Multiclass Semantic Segmentation of Mediterranean Food Images

Fotios S. Konstantakopoulos[1], Eleni I. Georga[1], and Dimitrios I. Fotiadis[1,2(✉)]

[1] Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, Ioannina, Greece
{fotkonstan,egeorga,fotiadis}@uoi.gr
[2] Biomedical Research Institute, FORTH, Ioannina, Greece

Abstract. With the continuous increase of artificial intelligence applications in modern life, the segmentation of images is one of the fundamental tasks in computer vision. Image segmentation is the key for many applications and is backed by a large amount of research, including medical image analysis, healthcare services and autonomous vehicles. In this study we present a semantic segmentation model for food images, suitable for healthcare systems and applications as major part of the dietary monitoring pipeline, trained on an annotation dataset of Mediterranean cuisine food images. To segment the images, we use for feature extraction the ResNet-101 CNN model pre-trained on the ImageNet LSVRC-2012 dataset as a backbone network and the Pyramid Scene Parsing Network - PSPNet architecture for food image segmentation. For the evaluation metric we use the Intersection over Union, where the proposed model achieves a meanIoU score 0.758 in 50 classes of the Mediterranean Greek Food image dataset and 0.933 IoU score in food/non-food segmentation. To evaluate the proposed segmentation model, we train and evaluate a U-Net segmentation model on the same dataset, which achieves meanIoU 0.654 and IoU score 0.901 in multiclass and food/non-food segmentation, respectively.

Keywords: Computer Vision · Image Segmentation · Semantic Segmentation · Deep Learning · Food Image Dataset · Dietary Assessment Systems

1 Introduction

In healthcare systems, image segmentation is used to segment biomedical images into different regions to assist physicians in diagnosing diseases [1]. Also, image segmentation can be used to dietary assessment applications, as part of the nutritional composition system, to assist individuals to follow a healthy diet, preventing chronic diseases such as obesity, diabetes, cardiovascular diseases (CVDs) and cancer [2]. Every year, millions of people are died from chronic diseases. For example, in 2021, diabetes was responsible for 6.7 million deaths worldwide [3]. A common factor that can affect the treatment of the above diseases is the management of the daily diet. Healthy habits are essential for the management of these diseases and, in some cases, changing the daily diet may be enough to control the disease.

A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 49–59, 2023.
https://doi.org/10.1007/978-3-031-34586-9_4

The image dataset is the key to creating a high-accurate model for a food image segmentation system. However, image datasets for deep learning segmentation models are hard to collect, because a lot of professional expertise is needed to label them. Moreover, the need for highly performance deep learning models requires the collection of large numbers of images. In dietary assessment systems there are a few datasets suitable for the training and evaluation of deep learning segmentation models. A food image dataset can be characterized by the total number of images they include, the number of food classes, the source of the food images, the type of cuisine and by the task they can be used (e.g., food segmentation, food classification or food volume estimation task). For example; Food524DB [4] represents a generic type of cuisine and consists of 247,636 food images with 524 food classes acquired from previous datasets; Vireo Food-172 [5] represents the Japanese cuisine and consists of 110,241 food images with 172 food classes downloaded from the web; while Food201-segmented [6] can be used for food image segmentation tasks.

The food image segmentation task plays an important role in AI applications for the daily management of nutrition [7]. These systems are divided in two main categories: (i) traditional machine learning segmentation approaches with handcrafted feature extraction [8], and (ii) deep learning segmentation approaches with automatic feature extraction [9]. Traditional machine learning approaches, use feature extraction algorithms, such as Gabor features and Speed-up robust features (SURF), to find and extract the features of the food image and then, the features are fed to a classifier, such as random forests, to segment the food [10]. In [11], an interactive food image segmentation algorithm has been proposed, where food parts are extracted based on user's inputs in the first step and then, a boundary detection and filling and the Gappy principal component analysis methods are applied to restore the missing information.

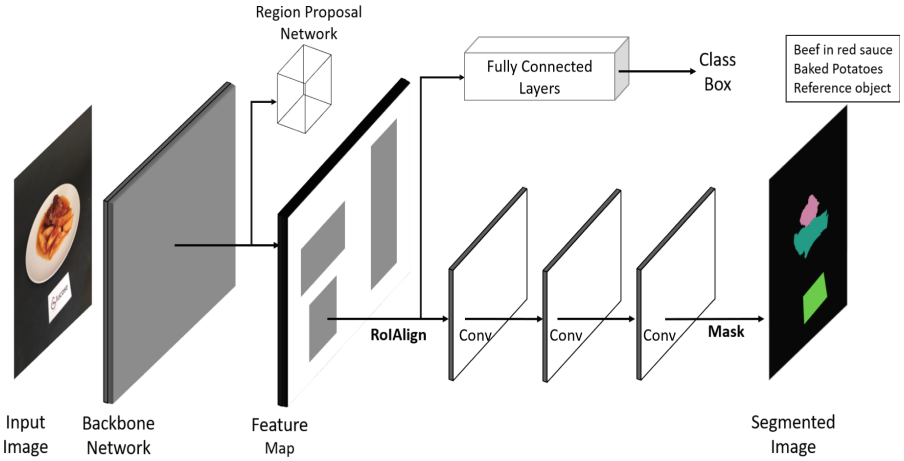

Fig. 1. An instance segmentation model for food items pixel classification.

Today, Convolutional Neural Networks (CNNs) a class of deep neural networks (DNN), are the state of the art methodology in food image segmentation systems. CNN

models are extremely accurate for computer vision tasks and surpass the traditional machine learning models in image segmentation Intersection over Union (IoU) metric. The deep learning techniques that are used for image segmentation are divided into semantic and instance segmentation techniques (Fig. 1). Semantic segmentation models provide segment maps as outputs that correspond only to the inputs they are fed, while instance segmentation models detect and delineate each distinct object of interest that appears in the image. In food image segmentation, semantic segmentation techniques are mainly used, which aim either to segment the food from the background, or to segment the different types of food contained in the image. For example, in [12] they proposed a semantic segmentation deep learning model which achieved 0.931 IoU score in food/no-food segmentation task using a DeepLab-V2 model in the UNIMIB-2016 [13] food image dataset, while in [9] they achieved 0.439 meanIoU for the segmentation of 103 food labels, by combining the proposed Recipe Learning Module (ReLe) and the Segmentation Transformer (SeTR) [14].

In this study, we propose a semantic segmentation network to segment images containing Mediterranean foods. Using a new food image annotated dataset, we present the architecture of the proposed segmentation model and its training pipeline. Our network is suitable for detecting specific food items as well as for separating food from the background using a semantic segmentation model. Although the segmentation step is not necessary in several dietary assessment systems, we observe that the studies using the segmentation step, result in better performance [15]. Relative to similar approaches, the innovation of this study is that it proposes a state of the art pre-trained DCNN model for food image segmentation using a novel annotated dataset of Mediterranean cuisine food images. While most related approaches focus on either handcrafted feature extraction or using existing annotated datasets [15], we propose a deep learning model for feature extraction and a new annotated food image dataset, which can be used in classification and volume estimation stages in dietary assessment systems. The proposed model can be part of dietary assessment systems and applications, by improving the accuracy of image classification and food volume estimation stages. In addition, it is suitable for healthcare systems that monitor patient malnutrition in hospitals, offering the ability to track the different food items served with the tray to the patient. Finally, to prove the dominance of the proposed segmentation model, we compare the performance of an additional segmentation model using a different architecture.

2 Methods

2.1 Food Image Dataset

In the present study, for the training and the evaluation of the segmentation model, we use two image datasets: (i) the ImageNet LSVRC-2012, and (ii) the MedGRFood[1] [16]. ImageNet is a large image dataset, that contains 1,431,167 images belonging to 1,000 object classes, such as bus, dog, puzzle etc. The MedGRFood is a new food image dataset, which contains 51,840 Mediterranean food images belonging to 160 classes appropriate for classification tasks and an additional 20,000 Mediterranean food images belonging to 190 classes appropriate for volume estimation tasks. All the images have been collected from the web and under a controlled environment along with their weight. For the proposed segmentation model, we annotated 5,000 food images of 50 classes from the MedGRFood dataset, with respect to the food category, the exact food name, the cuisine and the weight of food. Figure 2 shows images from the datasets ImageNet and MedGRFood.

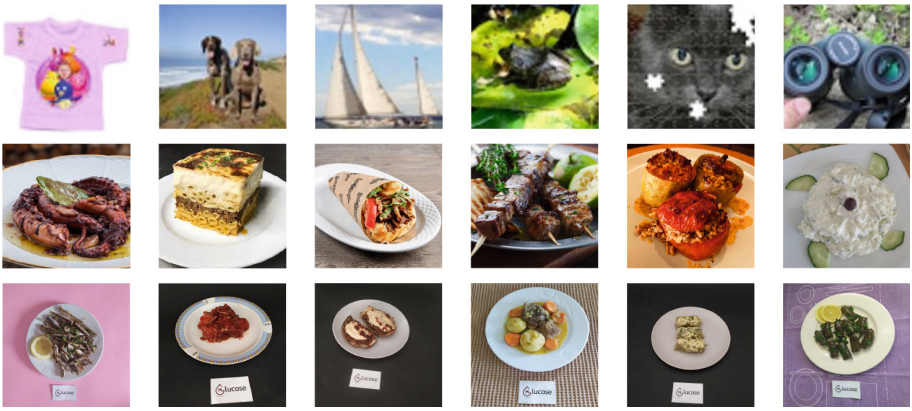

Fig. 2. Images from ImageNet LSVR and MedGRFood datasets. The first row shows images from the ImageNet dataset and the others rows show images from the MedGRFood dataset.

2.2 Segmentation Network Architecture

Knowing that most semantic segmentation models contain two main parts (i) an encoder and (ii) a decoder, for feature map extraction and pixel prediction, respectively, we choose the Pyramid Scene Parsing Network (PSPNet) [17] architecture for food image segmentation. Specifically, the proposed PSPNet encoder contains the ResNet-101 [18] model as backbone network with dilated convolutions along with the pyramid pooling module for feature extraction. ResNet-101 is a 101-layer deep CNN that democratizes the concepts of residual learning and skip the connections between some of the blocks.

[1] MedGRFood dataset is available for research purposes via the website: http://glucoseml.gr/.

We use a pretrained version of ResNet-101, trained on the ImageNet dataset. Knowing that the early layers on CNNs extract and learn general features (such as edges and simple textures) while the later layers extract and learn detailed or high-level features (such as more complex textures and patterns), we take advantage of the ImageNet dataset by transferring knowledge to our own task. In PSPNet architecture, the last layers of the backbone network replace the convolutional layers with dilated convolutional layers, which help the receptive field to grow. The dilated convolution layers are placed in the last two blocks of the backbone network with dilation values two and four, respectively, so that the features obtained at the end of the backbone contain richer features. The dilation value determines the sparsity when performing the convolution. The pyramid pooling module is the main part of the PSPNet architecture, which acts as an efficient global contextual prior, helping the model to classify the pixels based on the global information present in the image. Using a multi-level pyramid with four different scales (1×1, 2×2, 3×3 and 6×6), the pooling kernels cover different size portions of the image. After each pyramid level, we used a 1×1 convolutional layer to reduce the dimension of context to maintain the weight of global feature. Then, the upsampled maps are concatenated with the original feature map to pass to the decoder. The decoder takes the features of the encoder and turns them into predictions, by passing them into its layers. Finally, we use a convolutional layer followed by an 8x bilinear upasampling, as decoder for our segmentation network to recover the original size. The architecture of the proposed classification model is shown in Fig. 3.

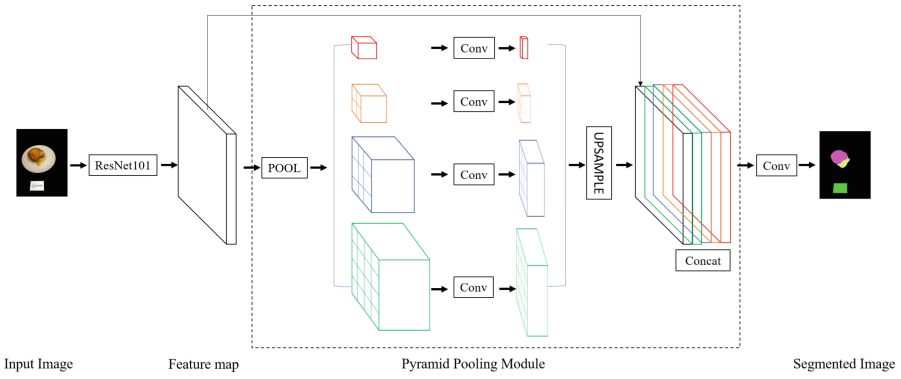

Fig. 3. The architecture of the proposed semantic segmentation model

2.3 Training and Testing

For the training phase, all images are resized to 240 × 240 × 3 resolution and the total number of the models' trainable parameters are 4,052,219. The MedGRFood dataset is partitioned into the training and validation set, using a ratio of 90:10 (90% is used for model training and 10% is used for model validation). In total, we used 4,500 food images and their masks for training, and 677 images and their masks for the validation set. Moreover, we chose a scaled learning rate with initial value 0.0001 and final value 0.0000001. The learning rate decreases by a factor 0.9 when the validation loss stops improving for three epochs. The model is trained for 50 epochs using the Adam optimizer [19], with an early stopping function if the validation accuracy stops improving for five epochs.

In several studies [20], pixel accuracy is chosen as the evaluation metric for the segmentation task. This metric can sometimes provide misleading results, when there are classes in the image with few pixel representations, as the measure will be biased in reporting how well you recognize the negative case. Here, we used the mean Intersection over Union score (meanIoU) to compute the accuracy of the proposed segmentation model. The IoU measures the similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets (Eq. 1). Then, to calculate the meanIoU of the 50 food classes we used Eq. 2:

$$IoU = \frac{Y_{true} \cap Y_{pred}}{Y_{true} \cup Y_{pred}}, \tag{1}$$

$$meanIoU = \frac{1}{50} \sum_{i=1}^{50} IoU_i, \tag{2}$$

where Y_{true} is the ground truth of the food image and Y_{pred} is the prediction mask.

We used the IoU loss function for our segmentation problem. The Eq. 3 shows the computation of IoU_{loss}:

$$IoU_{loss} = 1 - IoU. \tag{3}$$

2.4 Implementation

We used the python programming language to implement the semantic segmentation model in the Anaconda environment. Knowing the increased computing power requirements of CNN models, we also used the cuda toolkit, the cudnn and tensorflow libraries, for model training and validation of classification subsystem, through the Nvidia GeForce RTX 3080 graphic processing unit. Also, we used the opencv and the segmentation models libraries for the implementation of the proposed food segmentation models.

3 Results

To evaluate the proposed food semantic segmentation model, we further constructed and trained an additional segmentation model using the U-net architecture [21]. We trained the U-net model with exactly the same parametrization that we applied to the proposed model with the PSPNet architecture. Moreover, we built and trained two additional binary segmentation models for food and non-food segmentation. At these models we aimed to segment food regions from the background. Table 1 presents the segmentation results of the four models. We observe that the proposed model achieves a higher meanIoU score from the U-net model. The PSPNet architecture considers the global context of the image to predict the local level prediction and, therefore, gives better performance on the MedGRFood dataset. In addition, the difference in the total number of generated parameters between the two models is very large, which requires more training time for the U-net model.

Table 1. Segmentation Results Between the Four Models.

Model	MeanIoU	Loss	Training time (ms/step)	Number of parameters ($x10^6$)
Multiclass PSPNet	0.758	0.242	320	4.052
Multiclass U-net	0.654	0.346	370	51.512
Binary PSPNet	0.933	0.076	140	4.052
Binary U-net	0.901	0.099	201	51.512

In Fig. 4 we present the multiclass and binary segmentation results of the proposed model. We observe that the predicted mask is very close to the real mask of the test image in multiclass segmentation. We also notice that there is no wrong result in the class estimation, despite the fact that the number of 50 classes is quite large. We see that there is a slight deviation in the food mask prediction from its ground truth. Regarding the food segmentation from the image background, i.e., the binary segmentation, the extracted food mask is almost identical to its actual mask. This is a very crucial step in dietary assessment systems, because having the mask of the food we can use a classification model to predict its class very accurately. In Fig. 5 we present the multiclass and binary segmentation results of the U-net model. In multiclass semantic segmentation,

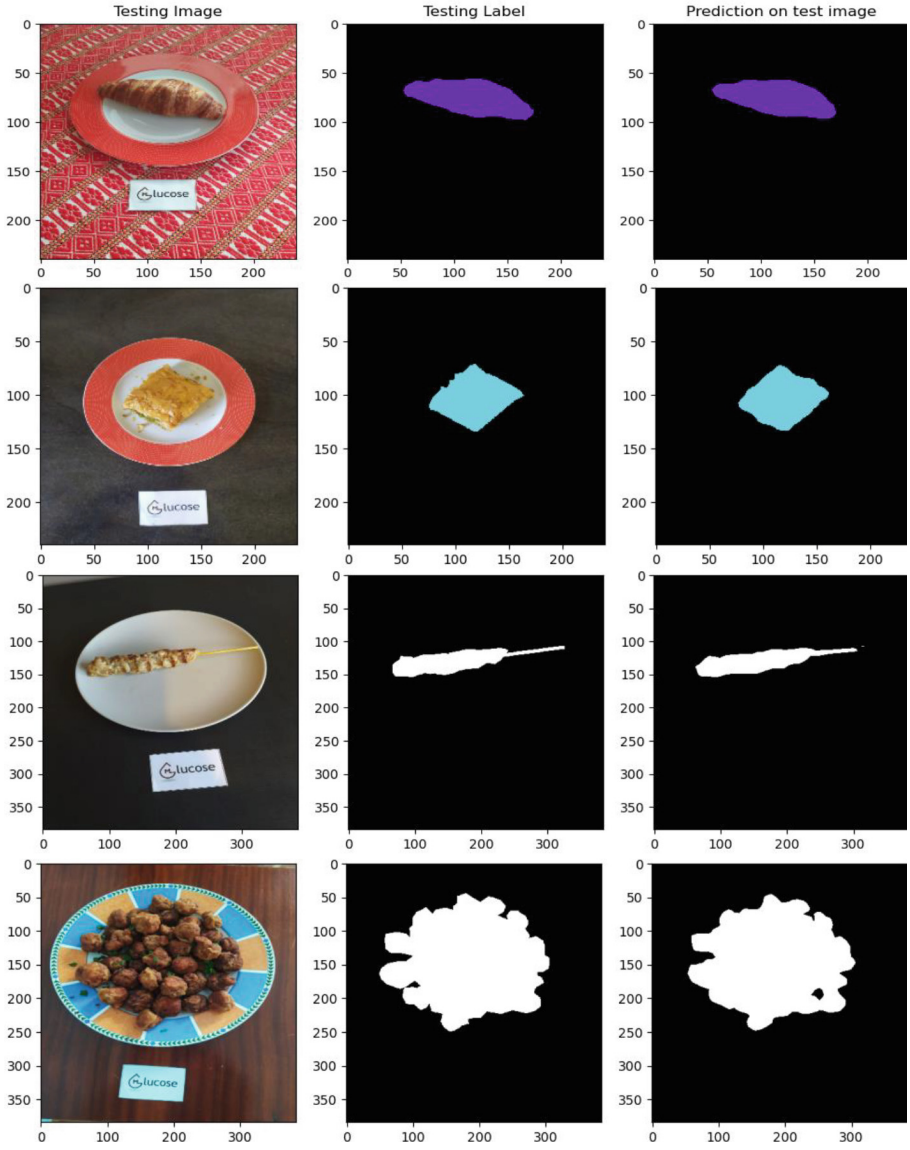

Fig. 4. Multiclass and binary semantic segmentation results of the proposed model.

we observe that the predicted mask has differences from the food ground truth. Moreover, we see that the U-net model incorrectly recognizes the food class in the second row. Comparing the results of the two semantic segmentation architectures, we can say that U-net does not perform as well as PSPNet, as it is not able to capture the context of the whole image. In predicted masks with differences with their ground truths, the application of morphological operations, such as erosion and dilation, could lead to the improvement of the results.

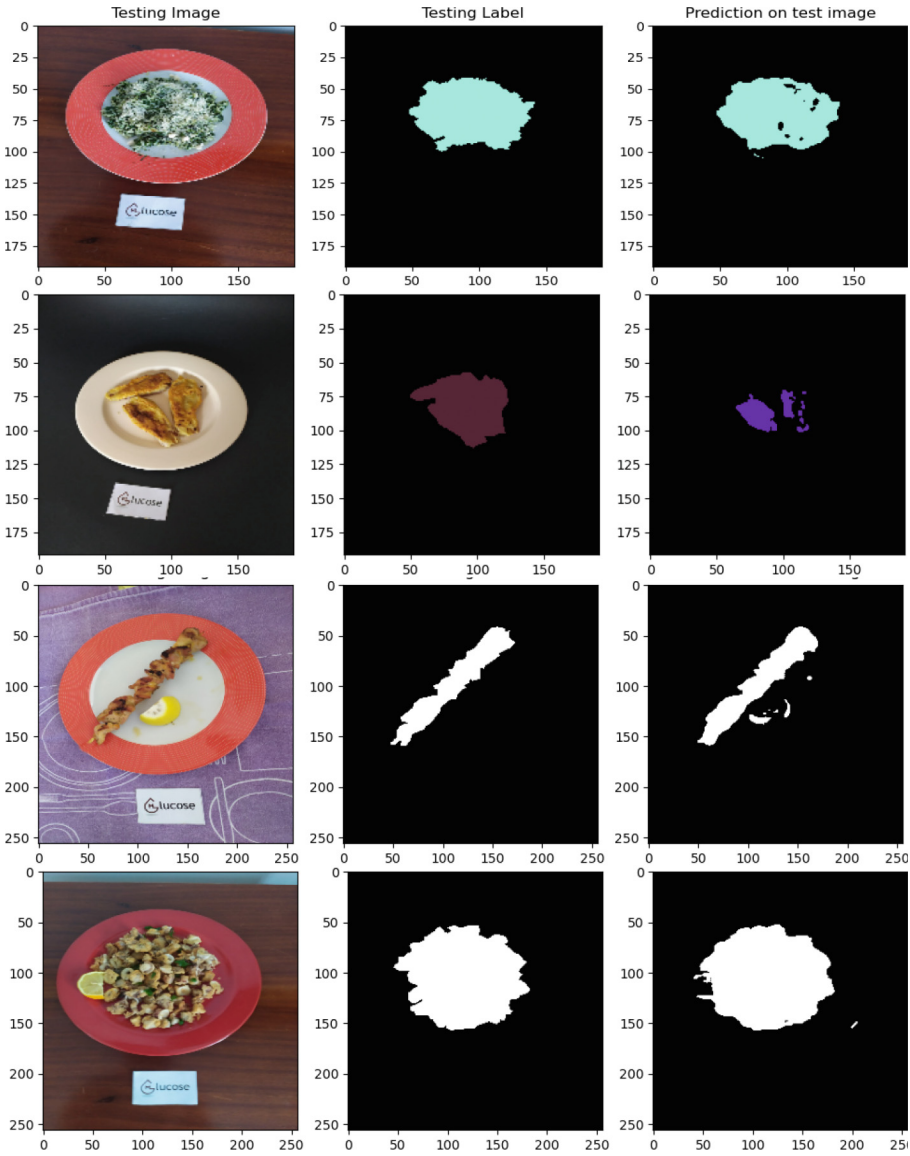

Fig. 5. Multiclass and binary semantic segmentation results of the U-net model.

4 Discussion and Conclusions

In food image databases, the use of deep learning techniques for food segmentation tends to create annotated databases with the largest possible number of images for each food class. However, the existing annotated databases are few and limited to the number of food classes they contain. Thus, there is a necessity to create a generic annotated food image database which covers as many food categories as possible and represents the

types of food from all cuisines. The collection of food images and the creation of food image databases is an easier task nowadays, due to the habit of capturing and posting images on social media. However, creating an annotated database of food images using their weight in addition to the type of food, remains a challenging task and will help build better and more accurate models for the segmentation and volume estimation steps in dietary assessment systems.

In automated food segmentation, the use of deep learning techniques has resulted in better performance compared to image processing techniques. Semantic segmentation and instance segmentation are techniques that have been used on a small scale in food image segmentation and could further improve the segmentation performance of dietary assessment systems. This presupposes the use of annotated food image databases, as it is a prerequisite to build segmentation models based on deep learning. In recent studies [22], the step of food image segmentation is omitted and in some others the performance of this step is not reported. In other studies, although the performance of the methods used to segment food images is high and improves the classification accuracy, there are still open issues related to cases where there are mixed foods. In these cases, the use of state of the art segmentation techniques, such as semantic and instance segmentation, can be used to improve the performance of this step and improve the efficiency to the classification step.

In this study, we presented a semantic segmentation model for multiclass and binary segmentation, using the pre-trained ResNet-101 as backbone network to the PSPNet architecture, applying the transfer learning technique from the ImageNet dataset. Comparing our results with related studies, we notice that the meanIoU score for multiclass segmentation has an excellent value, while the IoU score for food/non-food segmentation is one of the best results in the related literature [15]. To demonstrate the superiority of the proposed methodology, we built and trained an additional segmentation model based on the U-net architecture. The proposed model performs better and provides more accurate food segments in both multiclass and binary segmentation. This is due to the PSPNet ability to render the context of the whole image and to locate the objects of interest with higher accuracy. Open issues of this study are: (i) the ability of the segmentation model to separate complex foods, (ii) the segmentation of dishes containing two or more food items, (iii) the segmentation of dishes with overlaps between food items and, (iv) to calculate a good accuracy score of semantic segmentation models on an image with two or more food classes.

Acknowledgments. This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness. Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EDK-03185).

References

1. Minaee, S., Boykov, Y.Y., Porikli, F., Plaza, A.J., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: a survey. J. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 3523–3542 (2021)

2. Farràs, M., et al.: Beneficial effects of olive oil and Mediterranean diet on cancer physiopathology and incidence. Semin. Cancer Biol. **73**, 178–195 (2021)
3. https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html. Accessed 9 Dec 2021
4. Ciocca, G., Napoletano, P., Schettini, R.: Learning CNN-based features for retrieval of food images. In: Battiato, S., Farinella, G.M., Leo, M., Gallo, G. (eds.) ICIAP 2017. LNCS, vol. 10590, pp. 426–434. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70742-6_41
5. Chen, J., Ngo, C.-W.: Deep-based ingredient recognition for cooking recipe retrieval, pp. 32–41 (2016)
6. Meyers, A., et al.: Im2Calories: towards an automated mobile vision food diary, pp. 1233–1241 (2015)
7. Konstantakopoulos, F.S., et al.: GlucoseML mobile application for automated dietary assessment of mediterranean food, pp. 1432–1435. IEEE (2022)
8. Fang, S., Liu, C., Tahboub, K., Zhu, F., Delp, E.J., Boushey, C.J.: cTADA: the design of a crowdsourcing tool for online food image identification and segmentation, pp. 25–28. IEEE (2018)
9. Wu, X., Fu, X., Liu, Y., Lim, E.-P., Hoi, S.C., Sun, Q.: A large-scale benchmark for food image segmentation, pp. 506–515 (2021)
10. Pouladzadeh, P., Shirmohammadi, S., Bakirov, A., Bulut, A., Yassine, A.: Cloud-based SVM for food categorization. Multimed. Tools Appl. **74**(14), 5243–5260 (2015)
11. Inunganbi, S., Seal, A., Khanna, P.: Classification of food images through interactive image segmentation. In: Nguyen, N.T., Hoang, D.H., Hong, T.-P., Pham, H., Trawiński, B. (eds.) ACIIDS 2018. LNCS (LNAI), vol. 10752, pp. 519–528. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75420-8_49
12. Aslan, S., Ciocca, G., Schettini, R.: Semantic food segmentation for automatic dietary monitoring, pp. 1–6. IEEE (2018)
13. Ciocca, G., Napoletano, P., Schettini, R.: Food recognition: a new dataset, experiments, and results. IEEE J. Biomed. Health Inform. **21**(3), 588–598 (2016)
14. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, pp. 6881–6890 (2021)
15. Wang, W., et al.: A review on vision-based analysis for automatic dietary assessment. Trends Food Sci. (2022)
16. Konstantakopoulos, F., Georga, E.I., Fotiadis, D.I.: 3D reconstruction and volume estimation of food using stereo vision techniques, pp. 1–4. IEEE (2021)
17. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network, pp. 2881–2890 (2017)
18. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: revisiting the ResNet model for visual recognition. J. Pattern Recogn. **90**, 119–133 (2019)
19. Zhang, Z.: Improved Adam optimizer for deep neural networks, pp. 1–2. IEEE (2018)
20. Subhi, M.A., Ali, S.H., Mohammed, M.A.: Vision-based approaches for automatic food recognition and dietary assessment: a survey. IEEE Access **7**, 35370–35381 (2019). https://doi.org/10.1109/ACCESS.2019.2904519
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
22. Yunus, R., et al.: A framework to estimate the nutritional value of food in real time using deep learning techniques. J. IEEE Access **7**, 2643–2652 (2018)