets.

D. AccGap: Estimating Label Noise Ratio in Image Datasets

In this section, a procedure for filtering a percentage of samples known to contain noise and characterized by lower cAUM values is introduced. Let k denote the noise ratio, which is typically unknown during actual application processes. The estimation of k is illustrated in Algorithm 1.

To address issues associated with manual verification processes, an innovative method for estimating the noise ratio is proposed. This method utilizes a small clean subset, D_{test}, of the overall training data, while retaining a small fraction of the training data known to contain noise for later analysis. The entire dataset is represented as D_{train}, with the reserved noisy training data as D_r and the remaining dataset (excluded from D_r) as D_l.

To understand the relationship between the noise ratio and its impact on classification accuracy with respect to clean D_{test} and noisy D_r datasets, a model is trained on the remaining noisy dataset, D_l.

Initially, the difference in accuracy between noisy D_r and its correctly labeled counterpart D_r^* is examined. For each training sample $(x, \widetilde{y}) \in D_r$, there exists a corresponding correctly labeled sample $(x, y^*) \in D_r^*$, where $\widetilde{y}$ represents a noisy label and

TABLE II
A PROPOSED ANALYSIS OF POTENTIAL CASES BASED ON THE NOISY LABEL $\widetilde{y}$, THE TRUE LABEL y^*, AND THE PREDICTED LABEL y

Cases	$\widetilde{y} == y^*$	$y == y^*$	$y == \widetilde{y}$	δ_{cn}
1	Yes	Yes	-	0
2	Yes	No	-	0
3	No	Yes	-	1
4	No	No	Yes	-1
5	No	No	No	0

The difference in the number of correct instances between D_r and D_r^* is represented as δ_{cn}. "–" indicates scenarios where the model's prediction does not influence the δ_{cn}.

y^* symbolizes the unknown true label. Given a model with parameters Θ_l trained on D_l, the predicted class y for each training sample $(x, \widetilde{y}, y^*)$ can be derived. The relationships among $\widetilde{y}$, y^*, and y are investigated in terms of their impact on the accuracy gap, as summarized in Table II.

As shown in Table II, in cases 1 and 2, the model's prediction aligns with the noisy label; therefore, the δ_{cn} remains unaffected. Conversely, instances 3, 4, and 5 are influenced by label noise, leading to discrepancies. In the ideal scenario of case 3, the model accurately predicts the correct (clean) label, resulting in an increase in δ_{cn} equivalent to the number of noise instances. However, the worst-case scenario occurs in case 4, where the model's prediction aligns with the incorrect label, thus deflecting the noise in the δ_{cn}.

The likelihood of the model predicting the incorrect label in accordance with the noise, as in case 4, is relatively low. Therefore, case 4 is disregarded in the analysis. Noise in scenario 5 also remains uncaptured.

By excluding case 4 from the analysis, it is identified that the total number of correct predictions' difference between D_r and D_r^* represents the number of noise instances in case 3. This can be considered as a representation of the total noise number when the model predicts the correct label. Furthermore, it is assumed that the noise ratio is independent of the model's performance in terms of either correct or incorrect predictions. Consequently, the noise ratio pertaining to the correctly labeled section of D_r can stand as a representative of the noise ratio of the entire D_r dataset, and by extension, the entirety of D_{train}.

Let N_r denote the total number of samples in D_r and D_r^*, while N_{r_c} and $N_{r_c}^*$ represent the total number of samples predicted correctly by model Θ_l. The associated test accuracies are represented by acc_r and acc_r^*. Thus, the estimated noise ratio, $\widetilde{nr}$, can be formulated as:

$$\widetilde{nr} = \frac{acc_r^* - acc_r}{acc_r^*} \cdot \frac{N_r}{N_{r_c}^*}. \tag{11}$$

This formulation provides an estimation of the noise ratio based on the observed accuracies and the number of samples in the respective datasets. The term $\frac{N_r}{N_{r_c}^*}$ serves to scale the noise ratio based on the total number of samples in D_r relative to the number of correctly predicted samples in the correctly labeled dataset D_r^*.

In summary, the proposed method for estimating the noise ratio leverages the differences in classification accuracy between noisy and correctly labeled data. By focusing on the relationship between the predicted labels and the true labels, this approach effectively quantifies the extent of label noise present in the dataset. The outcomes of this analysis can significantly enhance the robustness of models trained on datasets impacted by label noise, leading to improved performance in real-world applications.

IV. EXPERIMENTS

A. Implementation Details

1) FoodnKB-Visual Dataset Construction: The FoodnKB-visual dataset is a visually-enhanced version of the original FoodnKB, based on the Chinese Food Composition Table of 2018 [37]. This table serves as the core of FoodnKB, providing over 2000 individual food items and their respective nutritional profiles. These food items are categorized as "atomic food," indicating they function as individual ingredients or components in more complex dishes.

To extend the database to include dishes, we sourced information from ChineseFoodWiki,[1] adding approximately 5,000 additional food items, each with unique nutritional information. The combination of atomic food items and dishes resulted in a comprehensive FoodnKB with a total of 7276 entries on Chinese food.

Identified VDIC FOOD: The visually discernible and instantly consumable (VDIC) foods identified by the LLMs achieved a kappa statistic of 0.834. Three human experts manually checked the disagreements between the two LLMs and made the final decisions. The visually-enhanced aspect of FoodnKB was achieved by associating 1251 categories of VDIC food images with 1342 food entries within FoodnKB. It is important to note that multiple food items can link to the same visual class, such as various types of filled buns or different kinds of milk.

Search Queries and Image Collection: The query expansion method generated an average of 0.8 synonyms for each identified VDIC food, resulting in a query size of 1.8 for each food. These queries were sent to three major search engines—Baidu.com, Bing.com, and Google.com—yielding a dataset of 565,452 images, with an average of 452 images per food item. We manually identified the noise labels of 118 foods as our experimental dataset.

2) VireoFood172: The Benchmark Chinese Food Dataset: In addition to the web-collected image dataset, we employed the benchmark Chinese Food Dataset, VireoFood172, which is generally considered to contain a cleaned food dataset with some potential label errors. VireoFood172 comprises 172 food categories, with no fewer than 100 images per category. The categories cover eight major food groups, including "Vegetables," "Soup," "Bean products," "Egg," "Meat," "Seafood," "Fish," and "Staple." We utilized the provided split, where 60% of the total images are used for training, 10% for validation, and the remaining for testing; however, we only used the training and test datasets.

3) Network Architecture and Training Procedure: Our methodology for noise label detection involves a streamlined three-stage process:

[1] [Online]. Available: www.chinesefoodwiki.org

1) *Initial Training:* We train the model on the full dataset, utilizing early stopping to prevent overfitting.
2) *cAUM Calculation:* In the second phase, we record each sample's cAUM (Cyclic-training-based Area Under the Margin) values within a cyclical training paradigm. Samples are ranked by their cAUM values, with the lowest-ranking $k\%$—those most likely to be noisy—filtered out. The $k\%$ is estimated using the proposed AccGap method.
3) *Refinement:* The final stage involves refining the model by training solely on the filtered samples.

For network architecture, our image classification system is built on a cleaned dataset of 1251 classes of VDIC food images, using the high-performing ResNet101 architecture as its core. This network was pre-trained on the ImageNet dataset, ensuring efficient learning from our diverse dataset.

For optimization during model training, we employed an Adaptive Moment Estimation (Adam) optimizer [38], recognized for effectively handling complex, noisy problems. Techniques to prevent overfitting included an L2 regularization term in our loss function and the inclusion of dropout layers in our network.

4) Evaluation Metrics: The comparison of our cAUM with established baselines revolves around two primary aspects:

- *Noise Label Identification:* Assessment is based on both the precision and recall achieved compared to other baselines.
- *Image Classification:* The accuracy of the final image classifier is taken into account. Consistent with previous works [39], peak accuracy—defined as the highest level of accuracy attained on the clean test set during training—is the fundamental evaluation metric.

The following baselines have been utilized in the evaluation:

- *Direct Training:* This basic baseline involves direct training of the image classifier on the original dataset despite the presence of noisy labels.
- *Co-teaching [9]:* This method involves two peer networks selecting a subset of clean samples for their partner based on each batch's loss metric per training cycle.
- *MentorNet:* Utilizes a data-driven curriculum to ascertain the difficulty of training samples.
- *CurriculumNet [33]:* Employs a density-based clustering algorithm to determine sample difficulty.
- *O2U [10]:* Introduces a cyclic loss metric intended to rank all training samples with noisy labels.
- *AUM [12]:* Introduces an AUM metric designed to average margins over training cycles and rank the noise levels of training samples.
- *cAUM:* Our proposed cyclic training area under the margin evaluation method leverages the AUM metric and cyclical training to identify noisy labels.

B. Benchmark of Noise Label Detection

We follow the three-stage training. First, we train a model on the whole training data following a standard training routine with early stopping. Then, we record the cAUM values of all training samples under a cyclical training stage. After ranking all samples by their cAUM values, a sample with a lower cAUM value is more likely to be a noisy sample with false labels. We filter $k\%$ noisy samples with lower cAUM values. k is our estimated noise ratio. Finally, we train a model only using the refined training samples.

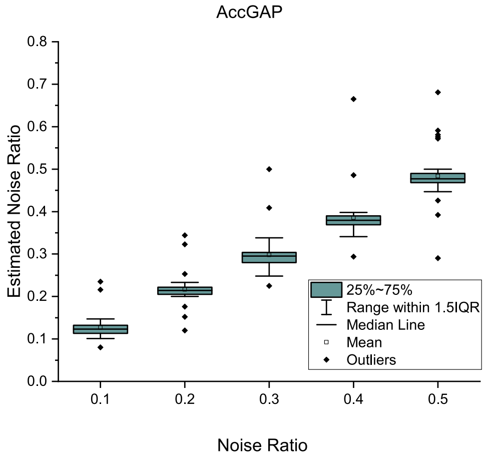


Fig. 2. Performance of AccGAP in estimating noise ratios across varying levels of noise.

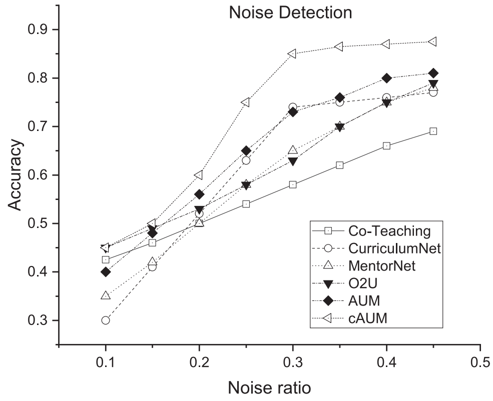


Fig. 3. Comparison of different methods for noise label detection accuracy.

1) The Performance of AccGap: We evaluated the performance of AccGap by conducting experiments following Algorithm 1. To mitigate the potential uneven sampling, the experiments were repeated 40 times each with a random split of the dataset. The results are presented in Fig. 2. We can see that on average, AccGAP well predicted the noise ratios at different given noise levels. While there are some outlines, the spread of predicted results is relatively small. Therefore, we conclude that AccGAP achieves the desirable performance of noise ratio estimation.

2) The Performance of cAUM: Fig. 3 presents the performance of various noise detection methods, namely Co-Teaching, CurriculumNet, MentorNet, O2U, AUM, and cAUM, at different levels of noise ratios. Firstly, we notice that all methods exhibit an overall upward trend in performance as the noise ratio increases. This implies that as the noise in the data becomes more pronounced, the methods are better able to detect and handle it, leading to improved results.

Among the methods, we can observe that cAUM consistently outperforms the other approaches across all noise levels.

TABLE III
COMPARISON OF FOOD RECOGNITION PERFORMANCE ACROSS VARIOUS METHODS AT DIFFERENT NOISE RATIOS ON FOODNKB-VISUAL

Comparing Methods	**Noise Ratio**							
	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
Direct Training	83.35	79.58	73.78	69.26	65.13	61.38	50.89	44.64
MentorNet	86.77	81.88	83.54	78.38	74.22	70.35	66.42	63.50
CurriculumNet	84.21	80.15	77.52	73.06	68.31	64.60	61.13	58.14
Co-Teaching	85.04	80.20	74.22	69.94	65.85	62.11	58.75	55.50
O2U	87.20	82.60	79.99	75.70	71.41	67.33	63.48	60.89
AUM	88.70	84.13	84.44	80.00	75.42	71.39	67.76	64.53
cAUM	90.50	89.51	88.10	85.01	84.49	81.59	80.14	79.65

TABLE IV
COMPARISON OF FOOD RECOGNITION PERFORMANCE USING VARIOUS NOISE REDUCTION METHODS ON THE VIREOFOOD172 AND FOODNKB-VISUAL DATASETS

	Direct Training	**O2U**	**AUM**	**cAUM**
FoodnKB-visual	65.13	71.41	75.42	84.49
VireoFood172	83.52	86.71	87.12	88.46

It demonstrates the highest accuracy in noise detection, indicating its robustness and effectiveness in handling noisy data. On the other hand, Co-Teaching and CurriculumNet display similar patterns of performance, following closely behind cAUM but consistently lagging behind. MentorNet, O2U, and AUM show a moderate level of performance, with AUM being the least effective among them. However, it is worth noting that even the methods with slightly lower accuracy still demonstrate substantial noise detection capabilities, especially at higher noise ratios. Overall, Fig. 3 reveals that cAUM consistently outperforms other approaches in dealing with noisy data.

C. Benchmark of Food Recognition With Noise Labeled Data

Table III presents the performance of various methods for handling noisy data across different noise ratios. Each row in the table represents a different method, and each column represents a specific noise ratio. The performance of methods generally degrades as the noise ratio increases. Handling noisy data is a challenging task, and most methods show some sensitivity to noise. MentorNet, CurriculumNet, and AUM exhibit relatively stable performance across different noise ratios, making them promising methods for handling noisy datasets. Co-Teaching and 02U-net perform well at lower noise ratios but seem to struggle as the noise ratio increases, indicating their limitations in high-noise scenarios. cAUM stands out as the top-performing method in this comparison, consistently achieving the highest accuracy across all noise ratios.

Table IV presents the performance of selected baselines on VireoFood172 and FoodnKB-visual. VireoFood172 adopts the original training set, with 5% of data considered potential noise. FoodnKB-visual adopts the 30% noise ratio setting. We can see that cAUM consistently outperforms O2U and AUM. The overall performance on VireoFood172 is better than on FoodnKB-visual, as the former is considered a clean training set while the latter is a naturally noise-labeled training set. Still, cAUM on FoodnKB-visual achieves satisfactory results. It shows that cAUM has the potential to improve the performance on normal clean datasets as well as noisy datasets.

D. Real-World Applications of FoodnKB-Visual

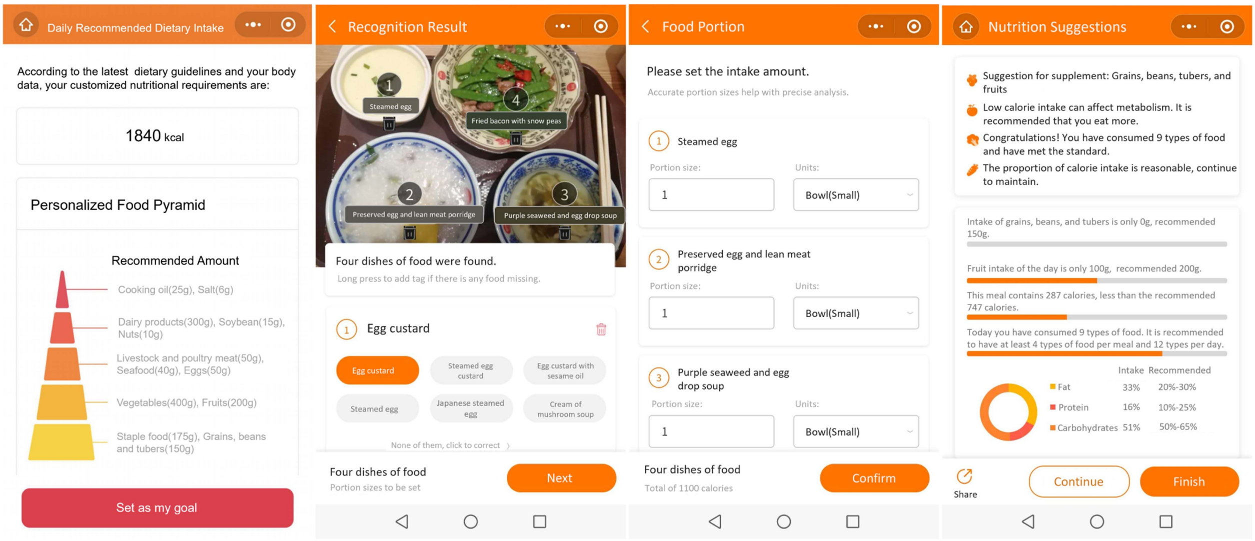
1) KnowFood Smart Nutritionist Application: We successfully applied the visually enriched FoodnKB and food recognition in a smart nutritionist mobile application, showcasing the real-world practicality of our approach. The application allows users to access food knowledge base entries through image recognition, effortlessly log their food intake, obtain accurate nutritional information, and make informed dietary decisions based on personalized recommendations (see Fig. 4). There are four key steps in the app.

The first step is the Daily Recommended Nutrition Intake (RNI) calculation. At the core of KnowFood-Smart Nutritionist lies the precise determination of each user's Daily Recommended Nutrition Intake (RNI). By considering essential user profile data, such as gender, age, weight, and height, the app employs validated nutritional algorithms and established guidelines to compute an individualized RNI. This personalized approach ensures that the nutritional goals align with the user's unique physiological requirements (see Fig. 4(a)).

The second step is visual food logging with image recognition. The app simplifies food logging through the integration of FoodnKB-visual linked image recognition techniques. Users capture images of their meals using the app's built-in camera feature. Employing state-of-the-art object detection models [40], KnowFood-Smart Nutritionist accurately identifies distinct food regions within the images. Subsequently, a food classification model trained on the FoodnKB-visual's images is applied to recognize specific food items present in the meal. To ensure precision, users are actively engaged in the process and are prompted to confirm the accuracy of the recognized food items (see Fig. 4(b)).

The third step is portion confirmation for accurate nutrition tracking. To provide users with comprehensive nutritional insights, KnowFood-Smart Nutritionist enables precise portion confirmation for each identified food item. Leveraging the FoodnKB-visual, the app retrieves nutrition information for each food entry. By integrating the confirmed portion sizes, the app accurately calculates the nutritional composition of the entire meal. This level of detail ensures that users have access to precise and relevant data to support their dietary decision-making process (see Fig. 4(c)).

The last step is the nutritional comparison and personalized feedback. KnowFood-Smart Nutritionist empowers users with a deeper understanding of their dietary habits. By comparing the actual food and nutrition intake with the personalized RNI, the



According to the latest dietary guidelines and your body data, your customized nutritional requirements are:

1840 kcal

Cooking oil(25g), Salt(6g)

Dairy products(300g), Soybean(15g), Nuts(10g)

Livestock and poultry meat(50g), Seafood(40g), Eggs(50g)

Vegetables(400g), Fruits(200g)

Staple food(175g), Grains, beans and tubers(150g)

Four dishes of food were found.

Long press to add tag if there is any food missing.

None of them, click to correct

Four dishes of food

Portion sizes to be set

Please set the intake amount.

Accurate portion sizes help with precise analysis.

1 Steamed egg
Portion size: 1
Units: Bowl(Small)

2 Preserved egg and lean meat porridge
Portion size: 1
Units: Bowl(Small)

3 Purple seaweed and egg drop soup
Portion size: 1
Units: Bowl(Small)

Four dishes of food

Total of 1100 calories

Suggestion for supplement: Grains, beans, tubers, and fruits

Low calorie intake can affect metabolism. It is recommended that you eat more.

Congratulations! You have consumed 9 types of food and have met the standard.

The proportion of calorie intake is reasonable, continue to maintain.

Intake of grains, beans, and tubers is only 0g, recommended 150g.

Fruit intake of the day is only 100g, recommended 200g.

This meal contains 287 calories, less than the recommended 747 calories.

Today you have consumed 9 types of food. It is recommended to have at least 4 types of food per meal and 12 types per day.

	Intake	Recommended
Fat	33%	20%-30%
Protein	16%	10%-25%
Carbohydrates	51%	50%-65%

(a) (b) (c) (d)

Fig. 4. Key Steps in the KnowFood-Smart Nutritionist App: **(a)** Calculation of Daily Recommended Nutritional Intake (RNI) based on user profile; **(b)** Visual food logging utilizing image recognition and user confirmation; **(c)** Portion confirmation for precise nutrition tracking with FoodnKB-Visual; **(d)** Nutritional comparison and personalized feedback.

TABLE V
COMPARISON OF VISUAL AND TEXT ACCESS IN THE SMART NUTRITIONIST APP BY AGE GROUP

Age Group (years)	18-24	25-35	36-55	56-65	66 and older
Ambiguity Level (Visual Search)	2.1	1.3	2.3	3.1	3.8
Ambiguity Level (Text Search)	4.2	2.8	3.3	4.6	4.8
Completion Time (Visual Search)	10.5	11.2	12.3	14.7	19.1
Completion Time (Text Search)	16.4	17.8	18.5	21.2	40.8

Ambiguity levels range from 1 to 10, with higher values indicating greater ambiguity. Completion times are measured in seconds.

app provides valuable feedback on the user's nutritional status. This feedback is presented through informative statistics and conveyed in natural language, making it easily comprehensible and actionable. Users gain insights into any gaps between their current nutritional intake and the recommended values, allowing them to make data-driven adjustments to their diet for improved health outcomes (see Fig. 4(d)).

2) User Study: Visual Vs. Text-Based Access in the Smart Nutritionist App: Table V indicates a comparison between visual and textual food search methods across different age groups integrated into the Smart Nutritionist App. The parameters taken into account are ambiguity levels and search completion time using both methods.

Primarily, it's imperative to approach the ambiguity levels directly reported by the users. Defined on a scale from 1–10, ambiguity alludes to the level of complexity or lack of clarity experienced by users within the search method. Across all age groups, the ambiguity level is noticeably lower when using visual search as compared to text-based research. This difference particularly amplifies in older age groups, such as 55–65 and 66 and up, where ambiguity levels reach 3.1 and 3.8, respectively, for visual search, as contrasted with high 4.6 and 4.8 levels for textual searches. While the ambiguity level slightly increases in visual search from 1.3 to 2.3 when age changes from 25–35 to 36–55, this can be offset against a higher escalation from 2.8 to 3.3 in text search, illustrating an overall advantage of the visual method.

The search completion time further reinforces the benefit of visual search across all age groups. Across all populations, the time to complete a search using visual methods is significantly shorter than using a text search. This trend is consistent across all age groups and is particularly pronounced in the 66 and up age category, where the completion time for a visual search is more than half the time required for a text search (19.1 seconds vs. 40.8 seconds, respectively). This indicates that older users may find a visual search to be a faster and more efficient method for retrieving food nutrition information.

While higher age groups tend to spend more time on both search methods, the time difference between visual and textual searches remains significant across all age groups. This suggests that the efficiency of visual search transcends the boundary of age and can potentially enhance the user experience for a wide demographic.

In summary, the data presented advocates that enhancing food nutrition knowledge bases with visually discernible and instantaneously consumable food images proves to be beneficial for all age groups. By harnessing the advantages of visual search, we can improve user efficacy and understanding while reducing the time and ambiguity associated with food nutrition search methods.

V. Conclusions, Limitations, and Future Directions

In conclusion, this work presents a pioneering approach to enhance the Food Nutrition Knowledge Base (FoodnKB) with visual content and enable image recognition, revolutionizing nutritional assessment. Through consensus-based Large Language Models representation, diverse web-based food image harnessing, and effective noise detection using cAUM and AccGap, we achieve superior image recognition accuracy and noise reduction. The visually enriched FoodnKB, integrated into a smart nutritionist mobile application, offers users a powerful tool for effortless food logging and access to precise nutrition information.

In addition to the advancements made, this work acknowledges several limitations that must be addressed in future iterations. Firstly, the reliance on user-generated food images may introduce variability in image quality and representation, potentially affecting the overall accuracy of the visual recognition system. Moreover, the current methods may struggle with diverse food presentations and cultural variations, leading to misclassifications. Additionally, while cAUM and AccGap enhance the recognition process, their effectiveness is contingent upon the availability of high-quality training data, which can be challenging to obtain. Addressing these limitations will require ongoing refinement of the image classification models and the incorporation of a broader range of food images to ensure robust performance across different user demographics and dietary practices.

For future work, we propose the selection of user-taken food photos to further enrich FoodnKB-visual, enabling a more extensive and diverse database. Continuous efforts to improve the food image classification model will enhance the app's performance and cater to evolving user needs. Therefore, accurately and efficiently updating FoodnKB-visual and the associated classification model is essential. Our work holds great potential in promoting informed dietary choices and healthier lifestyles worldwide, positioning us at the forefront of precision nutrition's cutting-edge developments.

References

[1] D. Aune et al., "Fruit and vegetable intake and the risk of cardiovascular disease, total cancer and all-cause mortality—A systematic review and dose-response meta-analysis of prospective studies," *Int. J. Epidemiol.*, vol. 46, no. 3, pp. 1029–1056, 2017.

[2] W. Wang et al., "A review on vision-based analysis for automatic dietary assessment," in *Proc. Trends Food Sci. Technol.*, vol. 122, pp. 223–237, 2022.

[3] J. K. Ahuja, A. J. Moshfegh, J. M. Holden, and E. Harris, "Usda food and nutrient databases provide the infrastructure for food and nutrition research, policy, and practice," *J. Nutr.*, vol. 143, no. 2, pp. 241S–249S, 2013.

[4] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," *ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–36, 2019.

[5] A. Worsley, "Nutrition knowledge and food consumption: Can nutrition knowledge change food behaviour?," *Asia Pacific J. Clin. Nutr.*, vol. 11, pp. S579–S585, 2002.

[6] Z. Zan, L. Li, J. Liu, and D. Zhou, "Sentence-based and noise-robust cross-modal retrieval on cooking recipes and food images," in *Proc. 2020 Int. Conf. Multimedia Retrieval,* 2020, pp. 117–125.

[7] G. Pruthi, F. Liu, S. Kale, and M. Sundararajan, "Estimating training data influence by tracing gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 19920–19930.

[8] E. Malach and S. Shalev-Shwartz, "Decoupling' when to update' from' how to update'," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 961–971.

[9] B. Han et al., "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 8536–8546.

[10] J. Huang, L. Qu, R. Jia, and B. Zhao, "O2u-net: A simple noisy label detection approach for deep neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.,* 2019, pp. 3325–3333.

[11] M. Toneva et al., "An empirical study of example forgetting during deep neural network learning," 2018, *arXiv:1812.05159*.

[12] G. Pleiss, T. Zhang, E. R. Elenberg, and K. Q. Weinberger, "Identifying mislabeled data using the area under the margin ranking," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17044–17056.

[13] H. Cheng et al., "Learning with instance-dependent label noise: A sample sieve approach," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–27.

[14] L. Jiang, D. Huang, M. Liu, and W. Yang, "Beyond synthetic noise: Deep learning on controlled noisy labels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4804–4815.

[15] M.-H. Wang et al., "Ontology-based multi-agents for intelligent healthcare applications," *J. Ambient Intell. Humanized Comput.*, vol. 1, pp. 111–131, 2010.

[16] N. Slimani et al., "The EPIC nutrient database project (ENDB): A first attempt to standardize nutrient databases across the 10 European countries participating in the epic study," *Eur. J. Clin. Nutr.*, vol. 61, no. 9, pp. 1037–1056, 2007.

[17] H. M. Krumholz, S. F. Terry, and J. Waldstreicher, "Data acquisition, curation, and use for a continuously learning health system," *Jama*, vol. 316, no. 16, pp. 1669–1670, 2016.

[18] Z.-Y. Ming et al., "Food photo recognition for dietary tracking: System and experiment," in *Proc. MultiMedia Model., 24th Int. Conf., MMM 2018, Bangkok, Thailand, Feb. 5-7, 2018, Proc., Part II 24*, 2018 , pp. 129–141.

[19] S. Jiang, W. Min, Y. Lyu, and L. Liu, "Few-shot food recognition via multi-view representation learning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 3, pp. 1–20, 2020.

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

[21] K. Aizawa and M. Ogawa, "Foodlog: Multimedia tool for healthcare applications," *IEEE MultiMedia*, vol. 22, no. 2, pp. 4–8, Apr.–Jun. 2015.

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

[23] W. Min et al., "Large scale visual food recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9932–9949, Aug. 2023.

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

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

[26] Z. Sun et al., "PNP: Robust learning from noisy labels by probabilistic noise prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5311–5320.

[27] Z. Zhang, H. Zhang, S. O. Arik, H. Lee, and T. Pfister, "Distilling effective supervision from severe label noise," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.,* 2020, pp. 9294–9303.

[28] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proc. 2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13723–13732.

[29] J. Han, P. Luo, and X. Wang, "Deep self-learning from noisy labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5138–5147.

[30] Y. Yao et al., "Jo-SRC: A contrastive approach for combating noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5188–5197.

[31] S. Rühling Cachay, B. Boecking, and A. Dubrawski, "End-to-end weak supervision," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 1845–1857.

[32] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, 1998, pp. 92–100.

[33] S. Guo et al., "CurriculumNet: Weakly supervised learning from large-scale web images," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 135–150.

[34] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proc. Int. Conf. Mach. Learn.,* 2018, pp. 2304–2313.

[35] X. Bi et al., "Deepseek LLM: Scaling open-source language models with longtermism," 2024, *arXiv:2401.02954*.

[36] H. Touvron et al., "Llama: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[37] Y. Yang, G. Wang, and X. Pan, *Chinese Food Composition Table*. Beijing, China: Peking University Medicine Publisher, 2018.

[38] N. Xiao, X. Hu, X. Liu, and K.-C. Toh, "Adam-family methods for nonsmooth optimization with convergence guarantees," *J. Mach. Learn. Res.*, vol. 25, no. 48, pp. 1–53, 2024.

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

[40] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," in *Procedia Comput. Sci.*, vol. 199, pp. 1066–1073, 2022.

Zhaoyan Ming received the Ph.D. degree in computer science from the National University of Singapore, Singapore. She is currently an Associate Professor with Hangzhou City University, Hangzhou, China, where she specializes in the application of artificial intelligence in multimedia and life sciences. She is actively involved in the academic community as an Executive Member of the Technical Committee on Natural Language Processing and the Large Language Model Forum under the China Computer Federation.

Zeyu Xie is currently a student with the College of Computer and Computing Technology, Hangzhou City University, Hangzhou, China. His research focuses on the field of multimodal large language models. He is a Member of the 2023 cohort's Distinguished Class in Computer Science and Technology.

Chao Zhang is the Founder of Beijing ZuoYi Technology Company Ltd., Beijing, China. He is engaged in the research and application of artificial intelligence technology in the field of healthcare, dedicated to building a leading AI doctor. Before founding the company, he was with the Natural Language Processing Department, Baidu, mainly focused on information extraction, knowledge graph, and related research.

Kui Su received the Ph.D. degree from the School of Computer Science and Technology, Zhejiang University, Hangzhou, China, in 2017. He was a Postdoctoral Researcher with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, during 2018–2021. He once worked with companies including Huawei, Sangfor Technologies, and Alibaba DAMO Academy. He has been a Research Faculty with Hangzhou City University, Hangzhou, since 2022. His research interests include cloud computing and artificial intelligence computing.

Changzheng Yuan is currently a research Professor with the School of Public Health, Zhejiang University School of Medicine, Hangzhou, China, and adjunct Assistant Professor with Harvard T.H. Chan School of Public Health, Boston, MA, USA. She has been engaged in research on nutritional epidemiology and has conducted a series of population-based empirical studies in the areas of dietary measurement, nutrition and health. She has authored or coauthored a series of articles in prestigious international journals, such as *Nature Aging*, *Nature Food*, *JAMA Psychiatry*, *Alzheimer's & Dementia*, *Neurology*, and *American Journal of Clinical Nutrition*. Her research focuses on epidemiological methods and nutritional cognitive neuroscience.

Tat-Seng Chua is currently the KITHCT Chair Professor with the School of Computing, National University of Singapore, Singapore. He is the Director of a joint research Center between NUS and Tsinghua (NExT) to research into big unstructured multi-source multimodal data analytics. His main research interests include multimedia information retrieval and social media analytics. In particular, his research focuses on the extraction, retrieval and question-answering of text, video and live media arising from the Web and social networks.