Semantic Food Segmentation Using Convolutional Deconvolutional Pyramid Network for Health Monitoring

Mazhar Hussain*, Alessandro Ortis, Riccardo Polosa, and Sebastiano Battiato

***Abstract*—This paper presents semantic food segmentation to detect individual food items in an image in the context of Food Recognition (FoodRec) project. FoodRec aims to study and develop an automatic framework to track and monitor the dietary habits of people, during their smoke quitting protocol. Studies have shown a strong correlation between dietary habits' changes of individuals and smoking cessation process. Abstinence from smoking is associated with several negative effects such as gain of weight, eating disorders, mood changes, and irritability during the initial period of smoke quitting. In this contribution, a novel Convolutional Deconvolutional Pyramid Network (CDPN) is proposed for food segmentation to understand the semantic information of an image at a pixel level. This network employs convolution and deconvolution layers to build a feature pyramid and achieves high-level semantic feature map representation. As a consequence, the novel semantic segmentation network generates a dense and precise segmentation map of the input food image. Furthermore, the proposed method achieved competitive results with 91.77% mean Intersection over Union (IOU) on TrayDataset and 77% mean IOU on MyFood dataset when compared to the state-of-the-art techniques.**

***Index Terms*—Artificial intelligence for health, dietary monitoring, food segmentation, food dataset**

I. Introduction

Food segmentation plays a key role in the context of food recognition technology for dietary monitoring [1] to predict and detect multiple food items present in an image. The output of a food segmentation system is a set of image regions associated to each detected food item to provide a semantic segmentation map of the input image. This accurate segmentation of food regions can be used to estimate the volume and hence the quantities of each food item detected within an image. This would allow people to estimate the assessment of calories and nutrients to track their food intake of what they consume to increase awareness of their daily diet by monitoring their eating habits, the type and amount of food, how often and what times the user eats a meal, how much time he spends eating in a day, advanced inferences performed can be compared to make correlations between eating habits and quitting process steps, bad habits, user's behavior, and mood changes [2] over time. The semantic organization of daily habits can help a doctor to have a better opinion with respect to the patient's behavior and habits changes, in the applications on quitting treatment response and health needs, smoke monitoring technology [3], dietary monitoring during smoking cessation and smoke quitting program [4]. Food plays a crucial role in human life that is strongly affected by diet [5]. Then, food recognition technology and its applications especially in the health department [6] for dietary and calorific monitoring motivated computer vision specialists to develop new methods in the areas such as food logging and automatic food dietary monitoring [7], food retrieval and classification [8], food recognition to monitor users' eating habits [9], and segmentation for food understanding and analysis.

In this study, particular efforts are devoted to the development of new segmentation algorithm for accurate food image analysis in the context of the FoodRec [1] project. Food segmentation aims to train a model that can look at the images of food items and infer semantic information to recognize individual food items present in an image. This would further allow people to estimate the assessment of calories and nutrients to track their food intake. This paper presents a novel Convolutional Deconvolutional Pyramid Network (CDPN) for semantic food segmentation. The proposed network employs convolution and deconvolution layers to generate a feature pyramid and achieves high-level semantic feature map representation. The deconvolution layers densify the feature map with learned filters to output upsampled and rich feature map representation. Experiments are conducted on two publicly available benchmark food datasets that reveals significant improvements in the results.

The rest of the paper is organized as follows: Section II presents the related works. Section III describes the proposed convolutional deconvolutional pyramid network for food segmentation. Next, experimental results on the food datasets are presented in Section IV. Finally, conclusions are drawn in Section V on the basis of results and evaluation.

II. Related Work

Computer vision and deep learning techniques have emerged as a powerful tool providing high levels of image analysis in the field of image segmentation. Lu *et al.* [10] proposed a system to estimate nutrient intake for hospitalized patients to reduce the risk of disease-related malnutrition by using RGB depth image pairs that are captured before and after meal consumption. The system consists of a multi-task contextual network for food item segmentation, classification with few-shot learning-based algorithms and 3D food surface extraction. This Artificial Intelligence-based system permits to estimate the nutritional intake automatically with

Manuscript received March 22, 2023; revised April 24, 2023; accepted August 17, 2023.

Mazhar Hussain is with the Department of Mathematics and Computer Science, University of Catania, Catania, Italy.

Alessandro Ortis and Sebastiano Battiato are with the Department of Mathematics and Computer Science and the Center of Excellence for the Acceleration of HArm Reduction (CoEHAR), University of Catania, Catania, Italy. E-mail: ortis@unict.it (A.O.), battiato@unict.it (S.B.)

Riccardo Polosa is with the Department of Clinical and Experimental Medicine and the Center of Excellence for the Acceleration of HArm Reduction (CoEHAR), University of Catania, Catania, Italy, and ECLAT Srl, Spin-off of the University of Catania, Catania, Italy. E-mail: polosa@unict.it (R.P.)

*Correspondence: mazhar.hussain@phd.unict.it (M.H.)

DOI: 10.7763/IJCTE.2023.V15.1346

sequential segmentation, recognition, and consumed volume estimation of each food. In addition, a new database of food images with related recipes and nutritional information is collected in real-world hospital settings. Sharma *et al*. [11] presented a GourmetNet for food segmentation with multi-scale feature representation. This network incorporates both spatial attention and channel attention using waterfall atrous spatial pooling module. Channel attention exploits high-level features to generate refined low-level features and spatial attention utilizes low-level features to produce refined high-level features. GourmetNet is inspired by the segmentation method [12] where stride spatial pyramid pooling is applied to get multiscale semantic information and dual attention decoder is used with a channel attention branch and a spatial attention branch to capture semantic feature map representation. Pfisterer *et al.* [13] introduced a fully automated food intake segmentation system using a macroarchitecture of the deep convolutional neural network that is a multi-scale encoder-decoder network for food intake tracking and estimation in long-term care homes. A residual encoder microarchitecture trained on the ImageNet dataset is used because of its discriminative feature learning ability. A pyramid scene parsing network is used as a decoder microarchitecture. This method obtained competitive results as compared to the semi-automatic graph cuts using monocular RGB images. Freitas *et al.* [14] focused on food segmentation and classification to recognize foods using deep learning techniques. This work introduced a Brazilian food dataset consisting of nine food classes to perform a comparative study on segmentation algorithms where Mask R-CNN [15] obtained better results. He *et al*. [16] presented a food image analysis in order to develop a dietary assessment system that monitors daily food intake. In this paper, food image segmentation and identification have been performed to identify the regions of food items in an image and to determine the food category. Then, weight is estimated using shape template and area-based weight estimation for foods to extract the nutrient content for dietary monitoring and assessment. Aguilar *et al*. [17] proposed Bayesian deep learning for food semantic segmentation and assessed the uncertainty in the predictions for the improvements of healthcare technologies and dietary monitoring. In this paper, Bayesian inference is approximated using the MC-dropout [18] method by placing a dropout layer after each residual block of the network. The uncertainty from the final prediction of the segmentation is captured with entropy and mutual information. Ramesh *et al.* [19] proposed detection and segmentation scheme to recognize food from egocentric camera images for efficient dietary monitoring. The YOLOv5 [20] is trained for the detection and localization of the food items and graph-cut method is applied for the food segmentation resulting in significant score.

Ronneberger *et al*. [21] introduced U-Net deep convolutional network architecture to perform biomedical image segmentation. It is comprised of a contracting path that follows the same fashion as the typical architecture of a convolutional network and an expansive path that is used for upsampling of the feature map. Zhou *et al*. [22] proposed the UNet++ architecture in which sub-networks encoder and decoder are connected through a series of nested and dense skip pathways for biomedical image segmentation that is more powerful as compared to the U-Net. Lin *et al.* [23] developed a feature pyramid network for segmentation and object detection. An architecture with skip connections is designed to extract semantic feature maps that involve a bottom-up pathway which computes a feature hierarchy, and a top-down pathway which computes stronger feature maps from higher pyramid levels enhanced with features from the bottom-up pathway. Feature pyramid network is used for land segmentation [24] with ResNet encoder that is pretrained on ImageNet dataset. Badrinarayanan *et al.* [25] introduced Segnet a convolutional encoder-decoder architecture for semantic segmentation and evaluated the performance of SegNet on two scene segmentation tasks, segmentation of the CamVid road scene which is currently of practical attention for autonomous driving activities, and segmentation of the SUN RGB-D indoor scene which is of direct concern for augmented reality applications. Experimental results show that SegNet provides good performance with significant results as compared to other architectures like fully convolutional networks for segmentation [26], DeconvNet [27], and also with DeepLab-LargeFOV [28] architectures. Zhao *et al.* [29] proposed a pyramid scene parsing network for semantic segmentation on PASCAL VOC 2012 benchmark and cityscapes benchmark datasets segmentation. It consists of a pyramid pooling module to exploit the local and global context information. Initially, a feature map is extracted from the last convolutional layer of the convolution neural network that is fed to the pyramid pooling module in order to produce different pooling pyramids. Further, upsampling and concatenation layers are used for the final feature representation. Then, the convolution layer is applied to obtain the pixel-wise prediction of the input image.

The feature pyramid plays an important role in recognition tasks. The proposed network for food segmentation employs convolution and deconvolution layers to develop a feature pyramid to generate a semantically strong and rich segmentation map of the input food image.

III. Proposed Convolutional Deconvolutional Pyramid Network

The food segmentation aims to develop a model which is able to extract and infer semantic information from the food images at pixel level to recognize different food items present in an image. This would further allow people to estimate the assessment of calories to track their food intake and to increase awareness of daily diet by monitoring their eating habits. In this context, a novel Convolutional Deconvolutional Pyramid Network (CDPN) is proposed for image semantic segmentation which takes food image as input and outputs segmentation map of the individual food items detected, as described in Fig. 1. A pretrained convolution neural network is used to harvest meaningful feature representation. Then, a feature pyramid is built with multi-scale feature maps representation. The proposed network employs convolution and deconvolution layers to generate a feature pyramid and achieves high-level semantic feature map representation. The deconvolution develops upsampling of the input features using learnable parameters to produce generalized upsampling of the feature

representation. As a result, the proposed segmentation network generates a dense and precise segmentation map of the input food image.

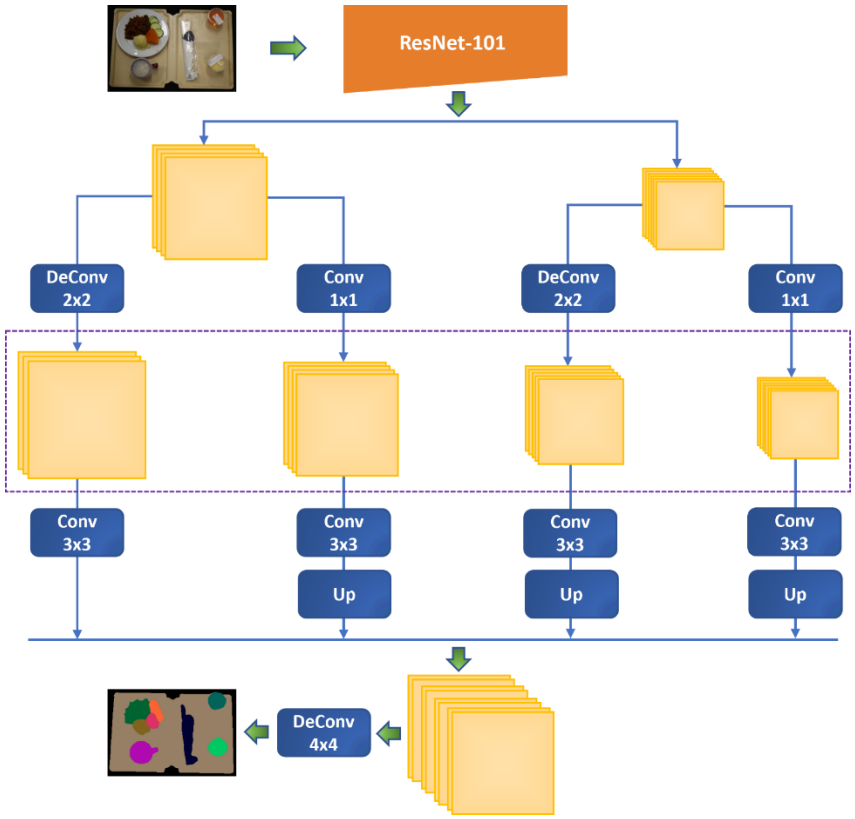

Fig. 1. The proposed Convolutional Deconvolutional Pyramid Network for semantic food segmentation. This network takes food image as input and outputs segmentation map of the individual food items present in the image. In the architecture, "Conv" represents the convolutional layer, "DeConv" represents the deconvolutional layer, and "Up" represents an upsampling layer. All convolutional layers with 3×3 kernel size are followed by group normalization layers and ReLU activation layers.

Initially, deep ResNet [30] architecture is applied with pre-activations due to its capability for learning highly descriptive feature representation by downsampling spatial resolution for complex scenes. The ResNet architecture with transferred weights from the ImageNet dataset extracts two feature set representations from the input RGB food image. More precisely, the proposed Convolutional Deconvolutional Pyramid Network obtains two discriminative feature sets of 512 channels with a downsampled spatial resolution (h/8, w/8), and 2048 channels with a downsampled spatial resolution (h/32, w/32) from an input image (height(h), width(w)) using Resnet-101 as the backbone network. The convolution coupled with upsampling layers is used to the low-resolution spatial information to produce high-resolution semantically strong segmentation map. Then, the convolution layer and deconvolution layer are employed to each feature set obtained by the ResNet to generate four high-level feature maps pyramid. The deconvolution layers densify the feature map with learned filters to output upsampled and rich feature map representation. As a result, a feature pyramid is produced with four multi-scale feature maps of sizes h/4 × w/4 × 256 channels, h/8 × w/8 × 256 channels, h/16 × w/16 × 256 channels, and h/32 × w/32 × 256 channels by applying convolution layer with 1×1 kernel size and deconvolution layer with 2×2 kernel size on each feature set obtained by the ResNet. Feature maps of each level are fused and concatenated through the convolution layer with 3×3 kernel size, group normalization, ReLU, and upsampling layer to resize the feature map using interpolation. Further, a deconvolutional layer is applied with 4×4 kernel size and output channels equal to the number of food categories followed by a softmax layer for the food segmentation pixel-wise predictions. The final output of the proposed CDPN network is a segmentation map of the same size as the input image that represents the probability of each pixel belongs to one of the food classes.

IV. Experimental Results

The proposed segmentation scheme is thoroughly tested using two publicly available benchmark food segmentation datasets which are TrayDataset [31] and MyFood dataset [32]. Further details are provided in this Section on these datasets. The performance of the proposed method is extensively evaluated and compared to the existing state-of-the-art approaches.

A. Performance Metrics

The results are compared using common performance metrics for segmentation to compare the quantitative performance of the proposed method with others. Semantic food segmentation results are evaluated using performance measures such as Intersection over Union (IOU) in Eq. (1) for each food category, mean Intersection over Union (mIOU) in Eq. (2) calculated by taking the average of all the food categories IOU, and pixel-level accuracy in Eq. (3) that represents the percentage of correctly classified pixels where TP, TN, FP, and FN represent true positives, true negatives,

false positives, and false negatives, respectively.

$$IOU = \frac{\text{target} \cap \text{predicted}}{\text{target} \cup \text{predicted}} \quad (1)$$

$$\text{m}IOU = \frac{1}{n} \sum_{i=1}^{n} IOU_i \quad (2)$$

$$PixelAccuracy = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (3)$$

B. Datasets

Tray food dataset: TrayDataset [31] is a food segmentation dataset comprised of 43 food classes. This database contains a total of 1241 food images with 17 unique trays where images are rotated, wrapped, and flipped versions of the unique trays. The dataset is composed of distinctive food classes e.g., bread, ham, custard, margarine, pumpkin, zucchini, milk, baked fish, creamed potato, orange juice, soup, carrot, vanilla yogurt, cucumber, broccoli, beef, etc. Tray food database is a well-defined publicly available with all food images and their respective ground truth segmentation masks on Kaggle.

MyFood dataset: MyFood dataset [32] is a well-defined publicly available database for food image segmentation. This dataset consists of the most consumed food types by the Brazilian population containing 1250 total images. The dataset is composed of nine food classes such as spaghetti, apple, beans, boiled egg, chicken breast, rice, salad, steak, and fried egg with an average of 125 food images in each class. For research and evaluation experiments, the dataset is divided into 60% for training, 20% for validation, and 20% testing. MyFood database is publicly available with all food images and their respective segmentation masks on the Zenodo website [33] with training, validation, and testing folder structure. It can be downloaded at the following link http://doi.org/10.5281/zenodo.4041488.

C. Evaluation on Tray Food Dataset

Experiments are conducted using TrayDataset [31] that consists of 43 distinctive food classes. The proposed Convolutional Deconvolutional Pyramid Network (CDPN) results are compared with other methods such as Feature Pyramid Network (FPN) [23], and Encoder-Decoder Food Network (EDFN) [13] using intersection over union and pixel accuracy. The EDFN [13] architecture is used for automatic semantic segmentation of tracking food and fluid Intake in long-term care homes. It is a deep convolutional encoder-decoder architecture for food pixel-wise segmentation. It employs ResNet architecture as an encoder to get 256 feature maps of the input image and a pyramid scene parsing [29] is used as decoder microarchitecture to decode the feature maps from the encoder. FPN [23] is designed for image segmentation and multi-scale object detection. This architecture is developed to extract semantic feature maps that involve a bottom-up pathway which computes a feature hierarchy, and a top-down pathway which computes stronger feature maps from higher pyramid levels. The proposed CDPN, FPN, and EDFN are trained using the same hyperparameters and settings for the comparative evaluation. All the networks are trained using the Adam optimizer and the standard Dice loss function. The learning rate is set to 0.0001. The batch size is set as 8 and networks are trained for 250 epochs. The networks are trained using ResNet-101 as the backbone network for the evaluation experiments.

TABLE I: CLASS-WISE INTERSECTION OVER UNION (IOU) FOR EACH FOOD ITEM OF TRAYDATASET [31]

Food Items	Class-Wise IOU (%)			Food Items	Class-Wise IOU (%)		
	Proposed CDPN	EDFN [13]	FPN [23]		Proposed CDPN	EDFN [13]	FPN [23]
Tray	**96.49**	83.95	89.28	**Pumpkin**	**88.17**	72.02	85.99
Cutlery	**87.96**	82.06	85.08	**Celery**	100.00	100.00	100.00
Bread	**93.67**	80.00	80.00	**Sandwich**	**90.59**	88.25	82.62
Straw	100.00	100.00	100.00	**SideSalad**	**92.86**	90.00	90.00
Custard	**93.09**	78.45	80.58	**TartareSauce**	85.00	85.00	85.00
Beef	100.00	100.00	100.00	**JacketPotato**	85.00	**91.00**	90.00
Roastlamb	85.00	85.00	85.00	**CreamedPotato**	**89.51**	72.00	72.00
BeefTC	83.74	75.35	**98.34**	**Form**	100.00	100.00	100.00
Ham	**98.64**	90.00	90.00	**Margarine**	80.00	78.00	**90.00**
Bean	90.00	90.00	90.00	**Soup**	**97.71**	83.61	92.45
Cucumber	**90.00**	**90.00**	86.79	**Apple**	100.00	100.00	100.00
Leaf	**97.31**	91.20	95.00	**CannedFruit**	82.85	**90.00**	**90.00**
Tomato	90.00	90.00	90.00	**Milk**	84.68	**90.00**	77.91
Boiledrice	**81.37**	80.00	80.00	**VanillaYogurt**	**88.23**	82.00	79.00
BeefMM	**99.47**	79.00	89.49	**Jelly**	**88.80**	70.00	73.95
SpinachPR	79.29	**90.00**	79.25	**Meatball**	100.00	100.00	100.00
BakedFish	75.08	**85.00**	**85.00**	**LemonSponge**	**99.34**	95.00	95.00
Gravy	89.83	86.78	**94.42**	**Juice**	100.00	100.00	100.00
Broccoli	100.00	100.00	100.00	**AppleJuice**	77.71	**90.00**	**90.00**
Carrot	100.00	100.00	100.00	**OrangeJuice**	**94.28**	90.00	76.00
Zucchini	**93.82**	75.50	88.11	**Water**	**97.14**	86.77	85.00

Note: "BeefTC" means Beef Tomato Casserole, "BeefMM" means Beef Mexican Meatballs, "SpinachPR" means Spinach and Pumpkin Risotto.

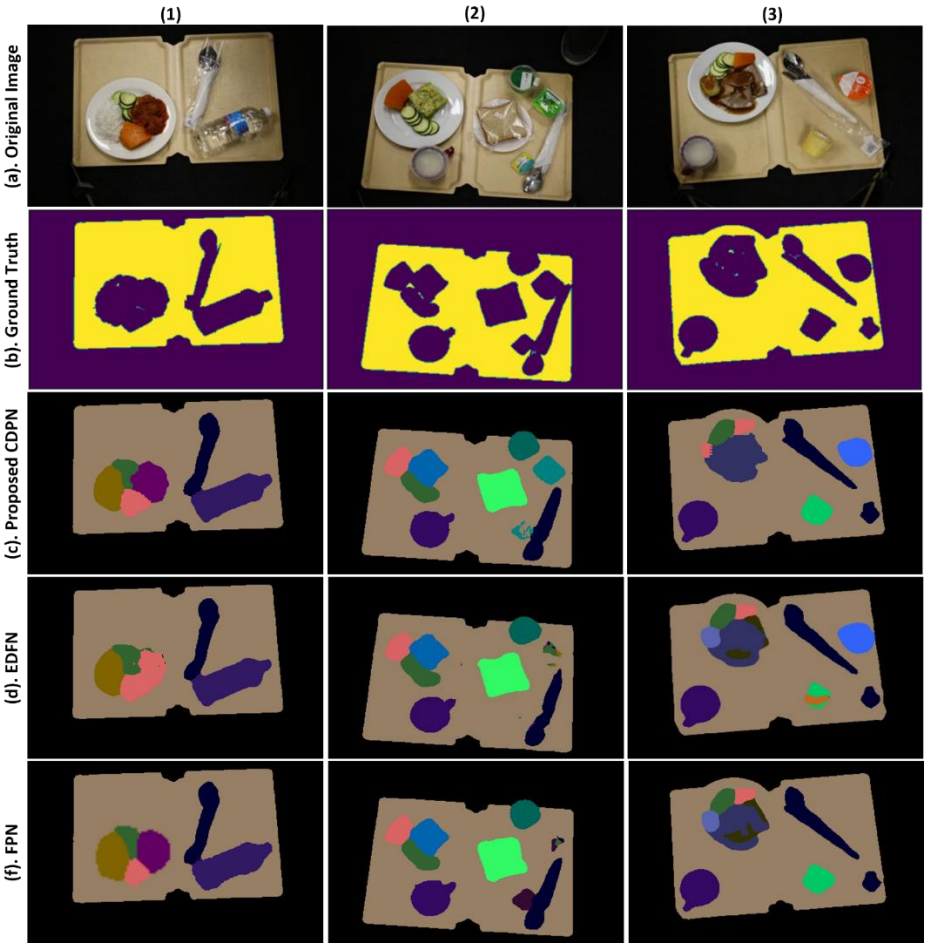

Fig. 2. Food segmentation results visualization of the proposed CDPN, FPN, and EDFN on TrayDataset [31]. (a) represents original images, (b) represents ground truths, (c) represents the proposed CDPN output segmentation maps, (d) represents EDFN output segmentation maps, and (f) represents FPN output segmentation maps.

On the TrayDataset [31], the experimental results of the proposed CDPN are compared with FPN and EDFN. The experimental results comparison of the networks using class-wise intersection over union is shown in Table I where the top IOU is represented in bold for each food category. Dataset also has the background class with IOU of 99.49% for the proposed CDPN, 99.28% for EDFN, and 98.52% for FPN. For most of the food classes, the proposed CDPN approach achieved higher class-wise IOU results as compared to others. The lowest IOU scores for the proposed CDPN, FPN, and EDFN are 75.08% for the bakedfish class, 72% for the creamedpotato class, and 70% for the jelly class, respectively. From the experimental results described in Table I, the proposed CDPN method achieved a competitive IOU score as compared to the FPN and EDFN.

The visual representation is presented in Fig. 2 of the original input image, ground truth, and output segmentation maps of the proposed network and baseline networks on TrayDataset. For example, consider the original input image (1) and its output segmentation maps generated by models, EDFN confuses beefmexicanmeatballs with pumpkin. The proposed CDPN and FPN predict correctly. Now, consider the original input image (2) and its output segmentation maps generated by models, EDFN mispredicts the vanillayogurt region by confusing most of the part of it with zucchini. The EDFN does not detect the margarine region as well. The FPN misclassifies the vanillayogurt region by confusing it with margarine and zucchini. The proposed CDPN detects vanillayogurt accurately but confuses margarine with vanillayogurt.

The results comparison of the proposed method CDPN with FPN and EDFN networks using mean intersection over union and global pixel level accuracy is presented in Table II. From the experimental results presented in Table I, II, and Fig. 2, the proposed CDPN method achieved competitive results. These experimental results show that the proposed Convolutional Deconvolutional Pyramid Network outperformed both EDFN and FPN.

TABLE II: THE PROPOSED CDPN METHOD RESULTS IN COMPARISON WITH EDFN AND FPN USING MEAN INTERSECTION OVER UNION (MIOU) AND PIXEL ACCURACY ON FOOD TRAYDATASET [31]

Method	Backbone	Mean IOU (%)	Pixel Accuracy (%)
Proposed CDPN	ResNet-101	91.77	98.90
FPN [23]	ResNet-101	89.30	98.49
EDFN [13]	ResNet-101	88.02	97.93

Split the image into pixel-level segments: The original image is divided into individual food pixel-level segments by setting the background to black on the basis of the segmented image map for further food analysis such as food annotation, classification, volume estimation, etc. In this process, the pixels from each segmented food item in the network-detected segmented image map are utilized to extract the corresponding pixels from the original image, and any residual pixels are set to zero. Finally, the individual food

segments of the original image are obtained with a black background for each food segment detected in the segmented image map. Fig. 3 displays the original image of TrayFood [31] together with its segmented image map. Fig. 4 displays each food segment that was extracted from the original image based on the segmented image map at the pixel level.

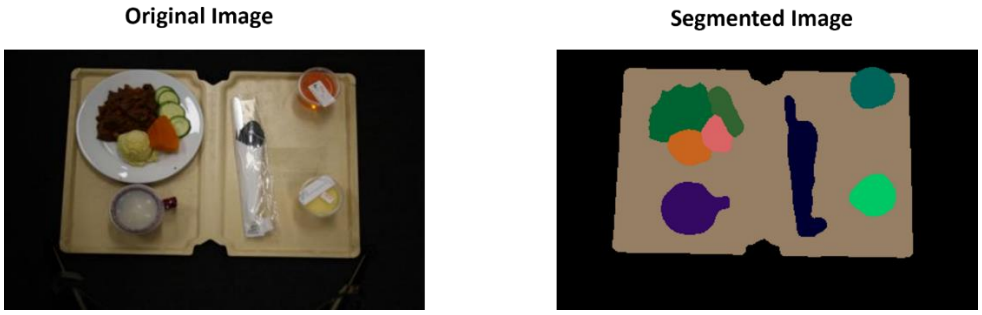

Fig. 3. Original image from TrayFood [31] dataset and its output segmented image map.

Fig. 4. Individual food pixel-level segments of the original image for each food segment detected in the segmented image map.

D. Evaluation on MyFood Dataset

On the MyFood [32] dataset, the proposed CDPN is compared with Mask R-CNN [15], Segnet [25], FCN [26], UNet++ [22], Enet [34], and DeepLabV3+ [35]. The Segnet [25] is a deep convolutional encoder-decoder architecture for semantic pixel-wise segmentation. The encoder architecture is topologically identical to the 13 convolutional layers of the VGG16 architecture [36]. To perform non-linear upsampling, the decoder utilizes pooling indices computed during the max-pooling step of the encoder. The mask R-CNN [15] architecture is an end-to-end convolutional neural network introduced by Facebook AI research group with accurate detection effect when it comes to target object instance segmentation. The mask R-CNN is the extended improvement of Faster R-CNN [37] with an object mask prediction branch in parallel with the existing bounding box detection branch. FCN [26] is a fully convolutional network for segmentation with skip architecture that combines layers of the feature hierarchy to produce refine segmentation. The classification networks GoogLeNet [38], VGG net [36], and AlexNet [39] are extended to fully convolutional networks by transferring their learning for the segmentation. The UNet++ [22] is the extension of U-Net [21] architecture in which sub-networks encoder and decoder are connected through a series of nested and dense skip pathways. It was proposed for biomedical image segmentation that is more powerful as compared to the U-Net. Enet [34] is a deep neural network for real-time segmentation performance on embedded platforms. It is heavily inspired by the ResNet [30] and Inception [40] architectures with the aim to perform large-scale computations efficiently. The DeepLabV3+ [35] is the extended version of DeepLabv3 [28] developed for semantic segmentation with the concept of atrous separable convolution. It employs encoder-decoder structure with atrous convolution comprised of a deep convolution and a clockwise convolution. The encoder is used to rich the contextual information and the effective decoder is used to refine the segmentation results.

TABLE III: HYPERPARAMETERS USED FOR TRAINING EACH MODEL ON MYFOOD [32] DATASET

Method	Optimizer	Learning Rate	Batch Size
Unet++	Adam	1E−4	8
Proposed CDPN	Adam	1E−4	8
Mask R-CNN	SGD	1E−3	2
FCN	SGD	1E−2	32
Segnet	SGD	1E−2	32
Enet	Adam	5E−4	10
DeepLabV3+	SGD	1E−2	32

The hyperparameters used are given in Table III where all the networks are trained for 100 epochs for the comparative evaluation. The parameters for SegNet, Mask R-CNN, FCN, Enet, and DeepLabV3+ are described in the research [14] using the MyFood segmentation dataset. The proposed CDPN and UNet++ are trained using the same hyperparameters with the Adam optimizer, the standard Dice loss function, the learning rate is set to 0.0001 and, the batch size is set as 8. The parameter used for training the proposed CDPN and other methods are listed in Table III for experiments evaluation on the MyFood dataset.

The experimental results comparison of the networks is shown in Figs. 5 and 6 using class-wise intersection over union. The results achieved by the proposed CDPN and UNet++ are presented in Fig. 5. For the class-wise IOU comparative evaluation with the proposed CDPN and UNet++, the results [14] are shown in Fig. 6 which shows IOU for Segnet, Mask R-CNN, FCN, Enet, and DeepLabV3+. Dataset also has the background class with IOU of 0.90 for the proposed CDPN, and 0.92 for UNet++. The proposed CDPN approach provided a competitive class-wise intersection over union score in comparison with other methods. According to Figs. 5 and 6, the chicken breast class had the lowest IOU score, with the proposed CDPN obtaining the highest IOU score of 0.62, UNet++ having an IOU score of 0.58, and Segnet, Mask R-CNN, FCN, Enet, and DeepLabV3+ having an IOU score less than 0.50. However, the proposed CDPN, Unet++, FCN, and Mask R-CNN produced results with comparatively high class-wise IOU scores.

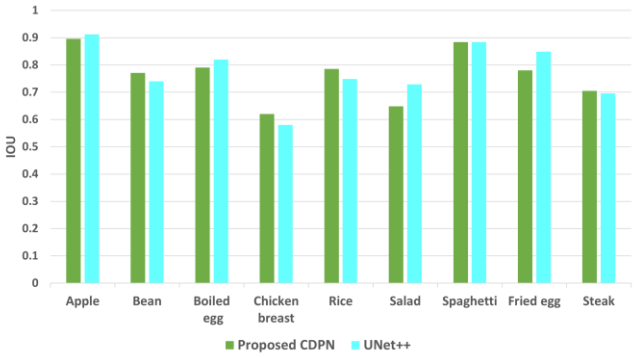

Fig. 5. The proposed CDPN method and UNet++ [22] class-wise intersection over union (IOU) results comparison on MyFood [32] segmentation dataset.

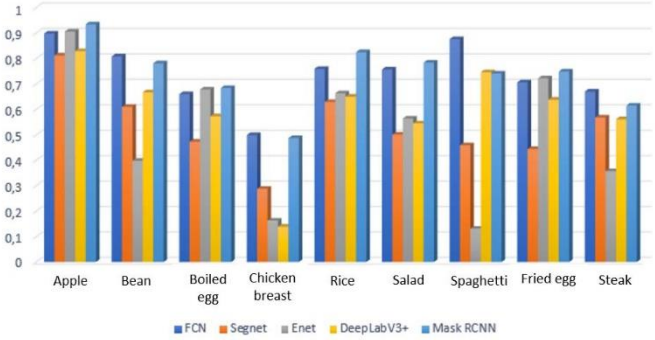

Fig. 6. Class-wise intersection over union (IOU) on MyFood [32] dataset. This figure is adopted from the research [14] that shows the comparison of the class-wise results for Segnet, Mask R-CNN, FCN, Enet, and DeepLabV3+.

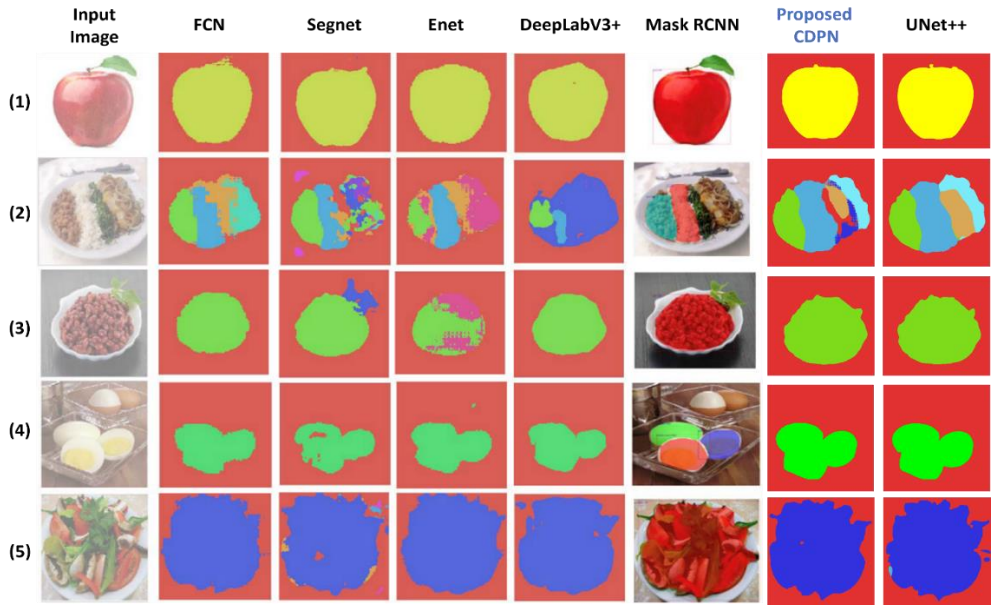

Fig. 7. Food segmentation results visualization of the proposed CDPN approach with other methods on MyFood [32] dataset. For instance, the input image (1) is an apple and its output segmentation maps generated by each network are presented here. The output segmentation maps of proposed CDPN and UNet++ are added together with the visualization of food image segmentation results described in research [14] for comparative evaluation with Segnet, Mask R-CNN, FCN, Enet, and DeepLabV3+.

The segmentation results on MyFood dataset are presented in Fig. 7 with the visual representation of the input image and output segmentation maps of the proposed CDPN, Segnet, Mask R-CNN, FCN, UNet++, Enet, and DeepLabV3+. In the case of a single food in an image, it can be noticed that most methods performed well to produce the output segmentation

maps of the input image. On the other hand, the proposed CDPN, UNet++, and FCN results are comparable when multiple food items are present in an image. In both cases, the proposed CDPN achieved better segmentation results compared to the Segnet, Mask R-CNN, FCN, Enet, and DeepLabV3+ as shown for input image (2) in Fig. 7. However, UNet++ produced better results than the proposed CDPN method.

On MyFood [32] dataset, the segmentation results of the proposed CDPN method are outstanding compared to the state-of-the-art approaches. The UNet++ [22] obtained higher results with 0.79 mean IOU, but there was very little marginal difference in mean IOU when compared to the proposed CDPN with 0.77 mean IOU. From the detailed experimental analysis presented in Figs. 5–7, and Table IV, the proposed CDPN approach provided competitive results. These experimental results show that the proposed CDPN outperformed other approaches [14] such as Segnet, Mask R-CNN, FCN, Enet, and DeepLabV3+ on the MyFood segmentation dataset.

TABLE IV: THE PROPOSED CDPN METHOD RESULTS IN COMPARISON WITH OTHER METHODS ON MYFOOD [32] DATASET

Method	Backbone	IOU
UNet++ [22]	ResNet-101	0.79 (0.11)
Proposed CDPN	ResNet-101	0.77 (0.09)
Mask R-CNN [15]	ResNet-101	0.70 (0.2)
FCN [26]	VGG16	0.70 (0.2)
Segnet [25]	-	0.52 (0.2)
Enet [34]	-	0.51 (0.3)
DeepLabV3+ [35]	MobileNet	0.50 (0.3)

V. CONCLUSION

Semantic segmentation is an important task in the field of computer vision and getting a lot of attention due to deep learning techniques providing a high-level of accuracy for image analysis. A new approach is proposed in the context of the FoodRec project towards the challenging task of semantic food segmentation in order to develop a system capable of producing state-of-the-art results. A novel Convolutional Deconvolutional Pyramid Network for food segmentation to infer semantic information from the food images at the pixel level and to recognize individual food items in the image. The network employs convolution and deconvolution layers to build a feature pyramid that generates a semantically strong and rich segmentation map of the input food image. Moreover, the detailed results were demonstrated on two benchmark food datasets for food segmentation performance evaluation and comparison of the proposed CDPN with the existing methods. The proposed approach produced comparatively higher results with 91.77% mean IOU on TrayDataset and 77% mean IOU on MyFood dataset. The proposed CDPN method achieved very competitive results as compared to the state-of-the-art approaches. The future plan is to extend this food segmentation research to a variety of applications including food volume estimation, nutrient intake monitoring, and calories estimation to track the food intake of the people involved in a smoke quitting program by simply taking a picture of what they consume to increase awareness of their daily diet.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Mazhar Hussain and Alessandro Ortis conducted the research and designed the architecture; Mazhar Hussain analyzed the data and wrote the first draft; Alessandro Ortis and Sebastiano Battiato supervised the research and revised the writing of the paper; Riccardo Polosa and Sebastiano Battiato supervised the research project; all authors had approved the final version.

FUNDING

This study was sponsored by ECLAT srl, a spin-off of the University of Catania, with the help of a grant from the private nonprofit Foundation for a Smoke-Free World Inc. (FSFW COE1-05).

ACKNOWLEDGMENT

The authors thanks to the ECLAT srl, a spin-off of the University of Catania for sponsorship, with the help of a grant from the private nonprofit Foundation for a Smoke-Free World Inc. (FSFW COE1-05) for the advancements and global progress in smoking cessation and harm reduction.

REFERENCES

[1] S. Battiato, P. Caponnetto, O. Giudice, M. Hussain, R. Leotta, A. Ortis, and R. Polosa, "Food recognition for dietary monitoring during smoke quitting," in *Proc. the International Conference on Image Processing and Vision Engineering (IMPROVE 2021)*, 2021, pp. 160–165.

[2] A. Ortis, G. M. Farinella, and S. Battiato, "Survey on visual sentiment analysis," *IET Image Processing*, vol. 14, no. 8, pp. 1440–1456, 2020.

[3] A. Ortis, P. Caponnetto, R. Polosa, S. Urso, and S. Battiato, "A report on smoking detection and quitting technologies," *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, 2614, 2020.

[4] G. Maguire, H. Chen, R. Schnall, W. Xu, and M. C. Huang, "Smoking cessation system for preemptive smoking detection," *IEEE Internet of Things Journal*, 2021.

[5] C. Nishida, R. Uauy, S. Kumanyika, and P. Shetty, "The joint WHO/FAO expert consultation on diet, nutrition and the prevention of chronic diseases: Process, product and policy implications," *Public Health Nutrition*, vol. 7, no. 1a, pp. 245–250, 2004.

[6] D. Allegra, S. Battiato, A. Ortis, S. Urso, and R. Polosa, "A review on food recognition technology for health applications," *Health Psychology Research*, vol. 8, no. 3, 2020.

[7] K. Kitamura, C. Silva, T. Yamasaki, and K. Aizawa, "Image processing based approach to food balance analysis for personal food logging," in *Proc. 2010 IEEE International Conference on Multimedia and Expo*., 2010, pp. 625–630.

[8] G. M. Farinella, D. Allegra, M. Moltisanti, F. Stanco, and S. Battiato, "Retrieval and classification of food images," *Computers in Biology and Medicine*, vol. 77, pp. 23–39, 2016.

[9] M. Hussain, A. Ortis, R. Polosa, and S. Battiato, "User-biased food recognition for health monitoring," in *Proc. Image Analysis and Processing - ICIAP 2022*, Springer, Cham, 2022, vol. 13233.

[10] Y. Lu, T. Stathopoulou, M. F. Vasiloglou, S. Christodoulidis, Z. Stanga, and S. Mougiakakou, "An artificial intelligence-based system to assess nutrient intake for hospitalised patients," *IEEE Transactions on Multimedia*, vol. 23, pp. 1136–1147, 2020.

[11] U. Sharma, B. Artacho, and A. Savakis, "Gourmetnet: Food segmentation using multi-scale waterfall features with spatial and channel attention," *Sensors*, vol. 21, no. 22, 7504, 2021.

[12] C. Peng and J. Ma, "Semantic segmentation using stride spatial pyramid pooling and dual attention decoder," *Pattern Recognition*, vol. 107, 107498, 2020.

[13] K. J. Pfisterer, R. Amelard, A. G. Chung, B. Syrnyk, A. MacLean, and A. Wong, "Fully-automatic semantic segmentation for food intake tracking in long-term care homes," arXiv e-prints, arXiv:1910, 2019.

[14] C. N. Freitas, F. R. Cordeiro, and V. Macario, "Myfood: A food segmentation and classification system to aid nutritional monitoring," in *Proc. 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, IEEE, 2020, pp. 234–239.
[15] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN. Facebook ai research (fair)," arXiv preprint, arXiv:1703.06870, 2018.
[16] Y. He, C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp, "Food image analysis: Segmentation, identification and weight estimation," in *Proc. 2013 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2013, pp. 1–6.
[17] E. Aguilar, B. Nagarajan, B. Remeseiro, and P. Radeva, "Bayesian deep learning for semantic segmentation of food images," *Computers and Electrical Engineering*, vol. 103, 108380, 2022.
[18] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. International Conference on Machine Learning*, 2016, pp. 1050–1059.
[19] A. Ramesh, V. B. Raju, M. Rao, and E. Sazonov, "Food detection and segmentation from egocentric camera images," in *Proc. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, pp. 2736–2740.
[20] Ultralytics/yolov5: v3.0. (Aug. 2020). [Online]. Available: https://doi.org/10.5281/zenodo.3983579
[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham., 2015, pp. 234–241.
[22] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. U. Liang, "A nested u-net architecture for medical image segmentation," arXiv preprint, arXiv:1807.10165, 2018.
[23] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2017 pp. 2117–2125.
[24] S. Seferbekov, V. Iglovikov, A. Buslaev, and A. Shvets, "Feature pyramid network for multi-class land segmentation," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 272–275.
[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
[27] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.
[28] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFS," arXiv preprint, arXiv:1412.7062, 2014.
[29] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[31] Tray food dataset for food image segmentation. [Online]. Available: https://www.kaggle.com/datasets/thezaza102/tray-food-segmentation
[32] C. N. Freitas, F. R. Cordeiro, and V. Macario, Myfood: MyFood Dataset (v1.0.0), 2020.
[33] Myfood dataset: Zenodo. [Online]. Available: https://doi.org/10.5281/zenodo.4041488
[34] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," arXiv preprint, arXiv:1606.02147, 2016.
[35] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. the European conference on computer vision (ECCV)*, 2018, pp. 801–818.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint, arXiv:1409.1556, 2014.
[37] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.
[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, A. Rabinovich *et al.*, "Going deeper with convolutions," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.