AI-Based Estimation from Images of Food Portion Size and Calories for Healthcare Systems

Akmalbek Abdusalomov[1](✉), Mukhriddin Mukhiddinov[2](✉), Oybek Djuraev[2], Utkir Khamdamov[2], and Ulugbek Abdullaev[3]

[1] Department of Computer Engineering, Gachon University Sujeong-Gu, Gyeonggi-Do, Seongnam-Si 461-701, South Korea
akmalbek@gachon.ac.kr

[2] Department of Communication and Digital Technologies, University of Management and Future Technologies, Tashkent 100208, Uzbekistan
{mukhiddinov,djuraev}@umft.uz

[3] Department of Hardware and Software of Management Systems in Telecommunication, Tashkent University of Information Technologies named after Muhammad al Khwarizmi, Tashkent 100084, Uzbekistan

Abstract. In the realm of nutrition science, it is well-recognized that individuals' dietary needs vary based on factors such as age, gender, and health status. This divergence in nutritional requirements is particularly critical for vulnerable groups, including newborns, the elderly, and individuals with diabetes, as their dietary choices can have profound implications for their health. Moreover, the dearth of Uzbek recipes in mainstream culinary literature, which predominantly focuses on Western cuisine, exacerbates the issue. To address these challenges, this study undertakes the ambitious task of constructing a comprehensive AI system, comprising both backend and frontend components, tailored to the nuances of Uzbek cuisine. The primary objectives encompass recipe classification, ingredient identification and localization, and the estimation of nutritional content and calorie counts for each dish. Convolutional Neural Networks (CNNs) are employed as the cornerstone of this computational solution, proficiently handling image-based tasks, including the recognition of diverse food items and portion size determination within Uzbek recipes. Food classification is executed using MobileNet, while the You-Only-Look-Once (YOLO) network plays a pivotal role in the dual functions of ingredient classification and localization within dishes. Upon rigorous training, testing, and system deployment, users can effortlessly capture images of food items through the smartphone application, facilitating the estimation of nutritional data and calorie counts. Ultimately, this vital information is presented to users via the smartphone interface, bridging the accessibility gap and enhancing comprehension of nutritional aspects within Uzbek cuisine.

Keywords: Food calories prediction · Deep learning · Uzbek dishes · Food classification · MobileNet · CNN · YOLO

B. J. Choi et al. (Eds.): IHCI 2023, LNCS 14532, pp. 9–19, 2024.
https://doi.org/10.1007/978-3-031-53830-8_2

1 Introduction

Contemporary generations exhibit a heightened consciousness regarding dietary choices and calorie intake due to the recognition of the adverse implications of excessive caloric consumption, particularly on body weight, a pressing public health concern. The integration of Artificial Intelligence (AI), specifically deep learning, presents a promising avenue for comprehending the nutritional content of consumed food items. Food consumption is an integral facet of daily existence, with profound repercussions on body weight. Alarming data from the World Health Organization (WHO) underscore the surge in global rates of overweight and obesity over the past four decades. In 2016, the global count of overweight adults exceeded 1.9 billion individuals, with over 650 million among them classified as obese. Closer examination of national statistics, such as those from Norway in the year 2017, as reported by the Norwegian Institute of Public Health (NIPH) in their late 2017 update, reveals that 25% of men and 20% of women in the country grappled with obesity. This epidemiological trend is of grave concern, as it substantially elevates the risk profile for debilitating diseases, including diabetes, cardiovascular conditions, musculoskeletal disorders, and various types of malignancies [1, 2].

AI holds promise as a tool for promoting healthier lifestyles. Employing sophisticated neural networks like ResNet, it becomes feasible to identify food items accurately. The process of segmentation, involving the subdivision of images into smaller, meaningful units, is facilitated through the application of the YOLO object detection, enabling the comprehensive identification of pertinent components within the images [3–5]. Subsequently, these segmented constituents of food can be leveraged to compute the weight of the food depicted in the image, achieved through the utilization of an inception network. This proposed solution marks a pioneering endeavor in the domain of estimating food weight solely from a single image. The outcomes gleaned from this thesis investigation underscore the network's capacity to make reliable weight predictions, demonstrating a standard error of 8.95 across all categories and a notably lower 2.40 for the most accurately predicted category, namely, bread.

While food classification has been extensively explored in prior research [6–8], the challenge of determining the nutritional content of food items depicted in images remains a significant gap in the field. A prominent hurdle in assessing nutritional levels from images is the inherent two-dimensional nature of images, making it intricate to accurately gauge the portion size of the food items contained within. Consequently, it becomes imperative to devise methodologies for quantifying the volume of food represented in the images. Furthermore, variations in food preparation methods introduce another layer of complexity. Discerning the nutritional content of an apple, for instance, is more straightforward compared to a complex dish like "plow" (as illustrated in the top-left corner of Fig. 1), where numerous ingredients may be obscured within the image, rendering precise estimation challenging, if not unattainable. Nevertheless, it is plausible to derive approximations based on known recipes for such intricate dishes, albeit with some degree of discrepancy. Conversely, simpler food items present a more tractable scenario for assessing nutritional levels, as the image can encompass all constituents, facilitating a more accurate estimation process.

This paper presents a foundational demonstration in the realm of estimating food calories from images. In summary, our contributions encompass the development of a method for approximating the weight of food within a single image, achieved through the training of a neural network on established weight data for various food items. Consequently, the weight of the food can be predicted exclusively from a single image, thereby enabling the computation of calorie content by combining this weight information with data sourced from a comprehensive food database. Additionally, this thesis introduces a novel dataset comprising weight annotations for each individual food element depicted in the images.

2 Dataset of Uzbek Foods

To effectively identify food items and derive calorie information from images, the necessity for comprehensive training data is evident. In this paper, we have leveraged three distinct datasets, comprising two publicly accessible ones and a novel dataset created specifically for this study. The most prevalent dataset utilized for food classification purposes is the Food-101 dataset [9], featuring 1000 images for each of the 101 most commonly encountered food categories, thereby amassing a total of 101,000 images. Our initial experimentation with this dataset revealed the formidable challenges associated with the project. Accurate calorie estimation hinges on precise food classification across a broader spectrum of food classes, necessitating a more extensive dataset. It became apparent that the 101 food categories provided insufficient coverage, and the utilization of a significantly larger dataset would introduce complexities in terms of training and model iteration. Despite the prevailing reliance on conventional approaches in existing research, where both food classification and calorie estimation revolve around such datasets, we opted for an alternative path due to the intricate nature of food classification within this context.

A novel dataset has been curated specifically for the purpose of calorie estimation within this research, hereafter referred to as the “Uzbek dishes” dataset. This dataset encompasses a collection of food images, accompanied by a scale or ruler for reference, facilitating the determination of food proportions and weights. In addition to the visual content, the dataset includes a document detailing the nutritional content for each food item depicted in the images. The “Uzbek dishes” dataset encompasses a total of 10 distinct categories of Uzbek cuisine, encompassing approximately 12,000 images, with a select few exemplified in Fig. 1. Each category is further subdivided into six distinct groups for estimating the quantity of food, denoted as E1, E2, E3, E4, E5, and E6, as illustrated in Fig. 2.

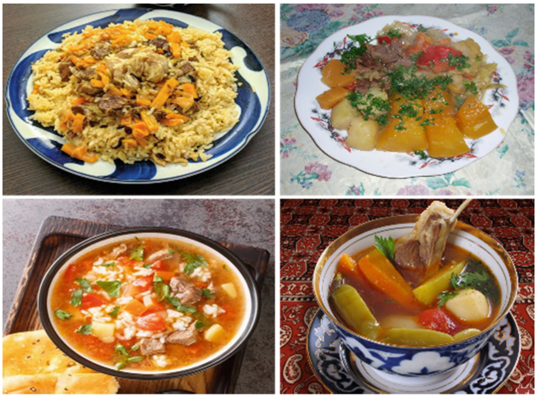

Fig. 1. Sample images from custom Uzbek dish dataset.

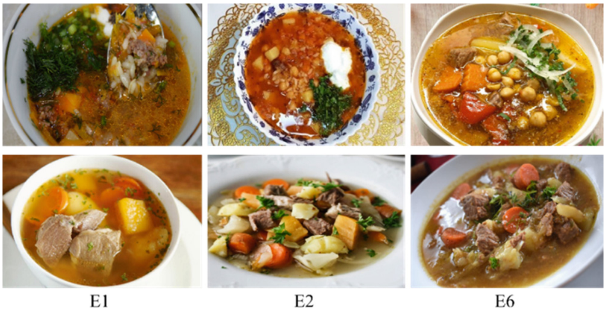

Fig. 2. Sample images for six amount estimation group such as E1, E2, and E6.

Our methodology comprises two primary components: Object and Food classification utilizing YOLO, and Food Quantity estimation employing Wide-Resnet50 [10]. To facilitate these tasks, we harnessed two distinct datasets tailored to each algorithm's specific requirements. Initially, a set of 12,000 images was employed solely for food item classification and detection, categorized into 10 distinct classes. These base images possessed dimensions of 4032 × 3024 pixels.

Subsequently, we adopted classical computer vision techniques with predefined parameters to segment and extract the ingredients within the images. These methods involve the computation of pixel intensities and their conversion into real-sized regions. Building upon our prior research efforts [11–13], we employed a range of image processing techniques to identify and crop objects from sequences of images. The pixel regions corresponding to non-uniform components were segmented, enabling the determination of nutritional information and calorie content through reference to a standardized information table [14].

In the context of this study, we developed a comprehensive, full-stack system that operates based on the RESTful protocol, utilizing JSON format for communication between a smartphone and an artificial intelligence server.

3 Proposed Method

The YOLOv3 model architecture incorporates the DarketNet53 backbone [15]. Our proposed methodology encompasses food detection, food quantity estimation, and object detection, which entails the following steps:

- Identification of regions corresponding to Uzbek dishes through YOLOv3.
- Subdivision of the detected regions into multiple segments for the purpose of object detection.
- Extraction of food images based on the bounding box boundaries.
- Classification of food quantity estimates using Wide-Resnet-50

The initial phase of our proposed approach involves object detection and object classification utilizing YOLOv3. This convolutional network concurrently predicts multiple bounding boxes and the associated class probabilities for these boxes, streamlining the detection process. YOLOv3 adopts a unique approach by training on entire images and directly optimizing overall performance. This integrated model boasts several advantages over conventional object detection techniques. Notably, YOLOv3 excels in terms of rapidity in object recognition. It conceptualizes detection as a regression problem, eliminating the need for intricate pipelines [16, 17]. During testing, object detection is achieved by running a neural network on the new image.

Following object detection, we encountered certain challenges, notably that the YOLOv3 object detection model struggled to identify all objects due to limitations in the dataset size [18]. To address this issue, we devised a solution by partitioning the input image into several segments.

A. **Quantity Estimation**

The network architecture we employed is based on the Wide-Resnet-50, which is a deep convolutional neural network (CNN). Wide-Resnet-50 is an extension of the ResNet architecture, known for its ability to handle very deep networks effectively. It consists of 50 layers and is wider than the original ResNet, making it suitable for the complexity of food quantity estimation tasks. For the training data, we collected a diverse and extensive dataset of food images with corresponding ground truth labels for food quantity. This dataset includes images of various food items, each annotated with the respective quantity or portion size. The dataset is essential for training the Wide-Resnet-50 model to learn how to estimate food quantities accurately (Fig. 3).

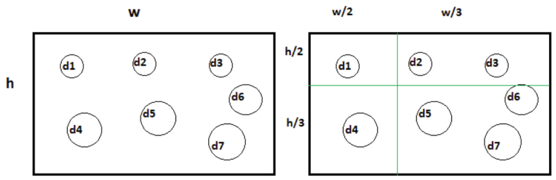

Fig. 3. Decreasing object with dividing the photo into several parts.

To enhance the effectiveness and resilience of our model, we employed data augmentation techniques. Specifically, we computed the mean and standard deviation for dataset normalization and incorporated methods such as Color-Jitter [19], RandomRotation, RandomAdjustSharpness, and RandomSharpness. Figure 4 provides visual representations of cropped images used for food quantity estimation, with E3, E6 representing 15%, 50%, and 100%, respectively. Additionally, Fig. 5 showcases the outcomes of data augmentation, illustrating the various augmentation types we previously mentioned.

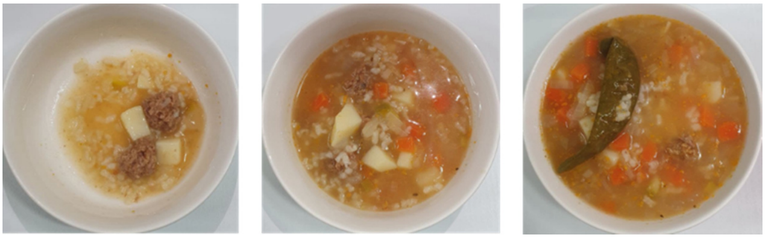

Fig. 4. Cropped images for a food amount estimation.

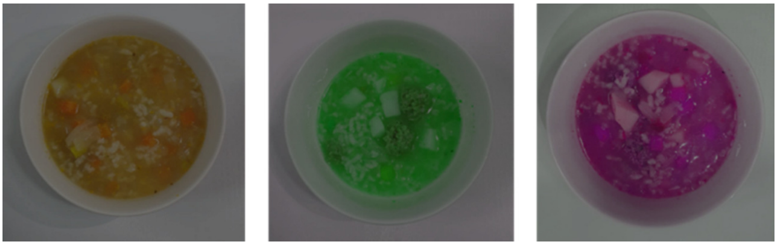

Fig. 5. The results of data augmentation.

As previously mentioned, we employed Convolutional Neural Networks (CNNs) for food quantity prediction. In our quest to select the most suitable CNN architecture for this purpose, we conducted training and testing across several CNN models, including MobileNetV2 [20], Resnet [21], EfficientNet [22], and Dilated CNN [23].

MobileNetV2, with a depth of 53 layers and comprising 3.5 million parameters, emerged as one of our candidate networks. It incorporates distinctive architectural features, including:

- Depth wise separable convolution
- Linear bottleneck
- Inverted residual

The concept of Depthwise Separable Convolution is visually depicted in Fig. 6.

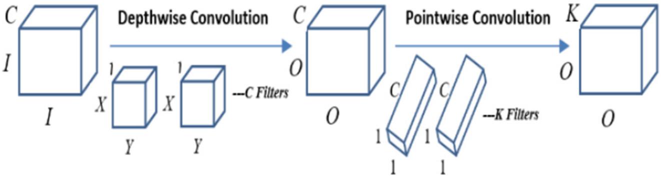

Fig. 6. The architecture of the Depthwise Separable Convolution.

MobileNetV2 employs a fusion of pointwise convolution and bottleneck structures, wherein pointwise convolution is employed to achieve the bottleneck effect. However, a challenge arises when applying ReLU activation after dimensionality reduction, resulting in the loss of valuable information. To mitigate this issue, a 1 × 1 convolutional layer, also known as a feature map pooling or projection layer, is introduced. This straightforward technique serves to reduce dimensionality by decreasing the number of feature maps while preserving their distinctive attributes. In our study, we utilized the max pooling technique for this purpose, and consequently, linear activation is implemented within the bottleneck to minimize information loss [25].

Furthermore, MobileNetV2 incorporates the concept of Inverted Residual blocks, following a pattern of narrow-wide-narrow configuration. This entails a sequence of 1 × 1 convolution, followed by 3 × 3 convolution, reducing parameter count, and concluding with another 1 × 1 convolution to further trim the number of channels. Importantly, shortcuts are established directly between these bottleneck stages [26].

4 Experimental Results and Analysis

In this section, we elucidate the experimental configuration and outcomes pertaining to the estimation of calories and quantity for Uzbek dishes using the YOLOv3 and MobileNetV2 models. The proposed method was implemented and carefully tested within Visual Studio 2019 C++, executed on a system equipped with an AMD Ryzen 9 5900X 12-Core Processor operating at 3.70 GHz, supported by 64 GB of RAM, and a NVIDIA RTX 3090 Graphics card. Our food dataset was employed for both training and testing purposes. Crucial parameters for our training experiments included a batch size of 32 pixels, an input image size of 416 × 416 pixels, a learning rate of 0.001, and a subdivision of 8. Our primary objective was to evaluate the classification performance, ensuring the reliable and accurate classification of Uzbek dishes. This study conducts a comprehensive analysis and comparison of various object detection and recognition techniques, encompassing YOLOv3 and MobileNetV2, in the training and evaluation of models for classifying Uzbek dishes. The findings underscore the precise detection capabilities of YOLOv3 in accurately identifying Uzbek dishes.

Our quantitative experiments entailed the application of object detection evaluation metrics, namely precision, recall, and accuracy, in line with our prior research endeavors [27]. These metrics serve as valuable tools for assessing and interpreting the outcomes. Precision gauges the classifier's capability to correctly identify relevant objects, quantifying the proportion of true positives detected. In contrast, recall measures the model's proficiency in identifying all pertinent instances, signifying the proportion of true positives among all ground truth cases. An effective model excels in recognizing the majority of ground-truth objects, showcasing high recall, while also limiting the inclusion of irrelevant objects, denoting high precision.

An ideal model would yield a false-negative value of 0 (recall = 1) and a false-positive value of 0 (precision = 1), signifying flawless performance. We computed precision and recall rates through the comparison of pixel-level ground-truth images with the outcomes generated by our proposed method. To calculate these metrics for food classification, we employed the following equations.:

$$\text{Accuracy} = \text{TP} + \text{TN}/\text{TP} + \text{TN} + \text{FP} + \text{FN} \tag{1}$$

$$\text{Recall} = \text{TP}/\text{TP} + \text{FN} \tag{2}$$

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP} \tag{3}$$

In Fig. 7, it is depicted that the YOLOv3 model achieved an accuracy rate of 98.2% and exhibited a loss of 0.2 when detecting objects.

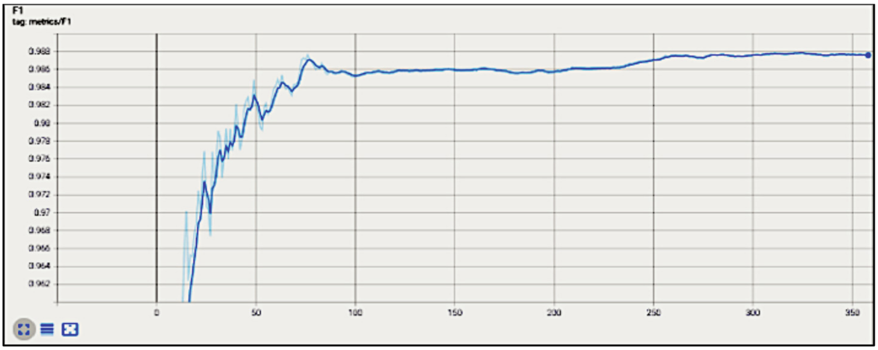

Fig. 7. The accuracy of the Yolov3 model.

We generated an additional dataset for quantity estimation purposes by resizing images to dimensions of 224 × 224 pixels for training with Wide-Resnet-50. Each of the food categories and quantity estimations comprises a set of 200 images. Consequently, this classification task incorporates a total of 5000 food images. The dataset was partitioned into training and validation subsets, with an 80% allocation to training and the remaining 20% allocated to validation. The results of the food quantity estimation using Wide ResNet-50 are shown in Table 1.

Table 1. Performance evaluation of the Wide ResNet-50 model for food quantity estimation.

Wide ResNet-50	Training	Testing
Accuracy	93%	90%
Recall	90%	88%
Precision	92%	91%

5 Limitation and Future Work

The proposed approach necessitates the incorporation of additional quantity estimation classes, as the utilization of just six such classes resulted in the omission of several categories by the MobileNetV2 model. In forthcoming research, we aspire to enhance our primary concept for quantity estimation. Additionally, we aim to optimize response times by deploying the service within a cloud computing environment, with the objective of reducing food identification duration [28].

Moreover, our future endeavors include the exploration of 3D modeling techniques for quantity estimation, which are expected to yield greater accuracy compared to the utilization of 2D images. We are currently in the process of acquiring expertise in the application of computer vision methodologies within the Unity environment. Our forward-looking plan entails the augmentation of algorithm complexity and the expansion of computer vision capabilities. For instance, our current approach involves the utilization of only ten quantity estimation positions per image; however, we aspire to incorporate a more extensive range of positions in the future.

6 Conclusions

This research introduces a Convolutional Neural Network (CNN) model designed to estimate the portion sizes of individual food items on a single plate based on single images. The envisioned outcome is the association of these portion sizes with a calorie database for each food unit. The primary objective of this paper is to leverage machine learning techniques for image classification, food segmentation, and the subsequent determination of weight and calorie content within the images. Additionally, we introduce a novel dataset termed “Uzbek dishes,” encompassing food images along with their corresponding weights. The findings from our testing phase affirm the capability of deep neural networks to effectively classify food images.

Future investigations will focus on enhancing the accuracy of our methodology through the utilization of advanced deep learning techniques. Furthermore, we intend to optimize cache memory utilization in multi-core processors to enhance the efficiency of food calorie detection and estimation [29].

References

1. World Health Organization: Obesity and Overweight (2018)
2. NIPH: Overweight and obesity in Norway. Tech. rep. (2014)
3. Mukhiddinov, M., Djuraev, O., Akhmedov, F., Mukhamadiyev, A., Cho, J.: Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people. Sensors **23**(3), 1080 (2023)
4. Mukhiddinov, M., Jeong, R.G., Cho, J.: Saliency cuts: salient region extraction based on local adaptive thresholding for image information recognition of the visually impaired. Int. Arab J. Inf. Technol. **17**(5), 713–720 (2020)
5. Mukhiddinov, M., Akmuradov, B., Djuraev, O.: Robust text recognition for Uzbek language in natural scene images. In: 2019 International Conference on Information Science and Communications Technologies (ICISCT), pp. 1–5. IEEE. (2019)

6. Sathish, S., Ashwin, S., Quadir, M.A., Pavithra, L.K.: Analysis of convolutional neural networks on indian food detection and estimation of calories. In: Materials Today: Proceedings, 16 Mar (2022)
7. Li, S., Zhao, Y., Liu, S.: How food shape influences calorie content estimation: the biasing estimation of calories. J. Food Qual. 24 May (2022)
8. Kumar, R.D., Julie, E.G., Robinson, Y.H., Vimal, S., Seo, S.: Recognition of food type and calorie estimation using neural network. J. Supercomput. **77**(8), 8172–8193 (2021)
9. Bossard, L., Guillaumin, M., Gool, L.V.: Food-101–mining discriminative components with random forests. In: European Conference on Computer Vision, pp. 446–461. Springer, Cham (2014)
10. Keras: Resnet-50 [Online]. Available: https://www.kaggle.com/keras/resnet50 (2017).
11. Rakhmatillaevich, K.U., Ugli, M.M.N., Ugli, M.A.O., Nuruddinovich, D.O.: A novel method for extracting text from natural scene images and TTS. Eur. Sci. Rev. **1**(11–12), 30–33 (2018)
12. Mukhamadiyev, A., Mukhiddinov, M., Khujayarov, I., Ochilov, M., Cho, J.: Development of language models for continuous Uzbek speech recognition system. Sensors **23**(3), 1145 (2023)
13. Abdusalomov, A.B., Mukhiddinov, M., Whangbo, T.K.: Brain tumor detection based on deep learning approaches and magnetic resonance imaging. Cancers **15**(16), 4172 (2023)
14. Khamdamov, U., Abdullayev, A., Mukhiddinov, M., Xalilov, S.: Algorithms of multidimensional signals processing based on cubic basis splines for information systems and processes. J. Appl. Sci. Eng. **24**(2), 141–150 (2021)
15. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
16. Ege, T., Yanai, K.: Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions. In: Proceedings on Thematic Workshops of ACM Multimedia, pp. 367–375 (2017)
17. Mukhiddinov, M., Abdusalomov, A.B., Cho, J.: Automatic fire detection and notification system based on improved YOLOv4 for the blind and visually impaired. Sensors **22**(9), 3307 (2022)
18. Mukhiddinov, M., Cho, J.: Smart glass system using deep learning for the blind and visually impaired. Electronics **10**(22), 2756 (2021)
19. Jalal, M., Wang, K., Jefferson, S., Zheng, Y., Nsoesie, E.O., Betke, M.: Scraping social media photos posted in Kenya and elsewhere to detect and analyze food types. In: Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management, pp. 50–59 (2019)
20. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, LC.: Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520 (2018)
21. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: revisiting the resnet model for visual recognition. Pattern Recogn. **90**, 119–133 (2019)
22. Koonce, B.: EfficientNet. In: Convolutional Neural Networks with Swift for Tensorflow, pp. 109–123. Apress, Berkeley, CA (2021)
23. Yuldashev, Y., Mukhiddinov, M., Abdusalomov, A.B., Nasimov, R., Cho, J.: Parking lot occupancy detection with improved mobilenetv3. Sensors **23**(17), 7642 (2023)
24. Abdusalomov, A., Mukhiddinov, M., Djuraev, O., Khamdamov, U., Whangbo, T.K.: Automatic salient object extraction based on locally adaptive thresholding to generate tactile graphics. Appl. Sci. **10**(10), 3350 (2020)
25. Chen, G., et al.: Food/non-food classification of real-life egocentric images in low-and middle-income countries based on image tagging features. Front. Artif. Intell. **4**, 644712 (2021)
26. Mukhiddinov, M.: November. Scene text detection and localization using fully convolutional network. In: 2019 International Conference on Information Science and Communications Technologies, pp. 1–5 (2019)

27. Khamdamov, U.R., Mukhiddinov, M.N., Djuraev, O.N.: An overview of deep learning-based text spotting in natural scene images. Problems of Computational and Applied Mathematics. Tashkent, **2**(20), 126–134 (2020)
28. Muminov, A., Mukhiddinov, M., Cho, J.: Enhanced classification of dog activities with quaternion-based fusion approach on high-dimensional raw data from wearable sensors. Sensors **22**(23), 9471 (2022)
29. Farkhod, A., Abdusalomov, A.B., Mukhiddinov, M., Cho, Y.I.: Development of real-time landmark-based emotion recognition CNN for masked faces. Sensors **22**(22), 8704 (2022)