Food Recognition with Visual Transformers

Simone Bianco, Marco Buzzelli, Gaetano Chiriaco, Paolo Napoletano, Flavio Piccoli
Department of Informatics, Systems and Communication
University of Milano-Bicocca
Milano, Italy
Email: {first_name.last_name}@unimib.it

***Abstract*—Food recognition is a major challenge in the field of computer vision, requiring models that can effectively handle the wide variability and complexity of food images. In this paper, we explore the use of vision transformers, a category of models based on self-attention mechanisms, to address the task of food recognition. We focus on training and fine-tuning different vision transformer architectures on Food2K, a large-scale dataset of food images with 2,000 categories. We compare the performance of vision transformers with convolutional neural networks (CNNs) on Food2K and Food101. In addition, we use state-of-the-art explainability techniques to highlight the regions of interest that vision transformers take into account when performing a prediction. Our results show that vision transformers can achieve competitive results on food recognition tasks, with the added benefit that pre-training on Food2K improve their generalization capabilities and interpretability. This study highlights the potential of vision transformers in food computing, paving the way for future research in this field.**

***Index Terms*—food recognition, vision transformers, ViT, CNNs**

I. Introduction

Food recognition is an essential task of computer vision that involves identifying and categorizing several types of foods using visual data, such as images or videos [1], [2].

Food recognition can have a major impact on our day-to-day life and improve or automate processes like nutrition tracking [3], food authentication [4] and food waste management [5]. The field of food recognition research encompasses a wide range of methods and approaches, making it a diverse and complex area of study. The used classifiers range from machine learning algorithms based on handcrafted features, to deep neural networks. An ideal classifier should be robust, meaning it can handle situations where the food dish deviates from its typical presentation. Food items within the same category can exhibit significant intra-class variation, caused by differences in cooking styles, plating and portion sizes [6].

In such a complex context, simple classification methods are generally not effective. At the time of writing, deep neural networks are the only valid options when choosing an algorithm to handle and classify food images. In particular, deep CNNs have demonstrated to be extremely powerful and to achieve competitive performance in most computer vision task. However, in the last years, transformer models, initially proposed to solve NLP tasks, have been adapted and used also in computer vision. Their demonstrated robustness to perturbations [7] and their ability to capture long-range dependencies can be particularly helpful in extracting robust representations of food images. On the other hand, vision transformers require a large amount of training data to work properly [8]. The field of food recognition is rich of datasets, however their dimensionality is not comparable to the datasets used to train transformers (e.g., ImageNet-21K [9]).

Nevertheless, a new large dataset on food images, called Food2K [10], was recently released, and large convolutional models trained on the proposed dataset have shown competitive results. This opened up the possibility of training vision transformers for the first time on large datasets containing only food images. Therefore, this study was conducted to answer the following research questions:

- Are vision transformers better than CNNs for Food2K recognition?
- Are predictors trained on Food2K better then ImageNet-21K weights on food-related tasks?
- Can explainability techniques be applied to vision transformers to identify the key regions of a dish without explicitly training for segmentation?

II. Related Works

With the continuous progress of computer vision, convolutional neural networks (CNNs) have emerged as the dominant model for food recognition. Recent popular models like EfficientNet [11] score a Top-1 Accuracy of 92.98% on Food101 with pre-training on ImageNet and 96.18% pre-training on JFT-300M. In order to further improve the results achieved on food datasets, PAR-Net [12], a convolutional network architecture using adversarial erasing and class activation maps, was introduced to integrate local and global features and to highlight discriminative regions in food images. In the last years, researchers began to explore the application of transformers in computer vision, i.e., vision transformers. In several computer-vision tasks, vision transformers have outperformed CNNs. In the context of food recognition, they have shown promising results.

The first transformer specifically designed for computer vision tasks is called Vision Transformer (ViT) [8]. The ViT architecture intentionally followed the original transformer model as closely as possible. It demonstrated that the attention mechanism, effective on the "tokens" that make up a document, can be applied in the same way to the "patches" that compose an image.

Recently, research on vision transformers has increased exponentially to exploit the strengths of ViT and to address its

limitation. The core component of transformers, the attention mechanism, has a quadratic computational complexity with respect to the dimensionality of the image, and vision transformer are "data-hungry" [8], i.e. they need a large amount of data to correctly generalize. One of the most popular and successful alternatives to the classical vision transformer proposed to solve its weaknesses is the Shifted-Window (SWIN) Transformer [13].

Swin transformers, using the shifted window attention approach, solve the computational complexity and fine-grained processing problems. Unfortunately, these types of model share the identical drawback of ViT, needing huge amount of data to function properly. Another very popular variant of vision transformers are Data-efficient vision transformers (DeiT) [14]. DeiT extend the original ViT architecture by incorporating a distillation token that interacts with class and patch tokens through self-attention layers. The objective of the distillation token is to reproduce the label predicted by a chosen teacher network, enabling DeiT models to achieve competitive accuracy even with small-sized datasets.

As with CNNs, researchers used the unique characteristics of food images to improve the performance of transformers. An example is the Semantic Center Guided Windows Attention Fusion Framework (SCG-WAFM) for food recognition [15]. In this case, the images are fed into a pre-trained Swin transformer, which assigns a label to the image. Then, using the self-attention mechanism, the discriminative region that should contain the dish is extracted and fed into the same Swin Transformer. The two predictions are then combined by a linear layer that produces the final prediction. The trained models achieved an accuracy of 93.48% on Food-101, surpassing the traditional Swin transformer architecture.

III. Datasets

One of the most comprehensive and challenging datasets recently presented is the Food2K [10], which contains $1,036,564$ color images of food divided in $2,000$ categories. For convenience, these categories are also grouped in super-classes such as vegetables, meat, barbecue and fried food. Some example categories from Food2k are shown in Figure 1. The number of images per category is in the range $[153, 1999]$, showing quite a larger class imbalance compared with existing food datasets, and image size varies from 220 pixels to 597 for both dimensions. A subsample of Food2K, Food1K, has been used for fine-tuning purposes on CNNs and continual learning. To our knowledge, this paper is the first to study the use of vision transformers on the whole Food2K dataset.

The Food-101 dataset [16] consists of a collection of food images belonging to 101 different food categories covering a wide range of dishes from various cuisines, including popular dishes like pizza, sushi, burgers, and salads, among others. Each food category contains ≈1,000 images, resulting in a total of ≈101,000 images.The images were collected from a wide range of sources, including popular cooking websites and photo-sharing platforms. One key aspect that distinguishes the Food-101 dataset is that it focuses on images containing a single dish, ensuring that each image primarily displays a specific food item without any significant clutter or interference from other elements.

Fig. 1. Example of images contained in Food2K from three different classes

IV. Proposed Approach

We propose a number of transformer models to be trained on Food2K, basing our experiments on the current studies on vision transformers models, and how to correctly fine-tune them. The following tasks are performed:

1) *Classification on Food2k dataset*: Starting from ImageNet-21k pre-trained weights, the vision transformers architectures ViT, Swin and DeiT are fine-tuned for a pre-determined number of epochs on Food2K in order to obtain the best possible accuracy on the $2,000$ available classes.
2) *Assessing the value of Food2K weights*: ImageNet-21k weights are publicly available for all the popular convolutional and transformer models and represent a good starting point for most computer vision tasks. One aim of this study is to verify if weights estimated by training on Food2K are a better starting point for food-related tasks then the commonly used ImageNet weights. To assess the viability of these weights, the previously cited transformer models are fine-tuned on the popular Food101 dataset [16], starting from ImageNet-21K and Food2K weights.
3) *Explainability for vision transformers*: Other than evaluating the models in terms of Top-1 and Top-5 accuracy, this study explores how vision transformer "see" and to which region of the dish they assign the highest importance. A Layer-wise Relevance Propagation (LRP)-based explainability method for transformers [17] is used to produce a relevancy map, that highlights the pixels of the image that contributed most to make the prediction. The aim is to verify if the relevancy maps obtained using vision transformers that were trained on a food-centric dataset, are preferable to the ones produced from a model trained on a generic dataset. Additionally, the relevancy maps are evaluated to verify if the trained Vision Transformer can also be used for food segmentation, other that only recognition.

The results obtained with the described experiments are compared with the benchmarks available on convolutional neural networks.

A. Data preprocessing and augmentation

Of the $1,036,564$ images in Food2K, $620,124$ are used for training and the rest for testing purposes. Each image is first resized to 224×224 pixels, and then normalized, subtracting each channel for the mean value and dividing it for the standard deviation. The performance of vision transformers is strongly influenced by augmentation and regularization [18]. To obtain the best possible performance in a pre-determined span of epochs whilst avoiding overfitting, data augmentation techniques were used. Random horizontal flip and color jitter are the most commonly used techniques when training vision transformers, while other more complex augmentation techniques are used in the following experiments, including:

- *MixUp* [19]: it blends pairs of input images and their labels using weighted averages. This encourages the model to learn smooth decision boundaries, improving generalization and robustness.
- *CutMix* [20]: it randomly crops patches from two images and pastes them together, blending the labels proportionally to the area of the patches. This reduces over-fitting and enhances model performance with limited data.
- *3-Augment* [21]: one augmentation between grayscale, solarization and gaussian blur is randomly applied to every image during training. The use of this augmentation has proven to be helpful when training ViT.

The preprocessing and augmentation transformations are applied in the same way on both Food2k and Food101, only changing the mean and standard deviation values in the normalization step.

B. Models and training scheme

The vision transformers that are here fine-tuned for food recognition are: Vision Transformer, Swin Transformer and Data-efficient image Transformer. Each of these models is used in the following experiments in their "base" form. The details of each model configuration are provided in Table I. The DeiT model is trained using ResNet152 as "teacher" model.

The Swin Transformer and Vision Transformer model are trained using two different configurations:

- **Base training**: the models are trained using a cross-entropy loss that compares ground truth label y with the corresponding predicted probability p:

$$L_{CE} = -\sum_{c=1}^{M} y_c \log(p_c). \tag{1}$$

- **BERT-assisted training**: the models are trained using both the information coming from the image representation, and the label associated to the image. Given an image X_k and its label Y_k which represents the name of the dish, the embedding model BERT [22] $e(\cdot)$ is used to produce a semantic embedding $t_k \in \mathbb{R}^d$ using the following formula:

$$t_k = \frac{1}{n}\sum_{i}^{n} e(y_{ki}), \tag{2}$$

Model	Emb. size	Patch Size	Layers	Params	IN Top-1 Acc.
ViT-B	768	16x16	12	86M	85.7%
DeiT-B	768	16x16	12	86M	84.2%
Swin-B	128	4x4	24	88M	86.4%

TABLE I
SPECIFICS OF THE TRAINED MODELS

where $y_{k1}, y_{k2}, \ldots, y_{kn}$ is a sequence of n tokens that composes the label Y_k. The Multi-Layer Perceptron layer j of the vision transformers produces an image feature x_{jk}, of the input image X_k. This hidden feature is a vectorial representation in R^d, just like the representation t_k given to the label. The distance between these two representations can be used to learn key semantic information contained in the food label that may guide the image classification. The distance is calculated as:

$$L_{emb} = ||x_{jk} - t_k||^2, \tag{3}$$

and can be added to the cross-entropy loss to influence the training process during the backpropagation, obtaining the following total loss:

$$L_{total} = \alpha L_{CE} + \beta L_{emb}, \tag{4}$$

where L_{CE} is the cross-entropy loss and α and β are the hyper-parameters that balance the influence of the two components. The aim is to obtain a similar latent representation for different images that share a similar name. In the following experiments α and β are set to 0.6 and 0.4 respectively. For ViT, BERT-base is used to produce embeddings of size 768, while for Swin Transformer BERT-large is used to obtain embeddings of size 1024.

Essentially, the trained models are the following:

- ViT-Base using cross-entropy loss (86 million parameters, 5 hours and 50 minutes per epoch)
- ViT-Base using cross-entropy + BERT loss (86 million parameters, 5 hours and 55 minutes per epoch)
- Swin Transformer using cross-entropy loss (88 million parameters, 6 hours per epoch)
- Swin Transformer using cross-entropy + BERT loss (88 million parameters, 6 hours and 5 minutes per epoch)
- DeiT using distillation loss and ResNet50 as "teacher" network (86 million parameters, 5 hours and 50 minutes per epoch)

Each of the previous models is trained using the same hyperparameters. The optimizer used is AdamW [23], using $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with a batch size of 16. Since ImageNet-21K weights are already a good starting point, the initial learning rate is set to the relatively low value of 2×10^{-4}, and divided by half every 7 epochs. All models were trained for 30 epochs, for a total of $1,162,860$ steps. To prevent over-fitting, label smoothing regularization is used and set to 0.1. Training and testing are performed with images of size 224×224 pixels.

The trained vision transformers models are compared with the only other available benchmarks provided by the creators

of Food2K. The available CNN benchmarks are trained for 200 epochs, using stochastic gradient descent as optimizer and a batch size of 2. The starting learning rate of their experiments is set to 1×10^{-2}, and divided by 10 after 30 epochs. Random horizontal flip and color jittering are used for data augmentation. The training and testing resolution is set to 224×224 pixels. To provide a meeting point between the different approaches, a ResNet50 model is trained using our configuration, and the configuration proposed in the Food2K paper [10] for 30 epochs.

V. Results

A. Performance on Food2K

Table II shows the performance obtained with the previously described vision transformers using our configuration. We can see that the recognition performance is slightly superior in terms of Top-1 and Top-5 accuracy using ViT-B compared to the other tested models. DeiT obtained an accuracy notably lower than the other two vision transformers. This is mainly caused by the lack of an available "good teacher" to guide the estimation of the model parameters. The ResNet152 model with pre-trained weights of Food2K only achieves an accuracy of 62.58% on Food2K test data, which makes this model not good enough to correctly guide DeiT model's training. The use of BERT-training did not have a positive effect on the performance of performance ViT, while it slightly increased Top-1 accuracy for Swin-B. Table III shows the performance of popular convolutional models on Food2K. All the listed baselines are superior in terms of Top-1 accuracy and achieve a similar Top-5 accuracy. To provide a way for comparing the results, Table IV shows a comparison of the performance of ResNet50 trained using the same configuration of the models from Table II and the configuration of the models from Table III for 30 and 200 epochs. The results clearly show that increasing the number of epochs can substantially improve the obtained performance. Despite vision transformers showing slightly lower accuracy, an adjustment in the architecture, training time and configuration can possibly bring the accuracy of these models on par with or beyond convolutional models.

Table V illustrates the food categories on which the ViT-B model performed worst. The top nineteen classes for lowest accuracy all have fewer than 500 images, which is lower than the average value of images per Food2K class. Food2K is a long-tailed dataset and this affects accuracy per class, with higher accuracy performance for categories that are represented by many images, and lower accuracy performance for less "populated" classes.

B. Generalization on Food101

The Food101 dataset is used to assess the generalization capabilities of vision transformers models, and to measure if fine-tuning on Food2K improves the classification ability of the models on Food101. ViT-B and Swin-B are fine-tuned on Food101, starting from ImageNet-21K weights and Food2k weights. The training configuration is similar to the one used in the previous section to fine-tune on Food2K. The vision transformers are fine-tuned for 10 epochs, using AdamW as optimizer and a batch size of 16. The starting learning rate is fixed, and set to 2×10^{-4}. Training resolution is set to 224×224 pixels, and the augmentation techniques randomly applied are 3-Augment, color jitter and random horizontal flip.

Table VI shows the performance obtained by fine-tuning vision transformers on Food101, while Table VII shows the generalization capabilities of popular convolutional baselines.

Two key conclusions can be drawn from the results:

- Fine-tuning on Food2k does lead to a significant increase in accuracy on Food101 for ViT-B model, while ImageNet-21K weights are a sufficiently good starting point for food recognition tasks when using Swin Transformer.
- Although CNNs had performed better on Food2K, when fine-tuned on Food101 they tend to perform markedly worse than Vision transformers. As reported in other fields, the generalization capabilities of vision transformers may prove to be superior to CNNs [24].

Models	Params	Top-1 Acc.	Top-5 acc.
ViT-B	86M	78.41%	96.33%
ViT-B + BERT	86M	74.82%	95.12%
Swin-B	88M	77.58%	96.17%
Swin-B + BERT	88M	78.52%	96.09%
DeiT-B	86M	73.45%	94.42%

TABLE II
Performance on Food2K of VT trained for 30 epochs

Models	Params	Top-1 Acc.	Top-5 Acc.
VGG16	136M	78.96%	95.26%
Inception v4	43M	82.02%	96.45%
ResNet50	26M	80.79%	95.74%
ResNet101	45M	81.28%	95.99%
ResNet152	60M	81.95%	96.57%
DenseNet161	29M	81.87%	96.53%
SENet154	256M	83.62%	97.22%

TABLE III
Performance on Food2K of CNNs trained for 200 epochs

Models	Epochs	Top-1 acc.	Top-5 acc.
ResNet50 (our config)	30	62.22%	87.81%
ResNet50 (Food2K config)	30	52.72%	81.24%
	200	80.79%	95.74%

TABLE IV
Comparison of ResNet50 performance on Food2K

Label	Errors	Accuracy	# of Images
Chocolate sundae	225	0.49	274
Stir-Fired pork	202	0.49	340
Artic bay sashimi	179	0.48	474
Fried rice	163	0.48	291
Chocolate cake	152	0.47	324

TABLE V
Worst classified categories of food.

Models	Top-1 Acc.	Top-5 Acc.
ViT-B (IN-21k)	88.46%	98.05%
+ FT on Food2k	90.38%	**98.51**%
+ FT on Food2k + BERT	**90.63**%	98.41%
Swin-B (IN-21k)	**92.67**%	98.95%
+ FT on Food2k	92.62%	98.96%
+ FT on Food2k + BERT	92.59%	98.98%

TABLE VI
PERFORMANCE ON FOOD101 USING PRE-TRAINED VIT

Models	Top-1 Acc.	Top-5 Acc.
VGG16 (IN-21k)	79.02%	93.78%
+ FT on Food2k	**80.68**%	94.45%
ResNet50 (IN-21k)	84.50%	96.18%
+ FT on Food2k	**85.89**%	96.66%
ResNet152 (IN-21k)	86.61%	96.95%
+ FT on Food2k	**87.58**%	97.28%
Inception V3 (IN-21k)	84.15%	96.11%
+ FT on Food2k	**87.61**%	97.25%
SENet154 (IN-21k)	88.62%	97.57%
+ FT on Food2k	**89.68**%	98.08%

TABLE VII
PERFORMANCE ON FOOD101 USING PRE-TRAINED CNNS

C. *Explainability for vision transformers*

Explainability tools can be used to produce relevancy maps that highlight pixels that most influenced the prediction of the model. The relevancy maps are produced adapting an LRP-based method for ViT [17] and applying it on the trained ViT-B variants. An ideal model should make a correct classification and produce a relevancy map that attributes a high importance to the main ingredients shown in the image. To assess relevancy maps produced by the explainability methods, we used the UECFoodPix dataset [25] that is composed of $10,000$ images. Mean Intersection over Union (mIoU) and Pixel Accuracy are estimated using the whole dataset.

Figure 2 shows two images from UECFoodPix, their ground truth, the relevancy maps produced by the vision transformer and the predicted masks using Otsu automatic thresholding method [26]. Predicted masks and ground truth masks are used to calculate IoU and Pixel Accuracy.

Figure 3 shows the results obtained by a ViT-B model trained on Food2K and one trained on ImageNet-21K. In complex images composed by multiple plates of food both models struggle to produce a sufficiently good mask, while in simpler images where there is only one dish the Food2K weights allow us to produce an accurate region that highlights the main dish.

Table VIII reports segmentation results of the ViT-B model used with different sets of weights. Figure 4 shows how respectively the mean IoU and Pixel Accuracy change choosing different thresholds. The mIoU value underlines the gain in performance obtained by fine-tuning on Food2K. We noticed that pre-training on food-image datasets, and especially on

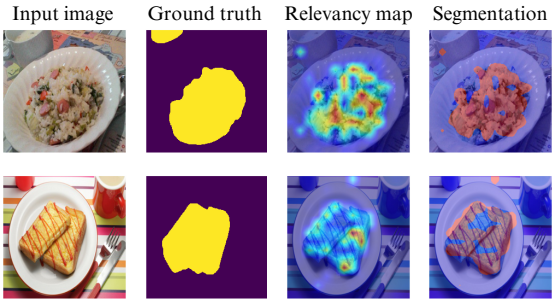

Fig. 2. Examples of input image, ground truth mask, relevancy map and segmentation obtained with Otsu's thresholding

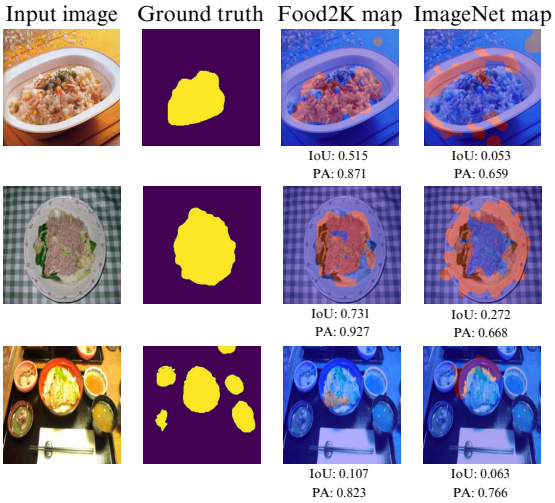

Fig. 3. Examples of input image, ground truth and predicted mask.

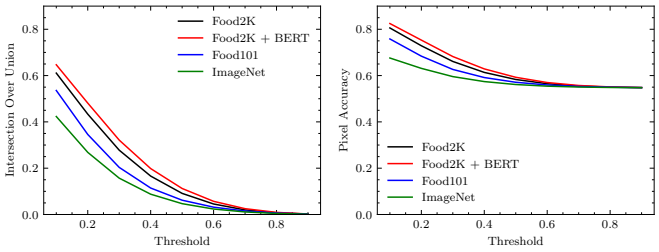

Fig. 4. Change in Intersection over Union value (left) and Pixel accuracy (right) when choosing a different threshold used to categorize a pixel as relevant or not. An higher threshold means that it is harder to classify a pixel as relevant.

Models	Otsu		Fixed at 0.1	
	mIoU	**mPA**	**mIoU**	**mPA**
ViT ImageNet-21K	0.257	0.639	0.412	0.670
ViT Food2K	0.402	0.722	0.604	0.802
ViT Food2K + BERT	0.438	0.735	0.623	0.816
ViT Food101	0.32	0.683	0.508	0.765

TABLE VIII
SEGMENTATION METRICS.

Food2K, produces relevancy maps that are closed to the ground truth masks. The mIoU increases by 156% and the mPA by 112% when using Food2K weights. The use of BERT-training had a positive effect on the produced relevancy maps, improving both mIoU and mPA.

Even though the ViT models presented were trained for classification purposes, the segmentation performance is close to methods specifically thought to work for segmentation tasks. We can conclude that these types of model can easily be adapted for food segmentation, and that pre-training on a food-image dataset, especially Food2K, is key to obtain even better performance.

VI. Conclusions

In this paper we demonstrated the effectiveness of vision transformers for food recognition tasks. We found that ViT-B outperforms other models on the Food2K dataset, particularly when fine-tuned on the same dataset. Additionally, vision transformers showcase superior generalization capabilities on the Food101 dataset compared to popular convolutional models. Furthermore, our study highlights the potential of vision transformers for food segmentation tasks through food recognition (i.e. without training for semantic segmentation). The explainability tools used to produce relevancy maps indicate their ability to identify regions of interest accurately, making them promising for segmentation purposes. Overall, vision transformers show promise in food computing applications, providing competitive performance and improved interpretability. This research paves the way for future investigations in utilizing vision transformers for various food-related tasks and advancing the field of computer vision in the context of food analysis.

Acknowlegment

Funder: Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 341 of 15 March 2022 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU;

Award Number: Project code PE00000003, Concession Decree No. 1550 of 11 October 2022 adopted by the Italian Ministry of University and Research, CUP D93C22000890001, Project title "ON Foods - Research and innovation network on food and nutrition Sustainability, Safety and Security – Working ON Foods".

References

[1] L. Zhou, C. Zhang, F. Liu, Z. Qiu, and Y. He, "Application of deep learning in food: a review," *Comprehensive reviews in food science and food safety*, vol. 18, no. 6, pp. 1793–1811, 2019.

[2] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–36, 2019.

[3] G. Ferrara, J. Kim, S. Lin, J. Hua, E. Seto *et al.*, "A focused review of smartphone diet-tracking apps: usability, functionality, coherence with behavior change theory, and comparative validity of nutrient intake and energy estimates," *JMIR mHealth and uHealth*, vol. 7, no. 5, p. e9232, 2019.

[4] M. Meenu, C. Kurade, B. C. Neelapu, S. Kalra, H. S. Ramaswamy, and Y. Yu, "A concise review on food quality assessment using digital image processing," *Trends in Food Science & Technology*, vol. 118, pp. 106–124, 2021.

[5] W. Lu and J. Chen, "Computer vision for solid waste sorting: A critical review of academic research," *Waste Management*, vol. 142, pp. 29–43, 2022.

[6] Y. Zhang, L. Deng, H. Zhu, W. Wang, Z. Ren, Q. Zhou, S. Lu, S. Sun, Z. Zhu, J. M. Gorriz, and S. Wang, "Deep learning in food category recognition," *Information Fusion*, vol. 98, p. 101859, 2023.

[7] S. Paul and P.-Y. Chen, "Vision transformers are robust learners," in *Proceedings of the AAAI conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2071–2081.

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

[9] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "Imagenet-21k pretraining for the masses," *arXiv preprint arXiv:2104.10972*, 2021.

[10] 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*, 2023.

[11] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020.

[12] 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*, 2022.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

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

[15] Y. Zhou, J. Chen, X. Zhang, W. Kang, and Z. Ming, "Semantic center guided windows attention fusion framework for food recognition," in *Pattern Recognition and Computer Vision*, 2022, p. 626–638.

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

[17] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 782–791.

[18] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.

[19] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[20] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[21] H. Touvron, M. Cord, and H. Jégou, "Deit iii: Revenge of the vit," in *European Conference on Computer Vision*, 2022, pp. 516–533.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[24] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 116–12 128, 2021.

[25] T. Ege, W. Shimoda, and K. Yanai, "A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice," in *Proceedings of the 5th international workshop on multimedia assisted dietary management*, 2019, pp. 82–87.

[26] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.