LOFI: LOng-tailed FIne-Grained Network for Food Recognition

Jesús M. Rodríguez-de-Vera
Universitat de Barcelona
Barcelona, Spain
j.molina.rdv@ub.edu

Imanol G. Estepa
Universitat de Barcelona
Barcelona, Spain
igonzaes42@alumnes.ub.edu

Marc Bolaños
AIGecko Technologies SL
Barcelona, Spain
marc.bolanos@aigecko.com

Bhalaji Nagarajan
Universitat de Barcelona
Barcelona, Spain
bhalaji.nagarajan@ub.edu

Petia Radeva
Universitat de Barcelona
Barcelona, Spain
petia.ivanova@ub.edu

Abstract

Food recognition plays a crucial role in several healthcare applications. Nevertheless, it presents significant computer vision challenges such as long-tailed and fine-grained distributions that hinder its progress. In this work, we propose LOFI, a Long-tailed Fine-grained Network aimed specifically at tackling these food recognition challenges by improving the feature learning capabilities of food recognition models. Specifically, we improve vanilla R-CNN architecture by tailoring it for food recognition. We design an efficient multi-task framework for fine-grained food recognition, which exploits the lexical similarity of dishes during training to improve the discriminative ability of the network. Secondly, we include a Graph Confidence Propagation module based on graph neural networks to aggregate the information of overlapping detections and refine the final prediction of the network. Extensive analysis and ablations of different components of LOFI highlight that it successfully addresses the targeted problems and leads to noticeable gains in performance. Remarkably, the proposed method achieves competitive results and outperforms the current state-of-the-art methods in three public food benchmarks: UECFood-256, AiCrowd Food Challenge 2022, and UECFood-100 segmented.

1. Introduction

Nutrition and well-being are closely interconnected and mutually influential [27, 53]. Recently, Food Computing [43] has gained a lot of research significance due to its potential applications in society. Food intake monitoring [27] promotes optimal health and helps individuals make informed decisions regarding their nutrition, which is particularly beneficial for people managing chronic conditions such as diabetes, hypertension and cardiovascular problems [46, 48]. Automatic food recognition is instrumental to most food computing tasks [18]. It leverages deep learning models for object detection and image segmentation tasks to common food recognition problems such as nutritional information estimation of dishes [56, 62] and smart-service restaurants [1, 26]. These applications commonly adopt generic models prevalent in the literature [74]. However, food images are highly complex, rendering them challenging to tackle only with general models.

Food images exhibit high intra-class variance and high inter-class similarity, highlighting a clear fine-grained nature [35]. Moreover, the food domain exemplifies a long-tailed distribution problem, where certain dishes are significantly less prevalent compared to others [29]. The combination of these challenges is inadequately addressed by the current generic recognition models like Mask R-CNN [31] and Cascade R-CNN [9]. As these models are trained on large object detection datasets [36, 45], they fail to take into account the long-tailed distributions and fine-grained complexities [24]. Specific approaches such as FGFR [59] and DoD [60] leverage subset learning strategies to address the fine-grained nature of food problems for image classification. However, these methods have been less explored in food recognition tasks.

In this work, we follow this line of subset learning strategies and efficiently apply them to food recognition tasks. Our approach, **LOFI** (**LO**ng-tailed **FI**ne-Grained Network for Food Recognition), emphasizes fine-grained classes and incorporates strategies to address the inherent long-tail distributions. LOFI focuses on increasing the precision of classification made over RoIs using a **multi-task fine-grained module**. We use lexical information to create multiple classification subheads that focus on a specific subset of samples (called *clusters*) [10, 60]. This module proves benefi-

cial in learning better fine-grained features and reduces the RoI classification error. We retain the original efficiency of the model by using this module only during training and removing it during inference. Secondly, we address scenarios in which multiple low-confidence yet correct detections are overshadowed by a single high-confidence detection. To address this, we implement a graph module called **Graph Confidence Propagation (GCP)**. The GCP module constructs a graph by connecting region proposals with edges that encode both spatial and lexical information. Thanks to this connection, we minimize these common scenarios. Finally, we address the long-tailed class imbalance by replacing the loss function and final classification layer with an equalization loss (EQLv2) [66] and a normed linear layer [69] respectively. Our novel modules and **smart replacements** boost the performance of traditional models and obtain SoTA performances on popular food recognition benchmarks, highlighting the effectiveness of handling these food-specific challenges. In summary, we outline our contributions as follows: (1) We present a novel multi-task-based framework, to address fine-grained food recognition. (2) We improve the confidence distribution of the predictions by integrating the Graph Confidence Propagation module. (3) We propose two different replacements for the loss function and final classification layer that empirically boost the performance of food recognition tasks. (4) LOFI improves previous state-of-the-art by **4.6**% and **2.2**% mAP on UECFood-256 and S-UECFood-100 datasets.

2. Related Works

2.1. Challenges in General Object Recognition

Long-tailed Distributions. Long-tailed distributions [75] are characterized by a few classes representing most instances (head), while most classes are underrepresented (tail). This imbalanced distribution is common in real-world situations. Generic datasets such as LVIS [28] are created to focus on addressing this challenge. Several approaches such as Seesaw loss calibration [68], IOF (Inverse Object Frequency Loss) [2], and Equalization losses (EQL) [65, 66] have been proposed to mitigate the long-tailed nature of datasets. Long-tailed object recognition also relies on class grouping: Forest R-CNN [70] clusters classes using their lexical embeddings. A classification head is added to each clustering to determine which cluster the object belongs to. The predicted probability for each cluster is used as a prior for inference. AHRL [38] creates clusters based on feature vectors generated by the model making it necessary to train the model twice. The normed linear layer [69] employs cosine similarity instead of the typical dot-product in the last classification layer. In contrast, our proposed LOFI addresses the long tail problems by a smart combination of EQL and normed linear layer, avoiding the downgrade of cluster methods that require a second training phase.

Post-processing of Detections. Post-processing methods are used to improve object detectors and instance segmentation models by removing duplicated detections from the models' outputs. Non-Maximum Suppression (NMS) [49] and the subsequent Soft-NMS [6] and Dual-NMS [39] are popular post-processing methods. Other works include Confidence Propagation Cluster (CPC) [63], which combined information from overlapping bounding boxes to refine the prediction of a single model. Of late, Graph Neural Networks (GNN) are used to refine object recognition predictions [71, 72]. Graphs are used to model the region proposals, enabling the combination of detections by relying on general priors that are not explicitly annotated. Most of these methods build edges based on the co-occurrence between categories [5, 16, 34] and lexical information of the labels [13, 14, 16]. It is also beneficial to encode additional spatial information in the edges between objects [13, 14]. Despite advancements in hand-crafted rule-based duplicate removal in object detectors and utilization of GNNs to model relations between objects, the combination of both remains unexplored to the best of our knowledge.

2.2. Food Image Recognition

Automated food recognition plays a pivotal role in various tasks such as dietary assessment [27], food perception [61, 64], and food recommendation [26]. Food recognition presents several unique challenges that are intrinsic to the nature of food images and datasets. High occlusion [55], fine-grained classes with high intra-class variance and inter-class similarity [44], highly imbalanced nature of food classes [35] constitute critical challenges in developing any food recognition model. Compared to food classification [43], food detection and instance segmentation are less explored tasks, because of their complexity and limited availability of public datasets. One of the common food datasets, UECFood-256 [36], has significantly fewer classes compared to real-world scenarios. BTBUFood-60 [8] consists of only 60 categories, which has minimal relevance to the fine-grained nature of food. Food detection literature often employs algorithms such as SSD [25], Faster-RCNN [41] and YOLO [47] on different food datasets. However, it does not propose solutions to tackle the said food-specific challenges. Creation of instance segmentation datasets [3, 45, 54] involves highly complex data collection process. Most of the existing food instance segmentation methods either use a limited variety of classes [54], or use simple baselines [20, 54] or "simply" focus on segmentation and mask quality, paying less attention to the classification (which is one of the main challenges of food recognition) [50–52]. In contrast, we address food recognition similar to general domains [28], considering not only the localization but also the classification of ingredients as a core task.

Food is Long-Tailed Fine-Grained. The differences between different food classes are subtle and there usually is a high imbalance between the number of samples of each class, making them both long-tailed and fine-grained [59]. Fine-grained recognition can be categorized into localization-classification sub-networks [32, 37], end-to-end feature encoding [4, 22], and subset learning [59, 60, 67]. Subset learning though less explored, are beneficial in food classification. In subset learning methods, the categories are split into groups of classes, and the network is forced to focus on distinguishing the classes within that group. Fine-grained expert learners exploit already existing multi-level hierarchies to train classifiers of different granularity [11, 76]. Subset learning has been previously employed in fine-grained food classification problems as in FGFR [59], ELFIS [67] and DoD [60]. These methods replicate the end of the backbone as many times as clusters that have been found and combine the output of these replicated blocks to produce the final output in inference. However, ELFIS [67] and FGFR [59] require multi-step training. To the best of our knowledge, no subset learning method exists for recognition tasks. Also, while some approaches tackle the fine-grained problem, there is a lack of literature that directly addresses the long-tail problem. Compared to the other subset methods, LOFI is trained end-to-end, incorporating model-agnostic clusters and smart modifications to tackle the inherent long-tailed distribution problem.

3. Our Proposal: LOFI

The increasing complexity and diversity of culinary dishes across different cultures present a unique challenge in the field of computer vision, particularly in food recognition (both food detection and instance segmentation). The main challenges of food recognition include: (1) food categories and ingredients are fine-grained, (2) food data is highly imbalanced and presents a long-tailed distribution of samples (with many classes underrepresented), (3) the visual arrangement of food dishes is non-uniform, contains occlusions and overlapping instances, as well as a large range of possible scales and orientations, In this section, we introduce our proposed **LOFI** (**LO**ng-tailed **FI**ne-Grained Network for Food Recognition), to address the aforesaid limitations in traditional recognition networks in the food domain.

An overview of LOFI is depicted in Figure 1. As seen in the figure, LOFI is a two-stage food recognition framework comprising a region proposal network (RPN) which identifies candidate regions of interest (RoIs) that likely contain objects. In the second stage, these RoIs are processed to classify the type of food present, refine the bounding box locations and, if dealing with instance segmentation, segment the object in the proposal. Classification of food items is affected by the fine-grained and long-tailed distribution of the data. To tackle the former, LOFI relies on an efficient multi-task fine-grained recognition framework, leveraging linguistic information, forcing the network to learn more discriminative features. To tackle the latter, LOFI replaces the traditional elements of recognition networks with an equalization loss and a normed linear layer to ensure balanced learning across all classes, regardless of their prevalence in the dataset. The framework also addresses the challenges of varying food object sizes, orientations, occlusions, and overlapping instances through the inclusion of Generalized Intersection over Union (GIoU) loss for bounding box refinement and a mask scoring head for improved segmentation accuracy. To solve the problem of overlapping predictions, LOFI implements Graph Confidence Propagation (GCP), using GNNs to aggregate information across predictions, enhancing decision-making accuracy. Together, these improvements address the nuanced challenges of food recognition, significantly advancing the state-of-the-art.

3.1. Food Classification

In two-stage networks, each RoI is independently classified by a classification branch. This classification is highly affected by the general challenges of food images: the fine-grained and long-tailed nature of the data.

Fine-grained Food Classification. The complexity of this task arises from the presence of closely related classes with subtle differences. To tackle this problem, we present an efficient multi-task fine-grained recognition framework designed to exploit non-visual cues (depicted in the red dashed box of Figure 1). This additional information improves the performance of food recognition models in fine-grained scenarios. More concretely, we leverage linguistic information to divide the classes into different clusters of similar categories. For each cluster, we force the network to learn the new task of classifying every proposal as belonging to one specific class in the cluster or to any "other" cluster. The rationale behind this is that it introduces additional non-visual information during the learning process.

Let $\mathcal{C}$ be the set of the categories of the dataset, with $|\mathcal{C}| = C$. We use a text encoder, $TE(\cdot)$, to encode every label in the data set $c \in \mathcal{C}$ into a fixed-length lexical embedding $l_c = TE(c)$. To create the clusters, we use cosine similarity between the embeddings. We then apply hierarchical agglomerative clustering using average linkage. Hierarchical clustering allows us to work with a previously unknown number of clusters, and the linkage helps us obtain clusters of balanced size and handle non-Euclidean distances. These clusters are used to build the multi-task component. Let $\mathcal{U} = \{U_1, U_2, \dots U_n\}$ be a set of clusters of the classes in the dataset, such that $U_i \cap U_j = \emptyset$ for all $1 \leq i, j \leq n$. For each cluster U_i, a new classification subhead $CLUS_i$ (blue and red dashed modules in Figure 1) is attached. These subheads are added as "sibling nodes" to the original classifi-

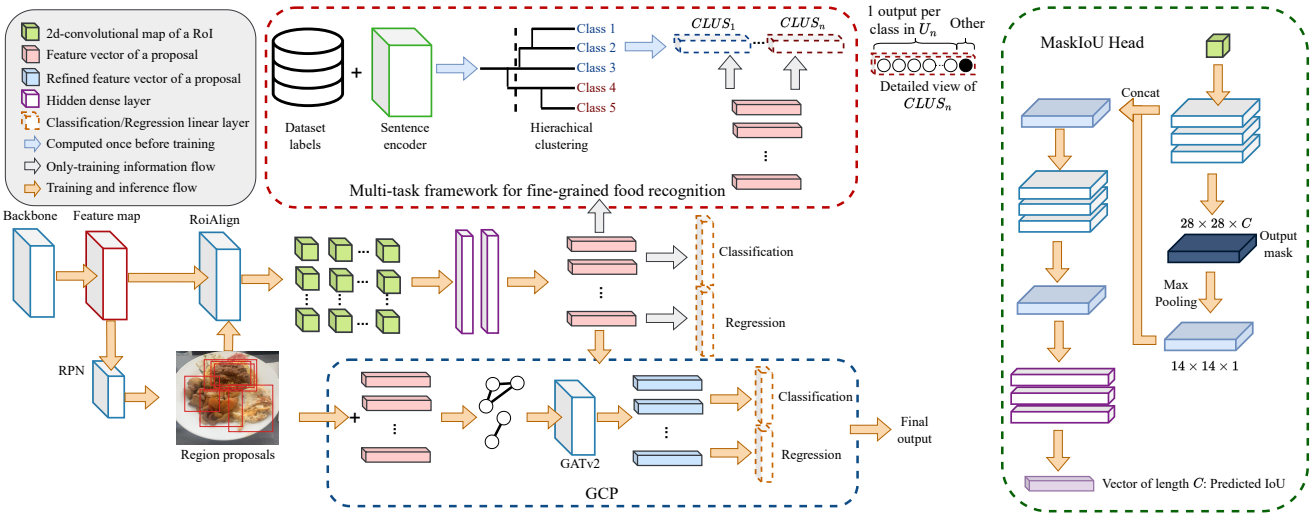


Figure 1. Schematic representation of LOFI, our proposed two-stage food recognition framework. The diagram illustrates the initial region proposal process, followed by the classification, detection, and segmentation stages. Key innovations include the integration of a multi-task fine-grained recognition framework, Equalization Loss v2 and a Normed Linear Layer for addressing the long-tailed nature of food data, and the application of GIoU loss and a mask scoring (or MaskIoU) head for improved bounding box refinement and segmentation accuracy respectively. The Graph Confidence Propagation module resolves overlapping predictions through graph neural networks.

cation head. The input for each subhead is the feature vector of every RoI. Each of these newly added subheads (as shown in Figure 1) is responsible for classifying the classes that belong to the cluster, as well as identifying categories from any other cluster. For example, if $|U_i| = n_i$, then $CLUS_i$ classifies an incoming 1024 vector into $n_i + 1$ categories: a particular class of the cluster U_i or "other" class. It is noteworthy that these heads are only used as guidance during training (removed for inference).

We reduce the imbalance in the subheads by using only the RoIs that have been matched with foreground objects. The proposed multi-task framework for fine-grained recognition can be used on any food recognition model. The computational overhead of this approach is minimal, as the class labels are processed only once to create the clusters, and the subheads are simple linear layers that are only used in training. The presence of these specialized heads during training introduces new tasks to the learning process, which leads to learning features that allow better separability of food classes and clusters in the latent space.

Long-tailed Food Recognition. The disproportionate distribution of food categories, where a small number of classes dominate the dataset while many others are underrepresented, presents a significant challenge often overlooked in the literature on food recognition [24]. To directly address the challenge of class imbalance inherent in food recognition datasets, our method prioritizes achieving a balanced learning environment where rare and common classes are treated equitably. To achieve this, we modify two key components of the classification head: the loss function and the final classification layer. Traditional classification losses such as cross-entropy favor the focus of the network on the most common categories. Thus, we replace it with the equalization loss v2 (EQLv2) [66], which is a hyper-parameter free loss that automatically balances the loss penalty of different losses according to their accumulated gradient (which is used as an indicator of imbalance). This component is particularly valuable in addressing the imbalanced nature of food datasets, as it ensures that the model does not favor the dominant classes while neglecting the rare ones, thus achieving a more balanced and robust performance across all classes. On the other hand, when the imbalance is very high, the weights of the final classification layer are commonly biased towards the most frequent classes (with higher magnitude for the most common categories). We solve it using a normed linear layer [69] to replace the traditional scalar product in the final classification layer with cosine similarity. This adjustment ensures the uniform treatment of all categories, regardless of their frequency in the training dataset. This leads to improved classification performance even for underrepresented classes.

3.2. Food Detection and Segmentation

In response to the unique challenges presented by food recognition, including the variability in sizes and orientations of food items and the occlusions and overlapping instances, we introduce modifications to both the detection and segmentation modules. These changes refine our approach to more effectively handle the intricate aspects of

food detection and segmentation. In two-stage detectors, the positions and dimensions of the predicted bounding boxes are refined to more accurately encompass the detected food items. To improve the bounding box location refinement, we replace the traditional $L1$, $L2$ or L_{IoU} regression losses with the **generalized intersection over union** (GIoU) loss [58]. This loss offers a superior approach for bounding box regression, and it inherently addresses the issue of varying scales and aspect ratios, thus enhancing the overall accuracy of food recognition models.

When dealing with segmentation, we propose using the **mask scoring head** [33], which provides a more accurate evaluation of predicted masks, refining instance-level recognition by explicitly learning the quality of predicted masks and adjusting the corresponding scores. The architecture of this component is depicted in the green box of Figure 1, and it receives as input the feature map of each RoI. As we can see, apart from predicting the mask (as usually done by other food segmentation methods), it also outputs a vector of size C (one per class), containing a prediction of the IoU between the output mask and the ground truth (self-evaluation). The inclusion of this head is particularly important for food instance segmentation, as it improves the model's ability to distinguish and accurately segment overlapping instances of food.

3.3. Post-processing and Refinement

In food recognition, a significant challenge arises where multiple predictions of different classes with differing confidence levels persist, even when using a low confidence threshold. We refer to these overlapping predictions as "islands". To overcome this limitation, we model the predictions as graphs and use Graph Neural Networks (GNNs) to consider the relationships between multiple predictions and make more accurate decisions based on the aggregated information. Thus, we enable the network to reason globally. We introduce the Graph Confidence Propagation (GCP) module (blue dashed box of Figure 1) to specifically address situations where multiple low-confidence predictions of one class may collectively indicate a higher likelihood of that class being present than a single higher-confidence prediction of another class, thereby facilitating more accurate information aggregation.

Given a set of N region proposals, we construct a graph $\mathbb{G} = (V, E)$, where the nodes $v_i \in V$ are the region proposals (represented by the RoI features) and the edges $e_{ij} \in E$ are defined based on the relationship between v_i and v_j. Particularly, an edge between the nodes v_i and v_j is created if $IoU_{ij} = IoU(v_i, v_j) \geq t$. This "sparsification" threshold t makes it easier for the network to focus on dealing with the island, since we substantially limit the number of input edges. Each edge consists of a 6-dimensional vector, and it encodes spatial and lexical information along with the IoU information $e_{ij} = [IoU_{ij}, S_{ij}, Lij]$. **Spatial relationship** [15] corresponds to the spatial features which represent all the relative position information of two proposals:

$$S_{ij} = \left[\log \frac{(x_i - x_j)^2}{w_i^2}, \log \frac{(y_i - y_j)^2}{h_i^2}, \log \frac{w_i}{w_j}, \log \frac{h_i}{h_j}\right]$$

where v_i and v_j are of size $w_i \times h_i, w_j \times h_j$ centered in $(x_i, y_i), (x_j, y_j)$, respectively. **Lexical relationship** is computed as the similarity between two proposals in the semantic space. For each category, as well as the additional background class, we compute the feature vector using $TE(\cdot)$ as described earlier. Then, we can compute a feature representation of the i^{th} proposal in the lexical space as $\hat{l}_i = \sum_{k=1}^{C+1} c_{ik} \cdot l_k \in \mathbb{R}^{256}$, where c_{ik} is the classification score for the k^{th} class prior to the GNN. L_{ij} is defined as the cosine similarity between $\hat{l}_i$ and $\hat{l}_j$. $C + 1$ indicates the number of classes plus background.

Once the graph is constructed, we use a GATv2 layer [7] to refine the representation of each node v_i by aggregating information from its neighbors $\mathcal{N}(i)$, obtaining v_i'. The refined RoI/proposal representation v_i' is finally passed to a classification layer and a regression layer to provide the final output of the detector. This GCP approach helps to mitigate the issue of overlapping predictions, improve the overall performance of the detector, and enhances the model's confidence in the presence of different classes.

3.4. Final Loss

The final training loss is computed as $\mathcal{L} = \mathcal{L}_{RPN} + \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{mask}$, where $\mathcal{L}_{RPN}$ refers to the region proposal loss, $\mathcal{L}_{cls}$ refers to the classification loss, $\mathcal{L}_{reg}$ is the localization loss of the bounding boxes, and $\mathcal{L}_{mask}$ is the loss associated with the segmentation masks.

$$\mathcal{L}_{cls} = EQLv2_{orig} + \frac{1}{n}\sum_{c=1}^{n} CE_{CLUS_c} + CE_{GCP} \quad (1)$$

$$\mathcal{L}_{reg} = GIoU_{orig} + GIoU_{GCP} \quad (2)$$

$$\mathcal{L}_{mask} = BCE_{GCP} + MaskIoU_{GCP}. \quad (3)$$

The sub-indexing scheme employed in Eqs. (1) to (3) denotes which model heads are updated by that specific loss component. More concretely, $EQLv2_{orig}$ and $GIoU_{orig}$ refer to the original classification and regression heads of the R-CNN architecture, respectively. CE_{CLUS_c} refers to the classification loss of the c^{th} cluster, and CE_{GCP} refers to the classification loss of the GCP head. $GIoU_{GCP}$ refers to the regression loss of the GCP head. Finally, BCE_{GCP} and $MaskIoU_{GCP}$ refer to the binary cross-entropy and the mask IoU losses of the GCP head, respectively. Note that $\mathcal{L}_{mask}$ is only used when dealing with instance segmentation, and $\mathcal{L}_{RPN}$ is not modified.

Table 1. mAP Comparison between SoTA methods and LOFI with ResNet-50. *'-' denotes not implemented for the given task.*

Method	UEC 256	S-UEC 100
Faster/Mask R-CNN [31, 57]	46.9	57.3
ForestDet [70]	49.0	61.9
QueryInst [23]	47.8	57.3
IOF [2]	50.3	64.1
DINO [73]	50.4	-
SparseInst [17]	-	60.6
LOFI (Ours)	**55.0**	**66.3**

Table 2. Comparison between base R-CNN architectures and LOFI on various tasks and architectures.

	UEC 256		AiCrowd	S-UEC 100
	Faster	Cascade	Mask	
	(ResNet-50)			
Base	46.9	54.1	19.3	57.3
LOFI	**55.0**	**59.4**	**24.5**	**66.3**
	(Swin-T)			
Base	56.4	62.1	25.1	68.2
LOFI	**56.7**	**63.0**	**26.2**	**69.7**

4. Validation

4.1. Setup

Datasets. We use three public datasets of varying sizes for food recognition tasks. **UECFood-256** [36] is a food detection dataset with a total of 29,774 images, composed of 256 different Asian dishes. We use an 80-20 training-test split for UECFood-256 experiments. **Segmented UECFood-100** [3] is a more recent food instance segmentation dataset, which provides instance-level annotations for the well-known UECFood-100 [42] dataset. The dataset consists of 12,740 images from 100 categories. We create a stratified 80-20 split for our experiments. **AiCrowd Food Recognition Challenge 2022** [45] is an extension of MyFoodRepo-273 benchmark, corresponding to the last edition of the food recognition challenge. It has 54,392 images containing 323 categories of food. Since the test set annotations are private, we split the training dataset into training and test sets in a multi-label stratified fashion.

Implementation Details. For all experimentation and testing, we utilize the PyTorch-based *mmDetection* framework v3.3.0 [12]. We use the Universal Sentence Encoder (USE) [10] as $TE(\cdot)$. We demonstrate the effectiveness of our method across various architectures using ResNet-50 [30] and Swin Transformer (Swin-T) [40] backbones. To further enhance its performance, we have upgraded the standard convolutions in ResNet-50 to deformable convolutions [19], providing additional flexibility and adaptability. This is especially beneficial for handling the intricate variations of food items in terms of shape, size, and texture. Both backbones are initialized with ImageNet-1K [21] pretrained weights. For LOFI, we adopt the default hyperparameters of the R-CNN counterpart, ensuring a fair comparison and avoiding over- or under-tuning of hyperparameters when comparing with other methods. In all considered methods, we use the default hyperparameters of the shortest scheduler recommended by *mmDetection*. We use the **mean average precision (mAP)** as the evaluation metric: box-based mAP for object detection and mask-based mAP for instance segmentation. This metric allows us to assess the quality of our model in a standardized way, facilitating comparisons with existing methods.

4.2. State-of-The-Art Comparisons

We present the performance of LOFI with ResNet-50 backbone in Table 1, showcasing its competitive analysis in different data sets against a variety of state-of-the-art methods under identical experimental conditions to ensure fairness. Specifically, our comparison includes Faster R-CNN [57] and Mask R-CNN [31] as baselines, alongside methods ForesDet [70] and IOF [2] that address challenges in imbalanced datasets. Additionally, recent advancements namely, QueryInst [23], DINO [73], and SparseInst [17] are also evaluated. LOFI achieves substantial improvements across datasets and tasks, showing its effectiveness and generalization ability. LOFI is the best-performing method across both datasets. Its performance gain over the second-best state-of-the-art method is 4.6% for UECFood-256 and 2.2% for Segmented UECFood-100. These results show the competitiveness of the proposed approach. IOF [2], the second-best performing approach in the considered cases, is inferior in general domain recognition compared to other methods in the list. However, IOF focusses on long-tailed recognition. These results highlight the importance of considering this challenge when designing solutions for food detection and segmentation, supporting our idea and reinforcing the key decisions made in the design of LOFI.

4.3. Discussions

4.3.1 Performance Comparison with R-CNNs

We report the mAP achieved by base R-CNN architectures and LOFI with two different backbones in Table 2. To explore LOFI's adaptability, we examine cascade architectures by comparing the performance of Cascade R-CNN [9] with *Cascade LOFI*, a version enhanced with all our proposed improvements. LOFI achieves noticeable improvements across datasets, tasks, and architectures, which highlights its effectiveness and generalization ability. More con-

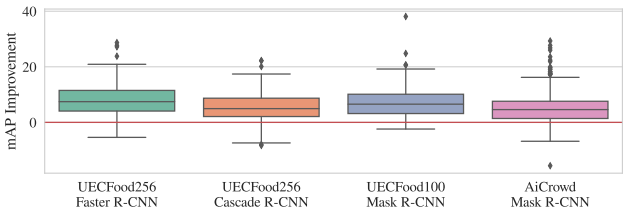


Figure 2. Class-wise mAP improvement for different datasets and architectures. The red line represents 0 difference (no change).

cretely, the performance gains range from +5.3% in the AiCrowd dataset to +8.1% in the UECFood256 dataset with Faster R-CNN. Notably, LOFI exhibits the capability to enhance the performance of more complex architectures such as Cascade R-CNN and Swin-T backbones, which are known for their high capabilities. This highlights the remarkable ability of LOFI to further optimize and refine the results obtained from these complex architectures, thereby achieving better overall performance.

The benefits of using LOFI can be further seen in the delta distribution of the class-wise mAP for the considered benchmarks as shown in Figure 2. LOFI shows an improvement in results for most of the categories. However, there is a small portion of classes that suffer a performance drop. This decrease is very slight in all cases, except one outlier in AiCrowd. Regarding qualitative results, in Figure 3 we visualize some examples of images from Segmented UECFood-100 in which LOFI outperforms the baseline. Three different kinds of situations, where the baseline fails in some way, but LOFI succeeds, are presented (in this order): correct classification, wrong segmentation; incorrect classification, correct segmentation; wrong for both tasks.

4.3.2 Effect of Multi-task Fine-grained Framework

To better understand the influence of the newly added subheads, we provide a deeper analysis using the changes in the inter- and intra-cluster confusion (Figure 4a). The inter-cluster confusion of the k^{th} cluster measures the proportion of detections whose ground truth is a class of the k^{th} cluster, but whose predicted label belongs to any other cluster. The intra-cluster confusion of the k^{th} cluster represents the percentage of predictions whose ground-truth label belongs to the k^{th} cluster, but the model has predicted a different class of the same cluster. According to Figure 4a, there is a noticeable drop in the confusion between clusters in both datasets after including the fine-grained module. Thanks to the output of "others" in the cluster subheads (described in Figure 1), the backbone is forced to learn features that allow the network to discern the belonging of an object to every cluster more effectively. Similarly, there is also a significant improvement in confusion between classes of the same cluster when using the fine-grained framework. In some cases, there is a slight increase in misclassification. However, this does not lead to a decrease in the overall performance.

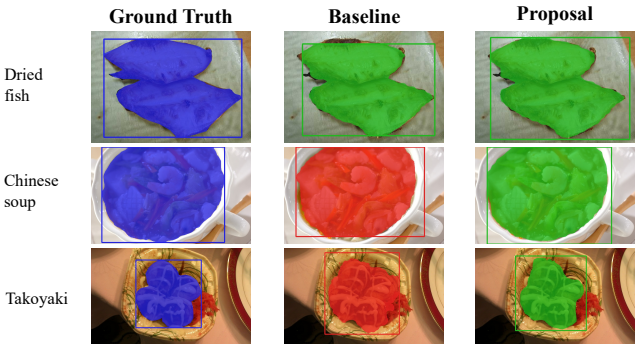


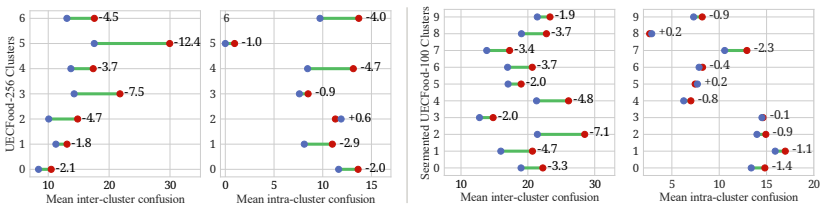
Figure 3. Comparison of base R-CNN and LOFI predictions on Segmented UECFood-100 dataset images. Blue indicates ground truth, green for correct class predictions, and red for wrong ones.

4.3.3 Effect of Graph Confidence Propagation Module

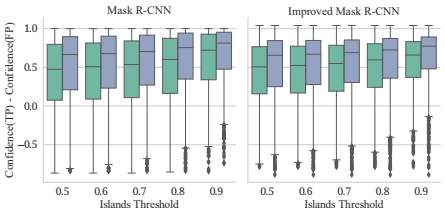
We provide in-depth analysis to understand whether the GCP helps the model to address the problem we are trying to tackle: the presence of "islands" of detections for the same object in which the most confident prediction is not a true positive. To this end, we present in Figure 4b the distributions of the differences between the confidence of the true positive and the confidence of the most confident false positive for every island in Segmented UECFood-100. A higher value indicates a better ability of the model to identify the proper label of an object. Since the idea of highly overlapping detections is not well defined, one natural question is *"When a group of detections should be considered an island?"*. Following a philosophy similar to that behind the mAP metric, we analyze the islands defined using different IoU thresholds. For a given threshold t_I, an island is formed by all the predictions that can be connected by IoU values above t_I. From Figure 4b, we can infer that using GCP results in an improvement in this aspect in all thresholds and all the situations considered. The GCP module helps the model to better identify the appropriate labels for objects within these "islands". This way, we address the originally targeted problem, improve the prediction confidence distributions and the overall performance of the detector.

4.4. Ablation study

In Table 3, we show results for different combinations of LOFI components (ResNet-50) for UECFood-256 and Segmented UECFood-100. More concretely, we evaluate all the modules independently and in conjunction with each other. The proposed smart replacements (SR) to the R-CNN architecture (losses and layer changes) provide a significant performance boost, especially when combined with



(a) Mean inter- and intra-cluster confusion comparison for UECFood-256 (left) and Segmented UECFood-100 (right) using base R-CNN ( ) and our multi-task fine-grained recognition framework ( ). Negative numbers indicate improvement.



(b) Distribution of confidence difference between true positives and the most confident false positives within each *island*. =No GCP, =GCP.

Figure 4. Analysis of the impact of the multi-task fine-grained framework and GCP.

Table 3. Ablation of different components of LOFI.

SR	FG	GCP	UEC 256	S-UEC 100
-	-	-	46.9	57.3
✓	-	-	52.6	63.6
-	✓	-	51.3	63.3
-	-	✓	48.4	62.9
✓	✓	-	54.3	64.8
✓	-	✓	54.4	65.6
✓	✓	✓	**55.0**	**66.3**

the other elements. In terms of mAP, the multi-task fine-grained framework (FG) shows noticeable benefits when used independently or in combination with the other modules. The GCP module also exhibits similar improvements when added to the pipeline. Combining all the modules delivers the best results, showing the benefits of LOFI.

4.5. Limitations

Despite the promising results and improvements in all the considered metrics and scenarios, we carefully elucidate the potential limitations of LOFI that can serve as future directions. (1) The clustering technique relies on the lexical embeddings of the category labels. This limits the applicability of the method to cuisines for whose language there is not a robust text encoder. (2) The improvement in confidence distribution brought by GCP might not always be reflected in the final performance. This is because sometimes the GCP priors might mislead the module, leading to an increase in the confidence of some false positives. (3) Although good results have been achieved without tuning, the presence of several losses might require extensive testing to obtain the optimal performance.

5. Conclusions and Future Lines

In this work, we introduce **LOFI**, a novel framework tailored for the intricate task of food recognition, which surpasses the performance of general state-of-the-art networks in this domain. Through our comprehensive evaluations across diverse datasets and architectures, we demonstrate the exceptional ability of LOFI to address the unique challenges of food object recognition ranging from long-tailed data distribution to the diverse shapes and sizes of food items. A key aspect of LOFI is the strategic use of cross-modal information, specifically leveraging linguistic similarities among food categories to refine and guide the model's learning process. Furthermore, our utilization of the GCP module further emphasizes the value of considering neighbouring predictions to refine detection outputs. With these enhancements, LOFI establishes a new benchmark for food recognition, underscoring the importance of focusing on food-specific challenges.

Future Lines. Exploring more advanced lexical models such as LLMs could refine our clustering approach, potentially offering better discrimination. The usage of GNNs can be further explored to obtain solutions that include other information (e.g. from ontologies). Additionally, our research represents a notable advancement in food monitoring technology and encourages both experienced and inexperienced users to interact with these innovations, closing the gap for more user-friendly and efficient food tracking.

Acknowledgements

This work was partially funded by the EU project MUSAE (No. 01070421), 2021-SGR-01094 (AGAUR), Icrea Academia'2022 (Generalitat de Catalunya), Robo STEAM (2022-1-BG01-KA220-VET-000089434, Erasmus+ EU), DeepSense (ACE053/22/000029, ACCIÓ), DeepFoodVol (AEI-MICINN, PDC2022-133642-I00), IDEATE (AEI-MICINN, PID2022-141566NB-I00), A-BMC (AEI-MICINN, CNS2022-135480) and CERCA Programme/Generalitat de Catalunya. B. Nagarajan acknowledges the support of FPI Becas, MICINN, Spain. J. M. Rodríguez-de-Vera acknowledges the support of FPU Becas, MU, Spain. The authors thankfully acknowledge the computer resources at FinisTerrae III and the technical support provided by the Galician Supercomputing Center (CESGA) (RES-IM-2023-2-0025).

References

[1] Eduardo Aguilar, Beatriz Remeseiro, Marc Bolaños, and Petia Radeva. Grab, pay, and eat: Semantic food detection for smart restaurants. *IEEE Transactions on Multimedia*, 20 (12):3266–3275, 2018. 1

[2] Konstantinos Panagiotis Alexandridis, Shan Luo, Anh Nguyen, Jiankang Deng, and Stefanos Zafeiriou. Inverse Image Frequency for Long-tailed Image Recognition, 2022. 2, 6

[3] Elena Battini Sönmez, Sefer Memiş, Berker Arslan, and Okan Zafer Batur. The segmented uec food-100 dataset with benchmark experiment on food detection. *Multimedia Systems*, pages 1–9, 2023. 2, 6

[4] Ardhendu Behera, Zachary Wharton, Pradeep RPG Hewage, and Asish Bera. Context-Aware Attentional Pooling (CAP) for Fine-Grained Visual Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 929–937, 2021. 3

[5] Aritra Bhowmik, Martin R Oswald, Yu Wang, Nora Baka, and Cees GM Snoek. Detecting objects with graph priors and graph refinement. *arXiv preprint arXiv:2212.12395*, 2022. 2

[6] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017. 2

[7] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022. 5

[8] Qiang Cai, Jing Li, Haisheng Li, and Yunxuan Weng. Btbufood-60: Dataset for object detection in food field. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–4. IEEE, 2019. 2

[9] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 1, 6

[10] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174, 2018. 1, 6

[11] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your "Flamingo" is my "Bird": Fine-Grained, or Not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11476–11485, 2021. 3

[12] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[13] Shengjia Chen, Zhixin Li, Feicheng Huang, Canlong Zhang, and Huifang Ma. Improving object detection with relation mining network. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 52–61. IEEE, 2020. 2

[14] Shengjia Chen, Zhixin Li, and Zhenjun Tang. Relation r-cnn: A graph based relation-aware network for object detection. *IEEE Signal Processing Letters*, 27:1680–1684, 2020. 2

[15] Shengjia Chen, Zhixin Li, Feicheng Huang, Canlong Zhang, and Huifang Ma. Object detection using dual graph network. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3280–3287, 2021. 5

[16] Shengjia Chen, Zhixin Li, Feicheng Huang, Canlong Zhang, and Huifang Ma. Object detection using dual graph network. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3280–3287. IEEE, 2021. 2

[17] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Wenqiang Zhang, Qian Zhang, Chang Huang, Zhaoxiang Zhang, and Wenyu Liu. Sparse instance activation for real-time instance segmentation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022. 6

[18] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. Food recognition: a new dataset, experiments, and results. *IEEE journal of biomedical and health informatics*, 21 (3):588–598, 2016. 1

[19] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017. 6

[20] Yanyan Dai, Subin Park, and Kidong Lee. Utilizing mask r-cnn for solid-volume food instance segmentation and calorie estimation. *Applied Sciences*, 12(21):10938, 2022. 2

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[22] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-Entropy Fine Grained Classification. *Advances in neural information processing systems*, 31, 2018. 3

[23] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6910–6919, 2021. 6

[24] Jixiang Gao, Jingjing Chen, Huazhu Fu, and Yu-Gang Jiang. Dynamic mixup for multi-label long-tailed food ingredient recognition. *IEEE Transactions on Multimedia*, 2022. 1, 4

[25] Xiaoyan Gao, Xiangqian Ding, Ruichun Hou, and Ye Tao. Research on food recognition of smart refrigerator based on ssd target detection algorithm. In *Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science*, pages 303–308, 2019. 2

[26] M Gerasimchuk and A Uzhinskiy. Food recognition for smart restaurants and self-service cafes. *Physics of Particles and Nuclei Letters*, 21(1):79–83, 2024. 1, 2

[27] Tonmoy Ghosh, Yue Han, Viprav Raju, Delwar Hossain, Megan A McCrory, Janine Higgins, Carol Boushey, Edward J Delp, and Edward Sazonov. Integrated image and sensor-based food intake detection in free-living. *Scientific Reports*, 14(1):1665, 2024. 1, 2

[28] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5356–5364, 2019. 2

[29] Jiangpeng He, Luotao Lin, Heather A Eicher-Miller, and Fengqing Zhu. Long-tailed food classification. *Nutrients*, 15(12):2751, 2023. 1

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 6

[32] Xiangteng He and Yuxin Peng. Weakly Supervised Learning of Part Selection Model with Spatial Constraints for Fine-Grained Image Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), 2017. 3

[33] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6409–6418, 2019. 5

[34] Chenhan Jiang, Hang Xu, Xiaodan Liang, and Liang Lin. Hybrid knowledge routed modules for large-scale object detection. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[35] Parneet Kaur, Karan Sikka, Weijun Wang, Serge Belongie, and Ajay Divakaran. Foodx-251: A dataset for fine-grained food classification. *arXiv e-prints*, pages arXiv–1907, 2019. 1, 2

[36] Y. Kawano and K. Yanai. Foodcam: A real-time food recognition system on a smartphone. *Multimedia Tools and Applications*, 2014. 1, 2, 6

[37] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-Grained Recognition without Part Annotations. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5546–5555, 2015. 3

[38] Banghuai Li. Adaptive hierarchical representation learning for long-tailed object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2303–2312, 2022. 2

[39] Zhiyuan Lin, Qingxiao Wu, Shuangfei Fu, Sikui Wang, Zhongyu Zhang, and Yanzi Kong. Dual-nms: A method for autonomously removing false detection boxes from aerial image object detection results. *Sensors*, 19(21):4691, 2019. 2

[40] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[41] Runyu Mao, Jiangpeng He, Zeman Shao, Sri Kalyan Yarlagadda, and Fengqing Zhu. Visual aware hierarchy based food recognition. In *International conference on pattern recognition*, pages 571–598. Springer, 2021. 2

[42] Y. Matsuda, H. Hoashi, and K. Yanai. Recognition of multiple-food images by detecting candidate regions. In *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, 2012. 6

[43] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A survey on food computing. *ACM Computing Surveys (CSUR)*, 52(5):1–36, 2019. 1, 2

[44] Weiqing Min, Linhu Liu, Zhengdong Luo, and Shuqiang Jiang. Ingredient-guided cascaded multi-attention network for food recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1331–1339, 2019. 2

[45] Sharada Prasanna Mohanty, Gaurav Singhal, Eric Antoine Scuccimarra, Djilani Kebaili, Harris Héritier, Victor Boulanger, and Marcel Salathé. The food recognition benchmark: Using deep learning to recognize food in images. *Frontiers in Nutrition*, 9, 2022. 1, 2, 6

[46] Alan Renier Jamal Occhioni Molter, Naise Oliveira da Rocha Carvalho, Paloma Ribeiro Torres, Marlete Pereira da Silva, Patrícia Dias de Brito, Pedro Emmanuel Alvarenga Americano do Brasil, Claudio Fico Fonseca, and Adriana Costa Bacelo. Development of a mobile application to represent food intake in inpatients: dietary data systematization. *BMC Medical Informatics and Decision Making*, 24 (1):28, 2024. 1

[47] Roberto Morales, Juan Quispe, and Eduardo Aguilar. Exploring multi-food detection using deep learning-based algorithms. In *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–7, 2023. 2

[48] Bhalaji Nagarajan, Rupali Khatun, Marc Bolaños, Eduardo Aguilar, Leonardo Angelini, Mira El Kamali, Elena Mugellini, Omar Abou Khaled, Noemi Boqué, Lucia Tarro, and Petia Radeva. Nutritional Monitoring in Older People Prevention Services. In *Digital Health Technology for Better Aging: A multidisciplinary approach*, pages 77–102. Springer International Publishing, Cham, 2021. 1

[49] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, pages 850–855. IEEE, 2006. 2

[50] Huu-Thanh Nguyen and Chong-Wah Ngo. Terrace-based food counting and segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2364–2372, 2021. 2

[51] Huu-Thanh Nguyen, Chong-Wah Ngo, and Wing-Kwong Chan. Sibnet: Food instance counting and segmentation. *Pattern Recognition*, 124:108470, 2022.

[52] Huu-Thanh Nguyen, Yu Cao, Chong-Wah Ngo, and Wing-Kwong Chan. Incremental learning on food instance segmentation. *arXiv preprint arXiv:2306.15910*, 2023. 2

[53] Lauren Owen and Bernard Corfe. The role of diet and nutrition on mental health and wellbeing. *Proceedings of the Nutrition Society*, 76(4):425–426, 2017. 1

[54] Deokhwan Park, Joosoon Lee, Junseok Lee, and Kyoobin Lee. Deep learning based food instance segmentation using synthetic data. In *2021 18th International Conference on Ubiquitous Robots (UR)*, pages 499–505. IEEE, 2021. 2

[55] Kaylen Pfisterer, Robert Amelard, Jennifer Boger, Heather Keller, Audrey Chung, and Alexander Wong. Enhancing food intake tracking in long-term care with automated food imaging and nutrient intake tracking (afini-t) technology:

Validation and feasibility assessment. *JMIR aging*, 5(4): e37590, 2022. 2

[56] Parth Poply and Angel Arul Jothi J. An Instance Segmentation approach to Food Calorie Estimation using Mask R-CNN. In *Proceedings of the 2020 3rd International Conference on Signal Processing and Machine Learning*, pages 73–78, New York, NY, USA, 2020. Association for Computing Machinery. 1

[57] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6

[58] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[59] Javier Ródenas, Bhalaji Nagarajan, Marc Bolaños, and Petia Radeva. Learning multi-subset of classes for fine-grained food recognition. In *Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management*, pages 17–26, 2022. 1, 3

[60] Jesús M Rodríguez-de-Vera, Pablo Villacorta, Imanol G Estepa, Marc Bolaños, Ignacio Sarasúa, Bhalaji Nagarajan, and Petia Radeva. Dining on details: Llm-guided expert networks for fine-grained food recognition. In *Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management*, pages 43–52, 2023. 1, 3

[61] Nareen O. M. Salim, Subhi R.M. Zeebaree, Mohammed A. M. Sadeeq, A. H. Radie, Hanan M. Shukur, and Zryan Najat Rashid. Study for Food Recognition System Using Deep Learning. *Journal of Physics: Conference Series*, 1963(1): 012014, 2021. 2

[62] Suriyakrishnan Sathish, S. Ashwin, Md. Abdul Quadir, and L. K. Pavithra. Analysis of Convolutional Neural Networks on Indian food detection and estimation of calories. *Materials Today: Proceedings*, 62:4665–4670, 2022. 1

[63] Yichun Shen, Wanli Jiang, Zhen Xu, Rundong Li, Junghyun Kwon, and Siyi Li. Confidence propagation cluster: Unleash full potential of object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1151–1161, 2022. 2

[64] Ghalib Ahmed Tahir and Chu Kiong Loo. A Comprehensive Survey of Image-Based Food Recognition and Volume Estimation Methods for Dietary Assessment. *Healthcare*, 9(12): 1676, 2021. 2

[65] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11662–11671, 2020. 2

[66] Jingru Tan, Xin Lu, Gang Zhang, Changqing Yin, and Quanquan Li. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1685–1694, 2021. 2, 4

[67] Pablo Villacorta, Jesús M Rodríguez-de Vera, Marc Bolaños, Ignacio Sarasúa, Bhalaji Nagarajan, and Petia Radeva. Elfis: Expert learning for fine-grained image recognition using subsets. *arXiv preprint arXiv:2303.09269*, 2023. 3

[68] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9695–9704, 2021. 2

[69] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9919–9928, 2020. 2, 4

[70] Jialian Wu, Liangchen Song, Qian Zhang, Ming Yang, and Junsong Yuan. ForestDet: Large-Vocabulary Long-Tailed Object Detection and Instance Segmentation. *IEEE Transactions on Multimedia*, 24:3693–3705, 2022. 2, 6

[71] Xiwei Yang, Xinfang Zhong, and Zhixin Li. Grdn: Graph relation decision network for object detection. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 2

[72] XiuTing You, He Liu, Tao Wang, Songhe Feng, and Congyan Lang. Object detection by crossing relational reasoning based on graph neural network. *Machine Vision and Applications*, 33:1–14, 2022. 2

[73] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2023. 6

[74] Yudong Zhang, Lijia Deng, Hengde Zhu, Wei Wang, Zeyu Ren, Qinghua Zhou, Siyuan Lu, Shiting Sun, Ziquan Zhu, Juan Manuel Gorriz, and Shuihua Wang. Deep learning in food category recognition. *Information Fusion*, 98:101859, 2023. 1

[75] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[76] Guohang Zhuang, Yue Hu, Tianxing Yan, and JiaZhan Gao. Gcam: Gaussian and causal-attention model of food fine-grained recognition. *arXiv preprint arXiv:2403.12109*, 2024. 3