Bayesian Deep Learning for Semantic Segmentation of Food Images

Eduardo Aguilar[a,*], Bhalaji Nagarajan[b], Beatriz Remeseiro[c], Petia Radeva[b,d]

[a]*Department of Computing and Systems Engineering, Catholic University of the North, Avenida Angamos 0610, Antofagasta, 1270709, Antofagasta, Chile*
[b]*Departament de Matematiques i Informatica, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, Barcelona, 08007, Barcelona, Spain*
[c]*Department of Computer Science, Universidad de Oviedo, Campus de Gijon s/n, Gijon, 33203, Asturias, Spain*
[d]*Computer Vision Center, Cerdanyola (Barcelona), Spain*

Abstract

Deep learning has provided promising results in various applications; however, algorithms tend to be overconfident in their predictions, even though they may be entirely wrong. Particularly for critical applications, the model should provide answers only when it is very sure of them. This article presents a Bayesian version of two different state-of-the-art semantic segmentation methods to perform multi-class segmentation of foods and estimate the uncertainty about the given predictions. The proposed methods were evaluated on three public pixel-annotated food datasets. As a result, we can conclude that Bayesian methods improve the performance achieved by the baseline architectures and, in addition, provide information to improve decision-making. Furthermore, based on the extracted uncertainty map, we proposed three measures to rank the images according to the degree of noisy annotations they contained. Note that the top 135 images ranked by one of these measures include more than half of the worst-labeled food images.

Keywords: Deep learning, Uncertainty quantification, Bayesian inference, Image segmentation, Food analysis

*Corresponding author
Email address: eaguilar02@ucn.cl (Eduardo Aguilar)

1. Introduction

Administrative policies around the globe, ranging from local counties to the World Health Organization, have been centered around better health care for everyone. Health is one of the core targets in the 2030 Agenda for Sustainable Development and is identified to impact each of the other goals in one way or another [1]. As outlined by the new Industry 5.0 paradigm, human-centric solutions are the future of things to come [2]. Industry 5.0, formally charted on January 2021, focuses on providing resilient, sustainable solutions that can be integrated into the existing social and environmental priorities [3]. Digitization of healthcare goes hand-in-hand with Industrial progress. Artificial Intelligence (AI) and Big Data have enabled the creation of better healthcare solutions that are cost-effective and efficient. Global healthcare has moved from being a centralized *organization* to a more personalized *service*. Recent decision-making in many of these customized approaches involves assistive technologies such as computer vision and machine learning solutions. Technologies are closer to the common person as never before, and have helped in improving the living conditions.

Expeditious growth in deep learning and computer vision algorithms, along with the availability of large-scale food datasets, serve as pillars in the development of food systems. Paying attention to the multi-class food segmentation problem, we can find some interesting works in the literature. Aslan et al. [4] proposed a method for dietary monitoring that solves the semantic food segmentation task employing a DeepLabv2 pre-trained on the MSCOCO dataset and fine-tuned on the UNIMIB2016 dataset. They also explored the binary food segmentation problem by applying Dense Conditional Random Fields (Dense CRF), obtaining more refined boundaries and fewer false positives. Chiang et al. [5] developed a system to analyze the composition of food images, in terms of calories and nutrients, which uses a Mask Region-based Convolutional Network method (mask R-CNN) with ad-hoc post-processing. Additionally, they introduced their dataset for food recognition, known as Ville Cafe. Also in the

context of monitoring diet and nutritional intake, Freitas et al. [6] built a dataset with Brazilian food and presented a comparative study with five segmentation models, including the aforementioned DeepLabv3+ and mask R-CNN. Another architecture used for semantic food segmentation is GourmetNet [7], which incorporates both channel and spatial attention through a multi-scale feature representation. The method achieved state-of-the-art performance on two datasets, UNIMIB2016 and UECFOODPIX. Wu et al. [8] provided FoodSeg103, a new dataset with 9,490 images annotated with 154 ingredient classes and pixel-wise masks. They also proposed ReLeM, a multi-modality learning approach that was evaluated on their dataset and compared with three well-known semantic segmentation methods.

Semantic segmentation encounters various challenges with respect to data collection [9]. The quantity of data needed to train any deep learning algorithm is typically huge. This is even more important for segmentation algorithms, which need information about each and every pixel. Collecting ground truth pixel labels for such volumes of data is expensive. For instance, in the Cityscapes dataset [10], widely used in training self-driving cars, it took close to 90 minutes to annotate a single image. With less accurate labels, the performance of the models deteriorates. Recent advances in active learning frameworks [11] have helped to some extent, whilst self-supervised learning algorithms have resulted in more confident predictions by using better label-error maps [12]. However, both active learning and self-learning approaches require high-quality labels to learn the unlabeled data. The problem is that annotation efforts, both manual and automated, lead to noisy labels due to the complexity of the task [13]. The most important aspect of learning from noisy labels is to accurately characterize the uncertainty of the label noise [14]. Confident Learning (CL) algorithms work on the assumption that label noise is often class-conditional and it can be learned directly from the class labels [15]. MultiNET [16], an improved version of the original CL algorithm, uses aggregated outputs from multiple deep networks followed by a detection threshold to improve the noise detection rate. Using the CL algorithm in a teacher network allowed the teacher model to identify

potential label noise and subsequently, by assigning soft-corrected masks, better students were created achieving more confident models [17].

Modern deep learning methods have reached real-world applications and, therefore, it is important for them to be *certain* about the predictions. In several critical applications, incorrect segmentation often has catastrophic results, such as cases involving self-driving cars [18]. Softmax probabilities were used as a measure of confidence. However, this is not always reliable [19] as seen in cases of adversarial examples. The Bayesian formulation is a popular way to estimate the model confidence in terms of uncertainty, which is treated as a measure of the trustworthiness of any deep learning algorithm. In fact, most of the semantic segmentation algorithms that implement uncertainty modeling are based on the Bayesian inference approach. Kendall et al. [20] presented Bayesian SegNet, the first probabilistic semantic segmentation approach using deep learning. The proposed method is an extension of SegNet, an encoder-decoder Neural Network (NN) architecture, to a Bayesian Convolutional Neural Network (CNN) that produces a probabilistic segmentation as output using MC-dropout [19]. Recently, Dechesne et al. [21] presented Bayesian U-Net, also based on MC-dropout [19] for uncertainty estimation but with the popular U-Net [22] for semantic segmentation. For their part, Mukhoti and Gal [23] proposed three metrics to evaluate Bayesian models designed for semantic segmentation, using also MC-dropout for experimentation purposes. All these works highlight the effectiveness of the Bayesian methods that provide accurate semantic segmentation and a reliable uncertainty map.

Food applications require a high level of certainty in their predictions since they are often used to make critical decisions such as what should constitute the next meal, what should be the quantity of certain food components, identify allergens of prepared foods, etc. However, to the best of our knowledge, there is no previous research work focused on uncertainty modeling applied to semantic food segmentation. Inspired by the good results obtained with the Bayesian methods described above, in this article we will extend the state-of-the-art semantic segmentation food models with a Bayesian approach. In addition, we

Figure 1: Noisy label present in a sample image from the UECFOODPIX dataset. From left to right: original image, ground-truth (GT) labels, predicted segmentation mask, and the uncertainty map. The patches in the original image (blue boxes) and the uncertainty map (red boxes) represent the noisy label where uncertainty in the prediction is present. Best viewed in color.

delve into the analysis of quantified uncertainty to assess the benefit of the uncertainty map in identifying noisy labels present in the data. The experimentation carried out shows how error predictions can be identified with these algorithms and how we take advantage of uncertainty map information to identify images with noisy labels (see Fig 1). There are two key contributions of this work. (1) We present the Bayesian version of two benchmark food segmentation methods: DeepLabv3+ and GourmetNet using MC-Dropout. Our proposal provides both segmentation results and uncertainty measures related to each prediction. (2) We analyzed the uncertainty maps to identify possible wrong or mislabeled data. The proposed methods were validated on several public food datasets, where superior performance was observed in all of them compared to the baseline architectures.

The rest of the paper is organized as follows. We explain the proposed method in Section 2. We detail the experimental setup in Section 3. We discuss the results in Section 4, followed by the conclusions in Section 5.

2. Bayesian Semantic Segmentation

In Bayesian learning, the posterior predictive distribution that we want to compute is given by:

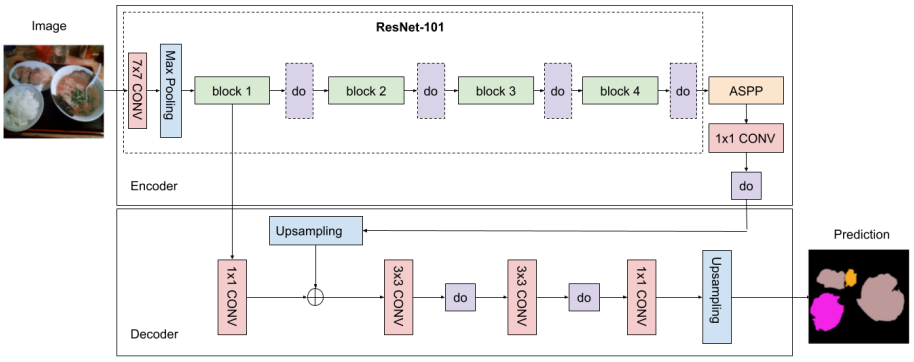
$$p(y|x,\mathcal{D}) = \int p(y|x,w)p(w|\mathcal{D})dw,$$

where $\mathcal{D}$ denotes the training set, x the input image, y the output class label, w the weights of the neural network, $p(y|x,w)$ the likelihood (e.g., the softmax output of the CNN), and $p(w|\mathcal{D})$ the posterior distribution.

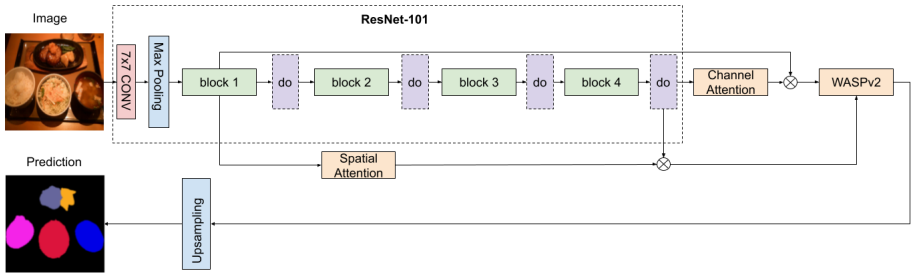
In particular, for deep learning methods, a full Bayesian inference is intractable. However, instead of calculating the exact posterior distribution, it can be approximated. Different methods have been proposed to approximate a Bayesian inference, including MC-dropout [19] that is the most popular strategy used in semantic segmentation [20, 21, 23]. MC-dropout can be interpreted as a variational inference technique where a Bernoulli distribution is placed on the weights (w) of the neural network. In variational inference, the real posterior distribution is approximated by minimizing its Kullback–Leibler (KL) divergence with respect to a variational distribution $q(w)$.

Interestingly, Gal and Ghahramani [19] showed that using a dropout layer with a dropout rate of p on a hidden layer has the same effect as placing a Bernoulli distribution with parameter p on the weights of that layer. In addition, they also found that minimizing the negative logarithm of the likelihood with the standard optimization algorithm provides an effect equivalent to minimizing the KL divergence. Thus, with MC-dropout (see Algorithm 1), Bayesian inference can be performed using the traditional training procedure by simply putting a dropout layer after each training layer (e.g., convolutional layer). Then, one can sample the results provided by the posterior distribution during the prediction phase by performing T forward passes for the same input data while keeping the dropout layer active. Finally, with the mean of the sampled results, the final prediction is calculated and the uncertainty is estimated with entropy and mutual information.

In order to perform Bayesian inference with MC-dropout, some adjustments to the architecture of the models are required; in particular, the incorporation of dropout layers after trainable layers. For the latter, previous Bayesian semantic



(a)



(b)

Figure 2: BayesianDeepLabv3+ (a) and BayesianGourmetNet (b) with ResNet101 as backbone. Dropout (do) layers added to both networks are marked by dashed lines.

segmentation deep learning approaches differ in criteria regarding the numbers of dropout layers and where these layers are placed. For encoder-decoder based segmentation methods, the dropout layer has been placed after all [20, 21] or some [20] convolutions and deconvolution blocks, or only in the middle flow of the network [23]. In [20], the authors noted in practice that using a dropout layer after each trainable layer was representing a too strong regularization, causing the network to learn very slowly. Alternatively, they explored a number of variants and found a good trade-off between accuracy and uncertainty quantification by placing dropout layers after four central encoders and decoders. More closely related to our work is the Bayesian semantic segmentation proposed in [23]. Here, the authors also used DeepLabv3+ as the base architecture, but with a different backbone, in this case the Xception network. They proposed to

Algorithm 1: Bayesian approximation using MC-dropout for image segmentation

input: Target image I, Trained model with dropout layers M, T forward passes;

output: Semantic segmentation SS, Entropy map E, Mutual information map MI;

$acc\_prob \leftarrow 0$;

$acc\_prob\_log\_prob \leftarrow 0$;

$t \leftarrow 0$;

while *t is less than T* **do**

  // Model prediction for input image using active dropout layers;

  $logits \leftarrow M.pred(I)$;

  $prob \leftarrow softmax(logits)$;

  $prob\_log\_prob \leftarrow prob \times log(prob)$;

  $acc\_prob \leftarrow acc\_prob + prob$;

  $acc\_prob\_log\_prob \leftarrow acc\_prob\_log\_prob + prob\_log\_prob$;

  $t \leftarrow t + 1$;

end

$mean\_prob \leftarrow \frac{1}{T} \times acc\_prob$;

// $mean\_prob$ *and* $acc\_prob\_log\_pro \in \mathbb{R}^{width \times height \times nclasses}$;

$E \leftarrow -sum(mean\_prob \times log(mean\_prob), axis = -1)$;

$MI \leftarrow E + \frac{1}{T} \times sum(acc\_prob\_log\_prob, axis = -1)$;

$SS \leftarrow argmax(mean\_prob, axis = -1)$;

return SS, E, MI;

add the dropout layer only in the backbone, after every four Xception modules in the middle of the network, arguing that higher level features in the deeper layer are better modeled using probabilistic weights.

As we mentioned before, the best results in semantic segmentation of food images have been obtained with the DeepLabv3+ and GourmetNet architec-

tures, both of them using ResNet101 [24] as the backbone. Inspired by the works discussed above, we propose to approximate a Bayesian inference in both networks by means of MC-dropout incorporating a total of four extra dropout layers placed after each residual block belonging to the backbone network (see Fig. 2). In this manner, it will be possible to obtain both the prediction and its corresponding uncertainty. A brief description of DeepLabv3+ and GourmetNet can be seen in the following subsections.

2.1. DeepLabv3+

DeepLabv3+ [25] is based on an encoder-decoder structure to perform the semantic image segmentation task (see Fig. 2(a)). On the encoder side, this architecture uses the original DeepLabv3 [26], which is composed of a backbone (e.g., Xception or ResNet101) for feature extraction and an Atrous Spatial Pyramid Pooling (ASPP) for capturing multi-scale information. Specifically, the encoding results correspond to the last feature map before the DeepLabv3 logits layer. On the decoder side, there is a simple, but effective approach to refine the segmentation mask by retrieving object segmentation details. First, the low-level features extracted from the backbone are convolved using a 1×1 convolution to reduce the number of channels, and then concatenated with the bilinear upsampled encoder output by a factor of 4. After that, some 3×3 convolutional layers are applied to the concatenation in order to refine the extracted features. Next, 1×1 convolutions are applied on top of the network where the number of convolutions equals the number of classes. Finally, a bilinear upsampling by a factor of 4 is used to provide the segmentation result.

2.2. GourmetNet

GourmetNet [7] is a segmentation architecture designed for food segmentation that incorporates a Waterfall Atrous Spatial Pooling (WASPv2) module to capture multi-scale features, coupled with dual attention modules to capture the context (see Fig. 2(b)). GourmetNet uses the ResNet101 as a backbone

Figure 3: Image samples with their respective annotations for the UECFOODPIX (left group), UNIMIB2016 (middle group), and Food201 (right group) datasets. Notice that in UECFOODPIX there are two annotations that correspond to the GrabCut version (first) and the manual annotation version (second) of this dataset.

architecture and modifies the last block to learn multi-scale features. The authors employed atrous convolutions instead of regular convolutions. On the other hand, GourmetNet uses spatial attention to capture the low-level features and channel attention to capture high-level features. Both the low-level and high-level features are refined in the attention modules and fed into the WASPv2 module. The WASPv2 module uses a waterfall-like configuration of atrous convolutional layers to increase the field of view. Utilizing the multi-scale representations from both the attention modules and increasing the field of view led to better features. The final layer of the module is similar to an in-built decoder, which produces the segmentation results.

3. Experiments

This section describes the datasets used for experimental purposes. After that, the experimental setup is explained in detail. Finally, the measures used to evaluate the performance are shown.

3.1. Datasets

Three public food datasets were selected to perform semantic segmentation of food images by means of the proposed approach and baseline architectures. These are: 1) UNIMIB2016 [27], an Italian food dataset composed of 1,027 food images and 73 food categories collected in a self-service canteen,

in a semi-controlled environment and using a smartphone camera; 2) UEC-FOODPIX [28, 29], a popular large-scale public dataset comprised primarily of Japanese foods consisting of 10,000 images and 102 food categories with annotations corresponding to plate masks automatically extracted from previously labeled bounding boxes (UECFOODPIX) or with pixel-wise annotation (UEC-FOODPIXComplete); and 3) Food201 [30], the first food dataset for semantic segmentation, but not widely used because annotations were not available until recently. The dataset is composed of a subset of images from Food101 [31], specifically 12,093 images which include pixel-wise annotations for 201 different foods that may be present in the images. Unlike UNIMIB2016, UECFOODPIX and Food201 contain images with noisy labels, different resolutions and large visual variations for the same foods.

3.2. Experimental setup

Two state-of-the-art semantic food segmentation methods were selected to perform the experiments, these are: DeepLabv3+ [25] and GourmetNet [7]. In both methods the ResNet101 [24] was used as the backbone. In addition, we adapt these methods to approximate Bayesian inference with the aim to extract and analyze the uncertainty in each given prediction in the segmentation of food images. For the latter, the backbone is adapted by adding a dropout layer with a probability of $p = 0.1$ after each residual block in order to estimate the uncertainty by means of the MC-dropout approach [19]. We named the Bayesian-based methods as BayesianDeepLabv3+ and BayesianGourmetNet.

For training we use the backbone pre-trained on ImageNet [32] and fine-tune the entire network for 100 epochs with a Cross-Entropy Loss and Stochastic Gradient Descent optimizer. For the optimization, we set a momentum of 0.9 and a weight decay of 0.0005. Regarding the learning rate (LR), it depends on the dataset and was set at 10 times higher for the weights of the layers located at the top of the backbone. Specifically, for UECFOODPIX and Food201 the base LR was 0.001 and for UNIMIB2016 it was 0.01. For DeepLabv3+-based models, the LR is decayed in a polynomial manner after each iteration with an

exponential factor of 0.9. For GourmetNet-based models, the LR is decayed in a multi-step manner at epochs 40 and 70 with a factor of 0.3.

For the Bayesian semantic segmentation models, MC-dropout was used with $T = 100$ during the prediction phase to compute the uncertainty, where T corresponds to the multiple forward passes through the model performed with the dropout turned on.

With respect to the data, for comparison purposes, a 320×320 input size was used for UECFOODPIX and Food201, and an input size of 480×360 for UNIMIB2016. The original images were resized according to the nearest filter and no other transformation was applied to them.

All experiments were performed on a server with a graphic card of 11 GB of VRAM, using Pytorch as the deep learning framework.

3.3. Validation

All the experiments were evaluated quantitatively and qualitatively. In the quantitative case, we based the evaluation in three standard semantic segmentation metrics: Accuracy (Acc), mean Accuracy ($mAcc$) and mean Intersection over Union ($mIoU$). In addition, we use two metrics proposed for uncertainty quantification [23]: predictive entropy ($\hat{\mathbb{H}}$), which captures both aleatoric and epistemic uncertainty; and mutual information ($\hat{\mathbb{I}}$), which only captures the epistemic uncertainty. The metrics mentioned above are formally detailed below:

$$\hat{\mathbb{H}}(y|x, \mathcal{D}_{train}) = -\sum_{c} (\frac{1}{T} \sum_{t} p(y = c|x, \hat{w}_t) \log(\frac{1}{T} \sum_{t} p(y = c|x, \hat{w}_t)))$$

$$\hat{\mathbb{I}}(y, w|x, \mathcal{D}_{train}) = \hat{\mathbb{H}}(y|x, \mathcal{D}_{train}) + \frac{1}{T} \sum_{c,t} p(y = c|x, \hat{w}_t) \log(p(y = c|x, \hat{w}_t))$$

where $p(y = c|x, \hat{w}_t)$ is the softmax probability of the input x being in class c, and $\hat{w}_t$ are the model weights for the forward pass t.

On the other hand, for the qualitative analysis, we will interpret the uncertainty estimated with the proposed approach with the intention of making a

visual interpretation of the meaning of the epistemic and aleatoric uncertainty present in the food images. Additionally, we will assess whether uncertainty information is useful in automatically determining mislabeled or incorrect data.

4. Results

This section reports the results obtained in the experiments using the traditional semantic segmentation measures (Acc, $mAcc$, $mIoU$) and also the analysis of the uncertainty associated with the predictions.

4.1. Results based on standard measures for Semantic Food Segmentation

To evaluate the results of the state-of-the-art models and the proposed methods in a comparative way, all of them were trained using the same hyperparameters, as explained in Section 3.2. Table 1 shows the results for the four datasets considered: UNIMIB2016, UECFOODPIX, UECFOODPIXComplete, and Food201. For all datasets, an improvement in the state-of-the-art models performance is evident when the dropout layers are incorporated into the backbone network (bb w/d) based on the proposed design. In particular, GourmetNet shows a bigger increase than DeepLabv3+ and surprisingly a great improvement is shown in UNIMIB2016, with about 4% more in terms of $mIoU$. The latter suggests that GourmetNet is more prone to strongly overfitting the network than DeepLabv3+ when training on a dataset with a small amount of data. Furthermore, it can be seen that the proposed Bayesian version of the state-of-the-art models not only allows us to obtain more information about the certainty of the segmented pixels, but also provides the best results regardless of the target dataset. BayesianGourmetNet shows better behavior with clean label datasets (UNIMIB2016 and UECFOODPIXComplete). On the contrary, BayesianDeepLabv3+ shows comparable or even better behavior than BayesianGourmetNet with more complex datasets (UECFOODPIX and Food201), which contain images of several food classes obtained in uncontrolled environments that are semi-automatically labeled and, therefore, may contain some degree of

noise in the labels. On the other hand, when comparing the results in terms of *Acc* and *mAcc*, it can be seen that the correct segmentation is mostly balanced between classes in most of the datasets, with the exception of Food201, where there is a clear bias towards the classes with the highest occurrence in the images (e.g., background). Likewise, it is observed that GourmetNet-based models tend to provide a better balance in performance between classes.

Table 2 shows the class-wise results provided by GourmetNet-based models on the UECFOODPIXComplete and Food201 datasets, respectively. Specifically, it shows the three classes that obtain the best (Top 3) and worst (Bottom 3) results, the mean, standard deviation (Std Dev), and median. All these metrics were calculated in terms of IoU. In the UECFOODPIXComplete case, it can be seen that there is coincidence in the classes with better (classes id 13 and 14) and worse (class id 45) results, and in all of them our approach provides better performance. Classes id 13 (*croissant*) and 14 (*roll bread*) can be considered easy food classes because the color, shape and texture are almost constant in the samples that belong to these two classes. For this reason, the model is able to learn them very well. On the contrary, class 45 (*fried fish*) can be considered as a fine-grained food class because the visual appearance is very similar to other fried foods present in the dataset, thus increasing the complexity of its segmentation. Overall, there is a noticeable improvement (more than 3% in terms of mean IoU and about 4.5% in terms of median IoU) when BayesianGourmetNet is compared to GourmetNet.

In the Food201 case, it is also observed that the proposed method provides an improvement in the prediction of the best segmented classes. Again, the best results are obtained in easy food classes, for example in *edamame* (class id 76) and *macarons* (class id 124). It is interesting to note that, in this dataset, there are several classes that have not been learned, that is, with IoU equal to 0. Specifically, 71 classes were not learned by GourmetNet and 77 by BayesianGourmetNet models. Most of these classes correspond to sauces, drinks (e.g., class id 208: *chocolate*) and ingredients (e.g., class id 68: *croutons*, class id 42: *onions*, class id 143: *orange slice*) that are underrepresented, so the model does

Table 1: Results for semantic food segmentation on four datasets. Best performances are shown in bold.

Dataset	Model	Acc	mAcc	mIoU
UNIMIB2016	BayesianDeepLabv3+	**0.9843**	**0.8271**	**0.7717**
	DeepLabv3+ (bb w/d)	0.9836	0.8243	0.7657
	DeepLabv3+	0.9822	0.8033	0.7469
	BayesianGourmetNet	**0.9861**	**0.8646**	**0.8076**
	GourmetNet (bb w/d)	0.9856	0.8606	0.8014
	GourmetNet	0.9829	0.8323	0.7685
UECFOODPIX	BayesianDeepLabv3+	**0.8408**	0.7012	**0.5907**
	DeepLabv3+ (bb w/d)	0.8375	**0.7020**	0.5810
	DeepLabv3+	0.8383	0.6961	0.5783
	BayesianGourmetNet	**0.8456**	**0.7196**	**0.5926**
	GourmetNet (bb w/d)	0.8419	0.7175	0.5867
	GourmetNet	0.8419	0.6947	0.5789
UECFOODPIXComplete	BayesianDeepLabv3+	**0.8729**	**0.7615**	**0.6421**
	DeepLabv3+ (bb w/d)	0.8664	0.7574	0.6311
	DeepLabv3+	0.8706	0.7571	0.6307
	BayesianGourmetNet	**0.8805**	**0.7817**	**0.6616**
	GourmetNet (bb w/d)	0.8767	0.7756	0.6506
	GourmetNet	0.8707	0.7587	0.6288
Food201	BayesianDeepLabv3+	**0.7696**	0.3296	**0.2520**
	DeepLabv3+ (bb w/d)	0.7659	**0.3341**	0.2464
	DeepLabv3+	0.7674	0.3290	0.2473
	BayesianGourmetNet	**0.7681**	**0.3366**	**0.2510**
	GourmetNet (bb w/d)	0.7629	0.3327	0.2449
	GourmetNet	0.7552	0.3155	0.2345

* bb w/d=backbone with dropout layer after each residual block

Table 2: Class-wise results in terms of IoU obtained by BayesianGourmetNet (BGNet) and GourmetNet (GNet) on the UECFOODPIXComplete (UFPC) and Food201 datasets.

Dataset	Model	Top 3 (class id)	Bottom 3 (class id)	Mean	Std Dev	Median
UFPC	BGNet	0.9566 (013)	0.1825 (045)			
		0.9381 (014)	0.2298 (089)	0.6616	0.1884	0.6794
		0.9330 (028)	0.2627 (090)			
	GNet	0.9565 (013)	0.1796 (045)			
		0.9220 (000)	0.2657 (101)	0.6288	0.1762	0.6348
		0.9182 (014)	0.2912 (056)			
Food201	BGNet	0.8829 (076)	0.0000 (208)			
		0.8604 (124)	0.0000 (142)	0.2510	0.2546	0.1920
		0.7997 (000)	0.0000 (143)			
	GNet	0.8806 (076)	0.0000 (208)			
		0.8318 (124)	0.0000 (068)	0.2345	0.2420	0.1677
		0.8268 (111)	0.0000 (142)			

not have enough information to learn them. Although BayesianGourmetNet has more unlearned food classes than GourmetNet, this method improves the overall results by more than 1.5% in terms of mean IoU and about 2.5% in terms of median IoU. From these results it can be inferred that the model discards the learning of classes with almost 0 IoU in favor of the rest of the classes.

4.2. Qualitative results based on the uncertainty map analysis

Creating pixel annotations for image segmentation is a time-consuming task. Common approaches start annotations with a semi-automatic technique (e.g., GrabCut) and then manually refine the labels to reduce the numbers of mislabeled or unlabeled data [29, 30]. Arguably, reviewing all images manually to make corrections is also time consuming. Keeping in mind the fact that uncertainty modeling allows us to improve our understanding of what the model learns from the data, we propose to take advantage of uncertainty map analysis to discover the data that should be reviewed in order to prioritize or select for

manual refinement. We propose to perform qualitative analysis of the uncertainty maps based on: (1) the segmented mask presented in the full image, (2) only considering the background mask, and (3) only using the foreground mask (any food).

4.2.1. Full image segmentation mask

The qualitative analysis was done using the full image segmentation mask in order to identify images annotated with noisy labels. In this case, the uncertainty present in the images was analyzed regardless of the nature of the noisy labels; that is, whether it occurs because the food categories are mislabeled or because the background images are labeled as some food category.

Let's consider $maxH = -ln(\frac{1}{C})$ the maximum $\hat{\mathbb{H}}$, where C corresponds to the number of classes. Then, the first evaluation measure (EM_1) proposed to prioritize the images in descending order is:

$$EM_1(x) = \sum_p ind(\hat{\mathbb{H}}(x_p) > maxH * \lambda)$$

where x_p corresponds to the p-th pixel of the input image x, $ind(*)$ is an indicator function that returns 1 when the condition is true and 0 otherwise, and λ is a threshold to determine when a pixel has an uncertain prediction. In our case, we use $\lambda = 0.1$.

The proposed metric EM_1 was used in the UECFOODPIX dataset to order the training data and then select some of the first images to visually analyze the obtained results. In most cases, we were able to identify images with a large number of incorrectly labeled pixels, mainly images fully labeled as a background even though food was present there. An example of the images obtained with our metric can be seen in the first row of Fig. 4. As can be observed, the model is forced to learn the labels as a background. However, a high $\hat{\mathbb{H}}$ is present in the areas where food appears, while a low $\hat{\mathbb{I}}$ appears throughout the image. Therefore, as expected, a high aleatoric uncertainty in $\hat{\mathbb{H}}$ is obtained when noisy labels are present.

4.2.2. Foreground mask

In this case, instead of using the full image segmentation mask, the uncertainty was calculated in those pixels that belong to the foreground mask. The aim is to discover those images that have food pixels mislabeled with some other food category or background. The high uncertainty present in this region is expected to allow us to identify those pixels with noisy labels.

The second evaluation measure (EM_2) proposed to prioritize the images in descending order is:

$$EM_2(x) = EM_1(\hat{x}),$$

where $\hat{x}$ corresponds to a subset of the image x placed in the foreground of the GT mask provided by UECFOODPIX.

When this measure was applied to order the images, we found that most of the first images contain a very small food region labeled with a large uncertainty. An example can be seen in the second row of Fig. 4. In this case, even if the image is mislabeled, if we focus on the foreground region, the labels are correct. The high $\hat{\mathbb{H}}$ in this region is not caused by noisy labels, but we assume that it is caused by epistemic uncertainty; that is, the absence of more correctly labeled data with the same visual content.

4.2.3. Background mask

The qualitative analysis of the uncertainty map was also performed using only the background mask. In this case, it is expected that those images whose pixels are confused with some food category will be identified by analyzing the high uncertainty present in the pixels that belong to the background mask.

The third evaluation measure (EM_3) proposed to prioritize the images in descending order is:

$$EM_3(x) = EM_1(\check{x}),$$

where $\check{x}$ corresponds to a subset of the image x placed in the background of the GT mask provided by UECFOODPIX.

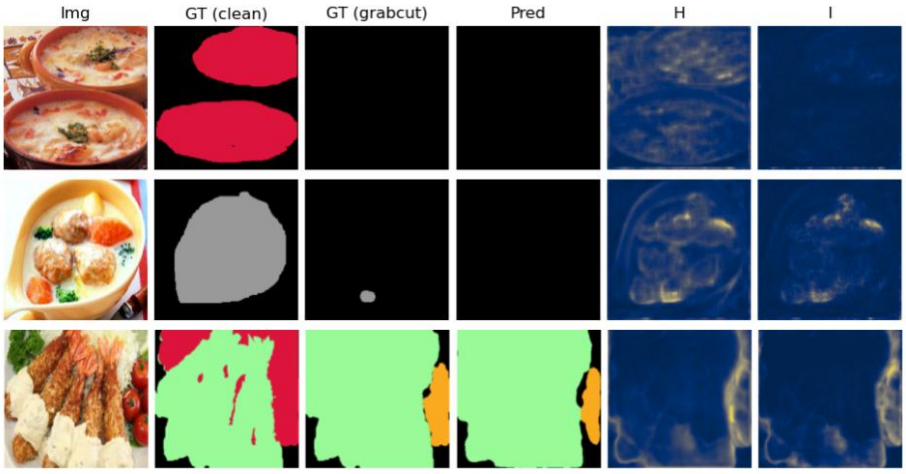


Figure 4: Qualitative results obtained in the training set of UECFOODPIX. The first, second, and third rows represent images selected using the criteria EM_1, EM_2, and EM_3, respectively. Note that Img, GT (clean), GT (GrabCut), Pred, H, and I correspond to the target image, the clean labels provided by UECFOODPIXComplete, the GrabCut labels provided by UECFOODPIX, the predicted mask, the uncertainty measured with the entropy, and the uncertainty measured with the mutual information.

Similarly to the foreground mask measure, after ordering the images with this measure, we observed that the first selected images contain a small background region with high uncertainty. An example can be seen in the third row of Fig. 4, where it is observed that there is a high uncertainty in the region poorly labeled as background. However, this same image does not present high uncertainties in the regions where mislabeling between different foods occurs. The reason could be that the difference between the background and some food category is large. However, since the data is tagged at the dish level and not at the ingredient level, the model expects that the ingredient may be part of different foods depending on the context and, therefore, it is not able to compute correctly in some cases the certainty between different food categories.

4.3. Quantitative results based on the uncertainty map analysis

In addition to the qualitative analysis, we performed a quantitative analysis regarding the capabilities of the proposed measures, specifically EM_1, to order

the images according to the degree of noise of the labels. For this purpose, we first computed the numbers of labels incorrectly annotated in the UECFOODPIX based on the annotations of UECFOODPIXComplete in all the training images. Next, we sorted the images in descending order considering the worst labeled first. Then, the intersection of the worst labeled images with respect to the images selected with the proposed measures was computed for the first 45, 90, 135, 180, and 225 images. Also, instead of using EM_1, the intersection was evaluated using Mean $\hat{\mathbb{H}}$ (mH) and Random Selection (RS). In the case of RS, we report the mean intersection achieved by 1000 random subsets. The results are reported in Table 3, where we show that our measure provides the highest degree of intersection. Specifically, more than 50% coincidence is evident when we analyze the first 135 images. These results support the effectiveness of our measure in finding images with noisy labels.

Table 3: Degree of intersection between a subset of the worst labeled images belonging to UECFOODPIX with respect to a subset of images selected by different criteria: Mean $\hat{\mathbb{H}}$ (mH), Mean EM_1 (mEM_1), and Random Selection (RS).

#Images	mH	mEM_1	RS
45	0.2444	0.3556	0.0051
90	0.3778	0.4556	0.0100
135	0.4593	0.5111	0.0150
180	0.4278	0.4444	0.0200
225	0.3733	0.3911	0.0249

4.4. Analysis of the successful and unsuccessful segmentation

Some successful and unsuccessful segmentation results on the test set of the three food datasets can be seen in Fig. 5 and Fig. 6, respectively. In general, across all datasets, the best results were achieved when the images contained few food instances and these instances correspond to food categories with few ingredients. Furthermore, we can see from Fig. 5 that the predictions contain a high degree of uncertainty in the contours of the food region (the boundary

between the food and the bottom) where a small error in the segmentation is detected. In particular, we can notice that the method loses the details in the contour part providing a smoother prediction. In the Food201 case (third row of Fig. 5), a high uncertainty is obtained in the region of the food produced mainly by the shadow that is placed on it. Regarding the unsuccessful segmentation results (Fig. 6), we observe in UNIMIB2016 and UECFOODPixComplete that errors occur because the model incorrectly segments food with other food categories and not with the background. In theses cases, high uncertainty is captured in the wrong prediction regions. A more complex case is evidenced in Food201, where the ingredients are mixed on the plate and do not separate well. In addition, we can observe noisy labels in the GT. Unlike the results of other datasets, the output is completely incorrect and highly uncertain for the segmentation of this image. Although UECFOODPIX has nearly the same strong imbalance of food instances contained in each image as Food201 (see Fig. 7), the latter is a more challenging dataset due to the large number of food categories and the complexity of the annotations themselves, where we noticed significant differences in performance from the rest of the tested datasets. On the contrary, UNIMIB2016 can be considered a less complex dataset due to its acquisition in a semi-controlled environment, which allows to easily differentiate the background with respect to different foods. However, there is still work to be done to avoid segmentation errors between different food categories.

4.5. Comparison with the results of the state-of-the-art methods

Finally, Table 4 includes a comparison between our proposal and the state-of-the-art models in three food datasets. We note that, despite using a standard training procedure without data augmentation, the proposed Bayesian methods outperform the results previously obtained on UNIMIB2016 and UECFOODPIXComplete, and provide comparable performance in Food201. As for the UECFOODPIXComplete dataset, we trained exactly the same model proposed by Sharma et al. [7] (GourmetNet), but achieved about 2.5% less $mIoU$ than results reported elsewhere. We believe that the performance difference is due

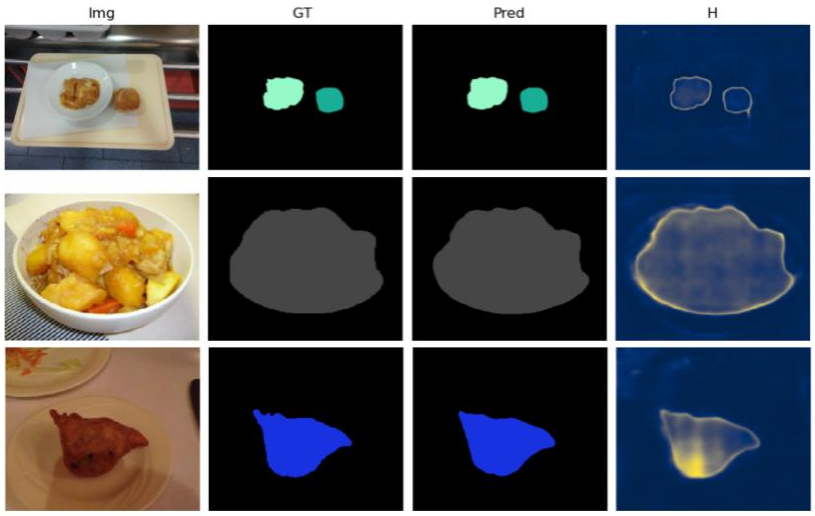


Figure 5: Images in the test set for the UNIMIB2016 (top), UECFOODPIXComplete (middle), and Food201 (bottom) datasets with the highest performance.

Table 4: Performance of semantic segmentation methods on three food datasets. Best performances are shown in bold.

Dataset	Method	Acc	mAcc	mIoU
UNIMIB2016	DeepLabv2 [4]	-	-	0.43
	SegNet [33]	-	-	0.44
	Sharma et al. [7]	-	-	0.72
	BayesianDeepLabv3+	0.98	0.83	0.77
	BayesianGourmetNet	**0.99**	**0.87**	**0.81**
UECFOODPIXComplete	Okamoto et al. [29]	-	0.67	0.56
	Sharma et al. [7]	-	-	0.65
	BayesianDeepLabv3+	0.87	0.76	0.64
	BayesianGourmetNet	**0.88**	**0.78**	**0.66**
Food201	Myers et al. [30]	0.76	0.33	**0.25**
	BayesianDeepLabv3+	**0.77**	0.33	**0.25**
	BayesianGourmetNet	**0.77**	**0.34**	**0.25**

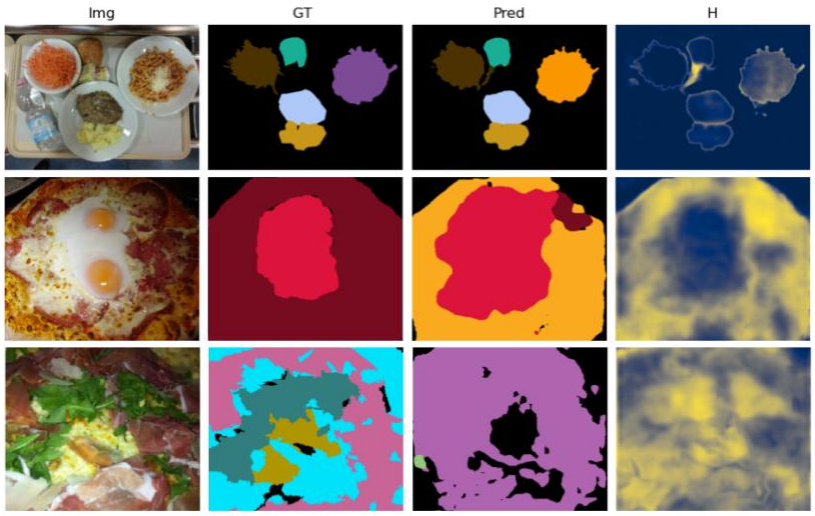


Figure 6: Images in the test set for the UNIMIB2016 (top), UECFOODPIXComplete (middle), and Food201 (bottom) datasets with the lowest performance.

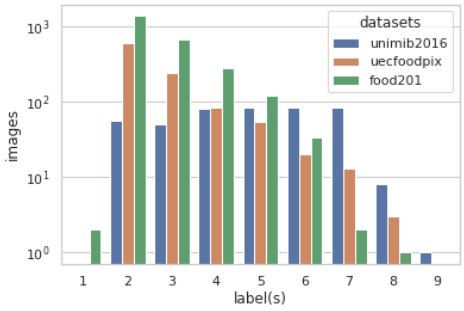


Figure 7: Total images according to the number of pixels with different labels contained.

to some hyperparameters used for training that are not properly reported in the article. Although the results of the base architecture were 2.5% lower than the reported ones, the proposed Bayesian version based on it outperformed the state-of-the-art by 1.0% in terms of $mIoU$. With respect to the Food201 dataset, Meyers et al. [30] proposed a more complex pipeline that requires training a multi-label classifier, extracting the context, and refining the CNN output before computing CRF. Also, the input data has 512 pixels for the maximum side,

which is higher than the one used for our experiments. Despite this, our proposed end-to-end method provides results comparable to those reported in this work.

5. Conclusions

We propose an adapted version of two state-of-the-art semantic segmentation methods used to segment food images in order to perform Bayesian inferences. Specifically, the backbone considered in both methods was adapted by incorporating a dropout layer after each residual block in order to approximate Bayesian inference using the MC-dropout technique. The resulting methods were named BayesianDeepLabv3+ and BayesianGourmetNet. Both of them outperformed the baseline results in terms of IoU on three public food datasets. Furthermore, these methods capture the uncertainty involved in the generated segmentation. The latter was useful to deepen the analysis of the uncertainty map by providing new measures to discover the incorrectly labeled images. The results of this analysis demonstrate the benefits of our approach both qualitatively and quantitatively, with special emphasis on food categories mislabeled as background. Finally, it is worth mentioning that this research work contributes to the generation of healthcare technologies, particularly in the improvement of food monitoring from images, which through human-machine collaboration facilitates the maintenance of a healthy diet and/or the prevention of health problems in those people who must limit monitor their diet due to allergies, intolerance or chronic diseases.

In future research, we plan to develop an algorithm that takes advantage of the proposed measures to select data labeled with noise and manage them during the training procedure minimizing the error incurred from this type of data.

Acknowledgments

This work was partially funded by 20211005001-VRIDT-UCN, TIN2018-095232-B-C21, PID2019-109238GB-C21, Measurer EIT Digital, Logmeal4Shape, and CERCA Programme/Generalitat de Catalunya. B. Nagarajan acknowledges the support of FPI Becas, MICINN, Spain. We acknowledge the valuable feedback of Prof. Ludmila Kuncheva on the research work. We acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPUs.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Acharya, V. Lin, N. Dhingra, The role of health in achieving the sustainable development goals, Bulletin of the World Health Organization 96 (9) (2018) 591.

[2] M. Breque, L. De Nul, A. Petridis, Industry 5.0: towards a sustainable, human-centric and resilient european industry, Luxembourg, LU: European Commission, Directorate-General for Research and Innovation (2021).

[3] P. K. R. Maddikunta, Q.-V. Pham, B. Prabadevi, N. Deepa, K. Dev, T. R. Gadekallu, R. Ruby, M. Liyanage, Industry 5.0: A survey on enabling technologies and potential applications, Journal of Industrial Information Integration 26 (2022) 100257.

[4] S. Aslan, G. Ciocca, R. Schettini, Semantic food segmentation for automatic dietary monitoring, in: IEEE 8th International Conference on Consumer Electronics-Berlin, 2018, pp. 1–6.

[5] M.-L. Chiang, C.-A. Wu, J.-K. Feng, C.-Y. Fang, S.-W. Chen, Food Calorie and Nutrition Analysis System based on Mask R-CNN, in: IEEE 5th International Conference on Computer and Communications, 2019, pp. 1721–1728.

[6] C. N. Freitas, F. R. Cordeiro, V. Macario, MyFood: A Food Segmentation and Classification System to Aid Nutritional Monitoring, in: 33rd SIBGRAPI Conference on Graphics, Patterns and Images, 2020, pp. 234–239.

[7] U. Sharma, B. Artacho, A. Savakis, Gourmetnet: Food segmentation using multi-scale waterfall features with spatial and channel attention, Sensors 21 (22) (2021) 7504.

[8] X. Wu, X. Fu, Y. Liu, E.-P. Lim, S. C. Hoi, Q. Sun, A Large-Scale Benchmark for Food Image Segmentation, in: 29th ACM International Conference on Multimedia, 2021, p. 506–515.

[9] Y. Guo, Y. Liu, T. Georgiou, M. S. Lew, A review of semantic segmentation using deep neural networks, International Journal of Multimedia Information Retrieval 7 (2) (2018) 87–93.

[10] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset, in: CVPR Workshop on the Future of Datasets in Vision, Vol. 2, 2015, pp. 1–4.

[11] Y. Siddiqui, J. Valentin, M. Nießner, Viewal: Active learning with viewpoint entropy for semantic segmentation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9433–9443.

[12] H. Chen, Y. Jin, G. Jin, C. Zhu, E. Chen, Semisupervised semantic segmentation by improving prediction confidence, IEEE Transactions on Neural Networks and Learning Systems (2021).

[13] D. Karimi, H. Dou, S. K. Warfield, A. Gholipour, Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis, Medical Image Analysis 65 (2020) 101759.

[14] C. Northcutt, L. Jiang, I. Chuang, Confident learning: Estimating uncertainty in dataset labels, Journal of Artificial Intelligence Research 70 (2021) 1373–1411.

[15] Z. S. H. Abad, J. Lee, Detecting uncertainty of mortality prediction using confident learning, in: 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, 2021, pp. 1719–1722.

[16] A. Popowicz, K. Radlak, S. Lasota, K. Szczepankiewicz, M. Szczepankiewicz, Combating label noise in image data using multinet flexible confident learning, Applied Sciences 12 (14) (2022) 6842.

[17] M. Zhang, J. Gao, Z. Lyu, W. Zhao, Q. Wang, W. Ding, S. Wang, Z. Li, S. Cui, Characterizing label errors: confident learning for noisy-labeled image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2020, pp. 721–730.

[18] Q. Rao, J. Frtunikj, Deep learning for self-driving cars: Chances and challenges, in: 1st International Workshop on Software Engineering for AI in Autonomous Systems, 2018, pp. 35–38.

[19] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, 2016, pp. 1050–1059.

[20] A. Kendall, V. Badrinarayanan, R. Cipolla, M. Leap, Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding, in: British Machine Vision Conference, 2017, pp. 1–12.

[21] C. Dechesne, P. Lassalle, S. Lefèvre, Bayesian U-Net: Estimating Uncertainty in Semantic Segmentation of Earth Observation Images, Remote Sensing 13 (19) (2021) 3836.

[22] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image Computing and Computer Assisted Intervention, 2015, pp. 234–241.

[23] J. Mukhoti, Y. Gal, Evaluating bayesian deep learning methods for semantic segmentation, arXiv preprint arXiv:1811.12709 (2018).

[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[25] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: European Conference on Computer Vision, 2018, pp. 801–818.

[26] L.-C. Florian, S. H. Adam, Rethinking atrous convolution for semantic image segmentation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 1–14.

[27] G. Ciocca, P. Napoletano, R. Schettini, Food recognition: a new dataset, experiments, and results, IEEE Journal of Biomedical and Health Informatics 21 (3) (2016) 588–598.

[28] T. Ege, W. Shimoda, K. Yanai, A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice, in: 5th International Workshop on Multimedia Assisted Dietary Management, 2019, pp. 82–87.

[29] K. Okamoto, K. Yanai, UEC-FoodPIX Complete: A Large-scale Food Image Segmentation Dataset, in: International Conference on Pattern Recognition, 2021, pp. 647–659.

[30] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, K. P. Murphy, Im2Calories: towards an automated mobile vision food diary, in: IEEE International Conference on Computer Vision, 2015, pp. 1233–1241.

[31] L. Bossard, M. Guillaumin, L. V. Gool, Food-101 -– Mining Discriminative Components with Random Forests, in: European Conference on Computer Vision, 2014, pp. 446–461.

[32] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems 25 (2012).

[33] S. Aslan, G. Ciocca, R. Schettini, Semantic segmentation of food images for automatic dietary monitoring, in: 26th Signal Processing and Communications Applications Conference, 2018, pp. 1–4.

Eduardo Aguilar is a Doctor in Mathematics and Computer Science from the University of Barcelona. He is currently an Assistant Professor in the Department of Computer and Systems Engineering at Universidad Católica del Norte. His main interest is in the research and application of robust Deep Learning algorithms for the semantic analysis of images.

Bhalaji Nagarajan is a PhD student at the University of Barcelona. His PhD thesis centers around sample analysis of deep learning (DL) training data and is focused on DL for food recognition. His research interests include uncertainty quantification, self-supervised learning, and learning from noisy labels.

Beatriz Remeseiro received her Ph.D. degree (2014) in Computer Science from the University of A Coruña. She is currently a tenured Associate Professor at the University of Oviedo and a board member of the Spanish Association for Artificial Intelligence (AEPIA). Her main research interests include computer vision and deep learning, mainly applied to real-world problems.

Petia Radeva is a Full Professor at the Universitat de Barcelona (UB), PI of the Consolidated Research Group "Computer Vision and Machine Learning at UB", IAPR Fellow, and Senior researcher in the Computer Vision Center. Associate editor of Pattern Recognition and Journal of Visual Communication and Image Representation, PI of European and national projects on Computer Vision food intake monitoring.