Jurnal Teknokes
Homepage: teknokes.org; Vol. 18, No. 4, pp. 337-346, December 2025;
e-ISSN: 2407-8964
p-ISSN: 1907-7904

RESEARCH PAPER
OPEN ACCES

Food Detection to Estimate Calories Using Detection Transformer

Joshua Putra Fesha Kristanto, Dedy Agung Prabowo, and Yohani Setiya Rafika Nur
Faculty of Informatics, Telkom University Purwokerto, Purwokerto, Indonesia

ABSTRACT

Accurately estimating calorie intake remains a common challenge, as many individuals have limited understanding of portion sizes and the caloric content of foods. This lack of nutritional knowledge is a major cause of both over- and under-calorie consumption and contributes to significant public health problems, including obesity, cardiovascular disease, and chronic metabolic disorders. Although computer vision–based approaches for dietary assessment have advanced, many methods still rely on handcrafted features, anchor-based CNN detectors, or controlled geometric assumptions. This indicates a practical gap in developing a fully functional system that operates on basic RGB images captured under everyday conditions. This study aims to develop an end-to-end food detection and calorie estimation system using the Detection Transformer (DETR) to predict calorie values directly from food images. The main contributions of this study include: (1) employing DETR to address non-maximum suppression limitations and improve the stability of multi-food recognition; (2) using a bounding box area-to-weight ratio as a low-complexity alternative to segmentation-based food portion estimation; and (3) developing a user-friendly interface for output visualization that displays detected food items and their estimated calorie values in real-world scenarios involving irregular food shapes and varying focal lengths. A DETR-based detector was trained using 2,228 COCO-formatted images across six distinct food classes. Calorie values were estimated by predicting food weight based on bounding box measurements, followed by calorie calculation using standardized reference weights. The method assessed robustness by evaluation on both controlled and real-life food images. Experimental results demonstrated moderate performance, with 0.617 mean Average Precision (mAP) and 0.656 mean Average Recall (mAR). The weight prediction module served as the primary estimation component, achieving a mean absolute residual of 8.7. These findings suggest that bounding box area is a reliable estimator of serving size. This study serves as a proof of concept for monitoring individual food intake and provides a foundation for further improvement in sub-item recognition, three-dimensional volume estimation, and the inclusion of broader food classes.

PAPER HISTORY

Received October 05, 2025
Revised November 01, 2025
Accepted December 01, 2025
Published December 30, 2025

KEYWORDS

Computer vision;
Deep learning;
Food detection;
Detection transformer;
Calories;

CONTACT:
joshuak@student.telkomuniversity.ac.id
dedyaprabowo@telkomuniversity.ac.id
yohani@ittelkom-pwt.ac.id

I. INTRODUCTION

Excessive caloric intake increases the risk of metabolic and various cardiovascular diseases [1], [2], [3], [4], while insufficient intake may lead to fatigue and impaired concentration [5], [6]. In 2022, the World Health Organization reported 2.5 billion overweight adults, including 890 million individuals with obesity, and 390 million underweight cases worldwide [7]. These conditions highlight the fact that manual calorie estimation remains impractical for most individuals, thereby creating the need for automated, efficient, and minimally processed visual-based measurement approaches that can operate reliably in real-world scenarios. Computer vision has emerged as a feasible approach for food detection and calorie estimation. Prior studies have largely used the Convolutional Neural Network (CNN) family [8], [9], [10], You Only Look Once (YOLO) [11], and other algorithms [12], [13]. Despite achieving high accuracy, many detectors require extensive post-processing, such as Non-Maximum Suppression (NMS), which introduces latency, duplicate suppression failures, and the absence of true end-to-end optimization. These limitations reduce computational efficiency, particularly for deployment in low-power or high-throughput environments. Food characteristics in Indonesia further exacerbate these challenges due to strong visual similarities arising from the predominance of fried dishes, which often exhibit similar golden-brown color distributions and irregular surface contours [14]. This visual homogeneity increases the likelihood of duplicate or incorrect object localization in many algorithms that rely on local receptive fields and NMS-based suppression. Transformer-based global attention mechanisms, such as those used in Detection Transformer (DETR), provide a more suitable modeling approach for distinguishing visually similar food items within a unified end-to-end inference pipeline. DETR introduces self-attention mechanisms that model spatial relationships globally and eliminate most post-processing dependencies [15], [16], [17]. Although DETR

Corresponding author: Dedy Agung Prabowo, dedyaprabowo@telkomuniversity.ac.id, Faculty of Informatics, Telkom University Purwokerto, Jl. DI Panjaitan No. 128, Purwokerto Selatan, 53147, Purwokerto, Indonesia).
DOI: https://doi.org/10.35882/teknokes.v18i4.132

has demonstrated strong generic object detection capabilities [18], only a limited number of studies have critically examined its applicability to food detection or evaluated bounding box area as a feature for food weight estimation. Existing food calorie estimation studies rarely assess whether bounding box area outputs from transformer-based detectors preserve sufficient spatial information to reliably estimate real-world food weight without segmentation masks. This raises a fundamental question: the challenge is not only whether DETR can accurately detect food items, but also whether its bounding box outputs can function as informative proxies for weight-based calorie estimation under practical constraints. This study addresses this gap by fine-tuning a DETR model on a multi-class food dataset labeled in COCO JSON format obtained from Roboflow and benchmarking the bounding box area feature for calorie inference. The dataset selection is justified by its inclusion of multiple food items per image, natural labeling conditions, and diverse serving geometries, which are essential for evaluating the transferability of bounding box area to food weight estimation. The study formulates the hypothesis that a DETR detector fine-tuned on Indonesian food images can maintain competitive detection accuracy while improving inference efficiency by eliminating dependency on NMS and producing bounding box area features that are sufficiently correlated with food weight to enable practical calorie estimation within an end-to-end pipeline.

This study has three main contributions, which are reframed as follows:

a) Applying the Detection Transformer (DETR) model to remove the Non-Maximum Suppression (NMS) bottleneck, thereby improving inference efficiency during multi-food recognition.
b) Analyzing the bounding box area-to-weight ratio as a lightweight alternative to segmentation masks.
c) Developing an interface visualization that displays detected food items and corresponding calorie values while considering real-world limitations, including irregular food shapes, variable camera distances, and bounding box sampling.

This study is structured as follows: Section II describes the dataset used, the proposed methods, and the proposed training and testing schemes. Section III presents the results of DETR accuracy on real food images. Section IV discusses the interpretation of the results, comparisons with other methods, and study limitations. Section V presents the conclusions, which restate the objectives, summarize the main findings, and outline future work.

II. MATERIALS AND METHODS

A. Dataset

This study utilized a collection of annotated food images that comply with the COCO JSON standard to enable integration with the DETR pipeline. The dataset includes six food classes: burger, fried rice, white rice, fried tofu, fried tempeh, and fried egg. The decision to focus on six classes was driven by measurement granularity and caloric unit consistency. The annotations are divided into representation groups based on food volume and serving characteristics as follows:

a) Fried tempeh and tofu are served as individual items but are commonly presented in portioned meals.
b) Burgers have unique shapes and stacked layers of varying sizes that influence food weight.
c) Fried eggs exhibit variable surface shapes and color distributions.
d) Fried rice and white rice are typically presented in a spread-out form, resulting in irregular outlines on the plate.

This classification scheme ensures that all food classes are accurately represented in terms of calories per gram and that serving size behavior is appropriately captured for bounding box area estimation. The total number of images was 2,228, of which 2,035 were allocated for training, 140 for validation, and 53 for testing. The majority of the data were assigned to the training set because DETR requires a large amount of training data [17]. Within the training dataset, the class distribution consisted of 522 fried tempeh images, 255 fried rice images, 444 burger images, 256 fried egg images, 345 fried tofu images, and 213 white rice images. The mean class can size is calculated with Eq. (1):

$$mean = \frac{training\_images}{classes\_total} \quad (1)$$

Within the dataset, a class size gap of 12.3% (267 samples) was observed when compared to the mean class size of 339.17. However, the Detection Transformer (DETR) framework fundamentally relies on a one-to-one Hungarian assignment to enforce set-based prediction consistency [19]. Although class imbalance exists, its extent is modest. With appropriate remapping of the dataset to densely pack label indices and the use of moderate batch sizes, the framework remains reliable and stable. Most issues related to slow convergence in DETR's matching scheme can be attributed to the instability of bipartite matching during the early stages. With appropriate strategies, such as denoising, convergence can be accelerated while maintaining matching robustness [20].

B. Data Collection

The detection model was trained usingpublicly available annotated images, while calorie constants wereobtained independently. Sources for calorie reference values were obtained from FatSecret and provided as fixed scalar values per 100 grams, which were then converted to per-gram constants. In addition topublicly available images, real-world images captured using smartphones were included specifically for evaluating the accuracy of food weight and calorie estimation. These real-world images were not used during detector training and were reserved exclusively for post-training evaluation.

Corresponding author: Dedy Agung Prabowo, dedyaprabowo@telkomuniversity.ac.id, Faculty of Informatics, Telkom University Purwokerto, Jl. DI Panjaitan No. 128, Purwokerto Selatan, 53147, Purwokerto, Indonesia).
DOI: https://doi.org/10.35882/teknokes.v18i4.132

C. Data Processing

Normalization and resizing were applied prior to the start of training. Normalization was mathematically defined to replicate the stable activation regime of the pretrained backbone encoder, as shown in Eq. (2):

$$I'_c = \frac{I_c - \mu_c}{\sigma_c}, c \in \{R, G, B\} \quad (2)$$

Where μ and σ q follow ImageNet priors. This approach maintains the stability of transformer positional encodings and mitigates scale drift, which can occur in the relationship between bounding box area and predicted weight [19]. Data augmentation was limited to non-deformative transformations that support efficient learning of bounding box representations without introducing geometric distortions that could reduce spatial reliability. These transformations included horizontal flipping with a probability of 0.5 and brightness–contrast jitter within ±10 percent. More aggressive augmentations, such as non-uniform scaling and perspective warping, were not applied because they introduce bounding box area prior inconsistencies that could bias weight estimation. The original category identifiers (IDs) were sequentially remapped to indices ranging from 0 to 5 using a deterministic lookup table. This remapping was performed to align with the transformer query embeddings and to eliminate sparse identifier sampling noise that degrades query learning. By embedding generalized class indices, geometric priors remain unaffected, while spatial detection is learned from bounding box features rather than category identifiers.

D. Calorie Estimation

Calorie computation in this study depends entirely on the predicted food weight. Calorie values obtained from FatSecret were converted to a per-gram basis [21] using Eq. (3):

$$cal_g = \frac{cal_{100g}}{100} \quad (3)$$

To estimate how heavy each detected food item might be, the relative size of its bounding box with respect to the total area of the input image is considered. The input image has a width W and height H, and the image area is therefore defined as Eq. (4):

$$A_{img} = W \times H \quad (4)$$

A_{img} denotes the total area of the input image in pixels. Using the bounding box definition [22] for all detected objects as: $(x_{min}, y_{min}, x_{max}, y_{max})$ the bounding box area (A_{box}) can be calculated using Eq. (5):

$$A_{box} = \max(0, x_{max} - x_{min}) \times max(0, y_{max} - y_{min}) \quad (5)$$

the relative area ratio (r) between the food object and the image is then calculated as:

$$r = \frac{A_{box}}{A_{img}} \quad (6)$$

Finally, each food class, denoted by c, is assigned an average reference weight w_c that corresponds to the weight of one serving of that specific food type. The weight of a detected instance is estimated using the following formula Eq. (7)::

$$weight = w_c \times r \times k \quad (7)$$

Where,

a) w_c= average reference weight associated with class c,
b) r= relative area ratio of the food item,
c) k= empirically determined correction factor.

throughout this study, the correction factor was set to a value of 4 and was used to reduce the estimation error arising from differences between two-dimensional image area measurements and the actual mass of food items. Combining the components described above yields the final weight estimation formulation adopted in this study Eq. (8):

$$weight = w_c \times \left(\frac{A_{box}}{W \times H}\right) \times k \quad (8)$$

using the above-referenced formula, the system approximates food weight based on the ratio of the bounding box projected area to the total image area. Nutritional constants are obtained from an off-detector pipeline using FatSecret per-100-gram scalar values, which are then converted to per-gram constants to align with the projected predicted weights. Final calorie estimation for each detected food instance is calculated through a linear weight projection, as shown in Eq. (9):

$$C = weight \times cal_g \quad (9)$$

where C represents the estimated calorie content of a detected food item. The total calorie estimation for an image containing multiple food items is calculated as Eq. (10):

$$C_{total} = \sum_{i=1}^{N} C \quad (10)$$

E. Training Method

Training employed the official pretrained Detection Transformer (DETR) model with a ResNet-50 backbone and COCO-pretrained weights from Facebook AI Research. Optimization dynamics are strongly influenced by limited parallel gradient accumulation in CPU-only environments and generally slower training throughput [23]. To mitigate these constraints, two deliberate strategies were adopted:

a) Training with a small batch size (batch = 2) to stabilize bipartite matching
b) Extending training duration to e 44 epochs-to compensate for limited computattional parallelism.

The optimization process used the AdamW optimizer with a learning rate of 1e-4, consistent with empirically established DETR stability ranges for low-resource transformer training between 1e-4 and 3e-4. A constant learning rate was maintained to prevent instability in bounding box regression loss, which could violate the

Corresponding author: Dedy Agung Prabowo, dedyaprabowo@telkomuniversity.ac.id, Faculty of Informatics, Telkom University Purwokerto, Jl. DI Panjaitan No. 128, Purwokerto Selatan, 53147, Purwokerto, Indonesia).
DOI: https://doi.org/10.35882/teknokes.v18i4.132

spatial area proportionality assumptions required for weight estimation. The training configuration summarized in Table 1 includes a CPU-based Google Colab environment, the PyTorch framework, and the official DETR model architecture (facebook/detr-resnet-50) as the base model. The entire implementation was developed in Python, and the dataset adhered to the COCO JSON annotation standard.

Table 1. Training Configuration

Parameter	Description
Platform	Google Collab
Environment	CPU
Framework	PyTorch
Base Model	DETR (facebook/detr-resnet-50)

III. RESULTS

A. Training Convergence and Loss Interpretation

The training loss outcomes are summarized in Table 2. The loss continuously decreased from 2.19 at epoch 1 to 0.95 at epoch 44, representing a reduction of approximately 56.58%, as computed using Eq. (11):

$$loss\ reduction(\%) = \frac{(x-y)}{x} \times 100 \quad (11)$$

Where,

a) x= first epoch,
b) y= last epoch.

This decrease indicates that the Detection Transformer (DETR) model successfully adapted pretrained weights to the custom food dataset. After the 35th epoch, the rate of improvement became marginal, indicating that the model reached a learning plateau, beyond which additional epochs contributed minimal improvement. This stabilization trend suggests that:

a) Underfitting is unlikely, as the model continued learning until the mid-stage of training.
b) Overfitting is limited, as the validation loss followed a similar downward trend without divergence.

Despite training with a batch size of 2 on a CPU-based environment, bipartite matching remained stable. This implies that pairwise assignment costs did not collapse, supported by evidence indicating that although CPU-based training slows convergence for deep models, simpler regression objectives can still converge when training exposure is increased across epochs [23].

Table 2. Training Loss per Epoch

Epoch	Average Loss
1	2.1955
5	1.4135
10	1.2845
15	1.2139
20	1.1437
25	1.0854
30	1.0358
35	1.0039
40	0.9663
44	0.9532

B. Detection Model Reliability

Based on the COCO evaluation protocol, the DETR model achieved a mean Average Precision (mAP) score of 0.617 across Intersection over Union (IoU) thresholds ranging from 0.5 to 0.95, along with a mean Average Recall (mAR) score of 0.656. The mAP score reflects the model's capability to detect food items by measuring both classification correctness and localization accuracy. The COCO evaluation framework measures Average Precision (AP) at multiple IoU thresholds (0.50–0.95), with higher thresholds requiring tighter bounding box alignment. The mAR score reflects the model's ability to detect the presence of objects across all classes, regardless of classification confidence. The reported mAR value indicates that the model successfully detected 65.6% of all ground truth food items present in the images. This performance is particularly relevant in meal-level analysis, as failure to detect any food item directly results in underestimation of total caloric intake. Figure 1 illustrates the detection and labeling results produced by the model. The figure suggests that:

a) Bounding-box generation remains spatially consistent across varying serving area dispersion.
b) Class label assignments align with visual priors
c) Confidences scores (are produced for detected food items)
d) Weight estimation outputs are generated following the detection stage

Fig 1. DETR model visualization result (a) fried rice,

(a) (b)

(b) burger

C. Per-Class Detection Performance

As shown in Table 3, detection performance varied significantly across food categories. This variation was influenced by two main factors:

a) the dataset exhibited a class imbalance of 12.3%, with the effect being more pronounced in classes with limited training samples;
b) class feature overlap contributed to confusion, particularly between fried egg, fried tofu, and fried tempeh, which share similar frying-related texture activations.

Corresponding author: Dedy Agung Prabowo, dedyaprabowo@telkomuniversity.ac.id, Faculty of Informatics, Telkom University Purwokerto, Jl. DI Panjaitan No. 128, Purwokerto Selatan, 53147, Purwokerto, Indonesia).
DOI: https://doi.org/10.35882/teknokes.v18i4.132

e-ISSN: 2407-8964
p-ISSN: 1907-7904

Table 3. Per-Class Performance result

Category	Precision	Recall	F1-Score
Burger	0.8089	0.8977	0.851
Fried Tofu	0.4102	0.6796	0.5116
Fried Rice	0.3363	0.76	0.4663
White Rice	0.4328	0.5053	0.4662
Fried Tempe	0.2761	0.5932	0.3768
Fried Egg	0.1837	0.4895	0.2671

The best-performing classes were burger, fried tofu, and fried rice. White rice achieved a moderate F1-score despite having fewer training samples than other classes, as its texture and color distribution across the image set was relatively uniform, thereby reducing cross-class intra-variability. Fried egg exhibited the lowest F1-score, as this class presents greater variability in yolk size, frying consistency, and surface texture, necessitating a larger dataset for robust learning. Fried tempeh also demonstrated weaker performance, which can be attributed to limited sample coverage and the fine-grained texture of the food item. Insufficient sampling reduces class visibility and limits the model's ability to internalize discriminative characteristics. In addition, external factors such as camera distance, lighting conditions, and background flatness in the test samples further degraded detection performance in the lower-performing classes.

D. Calorie Estimation Accuracy

Following bounding box estimation, the weight estimation module predicted the weight each detected food region using a regression-based approach. Table 4 presents the actual and estimated weights of the food instances. The reported estimation error across the full dataset was 7.47%. This value was calculated as the mean of per-instance percentage errors using a standard percentage-based regression error formulation[24], as shown in Table 4 Eq. (12):

$$Avg\ Error = \frac{1}{N}\sum_{i=1}^{N}\frac{W_{pred,i} - W_{real,i}}{W_{real,i}} \times 100\% \quad (12)$$

N denotes the total number of samples with valid predictions. The error values were averaged across food items, and the combined estimation error for burger, fried rice, fried tempe, fried tofu, and white rice was 7.47%. This result indicates that bounding box area serves as a reliable approximation of actual portion size when class-specific identification is sufficiently accurate. Among the evaluated classes, burger and white rice exhibited the lowest error ranges (0.14–10.17%), which can be attributed to their relatively uniform shapes and consistent surface structures. Fried rice also demonstrated a comparatively low prediction error range (0.50–10.00%), as the regularity of its plate-level distribution enabled more accurate estimation of total area relative to actual mass. In contrast, fried tempeh and fried tofu exhibited higher percentage errors

Table 4. Food Weight Prediction Result

Food Class	Real Weight	Predicted Weight	Error Precentage(%)
Burger	158gr	165gr	4.43
	151gr	141.8gr	6.09
	235gr	217gr	7.66
Fried Rice	301gr	299.5gr	0.50
	325gr	315.6gr	2.89
	273gr	245.7gr	10.00
Fried Tempe	8gr	9.1gr	13.75
	10gr	10.1gr	1.00
Fried Tofu	9gr	10.9gr	21.11
	8gr	6.6gr	17.5
Fried Egg	-	-	-
White Rice	141gr	138.4gr	1.84
	177gr	195gr	10.17
	143gr	143.2gr	0.14

Corresponding author: Dedy Agung Prabowo, dedyaprabowo@telkomuniversity.ac.id, Faculty of Informatics, Telkom University Purwokerto, Jl. DI Panjaitan No. 128, Purwokerto Selatan, 53147, Purwokerto, Indonesia).
DOI: https://doi.org/10.35882/teknokes.v18i4.132

e-ISSN: 2407-8964
p-ISSN: 1907-7904

primarily due to their small physical sizes. Even minor deviations between predicted and actual weights resulted in disproportionately large percentage errors, making these classes more sensitive to small prediction inaccuracies.

Taken together, these findings indicate that weight prediction accuracy is strongly dependent on class separability and visual uniformity. Lower error rates were achieved for food classes with stable shapes and consistent textures, whereas higher errors were observed for small, irregular, or fine-grained food classes.

E. Fried Egg Prediction as a Methodological Gap

Fried egg data were not included in Table 4 due to limitations in the current classification methodology. The Detection Transformer (DETR) model frequently misclassified fried egg as fried tofu or fried tempeh because of similarities in surface-fried texture and color uniformity. As a result, the transformer attention mechanism failed to generate stable bounding boxes labeled as "fried egg," preventing the use of anchored bounding boxes for weight prediction. Consequently, the error propagation follows the sequence:

$$detection\ errors \rightarrow missing\ weight\ prediction \rightarrow calorie\ estimation\ errors.$$

Calorie estimation is directly dependent on accurate detection; therefore, visually ambiguous or highly similar food items remain challenging. Errors may arise in the following aspects:

a) Localization resolution
b) Category description (class assignment)
c) delineation of the object field through bounding-box geometry

Weight estimation and subsequent calorie estimation represent stages that inherit errors originating from food detection. Detection gaps can be attributed to external conditions, such as excessive lighting, variations in camera angle, and visual variability in fried food appearance. These factors also reduce the effective number of detectable food classes. The fried egg class contained only 256 training images, which is much less than classes like burger or fried tempeh, which are visually dominant, and likely contributed to this misclassification. Due to limited intra-class variability, the model has difficulty learning reliable visual boundaries to determine differences between fried egg and other fried foods. The lack of representative examples also hindered the formation of robust class embeddings.

Therefore, increasing the number of annotated fried egg images and incorporating class-specific visual priors are expected to improve detection performance for this category.

IV. DISCUSSION

A. Food Calorie Estimation Performance

The use of DETR enables integrated food item recognition and calorie estimation within a unified framework. This architecture allows the recognition module to detect food items consistently across variations in lighting conditions and background environments [25]. Following detection, the calorie estimation module performs prediction based on estimated food weight and standardized calorie-per-gram reference values. Because calorie values are derived linearly from estimated weight, the accuracy of calorie estimation is directly dependent on the precision of weight prediction [26], [27], [28]. Despite these advantages, global self-attention mechanisms exhibit notable limitations in discriminating visually similar food categories [29]. Transformer-based attention analyzes images holistically, which can result in insufficient emphasis on fine-grained, texture-level information. This limitation is particularly evident in fried foods that share similar color distributions and surface textures, where overlapping attention responses hinder discrimination of subtle visual differences. From a representational perspective, this limitation stems from DETR's reliance on global token-to-token interactions without explicit convolutional inductive biases. In contrast to CNN-based detectors that progressively encode local texture information through hierarchical receptive fields, DETR distributes attention more uniformly across image regions. Consequently, sensitivity to high-frequency texture cues that are critical for differentiating visually similar fried foods is reduced. The influence of bounding box scale on weight estimation, affected by camera position, viewing angle, and illumination conditions, was also observed. Preliminary observations indicated that bounding box area may vary by approximately ±18% when camera distance changes by ±10 cm, resulting in corresponding deviations in predicted weight. Although these observations were not obtained under a fully controlled experimental protocol, they function as a preliminary sensitivity analysis, demonstrating how small geometric variations can propagate into substantial weight and calorie estimation errors. This observation is consistent with previous vision-based food estimation studies, which identify camera geometry and scale as major sources of uncertainty in portion size estimation.

In addition, the proposed system does not account for changes in caloric density resulting from cooking processes. Nutritional studies have shown that frying alters caloric content through mechanisms such as oil absorption and moisture loss, factors that cannot be inferred from monocular RGB images alone. This limitation is not unique to the proposed framework, as modeling invisible physicochemical changes in food remains a major challenge in image-based calorie estimation without additional sensing modalities. Finally, the system produces estimates of total caloric content without providing information on macronutrient composition, which further limits its applicability for comprehensive dietary monitoring.

B. Performance Comparison with Mask RCNN

A comparison between DETR and Mask R-CNN was conducted using the same dataset, training configuration, and evaluation metrics. Table 5 summarizes detection performance across different IoU

Corresponding author: Dedy Agung Prabowo, dedyaprabowo@telkomuniversity.ac.id, Faculty of Informatics, Telkom University Purwokerto, Jl. DI Panjaitan No. 128, Purwokerto Selatan, 53147, Purwokerto, Indonesia).
DOI: https://doi.org/10.35882/teknokes.v18i4.132

thresholds and object scales. The results in Table 5 reveal that Mask R-CNN consistently achieved higher mean Average Precision (mAP) scores than the other evaluated models at IoU thresholds of 0.5 and 0.50–0.95. This performance can be attributed to the presence of a Region Proposal Network, which performs iterative proposal refinement to enhance localization accuracy. In contrast, the results in Table 5 indicate that DETR achieved the highest mean Average Recall (mAR), particularly for medium and large object categories. This suggests that transformer-based architectures with global attention mechanisms are more effective at leveraging spatial relationships between objects, thereby reducing false negatives caused by missed detections.

This distinction is particularly important for calorie estimation, as failing to detect a food item (low recall) is more detrimental than moderate inaccuracies in bounding box localization. As demonstrated by the results in Table 6, DETR's higher recall rate and stable bounding box generation resulted in more accurate weight estimation outcomes. On average, DETR achieved a substantially lower weight estimation error (7.47%) compared with Mask R-CNN (36.81%).

Table 6. Comparison of Food Weight Estimation Results

Model	Average Weight Error (%)
DETR (Proposed Method)	7.47
Mask R-CNN	36.81

C. Comparison with Recent Studies

Recent studies on food detection and calorie estimation have explored several alternative approaches, each with distinct strengths and limitations. Convolutional neural network (CNN)–based methods remain widely used. For example, Li Ki Seung (2023) utilized a CNN integrated with multiple UV/VIS/NIR light sources to enhance feature extraction for food classification and caloric estimation [9]. Although the reported accuracy is high, the requirement for specialized multispectral hardware limits real-world applicability and diverges from the single-image, consumer-level scenarios targeted in the present study. Similarly, Haque et al. (2022) adopted a CNN architecture for food identification and nutritional estimation; however, the proposed system relied solely on image appearance and predefined calorie datasets without incorporating volume or weight estimation. This limitation reduces its ability to provide accurate calorie values in real-world scenarios where food portion size or weight affects caloric content [8]. These constraints underscore the necessity of integrating geometric or area-based estimation, which is addressed by the proposed DETR model.

Other studies attempted to estimate food dimensions using geometric assumptions. For instance, Kalivaraprasad et al. (2024) proposed a method that infers physical food measurements from pixel dimensions with the assistance of a reference object [10]. While effective under controlled conditions, reliance on an external object makes the method incompatible with casual daily use, an issue avoided in the present study by eliminating the need for calibration items. Another active line of study involves YOLO-based detectors. Huang et al. (2022) used YOLOv5 for food detection and weight estimation but reported that obtaining accurate calorie values is challenging due to factors such as oil absorption and inconsistencies in food preparation [11]. In addition, YOLO architectures rely on anchor boxes and non-maximum suppression, thereby increasing complexity in tuning and post-processing. YOLOv5 was also used by Jubayer F. et al. (2021) for detecting mold on food, demonstrating strong detection performance [30]. Compared with these anchor-based approaches, the end-to-end nature of DETR eliminates hand-tuned anchors and non-maximum suppression, reducing engineering overhead and improving output consistency, which is beneficial for downstream calorie estimation. Recent studies have incorporatedhybrid architectures, such as a study by Zhejun Kuang et al. (2025), which combined CNN and Vision Transformer (ViT) backbones to model both local texture features and global contextual information in food images [31]. Most of these studies, while effective for recognition, did not extend their systems to calorie estimation pipelines or still required additional components such as segmentation or external depth cues. Xinle et al. (2024)

Table 5. Detection Performance Comparison with Mask R-CNN

Metric	Threshold	Area	DETR (Proposed Method)	Mask R-CNN
mAP	IoU=0.50:0.95	All	0.408	0.443
	IoU=0.50		0.617	0.758
	IoU=0.75		0.438	0.444
	IoU=0.50:0.95	Large	0.409	0.445
mAR	IoU=0.50:0.95	All	0.568	0.426
	IoU=0.50:0.95		0.654	0.590
	IoU=0.50:0.95	Medium	0.300	0.212
	IoU=0.50:0.95	Large	0.656	0.601

Corresponding author: Dedy Agung Prabowo, dedyaprabowo@telkomuniversity.ac.id, Faculty of Informatics, Telkom University Purwokerto, Jl. DI Panjaitan No. 128, Purwokerto Selatan, 53147, Purwokerto, Indonesia).
DOI: https://doi.org/10.35882/teknokes.v18i4.132

proposed a high-accuracy food image classification method based on a Vision Transformer architecture enhanced with data augmentation and feature refinement mechanisms. The proposed AlsmViT integrates AugmentPlus, LayerScale, and a feature local enhancement multilayer perceptron to address common issues of overfitting and early saturation observed in standard ViT models. Experimental evaluations on the Food-101 and Vireo Food-172 datasets demonstrated strong performances ; however, the study focused exclusively on food image categorization and did not estimate portion size, weight, or caloric content[32]. Xin Chen et al. (2023) reported that vision-based systems are effective for food recognition, intake action classification, and food volume estimation, while fluid-related monitoring remains significantly underexplored. The study primarily focused on classifying food and non-food items, food types, and drinking actions [33]. Although these models improve recognition accuracy, they do not integrate calorie estimation within a single pipeline, and many continue to rely on segmentation, depth cues, or multi-view setups, introducing constraints not present in single RGB image applications.

Overall, the literature demonstrates a trade-off among accuracy, hardware requirements, and system complexity. CNN- and ViT-based models excel in classification but often neglect portion-size estimation; geometric methods require controlled environments; and YOLO-based detectors depend heavily on post-processing. In contrast, the proposed DETR-based approach contributes to bridging this gap by unifying detection and calorie estimation within a single end-to-end pipeline using only an RGB image. By leveraging global attention, the model maintains robust detection under varied backgrounds while enabling bounding box–based weight prediction without additional hardware or reference objects. This positions the proposed system as a more practical and generalizable solution for real-world dietary monitoring.

V. CONCLUSION

This study proposes an end-to-end food detection and calorie calculation system based on DETR and evaluates the applicability of DETR for single image–based nutritional analysis. The experimental results show that the model achieved an Average Precision of 0.617 and an Average Recall of 0.656 under the COCO metric, indicating a moderate detection capability for multiple food items. By adopting an end-to-end DETR architecture, anchor design and non-maximum suppression are eliminated, enabling stable bounding box estimation that is suitable for subsequent calorie estimation tasks. Weight approximation based on bounding box information yielded an average calorie estimation error of 7.47%, supporting its applicability as a lightweight alternative to segmentation-based methods under ideal image capture conditions. However, performance degradation was observed for visually similar food categories, such as fried eggs, fried tofu, and fried tempeh. In addition, the limited scope of the dataset restricts generalizability to broader food categorizations across diverse culinary cultures. In summary, this study demonstrates that transformer-based models can serve as a viable foundation for joint food detection and total calorie estimation from a single food image. Future studies are planned to enhance subcategory recognition and to incorporate feature representations that better capture scale and texture variations.

REFERENCES

[1] K. D. Hall *et al.*, “The energy balance model of obesity: beyond calories in, calories out,” *Am J Clin Nutr*, vol. 115, no. 5, pp. 1243–1254, May 2022, doi: 10.1093/AJCN/NQAC031.

[2] A. Afshin *et al.*, “Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017,” *The Lancet*, vol. 393, no. 10184, pp. 1958–1972, May 2019, doi: 10.1016/S0140-6736(19)30041-8.

[3] D. Liu *et al.*, “Calorie Restriction with or without Time-Restricted Eating in Weight Loss,” *New England Journal of Medicine*, vol. 386, no. 16, pp. 1495–1504, Apr. 2022, doi: 10.1056/nejmoa2114833.

[4] M. Mazur, A. Przytuła, M. Szymańska, and J. Popiołek-Kalisz, “Dietary strategies for cardiovascular disease risk factors prevention,” *Curr Probl Cardiol*, vol. 49, no. 9, p. 102746, Sep. 2024, doi: 10.1016/J.CPCARDIOL.2024.102746.

[5] D. M. Logue *et al.*, “Low energy availability in athletes 2020: An updated narrative review of prevalence, risk, within-day energy balance, knowledge, and impact on sports performance,” Mar. 01, 2020, *MDPI AG*. doi: 10.3390/nu12030835.

[6] H. Peña-Jorquera, V. Campos-Núñez, K. P. Sadarangani, G. Ferrari, C. Jorquera-Aguilera, and C. Cristi-Montero, “Breakfast: A crucial meal for adolescents’ cognitive performance according to their nutritional status. the cogni-action project,” *Nutrients*, vol. 13, no. 4, Apr. 2021, doi: 10.3390/nu13041320.

[7] World Health Organization, “Malnutrition,” https://www.who.int/news-room/fact-sheets/detail/malnutrition.

[8] R. U. Haque, R. H. Khan, A. S. M. Shihavuddin, M. M. M. Syeed, and M. F. Uddin, “Lightweight and Parameter-Optimized Real-Time Food Calorie Estimation from Images Using CNN-Based Approach,” *Applied Sciences (Switzerland)*, vol. 12, no. 19, Oct. 2022, doi: 10.3390/app12199733.

[9] K. S. Lee, “Multispectral Food Classification and Caloric Estimation Using Convolutional Neural

Corresponding author: Dedy Agung Prabowo, dedyaprabowo@telkomuniversity.ac.id, Faculty of Informatics, Telkom University Purwokerto, Jl. DI Panjaitan No. 128, Purwokerto Selatan, 53147, Purwokerto, Indonesia).
DOI: https://doi.org/10.35882/teknokes.v18i4.132

Networks," *Foods*, vol. 12, no. 17, Sep. 2023, doi: 10.3390/foods12173212.

[10] P. Mvd and N. Kishore Gattim, "Deep Learning-based Food Calorie Estimation Method in Dietary Assessment: An Advanced Approach using Convolutional Neural Networks," 2024. [Online]. Available: www.ijacsa.thesai.org

[11] J.-T. Huang Taipei Tech, C. E. Hsiao, and C.-H. Wang Taipei Tech, *Deep Learning-Based Food Identification and Calorie Estimation System*. 2022.

[12] K. S. Lee, "Multi-Spectral Food Classification and Caloric Estimation Using Predicted Images," *Foods*, vol. 13, no. 4, Feb. 2024, doi: 10.3390/foods13040551.

[13] P. Chotwanvirat, A. Prachansuwan, P. Sridonpai, and W. Kriengsinyos, "Advancements in Using AI for Dietary Assessment Based on Food Images: Scoping Review," 2024, *JMIR Publications Inc.* doi: 10.2196/51432.

[14] M. Mansouri, S. Benabdellah Chaouni, S. Jai Andaloussi, and O. Ouchetto, "Deep Learning for Food Image Recognition and Nutrition Analysis Towards Chronic Diseases Monitoring: A Systematic Review," Sep. 01, 2023, *Springer*. doi: 10.1007/s42979-023-01972-1.

[15] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 2023, [Online]. Available: http://arxiv.org/abs/1706.03762

[16] Y. Li, N. Miao, L. Ma, F. Shuang, and X. Huang, "Transformer for object detection: Review and benchmark," *Eng Appl Artif Intell*, vol. 126, p. 107021, Nov. 2023, doi: 10.1016/J.ENGAPPAI.2023.107021.

[17] L. Yu, L. Tang, and L. Mu, "A Review of DEtection TRansformer: From Basic Architecture to Advanced Developments and Visual Perception Applications," Jul. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/s25133952.

[18] Y. Li *et al.*, "End-to-end plant disease detection using transformers with collaborative hybrid assignment training," *Appl Soft Comput*, vol. 186, p. 114137, Jan. 2026, doi: 10.1016/J.ASOC.2025.114137.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," May 2020, [Online]. Available: http://arxiv.org/abs/2005.12872

[20] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR Training by Introducing Query DeNoising," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13609–13617. doi: 10.1109/CVPR52688.2022.01325.

[21] G. A. Tahir and C. K. Loo, "A comprehensive survey of image-based food recognition and volume estimation methods for dietary assessment," Dec. 01, 2021, *MDPI*. doi: 10.3390/healthcare9121676.

[22] W. Zafar *et al.*, "Enhanced TumorNet: Leveraging YOLOv8s and U-net for superior brain tumor detection and segmentation utilizing MRI scans," *Results in Engineering*, vol. 24, Dec. 2024, doi: 10.1016/j.rineng.2024.102994.

[23] D. Gyawali, "Comparative Analysis of CPU and GPU Profiling for Deep Learning Models," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2309.02521

[24] F. S. Konstantakopoulos, E. I. Georga, and D. I. Fotiadis, "A novel approach to estimate the weight of food items based on features extracted from an image using boosting algorithms," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-47885-0.

[25] G. A. Tahir and C. K. Loo, "A comprehensive survey of image-based food recognition and volume estimation methods for dietary assessment," Dec. 01, 2021, *MDPI*. doi: 10.3390/healthcare9121676.

[26] Y. Han, Q. Cheng, W. Wu, and Z. Huang, "DPF-Nutrition: Food Nutrition Estimation via Depth Prediction and Fusion," *Foods*, vol. 12, no. 23, Dec. 2023, doi: 10.3390/foods12234293.

[27] E. Shonkoff *et al.*, "AI-based digital image dietary assessment methods compared to humans and ground truth: a systematic review," 2023, *Taylor and Francis Ltd.* doi: 10.1080/07853890.2023.2273497.

[28] E. Robinson, M. Khuttan, I. McFarland-Lesser, Z. Patel, and A. Jones, "Calorie reformulation: a systematic review and meta-analysis examining the effect of manipulating food energy density on daily energy intake," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 19, no. 1, Dec. 2022, doi: 10.1186/s12966-022-01287-z.

[29] J. Maurício, I. Domingues, and J. Bernardino, "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review," May 01, 2023, *MDPI*. doi: 10.3390/app13095521.

[30] F. Jubayer *et al.*, "Detection of mold on the food surface using YOLOv5," *Curr Res Food Sci*, vol. 4, pp. 724–728, Jan. 2021, doi: 10.1016/J.CRFS.2021.10.003.

[31] Z. Kuang, H. Gao, J. Zhao, L. Wang, and L. Sun, "FFFNet: A Food Feature Fusion Model with Self-Supervised Clustering for Food Image Recognition," *Applied Sciences (Switzerland)*,

Corresponding author: Dedy Agung Prabowo, dedyaprabowo@telkomuniversity.ac.id, Faculty of Informatics, Telkom University Purwokerto, Jl. DI Panjaitan No. 128, Purwokerto Selatan, 53147, Purwokerto, Indonesia).
DOI: https://doi.org/10.35882/teknokes.v18i4.132

vol. 15, no. 17, Sep. 2025, doi: 10.3390/app15179542.
[32] X. Gao, Z. Xiao, and Z. Deng, "High accuracy food image classification via vision transformer with data augmentation and feature augmentation," *J Food Eng*, vol. 365, p. 111833, Mar. 2024, doi: 10.1016/J.JFOODENG.2023.111833.
[33] X. Chen and E. N. Kamavuako, "Vision-Based Methods for Food and Fluid Intake Monitoring: A Literature Review," Jul. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/s23136137.

AUTHOR BIOGRAPHY

Joshua Putra Fesha Kristanto is a student majoring in **Informatics Engineering at Telkom University Purwokerto**, Central Java, Indonesia. He has an interest in Artificial Intelligence, particularly in areas such as classification, forecasting, and predictive modeling. In addition, he is passionate about developing AI-integrated systems that can be applied across various sectors such as healthcare, education, business, and personal applications. He is also enthusiastic about the potential of Artificial Intelligence to improve quality of life, simplify processes, and enable faster and more accurate decision-making. He aims to contribute to innovative solutions by applying AI technologies to solve real-world problems. Furthermore, he has an interest in computer networks and actively participates in various networking competitions and certification programs to enhance his knowledge and skills.

Dedy Agung Prabowo is an instructor in the Informatics and Computer Science research group with the academic rank of Lecturer. He obtained his bachelor's and master's degrees in Informatics Engineering from Universitas Dian Nuswantoro in 2008 and 2010, respectively. His areas of expertise include Database Programming, Computer Security, Cryptography, and Artificial Intelligence. In addition to teaching, he is actively involved in research and community service activities supported by both internal and external funding sources. He completed his primary education in 1997 at SD Inpres 1 Kota Sorong, continued his junior secondary education at SMP Negeri 9 Kota Sorong from 1997 to 2000, and completed his senior secondary education at SMA Negeri 15 Kota Semarang from 2001 to 2003. He is currently a lecturer in the Informatics Engineering Study Program, Faculty of Informatics, Telkom University Purwokerto

Yohani Setiya Rafika Nur is a Lecturer in the Bachelor's Program in Information Technology at Telkom University Purwokerto. Her area of expertise focuses on intelligent systems and causal modeling. Her research interests include Decision Support Systems, Causal Modeling, Data Science, and Data Mining. In addition to actively participating in academic activities, she is also involved in various scientific programs and community service initiatives aimed at enhancing community empowerment through the development and utilization of technology for sustainable economic improvement.

Corresponding author: Dedy Agung Prabowo, dedyaprabowo@telkomuniversity.ac.id, Faculty of Informatics, Telkom University Purwokerto, Jl. DI Panjaitan No. 128, Purwokerto Selatan, 53147, Purwokerto, Indonesia).
DOI: https://doi.org/10.35882/teknokes.v18i4.132