DOI: 10.37943/BZIF1573

Ademi Sharipova
Bachelor's Degree, Junior Data Scientist, Institute of Smart Systems and Artificial Intelligence
ademi.sharipova@nu.edu.kz, orcid.org/0009-0000-7173-7518
Nazarbayev University, Kazakhstan
Baimyrza Kalmyrzayev
Master's Student, School of Engineering and Digital Sciences
baimyrza.kalmyrzayev@nu.edu.kz, orcid.org/0009-0000-7236-7213
Nazarbayev University, Kazakhstan
Aidana Mashrapova
Bachelor Student, School of Sciences and Humanities
aidana.mashrapova@nu.edu.kz, orcid.org/0009-0004-2358-0244
Nazarbayev University, Kazakhstan
Aigerim Yeleussizova
Bachelor Student, School of Sciences and Humanities
a.yeleussizova@nu.edu.kz, orcid.org/0009-0000-9969-6228
Nazarbayev University, Kazakhstan
Huseyin Atakan Varol
PhD, School of Engineering and Digital Sciences (SEDS)
ahvarol@nu.edu.kz, orcid.org/0000-0002-4042-425X
Nazarbayev University, Kazakhstan
Mei Yen Chan
Assistant Professor, Department of Biomedical Sciences, School of Medicine,
yen.chan@nu.edu.kz, orcid.org/0000-0003-2324-9338
Nazarbayev University, Kazakhstan

COMPARISON OF DIFFERENT OBJECT DETECTION TECHNIQUES FOR CENTRAL ASIAN FOOD RECOGNITION

Abstract: Automated food detection is a major component in building reliable dietary monitoring systems, particularly in regions where cuisine-specific datasets are limited. Central Asian food recognition presents unique challenges due to strong intra-class variation, visually similar dishes, multi-item meal compositions, and long-tailed class distributions. To address this problem, we present a comprehensive benchmarking analysis of five modern object detection models across three groups: two-stage (Faster R-CNN), one-stage anchor-based (RetinaNet), one-stage anchor-free (YOLOv12, YOLO26), and transformer-based (RT-DETR). Experiments are conducted on a unified dataset constructed by merging the Global Gastronomic Culinary Dataset (GGCD) and the Food Portion Benchmark (FPB), resulting in 48,228 images across 258 food classes. The dataset includes controlled single-portion photographs and multi-item meals captured across various backgrounds and from different viewing angles. The merged dataset therefore covers both isolated food portions and more complex scenes containing several dishes. We compare the models in terms of detection accuracy, computational requirements, and inference speed, and examine AP by food class across all models. Overall, YOLOv12 achieves the highest accuracy, with an mAP50 of 75.80 and an mAP50-95 of 67.90, while using fewer parameters and less computation than the heavier baselines. RT-DETR reaches an mAP50 of 72.90 but requires more computation, whereas YOLO26 records the shortest single-image inference time on both GPU and CPU. Class-wise analysis further shows high performance on visually distinctive categories, with near-perfect AP for ten top-detected classes. These findings provide a practical

Received: 01.06.2026 Accepted: 17.07.2026 Published: 30.06.2026

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

baseline for Central Asian food detection and show that future progress depends on improving robustness for long-tail and visually ambiguous classes.

Keywords: object detection; food recognition; Central Asian cuisine; dietary monitoring; Faster R-CNN; RetinaNet; YOLO; Detection Transformer; benchmark evaluation.

Introduction

Automated food recognition has become a significant focus in computer vision, motivated by the need for scalable solutions that support dietary monitoring, nutritional assessment, and public health efforts [1-2]. Image-based approaches provide a practical alternative to self-reported dietary intake, which is often inaccurate, time-consuming, and subject to recall bias [1, 3]. As a result, computer vision and deep learning methods are increasingly used in food recognition systems and nutrition-related applications [2-3].

Advances in deep learning, particularly convolutional neural networks (CNNs) and vision transformers, have led to notable improvements in the accuracy and robustness of food image classification and detection [4]. However, the performance of food recognition models depends heavily on the quality and diversity of training datasets. Many existing food datasets are region-specific. For instance, Food2K [5] primarily covers Western dishes, alongside regionally focused collections such as the ChineseFoodNet, Turkish-Foods-15, and Indian Food Database dataset [6].

In real-world dietary images, multiple foods often appear together, with complex backgrounds, partial occlusions, and varying distances between main and side dishes. These conditions require models to go beyond single-label classification and instead detect and identify each food item individually. Because annotating dense bounding boxes and masks is resource-intensive, early deep learning research aimed to reduce the annotation burden. Some methods used activation-based cues to extract candidate food regions with minimal supervision, eliminating the need for full bounding box or mask labels during training [3]. For example, a CNN-based approach generated candidate regions under limited supervision [7]. More recently, He et al. proposed an end-to-end food image analysis framework that integrates food localization, classification, and portion-size estimation, allowing multiple food items in an arbitrary dietary image to be localized and identified [8].

With advances in object detection models, research has shifted toward real-world applications. General-purpose detectors have been adapted for dietary analysis, including use cases with wearable cameras where images may be noisy, and food items are small or partially visible [9]. Recent studies have benchmarked modern detectors directly on food datasets. For example, Tan et al. [10] compared SSD, Faster R-CNN, RetinaNet, and YOLOv5 on public food benchmarks with bounding-box annotations, showing that YOLO-based models can achieve competitive accuracy while maintaining efficient inference. These results demonstrate that deep learning methods can localize food items effectively, but also underscore the importance of benchmark coverage, annotation style, and scene complexity. Careful evaluation of representative regional data remains essential.

Central Asian cuisine is a unique and relatively under-studied area in food computing. To address the lack of data, Karabay et al. introduced the Central Asian Food Dataset (CAFD), which contains over 16,000 images in 42 categories for image classification [11]. The Central Asian Food Scenes Dataset (CAFSD) expanded this work to include food localisation and detection, with 21,306 images and 69,856 annotated instances across 239 categories [12]. The Global Gastronomic Culinary Dataset (GGCD) further extended CAFSD by incorporating data from the Nutrition5k dataset, resulting in 34,145 images across 241 classes [13-14]. The Food Portion Benchmark (FPB) dataset provides 14,083 images in 138 classes with laboratory-measured weight annotations, supporting research on both detection and portion size estimation [15]. Together, these datasets provide a foundation for evaluating detection models in the context of Central Asian food.

DOI: 10.37943/BZIF1573

The importance of this work is underscored by the region's growing burden of diet-related non-communicable diseases. In Kazakhstan, for example, 25.3% of adult women and 21.4% of adult men are living with obesity, while diabetes affects an estimated 13.0% of women and 14.5% of men [16]. In this context, automated systems that can detect and quantify local dishes could support personalized dietary interventions and population-level nutritional monitoring, provided that models are trained and validated on data reflecting local cuisine, terminologies, and food portion sizes [17].

Object detection architectures have developed rapidly and can be grouped into three main categories. Two-stage detectors, such as Faster R-CNN-style models, first generate region proposals and then classify each candidate [18]. When combined with a Feature Pyramid Network (FPN) [19] and a ResNet-50 backbone, this approach remains a strong baseline, particularly for detecting small objects. One-stage detectors skip the proposal step and predict bounding boxes directly. Anchor-based one-stage detectors such as RetinaNet commonly use focal loss to address foreground-background class imbalance in dense prediction [20], while the YOLO series has focused on improving inference speed. Recent versions such as YOLOv12 include attention mechanisms like Area Attention and Residual Efficient Layer Aggregation Networks (R-ELAN) [21]. YOLO26 further advances this by using the MuSGD (multi-step stochastic gradient descent) hybrid optimizer and a design that avoids non-maximum suppression (NMS), reducing CPU-side latency [22]. Transformer-based detectors, beginning with DETR, treat detection as a set prediction problem using bipartite matching. While early transformer-based models had slow convergence, newer models such as Deformable DETR and RT-DETR have improved efficiency [23-24].

This paper presents a comprehensive benchmarking study comparing five object detection models across these paradigms on a combined Central Asian food dataset constructed from GGCD [13] and FPB [15]. The evaluated models include: (1) one-stage detectors, comprising the anchor-based RetinaNet and the real-time YOLO family (YOLOv12, and YOLO26); (2) a two-stage detector, represented by Faster R-CNN with a ResNet-50-FPN backbone; and (3) a transformer-based detector, RT-DETR with a ResNet-50 backbone. This study provides the first extensive performance analysis of these diverse architectures specifically for Central Asian cuisine. The main contributions of this work are as follows:

1. The first comprehensive benchmark of object detection architectures for Central Asian food recognition;
2. A systematic comparison of one-stage, two-stage, and transformer-based detectors under a common experimental setting;
3. Detailed class-wise analysis to identify strengths and limitations of current detection models on Central Asian food categories.

Methods and Materials

1. Dataset

This study uses a benchmark dataset built by combining two food image datasets. We merged them to improve regional representation and overall culinary variety. The first source is the Food Portion Benchmark (FPB) [15]. It is a controlled laboratory dataset which contains 14,083 high-resolution images from 138 food categories that are common in Central Asia. The dataset covers main dishes, salads, drinks, and traditional pastries. Foods were photographed as both single items and full meal sets (breakfast, lunch, dinner). Each item appears in three serving sizes: small, medium, and large. To capture visual variation, images were taken from several views, including two top-view and four side angles. All images have tight bounding box annotations created under a standard protocol. The training, validation, and test splits were separated by a collection session to reduce data leakage across partitions. More details about food selection, imaging setup, and annotation are provided by Sanatbyek et al. in their Food Portion Benchmark study [15].

The second source is the Global Gastronomic Culinary Dataset (GGCD) [13], which includes 241 globally representative food categories with bounding-box labels. GGCD combines

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

two datasets. One is Nutrition5k [14], whose images were re-annotated with bounding boxes. The other is the Central Asian Food Scenes Dataset (CAFSD) [12], which contains food scenes with globally common items such as fruits, vegetables, pastries, desserts, and fast food, together with traditional Central Asian dishes. Bringing these two sources together gives broad global coverage while keeping a strong representation of the less-studied Central Asian food domain.

After merging FPB and GGCD and unifying overlapping labels, the combined dataset contains 48,228 images across 258 food classes. The data were split into training (36,824 images), validation (5,721 images), and test (5,683 images) partitions to support consistent model development and fair benchmarking. Table 1 summarizes the dataset composition.

Table 1. Summary of the combined food detection dataset

Component	Source	Classes	Images	Annotation Type
FPB [15]	Laboratory collection	138	14,083	Bounding Box
GGCD [13]	Nutrition5k + CAFSD	241	34,145	Bounding Box
Combined	FPB+GGCD	258	48,228	Bounding Box

The class distribution in the combined dataset follows a long-tailed pattern, which is very common in real-world food recognition. In practice, some foods are seen much more often than others. For example, common items like bread, rice, and tomato appear in thousands of images, while rare dishes such as nauryz-kozhe and talkan-zhent appear in fewer than 100 samples. This means the model gets many chances to learn frequent classes, but only limited exposure to rare ones. As a result, detection performance is usually stronger on common foods and weaker on underrepresented classes. This imbalance is an important and persistent challenge for all detection models. Figure 1 shows samples from the combined dataset with bounding-box annotations.

Figure 1. Sample images from the combined dataset with ground-truth bounding-box annotations.

DOI: 10.37943/BZIF1573

2. Models

In this study, we evaluate five object detection models from three different detection paradigms. This provides a broad and fair comparison of modern detection approaches. Each model starts from weights pre-trained on the Microsoft COCO dataset, then is fine-tuned on our combined food detection dataset. This setup helps us compare model behavior under consistent training conditions. A summary of all models is presented in Table 2.

Table 2. Overview of evaluated object detection models

Model	**Paradigm**	**Backbone**	**Proposal Mechanism**
Faster R-CNN [18]	Two-stage	ResNet-50-FPN	Region Proposal Network
RetinaNet [20]	One-stage (anchor)	ResNet-50-FPN	Dense anchors + focal loss
YOLOv12 [21]	One-stage (anchor-free)	Custom (R-ELAN)	Grid-based prediction
YOLO26 [22]	One-stage (anchor-free)	Custom	Grid-based prediction
RT-DETR [24]	Transformer	ResNet-50 + Transformer	Learned object queries

Faster R-CNN [18] is a two-stage object detector. In the first stage, a Region Proposal Network (RPN) generates candidate object regions from multi-scale anchor boxes. In the second stage, a Fast R-CNN head assigns class labels and refines bounding box coordinates for each proposal. In this study, we use a ResNet-50 backbone with a Feature Pyramid Network (FPN) [19]. FPN produces multi-scale feature maps, which is important for food detection because object sizes vary widely, from small garnishes to large main dishes.

RetinaNet [20] is a single-stage, anchor-based detector. Specifically, it predicts object bounding boxes and class probabilities directly without a secondary proposal stage. One key challenge in dense detection is class imbalance: background examples are far more common than true objects. RetinaNet addresses this with focal loss. Easier examples are assigned less weight, while harder ones receive more attention during training. For fairness, we use the same backbone as Faster R-CNN: ResNet-50 with FPN. However, the detection process is different. RetinaNet uses parallel heads for classification and bounding box regression and applies them densely across each pyramid level. This comparison is useful. With training conditions kept the same, differences in performance mainly reflect architectural choice, that is, two-stage proposal-based detection versus single-stage dense prediction with focal loss.

YOLOv12 [21] is a one-stage detector with added attention modules. It uses Area Attention to focus on useful local regions in an efficient way. It also uses R-ELAN blocks to improve feature reuse across scales and support better gradient flow. In practice, this helps the model capture wider context while still running at real-time speed. We use YOLOv12m, the medium version, because it gives a good balance between accuracy and computation.

YOLO26 [22] is the latest generation of the YOLO family. It builds on YOLOv12 and introduces further architecture updates. To keep the comparison fair, we also use the medium version here, YOLO26m.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

RT-DETR (Real-Time DEtection TRansformer) [24] is an improved version of DETR that addresses the speed limitations of the original model. The architectural changes include efficient hybrid encoder to process multiscale features faster and efficient query selection for decoder to improve performance. It employs a ResNet-50 backbone for feature extraction, followed by a transformer encoder-decoder. A fixed set of learned object queries directly predict detections, matched to ground truth via bipartite (Hungarian) matching during training.

3. Training Protocol

Our training protocol was designed to keep the comparison fair, while still respecting real differences between model families and training frameworks. To ensure consistency, all models used the same training and validation splits, the same class labels, and the same annotation format. We selected checkpoints using the validation metric: mAP averaged across IoU thresholds from 0.50 to 0.95. Test performance was then reported once per model, using only the checkpoint chosen on validation data.

All models were implemented using PyTorch and initialized with COCO-pretrained weights. Experiments were conducted on a uniform compute node equipped with 4 NVIDIA H100 GPUs.

Table 3 summarizes the training specifications for the selected object detection models, including the optimizer, learning rate, weight decay, batch size, and number of training epochs.

Table 3. Training specifications of evaluated object detection models.

Parameter	**Faster R-CNN**	**RetinaNet**	**YOLOv12**	**YOLO26**	**RT-DETR**
Framework	Detectron2	Detectron2	Ultralytics	Ultralytics	HuggingFace
Optimizer	SGD (m=0.9)	SGD (m=0.9)	AdamW	AdamW	AdamW
Base LR	0.001	0.001	0.001	0.001	0.0001
Weight Decay	0.0001	0.0001	0.0005	0.0005	0.0001
Batch Size	16	16	16	16	16
Max Epochs	200 epochs (steps)	200 epochs (steps)	200 epochs	200 epochs	200 epochs

Where possible, we align training budget and optimization stability across models. In practice, full unification of optimizers and schedules across different frameworks can produce systematically suboptimal behavior for specific architectures, which would undermine the benchmark's validity. For this reason, we adopt a single best-practice yet controlled configuration: each model is trained using a widely accepted optimization recipe for its framework, and differences are permitted only when they are necessary for stable convergence or are standard for the model family.

DOI: 10.37943/BZIF1573

Detectron2-based models (Faster R-CNN and RetinaNet) were trained with SGD and momentum, which remains the standard optimizer choice for these architectures in this framework. Validation was performed periodically, and early stopping was driven by validation mAP50-95.

Ultralytics-based models (YOLOv12 and YOLO26) were trained with AdamW and cosine scheduling using the standard Ultralytics pipeline. Both models used a fixed input resolution of 640×640. Detectron2-based models (Faster R-CNN and RetinaNet) followed the default Detectron2 resizing policy: during training, images were resized while keeping the aspect ratio, and the shorter side was randomly sampled between 640 and 800 pixels. RT-DETR was trained at 640×640, consistent with its standard configuration. YOLOv12 and YOLO26 were evaluated in comparable model size variants to avoid conflating architecture changes with large differences in parameter count.

RT-DETR was trained using AdamW with a lower learning rate and gradient clipping, consistent with standard transformer training practice. A smaller learning rate was applied to the backbone than to the transformer layers to preserve pretrained features while allowing the transformer components to adapt.

Results

This section presents the quantitative evaluation of the five object detection models - Faster R-CNN, RetinaNet, YOLOv12, YOLO26, and RT-DETR which are trained and evaluated on the combined FPB-GGCD food detection dataset. Table 4 and Figure 2 present the performance results of five models evaluated on the combined food dataset.

Table 4. Performance results of evaluated object detection models

Model	mAP50	mAP50-95	GFlops	Number of parameters (M)	GPU latency (ms)	CPU latency (ms)
Faster R-CNN	66.73	54.21	134.38	42.6	20.31	1848.26
RetinaNet	66.46	56.99	151.54	41.6	32.09	3589.01
YOLOv12	**75.80**	**67.90**	**67.5**	**20.2**	**18.55**	**1324.57**
YOLO26	72.10	64.80	68.2	22.5	12.98	1138.43
RT-DETR	72.90	65.55	136	40.6	25.85	2053.98

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

Figure 2. Detection performance comparison across evaluated metrics (mAP50 and mAP50-95).

YOLOv12 achieves the best overall performance by a clear margin compared to other models. In addition, it has the smallest model size (20.2 million parameters), which demonstrates strong trade-off between performance and efficiency. YOLO26 and RT-DETR show competitive results, achieving 72.10 and 72.90 mAP50, respectively. However, both models remain several percentages below YOLOv12 in both evaluation metrics. Moreover, RT-DETR requires double the parameters compared to both YOLOv12 and YOLO26. Faster R-CNN and RetinaNet show similar performance in both metrics. Both models are heavily outperformed by the other three models.

To assess the practical inference cost of each detector, we additionally measured single-image latency at 640×640 input with 258 classes on an NVIDIA Tesla V100-SXM3-32GB GPU and on a single thread of an Intel Xeon Platinum 8168 CPU, reporting the median over 200 timed iterations on GPU and 30 on CPU after warm-up. As shown in Table 4, YOLO26 is the fastest model on both devices, reaching 12.98 ms (77.0 FPS) on the GPU and 1.14 s per image on a single CPU thread. YOLOv12 is the second fastest (18.55 ms GPU, 1.32 s CPU) while also producing the highest detection accuracy, confirming its strong accuracy–efficiency trade-off. Faster R-CNN and RT-DETR are noticeably slower, and RetinaNet is the slowest overall, in line with its higher GFLOPs.

Overall, the YOLOv12 model provides the best efficiency (the smallest size) and the best performance for the given dataset with YOLO26 coming second. This indicates that YOLO-based models are better suited for food image datasets.

Table 5 reports the top-performing classes for each detector. A clear pattern appears: all models perform strongly on visually distinctive and well-structured foods, while YOLO-based models usually achieve the highest scores. For example, YOLOv12 reaches 98.80 on quesadilla, and YOLO26 reaches 98.60 on grilled vegetables. RT-DETR is also competitive, including a perfect 100.0 on pumpkin seeds. In contrast, Faster R-CNN and RetinaNet generally remain lower on the same classes, although their scores are still high for several categories.

DOI: 10.37943/BZIF1573

Table 5. Top 10 best-detected food classes by mAP50-95 across evaluated models. Cells are color-coded from lower (red) to higher (green) mAP values

Rank	Class	Faster R-CNN	RetinaNet	YOLOv12	YOLO26	RT-DETR
1	pumpkin seeds	88.35	96.70	97.90	98.40	**100.00**
2	fried aubergine	94.92	92.31	95.50	**97.50**	96.69
3	grilled vegetables	86.16	94.63	97.10	**98.60**	96.89
4	udon	92.63	**95.92**	94.50	92.80	94.44
5	shelpek	90.13	92.30	**97.40**	95.10	94.24
6	olivie	88.50	94.72	**97.40**	96.20	91.60
7	quesadilla	87.28	86.53	**98.80**	96.40	98.34
8	Hachapuri	88.80	91.30	95.00	92.80	**96.17**
9	Meatball	90.90	92.78	90.90	91.30	**94.87**
10	Cheburek	82.81	86.71	92.20	94.30	**97.03**

Figure 3 illustrates the class-wise detection performance of Faster R-CNN, RetinaNet, YOLOv12, YOLO26, and RT-DETR by presenting the top 10 best-detected classes and the top 10 most challenging classes based on mAP50-95. The left chart shows that all models achieve high detection accuracy for visually distinctive food categories such as pumpkin seeds, fried aubergine, grilled vegetables, udon, and quesadilla. YOLO-based models often achieve the highest scores among the evaluated detectors, indicating strong performance on structured and visually separable dishes.

In contrast, the right chart presents the most challenging categories, where almost all models experience a significant drop in performance. Classes such as sweets, desserts, honey, and corn flakes show relatively low mAP scores, particularly for Faster R-CNN and RetinaNet. These results suggest that visually ambiguous, texture-similar, or less structured food items remain difficult for detection models. Although YOLO-based architectures still maintain higher scores compared to other methods, the performance gap between models becomes smaller for these challenging classes. Overall, the figure highlights the robustness of YOLOv12 and YOLO26 across both easy and difficult food categories, while also emphasizing the inherent complexity of certain Central Asian food classes.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

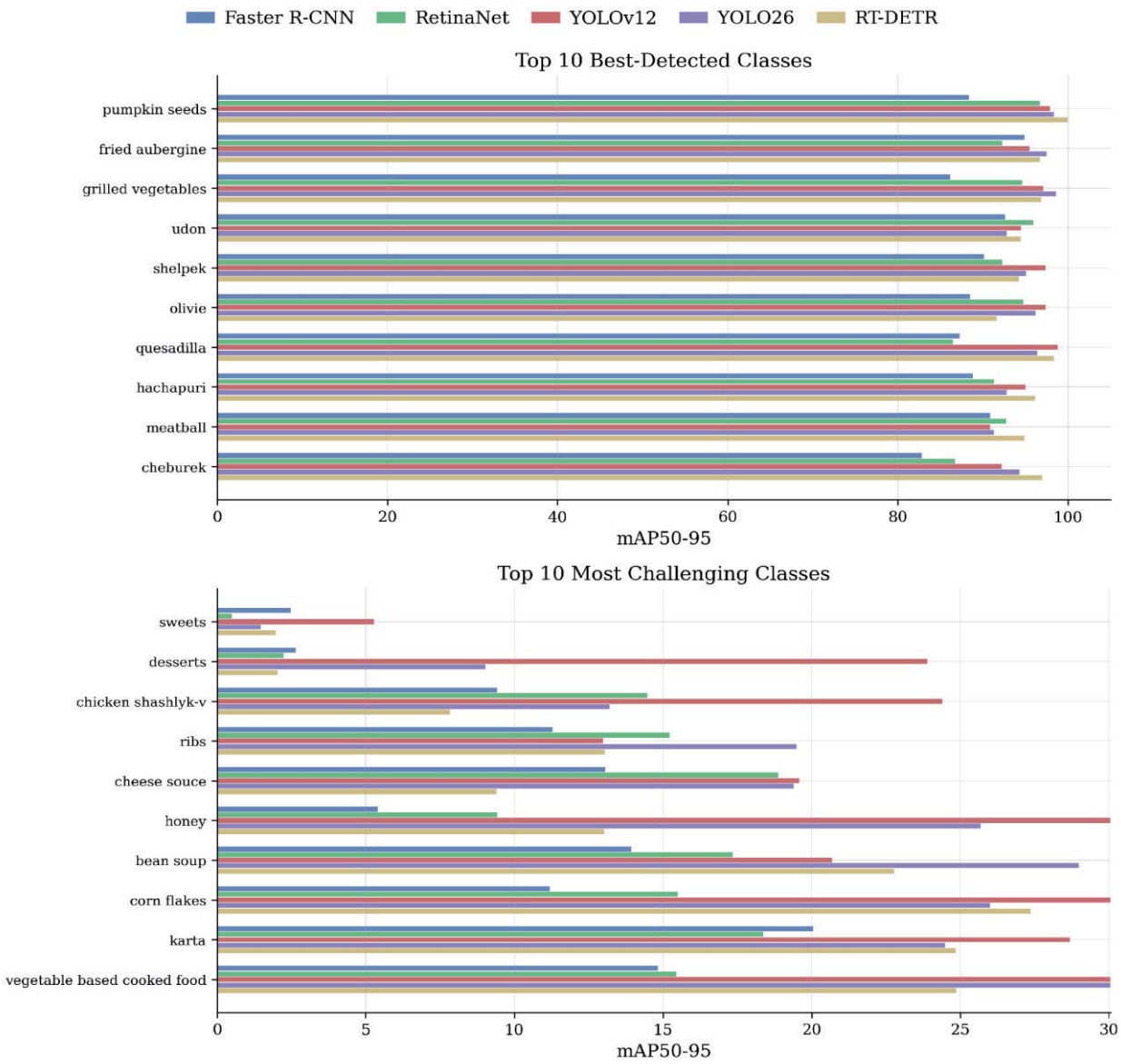

Figure 3. Class-wise detection performance comparison: Top 10 best-detected and top 10 most challenging food classes (mAP50-95).

Figure 4 shows model predictions on a multi-item tray scene. This type of image is challenging because several items are small, placed close together, and partially overlap. Some categories also share similar color and texture, which increases fine-grained confusion. In this example, Faster R-CNN detects several regions but shows label mismatches and misses some annotated items such as smoked fish and bell pepper. RetinaNet returns fewer detections and misses several annotated items. YOLOv12 and YOLO26 detect the main items, while some errors appear as confusion between visually similar categories or missed small side items. RT-DETR returns many detections, but some labels change between similar-looking classes when items are tightly placed. This example highlights common difficulties in dense meal scenes, where small objects and similar looking food items become harder to separate.

DOI: 10.37943/BZIF1573

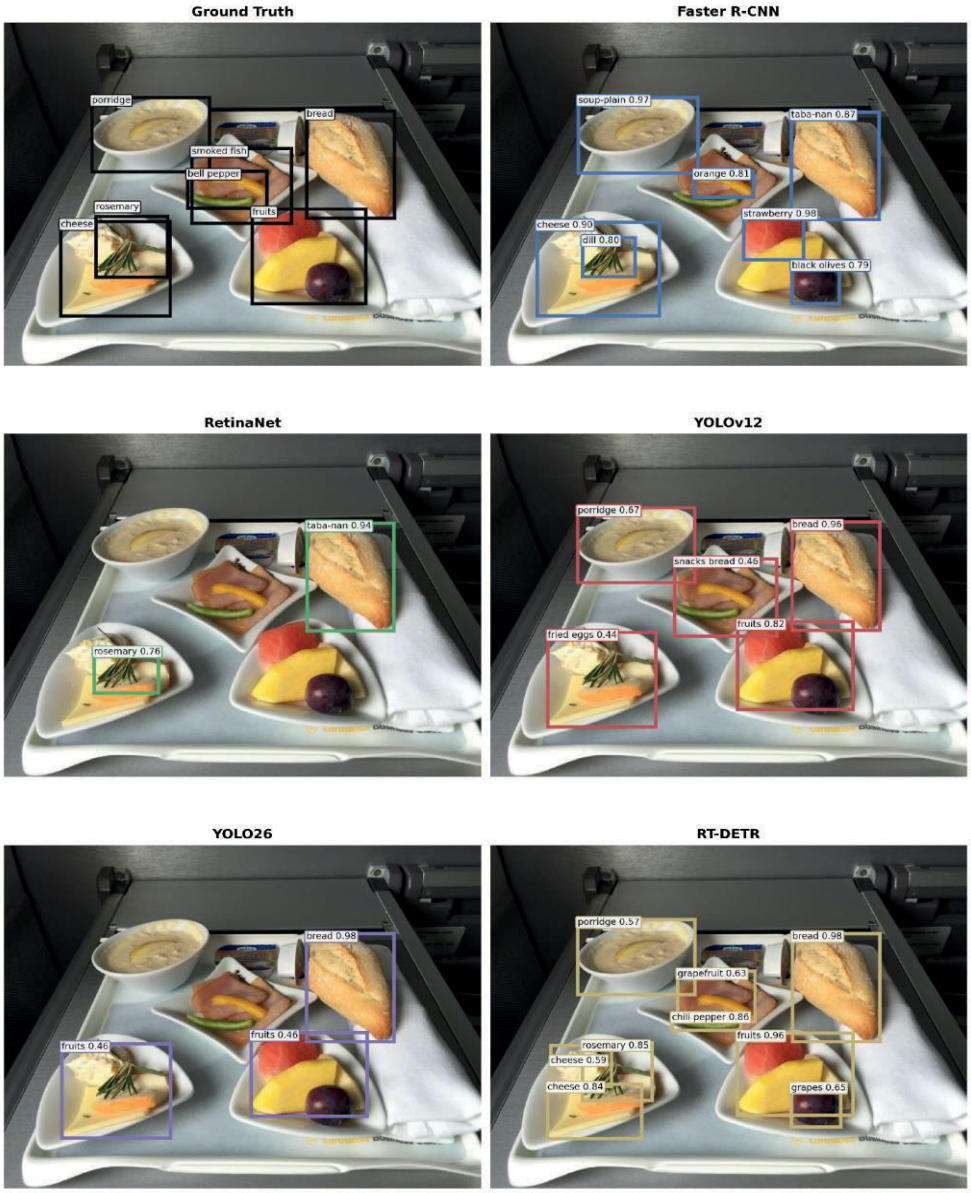

Figure 4. Visual comparison of detector outputs on a multi-Item meal scene.

Discussion

This study compared five detection models on a unified Central Asian food dataset with 258 classes. YOLOv12 gave the best overall performance, with the highest mAP50 and mAP50-95, while also having the smallest parameter count among the evaluated models. The comparison also highlights an important practical consideration. Higher model size did not automatically lead to better accuracy. RT-DETR achieved strong results and outperformed YOLO26 in accuracy, but it used a much larger model. In contrast, YOLOv12 reached higher accuracy with lower complexity. For deployment scenarios such as mobile-assisted dietary monitoring or clinic-side tools, this balance between quality and efficiency is critical. Class-level analysis gives more context. All models performed well on visually distinctive classes, where structure and shape are easier to separate. However, performance dropped for challenging classes with similar texture, mixed composition, or weak boundaries. This suggests that model architecture alone does not solve the full problem. Dataset characteristics still dominate hard cases, especially in food scenes where items overlap or look alike.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

Figure 4 helps interpret the numeric results by illustrating the kinds of errors that appear in realistic multi-item meals. In practice, overall mAP does not fully capture two recurring difficulties. One is coverage of small side items, which occupy limited image area and are easy to miss when many foods are packed into the same region. The other is category specificity: several classes differ only by small visual differences in color, texture, or plating, so predictions can shift to a closely related label even when the localization is reasonable. This pattern matches the class-wise results in Figure 3, where visually distinctive foods score highly across models, while foods with weaker boundaries or mixed composition remain challenging.

No class-balancing techniques were applied in this benchmark. Each model was trained with its framework's default loss and sampler, leaving the long-tailed class distribution unaddressed. This was intentional, so that the comparison reflects the architectures themselves rather than the effect of long-tail resolving. Class-balanced loss functions, rare-class oversampling, and targeted augmentation for underrepresented dishes are left to future work.

This work adds new evidence for Central Asian cuisine, a domain still underrepresented in large benchmark studies. The results therefore support both a practical recommendation (YOLOv12 as a strong default model) and a research direction (better handling of hard and rare classes). However, this study has its limitations. First, although the merged dataset is large, some classes remain sparse. Second, the benchmark focuses on bounding-box detection and does not evaluate downstream nutrition variables. Future work should test long-tail-aware training, targeted augmentation for difficult classes, and open-vocabulary or multimodal methods. It is also important to evaluate robustness under real user conditions, including lighting changes, occlusion, packaging, and mixed plates.

Conclusion

This study focuses on an important gap in food computing: reliable object detection for Central Asian cuisine in realistic, multi-item meal scenes. By evaluating multiple detector families on a unified benchmark, the work shows that strong food recognition performance is achievable for this domain and that model choice should consider both detection quality and practical deployment limits. The main value of this paper goes beyond reporting leaderboard numbers. It provides evidence about where progress is still needed for real-world use, especially for rare classes and visually similar dishes that remain difficult across architectures. These findings define a concrete direction for future research: improve representation of under-sampled foods, strengthen learning on difficult categories, and evaluate models in conditions closer to daily dietary monitoring.

Acknowledgements

This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23485288) as well as Nazarbayev University, under the Faculty Development Competitive Research Grant Program (Grant No. 201223FD2603).

CRediT authorship contribution statement

Ademi Sharipova: Methodology, Data curation, Formal analysis, Investigation, Validation, Visualization, Writing - original draft. **Baimyrza Kalmyrzayev:** Methodology, Writing - original draft, Writing - review & editing. **Aidana Mashrapova:** Data curation, Investigation, Writing - original draft, Writing - review & editing. **Aigerim Yeleussizova:** Data curation, Investigation, Writing - original draft, Writing - review & editing. **Huseyin Atakan Varol:** Funding acquisition, Project administration, Resources, Supervision, Writing - review & editing. **Mei Yen Chan:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing - review & editing.

References

[1] Liu, D., Zuo, E., Wang, D., He, L., Dong, L., & Lu, X. (2025). Deep Learning in Food Image Recognition: A Comprehensive review. *Applied Sciences*, 15(14), 7626. https://doi.org/10.3390/app15147626

[2] Kaushal, S., Tammineni, D. K., Rana, P., Sharma, M., Sridhar, K., & Chen, H. (2024). Computer vision and deep learning-based approaches for detection of food nutrients/nutrition: New insights and advances. *Trends in Food Science & Technology*, 146, 104408. https://doi.org/10.1016/j.tifs.2024.104408

[3] Tahir, G. A., & Loo, C. K. (2021). A Comprehensive Survey of Image-Based Food Recognition and Volume Estimation Methods for Dietary Assessment. *Healthcare*, *9*(12), 1676. https://doi.org/10.3390/healthcare9121676

[4] Zhang, Y., Deng, L., Zhu, H., Wang, W., Ren, Z., Zhou, Q., Lu, S., Sun, S., Zhu, Z., Gorriz, J. M., & Wang, S. (2023). Deep learning in food category recognition. *Information Fusion*, 98, 101859. https://doi.org/10.1016/j.inffus.2023.101859

[5] Min, W., Wang, Z., Liu, Y., Luo, M., Kang, L., Wei, X., Wei, X., & Jiang, S. (2023). Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8), 9932–9949. https://doi.org/10.1109/tpami.2023.3237871

[6] Niu, C., Ying, X., Pei, G., Hu, M., & Zhai, G. (2024). Review of the deep learning for food image processing. *International Journal of Agricultural and Biological Engineering*, *17*(5), 15–30. https://doi.org/10.25165/j.ijabe.20241705.8975

[7] Vlachopoulou, V., Sarafis, I., & Papadopoulos, A. (2023). Food Image Classification and Segmentation with Attention-Based Multiple Instance Learning. In *2023 18th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP)* (pp. 1–5). *IEEE*. https://doi.org/10.1109/SMAP59435.2023.10255199

[8] He, J., Mao, R., Shao, Z., Wright, J. L., Kerr, D. A., Boushey, C. J., & Zhu, F. (2021). An end-to-end food image analysis system. *Electronic Imaging, 33*(8), 285-1–285-7. https://doi.org/10.2352/ISSN.2470-1173.2021.8.IMAWM-285

[9] Ramesh, A., Raju, V., Rao, M., & Sazonov, E. (2021). Food Detection and Segmentation from Egocentric Camera Images. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 2736–2740). IEEE. https://doi.org/10.1109/EMBC46164.2021.9630823

[10] Tan, S. W., Lee, C. P., Lim, K. M., & Lim, J. Y. (2023, August). Food Detection and Recognition with Deep Learning: A Comparative Study. In *2023 11th International Conference on Information and Communication Technology (ICoICT)* (pp. 283–288). IEEE. https://doi.org/10.1109/ICoICT58202.2023.10262523

[11] Karabay, A., Bolatov, A., Varol, H.A., & Chan, M. (2023). A central Asian Food dataset for personalized dietary interventions. *Nutrients*, 15(7), 1728. https://doi.org/10.3390/nu15071728

[12] Karabay, A., Varol, H. A., & Chan, M. Y. (2025). Improved food image recognition by leveraging deep learning and data-driven methods with an application to Central Asian Food Scene. *Scientific Reports*, 15(1), 14043. https://doi.org/10.1038/s41598-025-95770-9

[13] Sanatbyek, A., Karabay, A., Varol, H. A., & Chan, M. Y. (2025, September). Deep Object Recognition-Based Analysis of Diverse Culinary Landscapes. In *2025 IEEE International Conference on Image Processing (ICIP)* (pp. 1127–1132). IEEE. https://doi.org/10.1109/ICIP55913.2025.11084477

[14] Thames, Q., Karpur, A., Norris, W., Xia, F., Panait, L., Weyand, T., & Sim, J. (2021). Nutrition5K: Towards Automatic Nutritional Understanding of Generic Food. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8903–8911). IEEE/CVF. https://doi.org/10.1109/CVPR46437.2021.00879

[15] Sanatbyek, A., Rakhimzhanova, T., Nurmanova, B., Omarova, Z., Rakhmankulova, A., Orazbayev, R., Varol, H. A., & Chan, M. Y. (2025). A multitask deep learning model for food

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

scene recognition and portion estimation—the Food Portion Benchmark (FPB) dataset. *IEEE Access*, 13, 152033–152045. https://doi.org/10.1109/access.2025.3603287

[16] Global Nutrition Report. (2025). Kazakhstan: Country nutrition profile. https://globalnutritionreport.org/resources/nutrition-profiles/asia/central-asia/kazakhstan/

[17] Omarova, Z., Nurmanova, B., Sanatbyek, A., Varol, H. A., & Chan, M.-Y. (2025). Digital Mapping of Central Asian Foods: Towards a Standardized Visual Atlas for Nutritional Research. *Nutrients*, 17(21), 3315. https://doi.org/10.3390/nu17213315

[18] Wang, Z., Chen, Y., Gu, Y., Liu, J., Zhu, X., & He, M. (2026). The evolution of object detection from CNNs to transformers and multi-modal fusion. *Scientific Reports*, *16*(1). https://doi.org/10.1038/s41598-026-37052-6

[19] Luo, Y., Cao, X., Zhang, J., Guo, J., Shen, H., Wang, T., & Feng, Q. (2022). CE-FPN: Enhancing Channel Information for Object Detection. *Multimedia Tools and Applications, 81*(21), 30685–30704. https://doi.org/10.1007/s11042-022-11940-1

[20] Oksuz, K., Cam, B. C., Kalkan, S., & Akbas, E. (2021). Imbalance Problems in Object Detection: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. https://doi.org/10.1109/tpami.2020.2981890

[21] Tian, Y., Ye, Q., & Doermann, D.S. (2025). YOLOv12: Attention-Centric Real-Time Object Detectors. In *Advances in Neural Information Processing Systems* (Vol. 38, pp. 78433–78457). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2025/hash/7103444259031cc58051f8c9a4868533-Abstract-Conference.html

[22] Jocher, G., & Qiu, J. (2026). Ultralytics YOLO26 (Version 26.0.0) [Computer software]. https://github.com/ultralytics/ultralytics

[23] Cao, J., Peng, B., Gao, M., Hao, H., Li, X., & Mou, H. (2025). Object detection based on CNN and vision-transformer: A survey. *IET Computer Vision*, 19(1), e70028. https://doi.org/10.1049/cvi2.70028

[24] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. (2024). DETRs beat YOLOs on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 16965-16974). https://doi.org/10.1109/CVPR52733.2024.01605