UEC-FoodPix Complete: A Large-scale Food Image Segmentation Dataset

Kaimu Okamoto and Keiji Yanai

The University of Electro-Communications, Tokyo
1-5-1 Chofugaoka, Chofu-shi, Tokyo, 182-8585 JAPAN
{okamoto-ka,yanai}@mm.inf.uec.ac.jp

Fig. 1. UECFood-100 images overlaid with segmentation masks annotated in "UEC-FoodPix Complete."

Abstract. Currently, many segmentation image datasets are open to the public. However, only a few open segmentation image dataset of food images exists. Among them, UEC-FoodPix is a large-scale food image segmentation dataset which consists of 10,000 food images with segmentation masks. However, it contains some incomplete mask images, because most of the segmentation masks were generated automatically based on the bounding boxes. To enable accurate food segmentation, complete segmentation masks are required for training. Therefore, in this work, we created "UEC-FoodPix Complete" by refining the 9,000 segmentation masks by hand which were automatically generated in the previous UEC-FoodPix. As a result, the segmentation performance was much improved compared to the segmentation model trained with the original UEC-FoodPix. In addition, as applications of the new food segmentation dataset, we performed food calorie estimation using the food segmentation models trained with "UEC-FoodPix Complete", and food image synthesis from segmentation masks.

Keywords: semantic segmentation, food image, calorie estimation, food image synthesis

1 Introduction

Nowadays the accuracy of image recognition has dramatically improved due to the development of deep learning, and excellent results have been achieved in various tasks such as image generation and semantic region segmentation. In supervised semantic segmentation by deep learning, a large-scale mask image dataset annotated for each pixel is required for training of segmentation models. PASCAL VOC 2012 [9] and MS COCO [13] are widely used as large-scale segmentation datasets, in which the annotated objects are generic objects such as animals and vehicles. MS COCO includes only a limited number of food categories.

Although there exists many food image datasets such as Food-101 [3] and VIREOFood-172 [4], most of them have only food category labels on each of the images. A few datasets have bounding box annotation or segmentation mask annotation. For example, UECFood-100 [15] has bounding box annotation for each of all the dishes. As a dataset annotated on segmentation masks, UEC-FoodPix [8] has been created by adding segmentation mask annotation to 10,000 images of UECFood-100 by Ege *et al.* However, UEC-FoodPix has the problem that it contains incomplete segmenation masks because they are semi-automatically annotated by GrabCut [20] based on the bounding boxes annotated in the UECFood-100 dataset. To enable accurate food segmentation, complete segmentation masks are desirable for training. Therefore, in this paper, we have updated UEC-FoodPix by manually modifying the incomplete segmentation masks. We call the updated food image segmentation dataset as "UEC-FoodPix Complete" (Figure 1).

By using the new food segmentation dataset, "UEC-FoodPix Complete" for training of the state-of-the-art semantic segmentation model, DeepLabV3+ [5], the segmentation performance was improved by 0.14 mIoU compared to the segmentation model trained with the original UEC-FoodPix. In addition, as applications of the new food segmentation dataset, we performed region-based food calorie estimation using the food segmentation models trained with "UEC-FoodPix Complete", and food image synthesis from segmentation masks employing SPADE [18].

2 Related Work

As bechmark datasets for semantic segmentation, PASCAL VOC 2012 [9] and MS COCO [13] are commonly used. PASCAL VOC is the dataset used in the competition held from 2005 to 2012, and the 2012 edition includes 9,993 images of 22 classes including airplanes and bicycles. MS COCO is a dataset provided by Microsoft and includes 330,000 images in 80 classes. Only 10 food classes such

as pizza and hot dogs are included in the COCO category. Therefore, it is not suitable as a training dataset for food image segmentation models.

Currently, there are a few large-scale open food image datasets with segmentation masks. The UNIMIB2016 dataset [7] provides food region information as polygons which are equivalent to segmentation masks. However, its scale is not so large (1027 multiple-dish images with 73 food categories), and the food images in UNIMIB2016 are biased and not unconstrained since all the food images were taken at the same canteen.

Lu *et al.* [14] proposed a food volume estimation method by extending Mask R-CNN [11] which extracts food regions from a given RGB-D image. To train the proposed model, they used the MADiMa17 dataset [1] which consists of 21 food categories with segmentation masks. However, all the images in the MADiMa17 dataset were taken in the laboratory environment which was different from uncontrolled real situations.

Okamoto *et al.* proposed a region-based food calorie estimation system running on a mobile phone [17]. They employed food image segmentation and estimated food calories based on the size of the reference card and food regions. However, at that time, no food image segmentation dataset on uncontrolled food images is available. Instead of the segmentation method which requires training data, they used GrabCut [20], which was a hand-crafted segmentation method that divides the foreground and background by graph-based reasoning.

To change this situation, Ege *et al.* created a large-scale food image segmentation dataset, which was called "the UEC-FoodPix dataset" [8]. They added pixel-wise annotation to 10,000 food images included in the UECFood-100 dataset [15]. Regarding 1,000 food images for testing, they added pixel-wise labels by hand, while for the other 9,000 images they created pixel-wise labels automatically by applying GrabCut [20] on each of the bounding boxes originally annotated in the UECFood-100 dataset. Before applying GrabCut, they verified if the bounding box annotations were enough correct one by one, and revised them if needed. In addition, Ege *et al.* [8] proposed a method to estimate actual size of foods without a reference card for estimating food calorie amounts of uncontrolled food images. To do that, they proposed to estimate actual size of foods in the image by using the size of rice grains as reference objects. Although this methods can be applied to food images containing rice, we can estimate real size of foods and their calories by combining food region segmentation without a reference card with this method.

However, since the UECFoodPix created by Ege *et al.* [8] generated pixel-wise annotations semi-automatically, it may contain noisy annotations, which is expected to be harmful for training of CNN models. Therefore, in this paper, we improve the UECFoodPix dataset for more accurate food image segmentation.

As the other dataset for unconstrained food images, Google Food-201 [16], SUEC Food Dataset [10], and Food segmentation benchmark [2] have been released so far. Food-201 [16] was created for the Im2Calories project by Google, and released to the public several years after the paper was published. They annotated 201 pixel-level labels to parts of the images in the ETH Food-101

Table 1. A list of the public food segmentation datasets on unconstrained food images.

Dataset name	release	#image	#class	annotation	Original dataset
Google Food-201 [16]	2017	12,093	208	crowdworker	ETH Food101 [3]
SUEC Food Dataset [10]	2019	28,897	256	auto (GrabCut)	UEC-Food256 [12]
Food segmentation benchmark [2]	2020	5,000	50	controlled	Chinese food 50 categories [6]
UEC-FoodPix [8]	2020	10,000	102	auto (GrabCut)	UEC-Food100 [15]
UEC-FoodPix Complete (this paper)	2021	10,000	102	controlled	UEC-Food100 [15]

dataset [3] with the help of the crowd-sourcing workers. SUEC Food Dataset [10] was created by GrabCut [20] based on the bounding box annotation of UEC-Food256 [12]. The dataset for segmentation benchmark created by the UMINIB group [2] contains 5,000 segmentation masks for all the images of 50-category Chinese food image dataset [6]. We listed the current public food segmentation datasets on unconstrained food images in Table 1.

3 Dataset Construction

Currently, many meal image datasets are open to the public, such as Food-101 [3], VIREO Food-172 [4], UECFood-100 [15] and UECFood-256 [12], for food image classification. They are commonly used as standard benchmark datasets. Only a few, such as UECFood-100/256, have bounding box annotations on food regions in all the images.

As a large-scale food image dataset with a segmentation mask, UEC-FoodPix [8] created by Ege *et al.* exists. However, the UEC-FoodPix dataset contains some incomplete segmentation masks on the boundaries of the food regions, since they were generated automatically from the bounding box annotations. Therefore, in this study, we created "**UECFoodPix Complete**" as a higher quality dietary image segmentation dataset by updating UECFoodPix manually.

We used the Web-based pixel-wise annotation tool implemented by Pongsate *et al.* [22]. This tool allows easy synthesis and separation of food regions with super-pixels. To make the annotation higher quality and more reliable, we did not used crowd-sourcing, instead he shared the jobs among the lab members and hired a limited number of bachelor students. To keep annotation quality, we set annotation rules on how to create food region masks for each of the food categories. After working with several people, the first author himself made the final confirmation of all of the food region masks of 10,000 images to keep annotation consistency in the dataset. The working period took about 4 months.

We show some examples modification on food region masks between UEC-FoodPix and UEC-FoodPix Complete in Figure 2. In the first row of the figure, in UEC-FoodPix (show in the second column) the region of "salad" was annotated as a "pork cutlet" region. We divides them into "salad" and "pork cutlet" regions in UEC-FoodPix Complete (showin in the third column). In the second rows, the"corn soup" region was incorrect in UEC-FoodPix, since its boundary is not circular. We revised it in UEC-FoodPix Complete.

Fig. 2. The differences on segmentation masks in the both datasets. (1st column: food images image, 2nd column: the corresponding segmentation masks in the orginal UEC-FoodPix, 3rd column: the corresponding segmentation masks in the renewed UEC-FoodPix Complete.)

4 Evaluation

The updated food image segmentation dataset is entitled "UEC-FoodPix Complete," and consists of 9000 training images and 1000 validation/testing images. We trained the state-of-the-art semantic segmentation method, DeepLab V3+ [5], to compare the model performance trained with the original UEC-FoodPix and the updated UEC-FoodPix Complete. In addition, we trained the DeepLab V3+ model with 2,000 UEC-FoodPix Complete images and 7,000 UEC-FoodPix images together as well. We evaluated both trained models with accuracy and mean Intersection over Union (mIoU).

The experimental results are shown in Table 2. The evaluation scores were improved by about 0.1 in Accuracy and about 0.14 in mIoU. In case of using 2,000 Complete images, the improvement was very limited. In fact, it was one of the reasons why we decided to update annotation on all the images. Figure 3 shows some segmentation results by both the UEC-FoodPix model. The results by the Complete model (shown in the fourth column) are similar to the groundtruth (shown in the second column), while the results by the original UEC-FoodPix were apparently irrelevant except for the first row.

Table 2. The accuracy and mean IoU scores on the three food segmentation models.

training dataset	Acc	mIoU
UEC-FoodPix (all automatic)	0.560	0.416
partial UEC-FoodPix Complete (2000 hand annotation)	0.597	0.436
UEC-FoodPix Complete (all 9000 hand annotation)	0.668	0.555

5 Application 1: Region-based Food Calorie Estimation

In this section, as one of the applications of UEC-FoodPix Complete, we explain the results applying region-based food calorie estimation method using rice grain as reference objects proposed by Ege *et al.* [8], and compare the results between the cases of using both segmentation models.

5.1 Method

As a utilization of the created dataset, food calorie amounts are estimated from the estimated food regions. In order to estimate the calorie amounts in consideration of the area of dishes, calorie values are estimated from the regression equation after estimating the dish category, extracting the food region, and estimating the actual size of the food region based on rice grains. The procedure is as follows, following Ege *et al.* [8].

1. Detect food bounding box using Faster-RCNN [19] from an input image.

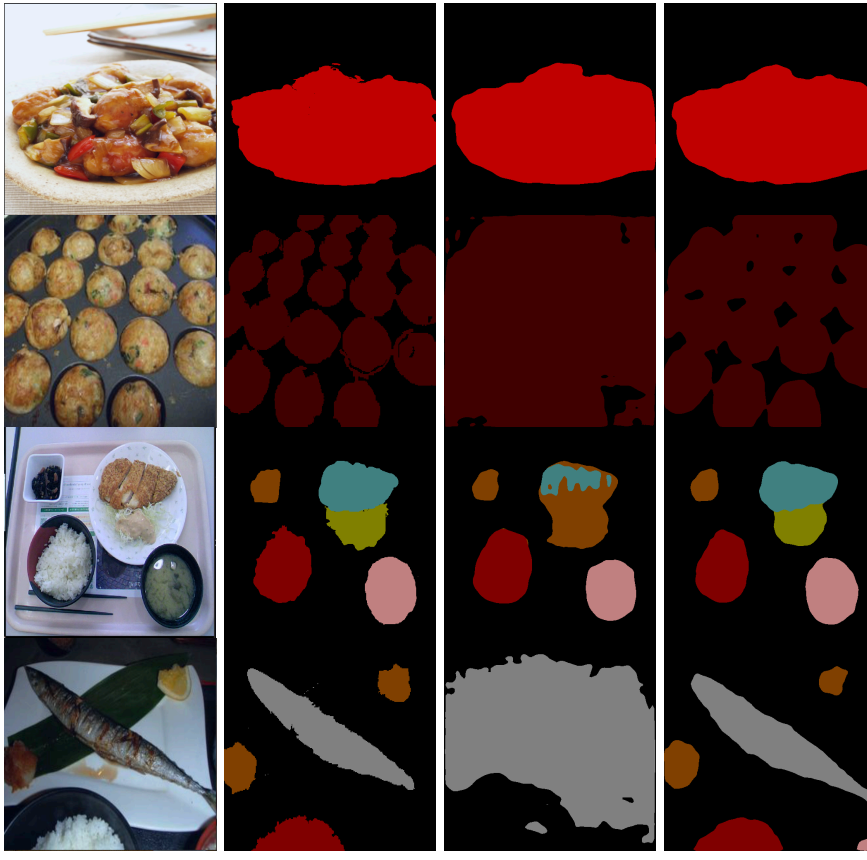

Fig. 3. Segmentation results by Deeplab V3+. (1st column: input images, 2nd column: groundtruth region masks, 3rd column: estimated masks by the model trained with "UECFoodPix," 4th column: results by the model trained with "UEC-FoodPix-COMPLETE.")

2. Estimate food regions by using Deeplab V3+ [5] in the bounding box.
3. Estimate the area of each of the detected food items by using the rice grain based system from the estimated food region of the cooked rice portion.
4. Estimate the calorie by using the calorie / area-regression formula created in advance for the estimated region.

Table 3. Evaluation on region-based food calorie estimation.

	area(rice)		area(multiple food)	
dataset	abs.err(cm^2)	rel.err(%)	abs.err(cm^2)	rel.err(%)
UECFoodPix [8]	7.21	8.73	**30.0**	**14.2**
COMPLETE	**3.03**	**3.67**	44.7	20.7

5.2 Experimental Results

First, a comparison was made up to area estimation. 51 images containing both actual rice and a reference card were used in the experiments. Since the actual size of the reference car is known, the actual size of the food region was annotated.

The evaluation was performed by both absolute and relative errors, and was performed only on images in which food was correctly detected. The experimental results are shown in Table 3 and in Figure 4. As a result, when comparing the areas using only rice, the accuracy was improved by $4.18cm^2$ in absolute error and 5.06% in relative error. In addition, from Figure 4, it was found that when UECFoodPix was used, accurate area estimation was not possible for some testing images, since the model failed to detect the region of streamed rice correctly. However, when comparing multiple dishes, both absolute and relative errors of the original UEC-FoodPix were lower. This is because the actual size estimation part has a greater influence on the area size estimation than the region estimation part. When the cooked rice mask estimated by Deeplab V3+[5] trained by UECFoodPix was used for the actual size estimation, it tended to estimate larger regions than actual food regions. Therefore, even if the other meal area than rice is somewhat lacking, a value close to the correct answer value is calculated. Therefore, it is necessary to improve the accuracy by improving the actual size estimation part.

Next, we estimated the calorie amounts from the size of food items and calorie density per unit size in the same way as Ege *et al.* [8]. The images were similarly performed using a reference card, and were estimated with the food regions estimated by the segmentation model trained with UEC-FoodPix Complete. The result is shown in Figure 5. Since the values close to the actual calories were calculated, it was shown that it can be fully utilized for application to calorie content estimation.

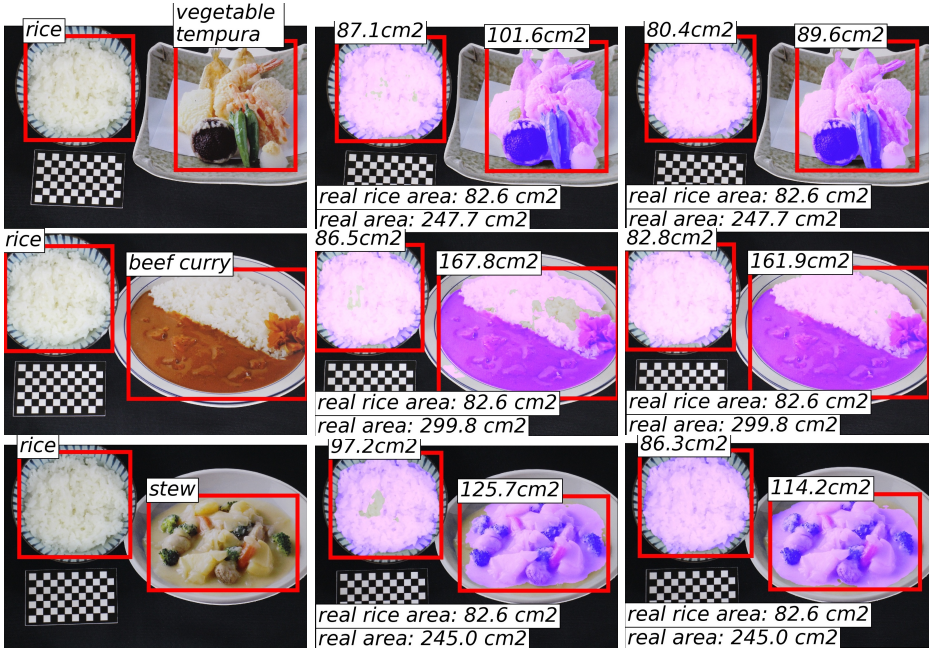

Fig. 4. Estimated results on the area size. (left: input images with food category labels, center: estimated results by "UECFoodPix," right: estimated results by "UEC-FoodPix Complete.")

6 Application 2: Mask-based Image Synthesis

As the second application of "UEC-FoodPix Complete," we performed mask-based image synthesis employing the state-of-the-art mask-based image generation method, SPADE [18]. SPADE is a GAN-based model for generating images from mask images, which spatially adapts the scaling and bias terms based on the given mask to reflect the semantic information effectively. Training of the SPADE model was performed using 9000 trained images and 1000 validation/test images.

Figure 6 shows the synthesized images in which the first and third rows represent input region mask images and the second and fourth rows represent the corresponding generated images. Rice bowls, noodles, and Japanese combo meals which look realistic were successfully generated. Interestingly, dish plates were generated around food regions naturally, although no region masks on plates were given. That is why distorted plates or bowls were sometimes generated like the images on 4th rows and 3th and 4th columns. In the case on 4th rows and 3rd column which was a tempura rice bowl, the generated bowl were deformed along the shrimp tempura, although the shrimp tempura is expected to stick out of the bowl. To solve this problem, pixel-wise plate region annotation is needed.

Fig. 5. The estimated result of food calorie amounts. (All the results were estimated from the segmentation masks produced by the segmentation model trained with "UEC-FoodPix Complete.")

In the case on 4th rows and 4th column which represents bread rolls, two pieces of bread rolls were concatenated, since input masks cannot represents multiple different instances of the same food category.

In the next experiments, we synthesized food images from the same mask images with different class labels. Figure 7 shows the results. In the second row, we generated a beef rice bowl, a chicken rice bowl, a chilled ramen noodle, and a ramen noodle from the same mask input. In the third row, we generated four multiple dish food images by changing only one dish on the left-bottom with the same food categories as the second row. In the fifth row, we generated multiple dish images as well by changing two dishes at the same time on the dishes on the left-bottom and middle-top. We confirmed that mask-based food image synthesis worked well even for multiple dish food images, although the shape of plates and bowls sometime were distorted and look unnatural.

Overall, we found that food image synthesis from region masks was possible. However, since our dataset, "UEC-FoodPix Complete" has no pixel-wise annotation on plate regions, we cannot control the shape of plates which are usually generated around food regions. Therefore, distorted plates tend to be generated, which made the synthesized images look unnatural. For future work, we think we need to add pixel-wise plate region annotation to the dataset for more natural food image synthesis. We plan to obtain plate annotation employing the unsupervised plate region estimation method [21] for the time being.

7 Conclusions

In this study, we created a new dataset, "UEC-FoodPix Complete," by updating the existing the food image segmentation dataset, "UEC-FoodPix [8]." We evaluated improvement on food semantic segmentation with the state-of-the-art segmentation model, DeepLabV3+ [5]. As a result of segmentation using the updated dataset, mean IoU was improved by 0.14 compared to the original dataset. In addition, as applications of the new food segmentation dataset, we performed region-based food calorie estimation using the food segmentation

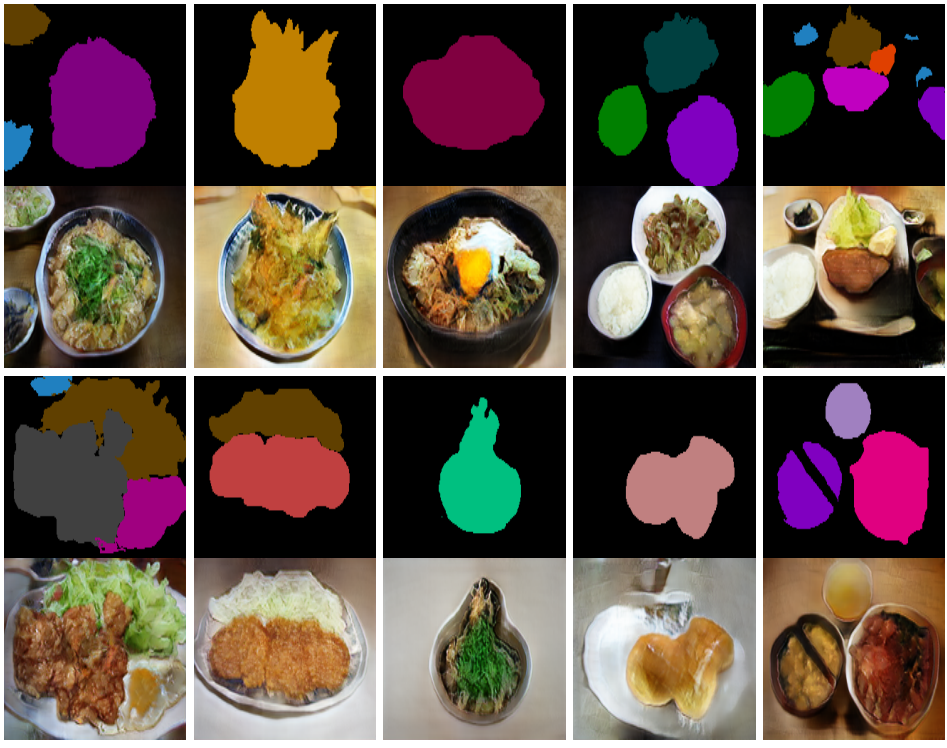

Fig. 6. The results of mask-based food image synthesis. (1st and 3rd rows：mask images, 2nd and 3rd rows: generated images by SPADE)

models trained with "UEC-FoodPix Complete" and food image synthesis from segmentation masks employing SPADE [18].

In fact, our motivation on updating the UEC-FoodPix dataset is to estimate food calories more accurately. Therefore, we like to improve region-based food calorie estimation. Our future works include improving the accuracy of actual size estimation using cooked rice, improving the regression equation to a more robust estimation, and dealing with images that do not contain cooked rice. In addition, as further development, we plan to introduce CNN-based architecture for the part of the calorie amount estimation as well.

The "UEC-FoodPix Complete" dataset can be downloaded from the following URL: http://mm.cs.uec.ac.jp/uecfoodpix/.

Acknowledgments We would like to thank all those who worked on pixel-wise annotation for creating "UEC-FoodPix Complete." This work was supported by JSPS KAKENHI Grant Number 17J10261, 15H05915, 17H01745, and 19H04929.

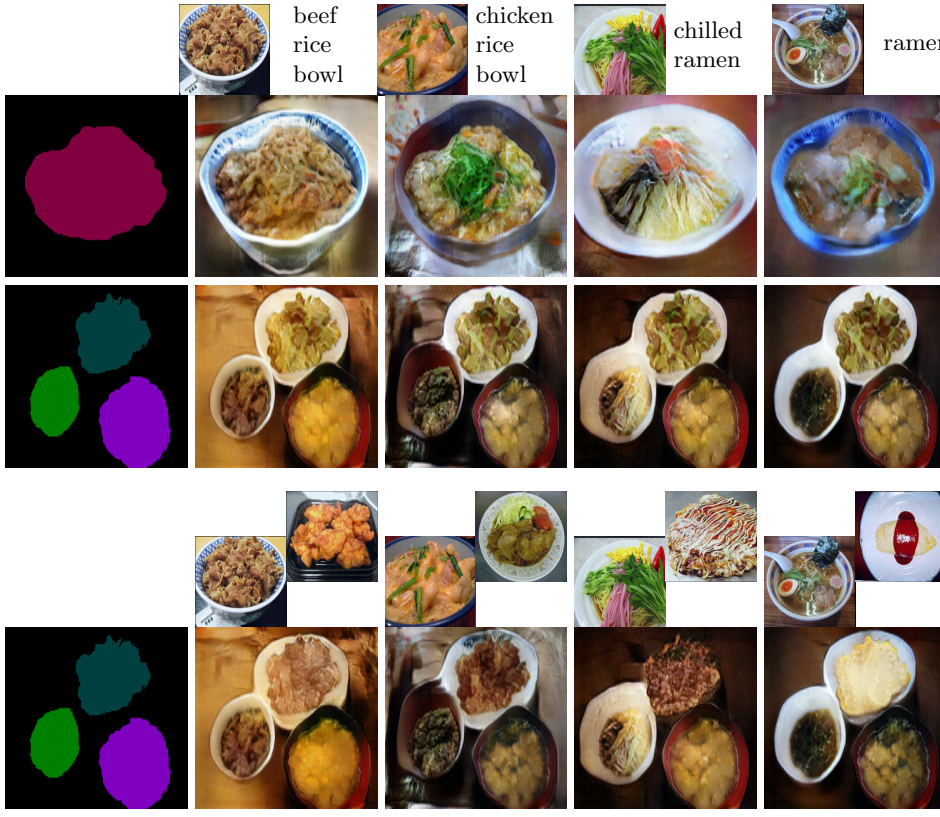

Fig. 7. Food image synthesis from the same region masks with different food category labels. (1st and 4th rows: class label images. 2nd row: single dish images. 3rd and 5th rows: multiple dish images.)

References

1. Allegra, D., Anthimopoulos, M., Dehais, J., Lu, Y., Stanco, F., Farinella, G.M., Mougiakakou, S.: A multimedia database for automatic meal assessment systems. In: Proc. of the ICIAP Workshop on Multimedia Assisted Dietary Management (2017)
2. Aslan, S., Ciocca, G., Mazzini, D., Schettini, R.: Benchmarking algorithms for food localization and semantic segmentation. International Journal of Machine Learning and Cybernetics **11**, 2827–2847 (2020)
3. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: Proc. of European Conference on Computer Vision (2014)
4. Chen, J., Ngo, C.W.: Deep-based ingredient recognition for cooking recipe retrival. In: Proc. of ACM International Conference Multimedia (2016)
5. Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc. of European Conference on Computer Vision (2018)

6. Chen, M.Y., Yang, Y.H., Ho, C.J., Wang, S.H., Liu, S.M., Chang, E., Yeh, C.H., Ouhyoung, M.: Automatic chinese food identification and quantity estimation. In: Proc. of SIGGRAPH Asia (2012)
7. Ciocca, G., Napoletano, P., Schettini, R.: Food recognition: a new dataset, experiments and results. IEEE Journal of Biomedical and Health Informatics **21**(3), 588–598 (2017)
8. Ege, T., Yanai, K.: A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice. In: Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management (2019)
9. Everingham, M., Eslami, S., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **88**(2) (2010)
10. Gao, J., Tan, W., Ma, L., Wang, Y., Tang, W.: MUSEFood: Multi-sensor-based food volume estimation on smartphones. arXiv:1903.07437 (2019)
11. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: Proc. of IEEE International Conference on Computer Vision (2017)
12. Kawano, Y., Yanai, K.: Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In: Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV) (2014)
13. Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.: Microsoft coco: Common objects in context. In: Proc. of European Conference on Computer Vision (2014)
14. Lu, Y., Allegra, D., Anthimopoulos, M., Stanco, F., Farinella, G.M., Mougiakakou, S.: A multi-task learning approach for meal assessment. In: Proc. of the IJCAI Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management. p. 46–52 (2018)
15. Matsuda, Y., Hajime, H., Yanai, K.: Recognition of multiple-food images by detecting candidate regions. In: Proc. of IEEE International Conference on Multimedia and Expo. pp. 25–30 (2012)
16. Myers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., Murphy, P.K.: Im2Calories: towards an automated mobile vision food diary. In: Proc. of IEEE International Conference on Computer Vision. pp. 1233–1241 (2015)
17. Okamoto, K., Yanai, K.: An automatic calorie estimation system of food images on a smartphone. In: Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management (2016)
18. Park, T., Liu, M., Zhu, J.: Semantic image synthesis with spatially-adaptive normalization. In: Proc. of IEEE Computer Vision and Pattern Recognition (2019)
19. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN:towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2017)
20. Rother, C., Kolmogorov, V., Blake, A.: "GrabCut": Interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. **23**(3), 309–314 (2004)
21. Shimoda, W., Yanai, K.: Predicting plate regions for weakly-supervised food image segmentation. In: Proc. of IEEE International Conference on Multimedia and Expo (2020)
22. Tangseng, P., Wu, Z., Yamaguchi, K.: Looking at outfit to parse clothing. arXiv:1703.01386 (2017)