ARANet: Adaptive Mining Region Feature Aggregation Network for Food Recognition

Yanling Li, Qingqi Liang, Wei He, and Mengmeng Liu

***Abstract*—Most existing food recognition techniques depend on convolutional neural networks (CNNs) for the purpose of extracting features directly from food images. However, these methods frequently fail to discriminate key details in food images with cluttered backgrounds. Here, we propose a novel network framework employing an adversarial erasing strategy via adaptive threshold segmentation for food recognition. The framework is designed to simultaneously learn global features and diverse, complementary local features. A residual network (ResNet) is employed in order to extract global features from the images under consideration, with the objective of generating class activation maps (CAMs) from the final convolutional layer. The model progressively mines discriminative food regions to capture diverse and complementary local feature representations. Adaptive threshold segmentation isolates discriminative regions within CAMs, eliminating cluttered backgrounds and enhancing key feature extraction. In conclusion, the framework integrates representations of the original input image and the identified regions for prediction. Extensive experimentation on five benchmark food recognition datasets has been conducted to demonstrate the superiority of the proposed approach. For instance, the efficacy of the proposed methodology is evidenced by the attainment of Top-1 and Top-5 accuracies of 90.4% and 98.7%, respectively, on the Vireo Food-172 dataset.**

***Index Terms*—Food recognition, Convolutional neural network, Adversarial erasing, Adaptive threshold**

I. Introduction

FOOD recognition has garnered significant attention in computer vision and related fields [1], [2], advancing the understanding of food from multidisciplinary perspectives, such as health [3], [4] and medicine [5], [6]. Food recognition holds significant potential for health-related applications, particularly in dietary assessment [7], offering critical insights for disease prevention.

Convolutional neural networks (CNNs) have been successfully used for food recognition, outperforming traditional methods [8], [9]. Several studies have employed features that have been extracted from pre-trained networks for the purpose of food recognition [10], [11]. For instance, McAllister et al. [12] employed pre-trained deep learning architectures, specifically GoogleNet [13] and residual network-152 (ResNet-152) [14] networks. Other approaches fine-tune existing deep networks [15], [16], [17]. For example, Tasci et al. [18] performed food image recognition by fine-tuning ResNet, GoogleNet, similar architectures. These methods rely on generic CNN architectures to extract features from food images, neglecting to design structures tailored to food characteristics. Consequently, opportunities remain to improve food image recognition accuracy. Some researchers have developed specialized deep neural networks for recognising food images [19], [20]. Elbassuoni et al. [21] employed a customized object detection model. Moreover, certain studies have approached the recognition of food images as a fine-grained visual recognition task [22], [23], focusing on mining and integrating discriminative regions within food images for enhanced recognition. Yang et al [24]. proposed a novel network capable of effectively capturing both global and local features from food images.

The aforementioned studies have significantly advanced the field of food recognition. However, food image recognition remains challenging. Similar to generic object recognition, its core challenge is extracting discriminative visual features from images. Food images often exhibit complex backgrounds that contain objects or visual noise unrelated to the target food. As illustrated in Figure 1, several food images from the ETH Food-101 [25] exemplify this complexity. Apparently, addressing this background complexity correctly can lead to more robust and accurate food recognition results.

To this end, a novel Adaptive Mining Region Feature Aggregation Network (ARANet) is proposed for the food recognition framework. The framework is primarily composed of two components: adaptive mining of discriminative regions and region feature fusion. The former adopts adversarial erasing (AE) [26] based on adaptive threshold segmentation to mine discriminative regions. While the latter combines the global and local features of the input image and the mined regions. In summary, our proposed ARANet model aims to reduce complex background interference in food recognition.

The paper is organised as follows: Section II is about our food recognition framework; Section III is about the experimental results and analysis; and Section IV is about the conclusion and future work.

II. Materials and Methods

ARANet consists of two core components: (1) adaptive discriminative region mining and (2) region feature fusion.

Manuscript received November 13, 2024; revised June 20, 2025.

This work was supported in part by the Science and Technology Project of Henan Province (252102211025), the Excellent Postgraduate Teaching Material Project of Henan Province (YJS2025JC30), and the Graduate Research Innovation Fund of Xinyang Normal University (2024KYJJ087).

Yanling Li is a Professor at the School of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, China (e-mail: lyl75@163.com).

Qingqi Liang is a postgraduate student at the School of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, China ((corresponding author to provide phone: +8618860274943; e-mail: qqliang2022@126.com).

Wei He is a Professor at the College of Artificial Intelligence, Guangzhou Maritime University, Guangzhou 510000, China, and the Center for Intelligent Information Processing and Applied Engineering, Guangzhou Maritime University, Guangzhou 510000, China (e-mail: violahw@126.com).

Mengmeng Liu is a postgraduate student at the School of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, China (e-mail: zerolmm@126.com).

Fig. 1. Some food images with complex backgrounds

The first component employs a Global Representation Network (G-Net), which classifies the original image to extract global features and identifies primary discriminative regions through adaptive threshold segmentation. Next, an Erased Image Classification Network (E-Net) classifies erased images to adversarially mine discriminative regions, whereas a Discriminative Region Classification Network (D-Net) processes cropped and upsampled regions to extract local features. The feature fusion component integrates a fully connected layer, which classifies concatenated global (original image) and local (mined regions) feature representations. The overall network architecture is depicted in Figure 2.

A. Global Feature Learning

Food images from different sub-classes exhibit significant visual differences, so can be better recognised using global representations. Inspired by [24], the G-Net is based on the existing ResNet model, and classifies the full input image I_i. Global Average Pooling (GAP) is employed on the output f_g of the final convolutional layer of the G-Net in order to extract the global features f_{glo}:

$$f_{glo} = \text{GAP}(f_g) \tag{1}$$

B. Adaptive Mining of Discriminative Regions

Food images often contain complex backgrounds, such as utensils or environmental textures, which introduce visual noise unrelated to the target food. Therefore, greater emphasis should be placed on fine-grained local features. In the proposed model, AE is iteratively employed to accomplish two key tasks: (1) a classification network is trained to localize discriminative regions, and (2) the mined regions are adversarially erased. In the AE method, an adaptive threshold segmentation is embedded to run on the associated upsampled class activation map (CAM) [27] that is generated from the final convolutional layer of some sub-networks in our model. The CAM is a visualization technique highlighting the regions which can assist the classification network to recognize the target class. The discriminative regions can be obtained by thresholding the upsampled CAM. The progressive mining of new discriminative regions is ensured by AE, enabling the model to progressively attend to distinct features of the target object across iterations.

First, the generated CAM is upsampled to produce a heatmap. Since the CAM originates in the sub-network's last convolutional layer, the process is defined as follows:

$$\text{CAM}(I_i, y_i) = \sum_{k=1}^{n} w_{k,c} \cdot \text{feature\_map}_k \tag{2}$$

where y is the target class of image I_i, feature_map_k is the k-th feature map from the last convolutional layer of some sub-networks, with n total feature maps, and $w_{k,c}$ represents the final linear transformation weights, indicating the importance of the k-th neuron in the GAP layer towards identifying y for I_i. During training, y is consistently used as the ground truth label associated with the image I_i. In contrast, during the testing stage, y refers to the predicted label assigned to the image I_i.

The inter-class variance $\sigma_B^2(t)$ is then computed to generate the heatmap. For a given threshold t, this metric measures the variance between background and foreground classes. The process is formulated as follows:

$$\sigma_B^2(t) = p_0(t) \cdot \mu_0(t) + p_1(t) \cdot \mu_1(t) \tag{3}$$

where t denotes the threshold, $p_0(t)$ and $p_1(t)$ are the proportions of pixels below and above the threshold t respectively. Similarly, $\mu_0(t)$ and $\mu_1(t)$ are the average grayscale values of the pixels below and above the threshold t respectively.

Next, inter-class variances are computed for all candidate thresholds, the maximum value is selected, and the corresponding optimal threshold t^* is determined, the foreground and background of the image are effectively separated, ensuring both accuracy and effectiveness in segmentation. The method for finding the optimal threshold is formulated as:

$$t^* = \arg\max_t \sigma_B^2(t) \tag{4}$$

Finally, by applying t^* to segment the heatmap into foreground and background, a binary image is generated. This method adaptively separates food regions from complex backgrounds, reducing background noise interference during the classification process and thereby obtaining discriminative regions.

Given the number of times M for discriminative region mining and the mining step m ($m \leq M$), the GAP layer is used to extract varied and compatible local features f_m^{loc}:

$$f_m^{loc} = \text{GAP}(f_m) \tag{5}$$

where f_m denotes the output from the mining step m of the D-Net.

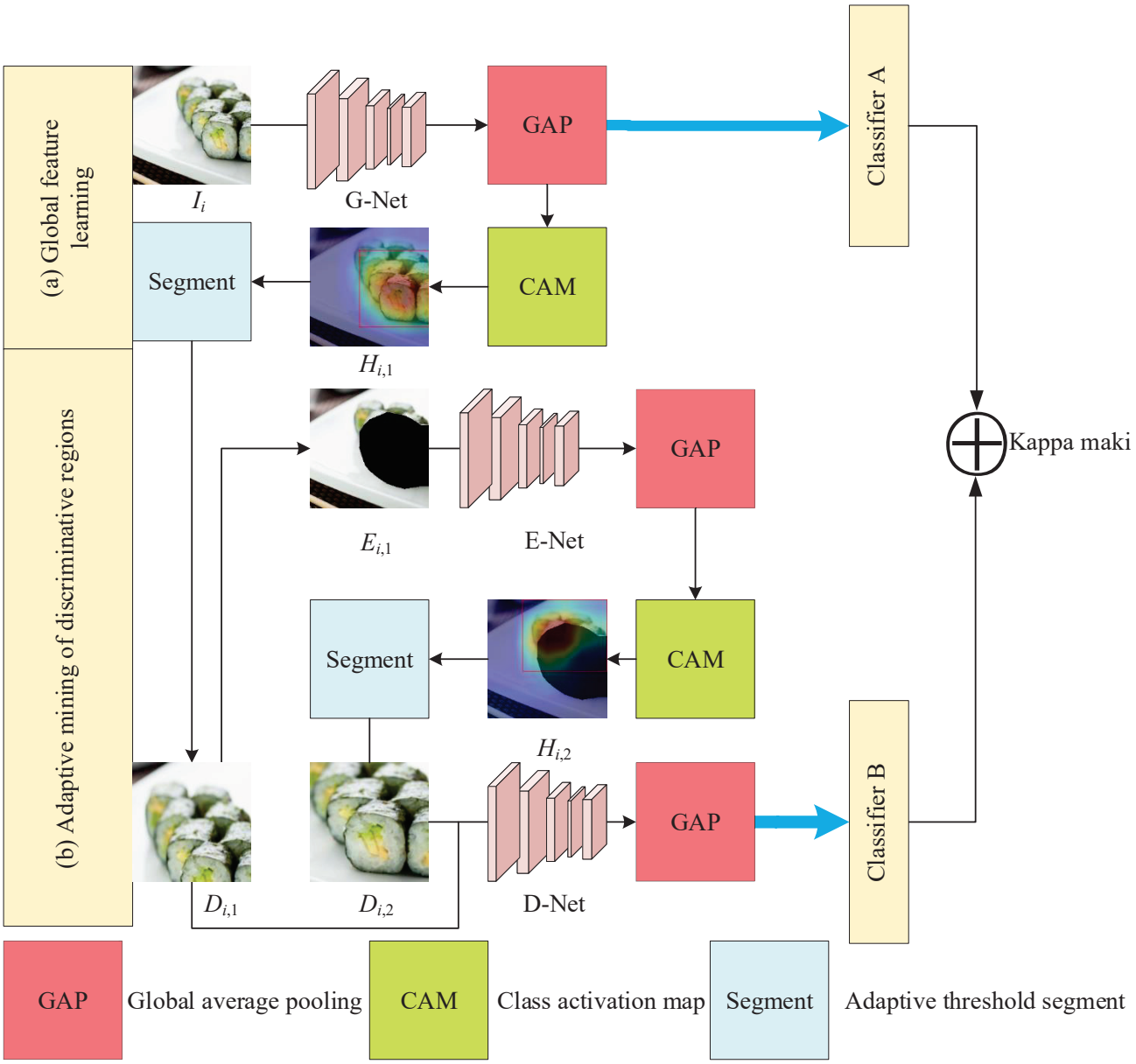

Fig. 2. The framework of ARANet. (a) Global Feature Learning branch extracts global features through a standard CNN backbone. (b) Adaptive Region Mining branch captures discriminative local features through a multi-stage attention mechanism. The predictions from both branches are fused through a learnable weighting layer for final classification.

C. Region Feature Fusion

In the training phase, global and local features are obtained and concatenated. These are then fed into a classification layer to produce the final output f_{concat} after feature fusion:

$$f_{concat} = \text{concat}(f_{glo}, f_m^{loc}) \tag{6}$$

In the testing stage, the same steps are followed.

D. Training and Testing

Throughout the training procedure, the cross-entropy loss function is employed to optimize all classification tasks.

$$L_{CE} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{7}$$

where N denotes the sample cardinality and C represents the category dimension, $y_{i,c}$ represents the ground truth label and $\hat{y}_{i,c}$ denotes the predicted probability for sample i belonging to class c.

Specifically, considering an image-label pair I_i, y_i as input along with the region mining iteration count M, I_i is processed through the G-Net to perform classification, with the corresponding loss represented as L_g. If $M \geq 1$, the G-Net computes CAM(I_i, y_i) and resizes it to produce a heatmap emphasizing distinctive regions. The CAM-based heatmap, extracted region, and erased image at mining step m ($m \leq M$) are designated as $H_{i,m}$, $D_{i,m}$, and $E_{i,m}$. Both $D_{i,m}$ and $E_{i,m}$ retain the original label y_i from I_i. Thus, following resizing, the initial heatmap $H_{i,1}$ is acquired, matching I_i in dimensions. The adaptive threshold segmentation method is initially employed on $H_{i,1}$, succeeded by connectivity examination to detect linked areas. For every area, the cumulative pixel intensity is determined, with the area showing the maximum intensity chosen as

the primary distinctive area in I_i. This primary distinctive area emphasized in $H_{i,1}$ is labeled as $D_{i,1}$, which might display an irregular shape. This region is then cropped using a tight bounding box. The resulting patch is upsampled to the dimensions of I_i via bilinear interpolation and subsequently processed by the D-Net for recognition.

When $M > 1$, the E-Net iteratively identifies subsequent discriminative regions by computing CAM($E_{i,m}, y_i$) for $m \in 1, \ldots, M-1$, while simultaneously classifying each $E_{i,m}$. Meanwhile, the D-Net continues to recognize the new discriminative regions $D_{i,m}$, $m \in \{2, \ldots, M\}$ fed from the E-Net. Simultaneously, during the first erasing operation, the pixels within $D_{i,1}$ in I_i are replaced with zeros, thereby yielding the initial discriminative erased image $E_{i,1}$. The image is then entered into the E-Net for classification. For every classification performed by the D-Net and E-Net at step m, the corresponding losses are designated as $L_{d,m}$ and $L_{e,m}$, respectively. Representations of the input image and each region are extracted from the GAP layers of the G-Net and D-Net. These representations are concatenated to form a robust and comprehensive feature representation, processed through an additional layer for classification. The resulting loss is denoted as L_{concat}. The total loss L is defined as the sum of L_g, $L_{d,m}$, $L_{e,m}$, and L_{concat}:

$$L = L_g + \sum_{m=1}^{M} L_{d,m} + \sum_{m=1}^{M-1} L_{e,m} + L_{concat} \tag{8}$$

As M increases, the number of remaining discriminative regions available to the E-Net for the purpose of identifying the correct class decreases. In this study, the number of mining iterations is set to $M = 3$, resulting in the discriminative region images $R_{i,1}$, $R_{i,2}$, and $R_{i,3}$, along with the erased images $E_{i,1}$ and $E_{i,2}$. It is therefore proposed that the total loss of the model be defined as follows:

$$L = L_g + \sum_{m=1}^{3} L_{d,m} + \sum_{m=1}^{2} L_{e,m} + L_{concat} \tag{9}$$

During the training process, ARANet employs label smoothing to adjust and optimize the model parameters, thereby mitigating class imbalance in the data and enhancing recognition performance. For a recognition task with C classes, given the true label y, the smoothed label y_{smooth} can be expressed as:

$$y_{smooth} = (1 - \epsilon) \cdot y + \frac{\epsilon}{C} \tag{10}$$

where ϵ is the smoothing coefficient, typically set to 0.1–0.6, and y is the original one-hot label.

III. Results and Discussion

All experiments are conducted using the PyTorch deep learning framework on a Linux platform. A hardware configuration consisting of four NVIDIA TITAN Xp GPUs (12GB VRAM each) is employed for parallel processing. Following [24], the model's performance is evaluated using both Top-1 and Top-5 classification accuracy metrics. Through epoch-wise analysis, superior recognition performance is consistently observed across all food recognition tasks. The model's effectiveness is further verified through accuracy curves generated during the testing phase.

A. Datasets

Experimental validation is conducted on five benchmark datasets to assess the method's effectiveness:

(1) Sushi-50 [28] is utilized as a fine-grained evaluation dataset, containing 3,963 Google-sourced sushi images distributed across 50 categories, with standardized 1:1 training-test splits being employed.

(2) The novel Central Asian Food Dataset (CAFD) [29] is introduced, comprising 16,499 carefully curated images spanning 42 unique Central Asian cuisine categories.

(3) Food-11 [30] is included as a baseline evaluation set, featuring 16,643 images representing 11 common daily food categories.

(4) ETH Food-101 [25] is incorporated as a large-scale benchmark, consisting of exactly 100,000 images, with a standardized split of 750 training samples and 250 test samples per class across 101 food categories.

(5) Vireo Food-172 [31] is adopted as the most challenging evaluation set, containing 110,241 professionally captured Chinese food images organized into 172 categories.

B. Implementation Details

The ARANet framework employs ResNet-based sub-networks, with various combinations of sub-networks at different ResNet depths investigated. During the training phase, input images are preprocessed through random cropping and resizing to 224×224 pixels, followed by horizontal flipping with a 50% probability. The experimental setup parameters for model training are tabulated in Table I.

TABLE I
Experimental parameter settings

Parameter	Value
Initial Learning Rate	0.001
Learning Rate Step	14
Learning Rate Decay Factor	0.1
Momentum	0.9
Weight Decay	0.0001
Batch Size	8
Number of Epochs	60
Optimizer	Stochastic Gradient Descent (SGD)
Loss Function	Cross Entropy Loss

C. Experimental Results on Five Datasets

To verify ARANet's generalization ability and robustness, comprehensive experiments were conducted across five benchmark datasets. The proposed model's performance was systematically compared against several state-of-the-art approaches, including ResNet-50, ResNet-101, YOLOv8, YOLOv11, and PAR-Net. Comprehensive benchmarks were conducted across five diverse datasets: Sushi-50, CAFD, Food-11, ETHZ Food-101, and Vireo Food-172, with detailed results documented in Table II.

As shown in Table II, ARANet exhibits remarkable performance across both evaluation metrics (Top-1 and Top-5 accuracy). For instance, on the Vireo Food-172 dataset, ARANet achieves improvements of 0.1% and 0.3% in Top-1 accuracy, and 0.7% and 0.3% in Top-5 accuracy, compared to the state-of-the-art YOLO-series models YOLOv8 and YOLOv11, respectively. When compared to the ResNet-50

TABLE II
EXPERIMENTAL RESULTS ON FIVE DATASETS

Method	Sushi-50		CAFD		Food-11		ETH Food-101		Vireo Food-172	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
YOLOv11	88.7	98.0	87.5	97.2	93.4	94.4	89.2	97.4	90.3	98.4
ResNet-50	89.0	98.5	82.0	97.3	93.6	98.0	87.1	97.3	88.0	97.3
ResNet-101	89.8	98.5	82.4	97.4	94.2	98.6	88.1	97.6	88.3	97.8
YOLOv8	91.3	98.7	86.8	96.8	93.4	99.0	88.3	96.8	90.3	98.0
PAR-Net	92.0	98.6	83.6	97.4	94.9	99.2	89.6	97.6	90.0	97.9
ARANet (Ours)	92.3	99.1	83.8	97.6	95.4	99.4	90.0	98.0	90.4	98.7

TABLE III
TESTING ACCURACIES (%) OF THE THREE RESNET COMBINATION MODELS ON THE FIVE EXPERIMENTAL DATASETS

Method	Sushi-50		CAFD		Food-11		ETH Food-101		Vireo Food-172	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
G50+E34+D50	90.2	98.9	83.0	97.3	94.9	98.8	89.5	97.6	89.9	98.0
G101+E34+D50	91.6	98.3	83.6	97.2	94.9	98.9	89.6	97.9	89.6	97.9
G101+E101+D101	92.3	98.8	83.6	97.3	95.0	99.0	89.9	98.0	90.3	98.5

TABLE IV
EXPERIMENTAL RESULTS OF ARANET UNDER DIFFERENT LABEL SMOOTHING PARAMETERS

ϵ	Sushi-50		CAFD		Food-11		ETH Food-101		Vireo Food-172	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
0	92.3	98.8	83.6	97.3	95.0	99.0	89.9	98.0	90.3	98.5
0.1	92.3	99.1	83.7	97.5	95.2	99.2	89.9	98.1	90.4	98.6
0.2	92.3	99.1	83.8	97.6	95.4	99.4	90.0	98.0	90.4	98.7
0.3	92.2	99.0	83.7	97.3	95.1	99.0	89.9	98.0	90.3	98.6
0.4	92.3	99.0	83.6	97.4	95.3	99.2	90.0	98.1	90.4	98.6
0.5	92.3	98.9	83.6	97.6	95.1	99.2	89.9	98.0	90.3	98.5
0.6	92.2	99.1	83.6	97.6	95.3	99.2	89.9	98.0	90.3	98.5

TABLE V
EXPERIMENTAL RESULTS OF ARANET UNDER DIFFERENT MINING STEP PARAMETERS M

M	Sushi-50		CAFD		Food-11		ETH Food-101		Vireo Food-172	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
2	92.0	98.6	83.6	97.2	95.0	98.0	89.7	97.6	90.0	98.2
3	92.3	99.1	83.8	97.6	95.4	99.4	90.0	98.0	90.4	98.7
4	92.3	99.0	83.9	97.6	95.3	99.4	90.0	97.8	90.2	98.8

and ResNet-101 models, ARANet exhibits enhancements of 2.4% and 2.1% in Top-1 accuracy, and 0.6% and 0.9% in Top-5 accuracy. The PAR-Net model, referenced in [28], employs a fixed-threshold segmentation method to extract discriminative regions. In comparison to PAR-Net, ARANet achieves a 0.4% enhancement in Top-1 accuracy and a 0.8% enhancement in Top-5 accuracy.

D. Results and Discussion of Combined Models, Sub-Networks, and Region Mining Iterations M

The proposed ARANet architecture comprises three principal sub-networks: (1) the G-Net, which performs classification on the full input image; (2) the E-Net, designed for processing erased images; and (3) the D-Net, specialized for classifying mined discriminative regions. To evaluate architectural variations, multiple ResNet configurations were implemented, including G50+E34+D50, G101+E34+D50, and G101+E101+D101 (where G50 denotes a G-Net based

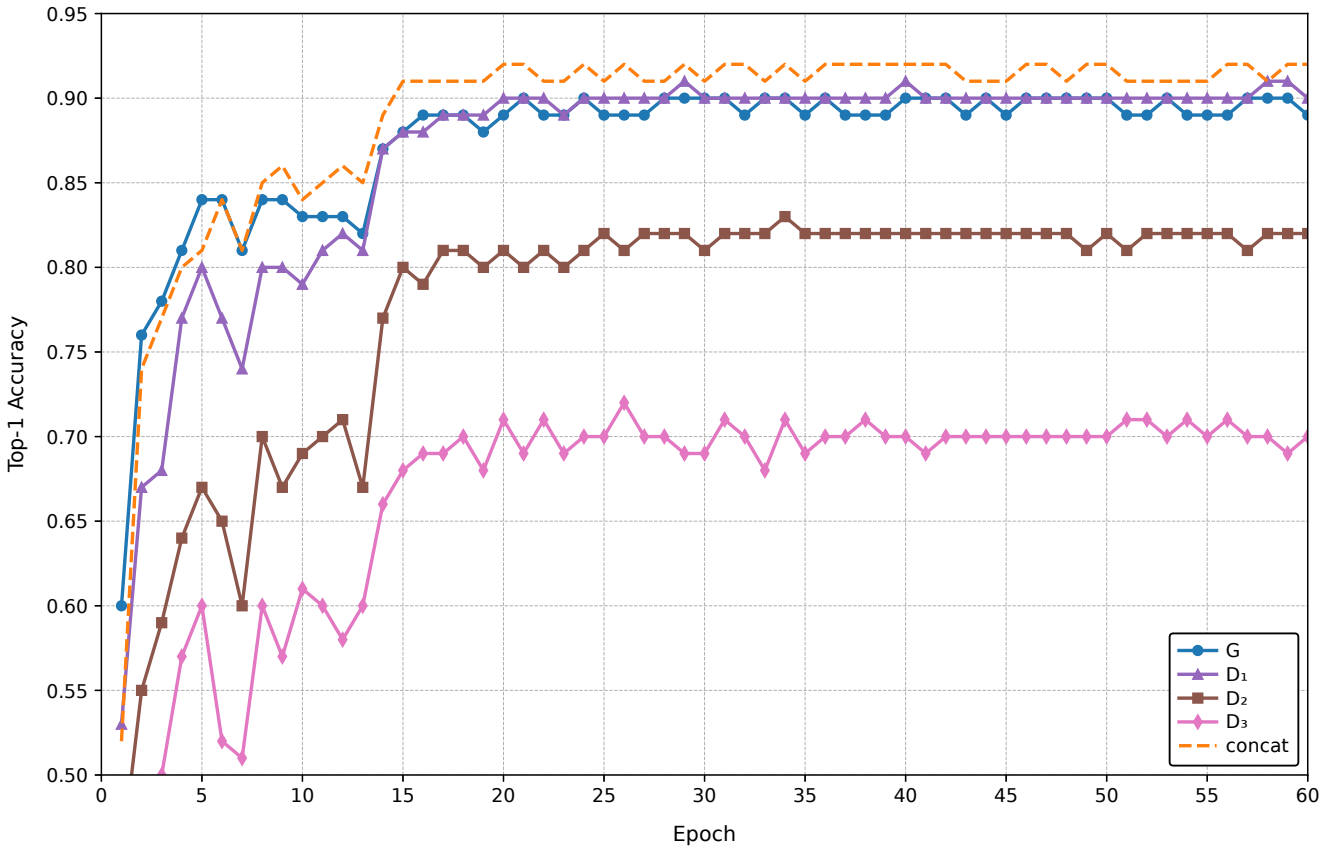

Fig. 3. Comparison of Top-1 accuracies of three sub-networks

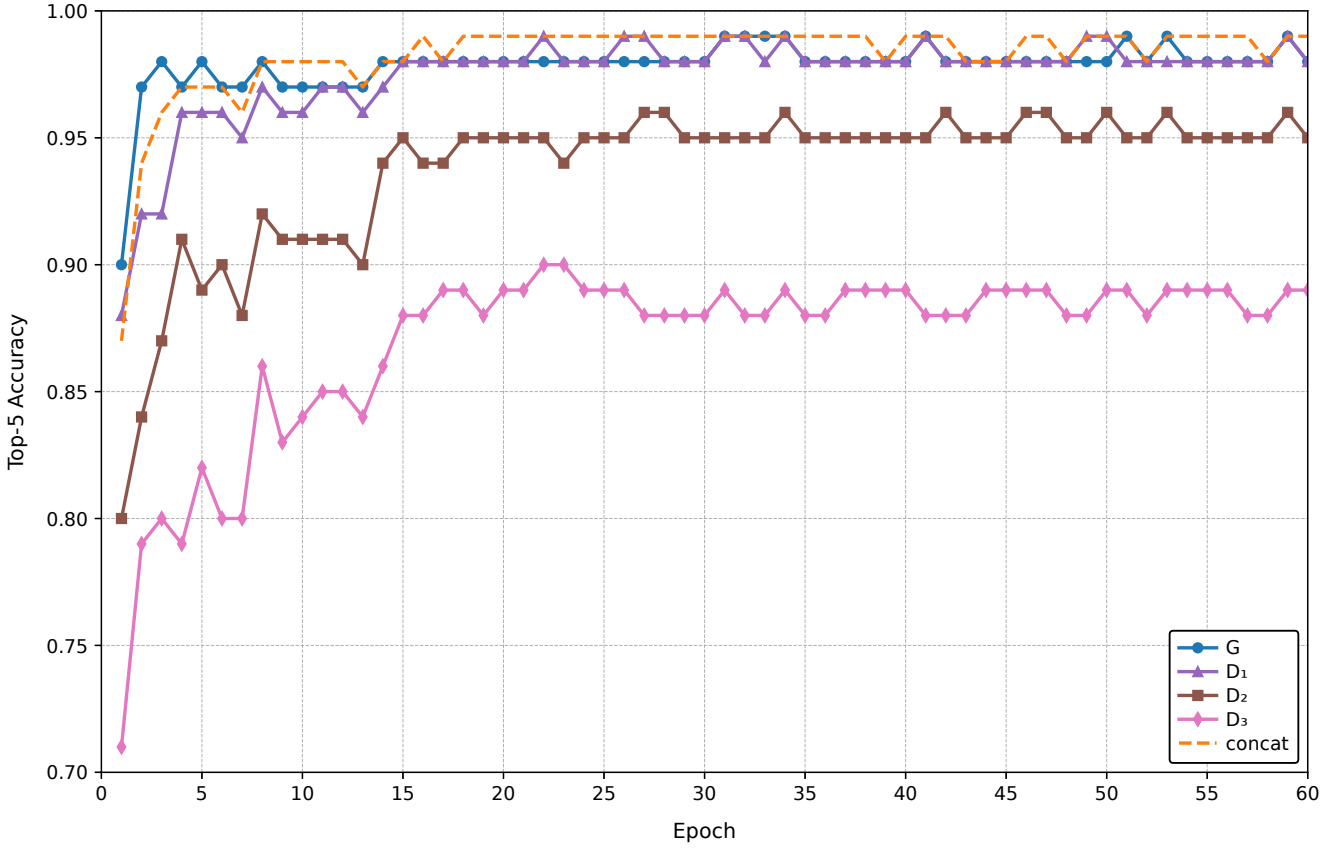

Fig. 4. Comparison of Top-5 accuracies of three sub-networks

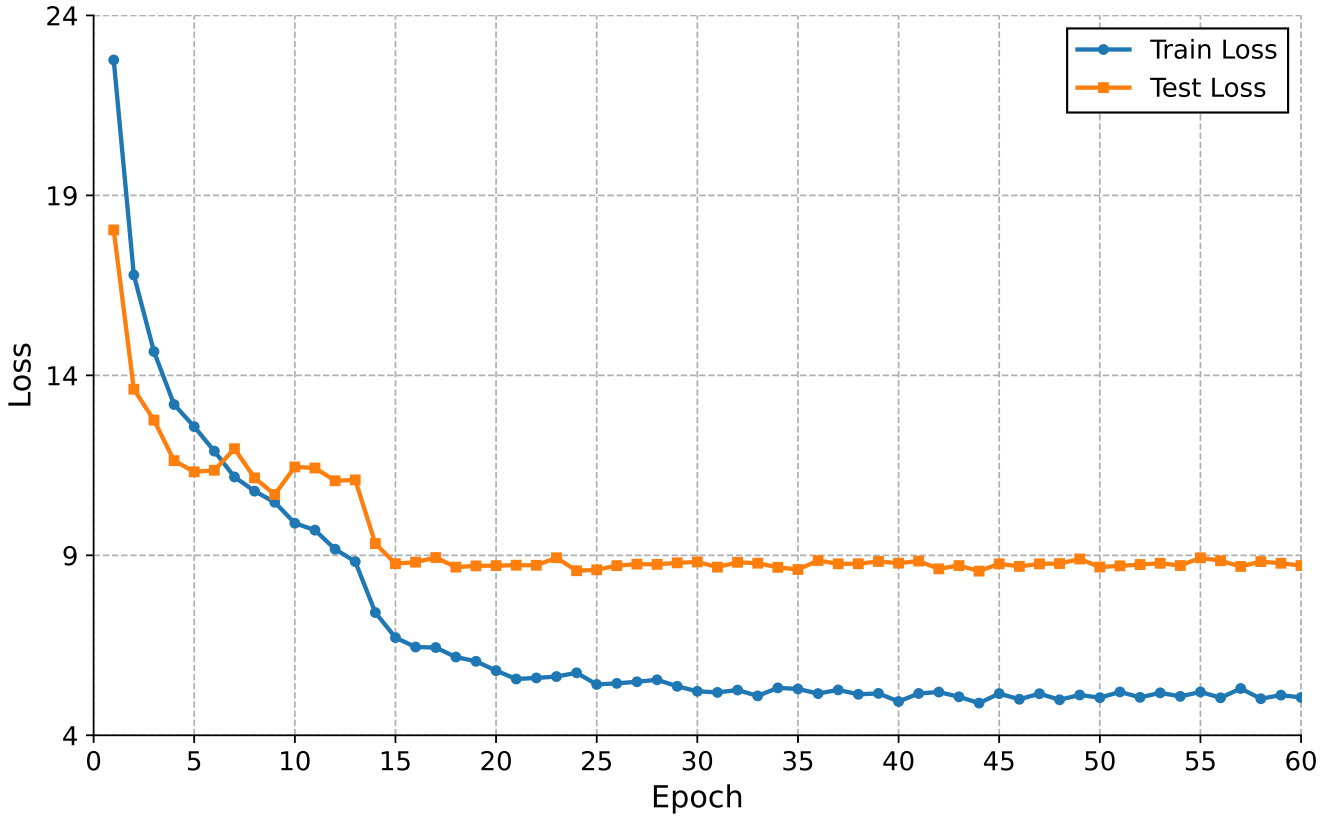

Fig. 5. The loss curve of ARANet for the Sushi-50 dataset

on ResNet-50, etc.). The results of these three models across the five experimental datasets are presented in Table III. As demonstrated in Table III, the ARANet configuration with G101+E101+D101 architecture achieves superior performance compared to other variants. Based on these results, this optimal configuration was selected for all subsequent experimental evaluations.

As demonstrated in Table 4, the experimental findings substantiate that the implementation of label smoothing enhances the performance of ARANet, validating its effectiveness and generalization capability. When $\epsilon = 0.2$, the model exhibits the most stable performance. On the Food-11 dataset, ARANet reaches peak Top-1 accuracy of 95.4% and Top-1 accuracy of 99.4% at $\epsilon = 0.2$, outperforming other ϵ values by 0.3% and 0.4%, Therefore, in subsequent experiments, ϵ is fixed at 0.2.

To comprehensively evaluate the performance improvements achieved by ARANet, an ablation study is conducted to analyze the contribution of the concatenated representation. The experimental results, presented in Figures 3 and 4, demonstrate the effectiveness of this architectural component. First, as mining continues, the classification performance for the mined regions declines as expected ($D_1 > D_2 > D_3$), because the later mined regions are less distinctive than the earlier ones. This trend is also observed in the erased images, where accuracy diminishes ($E_1 > E_2$) as the later images contain fewer characteristic regions compared to the prior ones. Secondly, the performance of the concatenated representation is consistently superior to that of either the original image or the mined regions alone. It confirms that the integration of global and local representations leads to a more robust and comprehensive feature representation, ultimately contributing to improved food recognition performance. Figure 5 illustrates the loss curve of ARANet on the Sushi-50 dataset. As shown in Figures 3, 4, and 5, as the training iterations progress, the loss metric steadily declines, while the primary and secondary classification accuracies consistently rise, demonstrating the enhanced performance of ARANet in food recognition tasks.

Several prediction results from the five datasets are visualized, with two examples selected from each dataset, as shown in Figure 6. The CAM-based heatmap $M_{i,1}$ has been projected. Each discriminative region $D_{i,m}$ has been extracted via the red annotation box and resized to align with the dimensions of the original image I_i. The masked zones in $M_{i,2}$ and $M_{i,3}$ represent the erased regions. I_i is classified by the G-Net and $D_{i,m}$ is classified by the D-Net. In I_i/Top-1, Top-1 is the prediction based on the G-Net. In $D_{i,m}$/Top-1, Top-1 is the prediction based on the D-Net. Concat. Top-1 is the prediction based on the concatenated representation. GT means the ground truth. The foods labeled by red color label are not correctly classified in the Top-1 results. In the first and tenth rows, both the input and the extracted regions are accurately identified by the G-Net and D-Net, respectively, leading to the target prediction after the concatenation of their representations. In the second to eighth rows, the input images are misclassified via the G-Net, whereas the extracted salient regions are accurately identified via the D-Net. Consequently, the target classification derived from the fused representation is achieved. Furthermore, Figure 6 highlights the diversity of discriminative regions mined

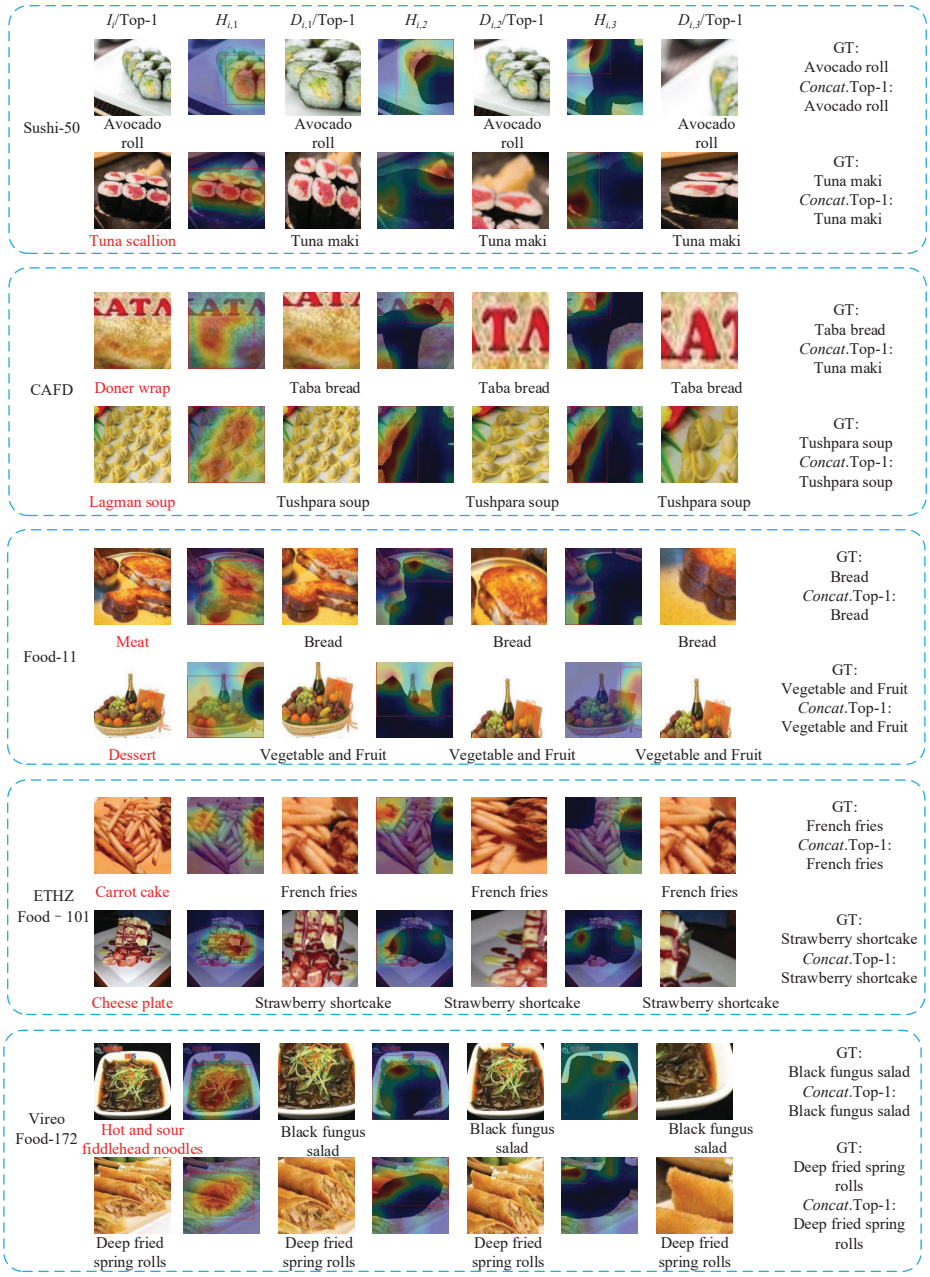

Fig. 6. Visualization of the predicted results of the ARANet on the five datasets

from the images, contributing to improved classification outcomes. These findings validate the utility of leveraging discriminative regions for food recognition improvement and showcase our approach's robust performance across diverse food datasets.

To investigate the optimal number of region mining iterations required for ARANet to achieve peak performance, systematic ablation studies are performed to assess the influence of different values for the number of mined regions M. As evidenced by Table V, optimal performance is achieved when mining three discriminative regions, with reduced accuracy observed for fewer regions. While a marginal improvement in accuracy is noted on the CAFD dataset when increasing to four regions, this configuration introduces significant computational overhead. Consequently, $M = 3$ was chosen as the best compromise between recognition performance and computational efficiency for all following experiments.

IV. Conclusions

In this paper, a novel food recognition model is proposed that employs an adaptive threshold segmentation method to progressively extract distinctive food regions. The representations of these regions are integrated with the global features of the full input image to achieve accurate predictions. This approach generates more robust and comprehensive representations, effectively addressing the distinctive visual complexities inherent in food images. Future work will incorporate ingredient information to further enhance model performance.

References

[1] M. Gerasimchuk and A. Uzhinskiy, "Food recognition for smart restaurants and self-service cafes," *Physics of Particles and Nuclei Letters*, vol. 21, no. 1, pp. 79–83, 2024.

[2] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5635–5662, 2024.

[3] M. Y. Ansari and M. Qaraqe, "Mefood: A large-scale representative benchmark of quotidian foods for the middle east," *IEEE Access*, vol. 11, pp. 4589–4601, 2023.

[4] A. K. Y. Chan, C. H. Chu, H. Ogawa, and E. H.-H. Lai, "Improving oral health of older adults for healthy ageing," *Journal of Dental Sciences*, vol. 19, no. 1, pp. 1–7, 2024.

[5] D. Mozaffarian, K. E. Aspry, K. Garfield, P. Kris-Etherton, H. Seligman, G. P. Velarde, K. Williams, E. Yang, A. P. of Cardiovascular Disease Section Nutrition, L. W. Group, and D. of Care Working Group, ""food is medicine" strategies for nutrition security and cardiometabolic health equity: Jacc state-of-the-art review," *Journal of the American College of Cardiology*, vol. 83, no. 8, pp. 843–864, 2024.

[6] R. Ramesh and G. A. Joseph, "The optimal control methods for the covid-19 pandemic model's precise and practical siqr mathematical model." *IAENG International Journal of Applied Mathematics*, vol. 54, no. 8, pp. 1657–1672, 2024.

[7] G. K. Folson, B. Bannerman, V. Atadze, G. Ador, B. Kolt, P. McCloskey, R. Gangupantulu, A. Arrieta, B. C. Braga, J. Arsenault *et al.*, "Validation of mobile artificial intelligence technology–assisted dietary assessment tool against weighed records and 24-hour recall in adolescent females in ghana," *The Journal of Nutrition*, vol. 153, no. 8, pp. 2328–2338, 2023.

[8] V. Bettadapura, E. Thomaz, A. Parnami, G. D. Abowd, and I. Essa, "Leveraging context to support automated food recognition in restaurants," in *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, HI, USA: IEEE, 2015, pp. 580–587.

[9] H. He, F. Kong, and J. Tan, "Dietcam: multiview food recognition using a multikernel svm," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 3, pp. 848–855, 2015.

[10] P. Pouladzadeh and S. Shirmohammadi, "Mobile multi-food recognition using deep learning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 3s, pp. 1–21, 2017.

[11] Z.-Y. Ming, J. Chen, Y. Cao, C. Forde, C.-W. Ngo, and T. S. Chua, "Food photo recognition for dietary tracking: System and experiment," in *24th International Conference on MultiMedia Modeling (MMM 2018)*. Bangkok, Thailand: Springer, 2018, pp. 129–141.

[12] P. McAllister, H. Zheng, R. Bond, and A. Moorhead, "Combining deep residual neural network features with supervised machine learning algorithms to classify diverse food image datasets," *Computers in Biology and Medicine*, vol. 95, pp. 217–233, 2018.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, 2015, pp. 1–9.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, 2016, pp. 770–778.

[15] A. Fakhrou, J. Kunhoth, and S. Al Maadeed, "Smartphone-based food recognition system using multiple deep cnn models," *Multimedia Tools and Applications*, vol. 80, no. 21, pp. 33 011–33 032, 2021.

[16] S. Alshomrani, L. Aljoudi, B. Aljabri, and S. Al-Shareef, "Food detection by fine-tuning pre-trained convolutional neural network using noisy labels," *International Journal of Computer Science & Network Security*, vol. 21, no. 7, pp. 182–190, 2021.

[17] M. Chun, H. Jeong, H. Lee, T. Yoo, and H. Jung, "Development of korean food image classification model using public food image dataset and deep learning methods," *IEEE Access*, vol. 10, pp. 128 732–128 741, 2022.

[18] E. Tasci, "Voting combinations-based ensemble of fine-tuned convolutional neural networks for food image recognition," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 30 397–30 418, 2020.

[19] H. Yang, S. Kang, C. Park, J. Lee, K. Yu, and K. Min, "A hierarchical deep model for food classification from photographs," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 14, no. 4, pp. 1704–1720, 2020.

[20] S. A. Ayon, C. Z. Mashrafi, A. B. Yousuf, F. Hossain, and M. I. Hossain, "Foodiecal: A convolutional neural network based food detection and calorie estimation system," in *2021 National Computing Colleges Conference (NCCC)*. Taif, Saudi Arabia: IEEE, 2021, pp. 1–6.

[21] S. Elbassuoni, H. Ghattas, J. El Ati, Z. Shmayssani, S. Katerji, Y. Zoughbi, A. Semaan, C. Akl, H. B. Gharbia, and S. Sassi, "Deepnova: A deep learning nova classifier for food images," *IEEE Access*, vol. 10, pp. 128 523–128 535, 2022.

[22] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, and S. Jiang, "Large scale visual food recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9932–9949, 2023.

[23] F. Shuang, Z. Lu, Y. Li, C. Han, X. Gu, and S. Wei, "Foodnet: Multi-scale and label dependency learning-based multi-task network for food and ingredient recognition," *Neural Computing and Applications*, vol. 36, no. 9, pp. 4485–4501, 2024.

[24] Y. Yang, W. Min, J. Song, G. Sheng, L. Wang, and S. Jiang, "Lightweight food recognition via aggregation block and feature encoding," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 10, pp. 1–25, 2024.

[25] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *13th European Conference on Computer Vision (ECCV 2014)*. Zurich, Switzerland: Springer, 2014, pp. 446–461.

[26] J. Ji, L. Jiang, T. Zhang, W. Zhong, and H. Xiong, "Adversarial erasing attention for fine-grained image classification," *Multimedia tools and applications*, vol. 80, pp. 22 867–22 889, 2021.

[27] Q. Chao, X. Wei, J. Tao, C. Liu, and Y. Wang, "Cavitation recognition of axial piston pumps in noisy environment based on grad-cam visualization technique," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 1, pp. 206–218, 2023.

[28] J. Qiu, F. P.-W. Lo, Y. Sun, S. Wang, and B. Lo, "Mining discriminative food regions for accurate food recognition," *arXiv preprint arXiv:2207.03692*, vol. arXiv:2207.03692, 2022.

[29] A. Karabay, A. Bolatov, H. A. Varol, and M.-Y. Chan, "A central asian food dataset for personalized dietary interventions," *Nutrients*, vol. 15, no. 7, pp. 1728–1729, 2023.

[30] A. Singla, L. Yuan, and T. Ebrahimi, "Food/non-food image classification and food categorization using pre-trained googlenet model," in *2nd International Workshop on Multimedia Assisted Dietary Management (MADiMa 2016)*. Amsterdam, Netherlands: ACM, 2016, pp. 3–11.

[31] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *24th ACM International Conference on Multimedia (MM 2016)*. Amsterdam, Netherlands: ACM, 2016, pp. 32–41.