[Demo] SetMealAsYouLike: Sketch-based Set Meal Image Synthesis with Plate Annotations

Yuma Honbu Keiji Yanai
The University of Electro-Communications, Tokyo, Japan
{honbu-y,yanai}@mm.inf.uec.ac.jp

ABSTRACT

By using semantic segmentation dataset with pixel-wise annotation for training GANs, image generation from a given mask image drawn by a user is possible. However, regarding mask-based food image synthesis, the existing food segmentation datasets have only food region masks and no plate region masks. When we train a mask-based image synthesis network with the datasets without plate mask annotation, the plate regions in the generated food images are uncontrollable by a user and tend to be distorted. To solve this problem, we use a Few-shot segmentation method to estimate the plate regions of the image in the existing food segmentation dataset using a limited number of plate region annotations, and add dish region masks to it. By using added plate masks as training data, we enable generating food images under the control of the shape of the plates. We have implemented the interactive food image drawing system in which we draw food masks as well as plate masks. In the demo, we demonstrate that we generate natural set meal images which include multiple dishes by the sketch interface easily.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Human-centered computing** → **Interactive systems and tools**.

KEYWORDS

food image synthesis, mask-based image generation, food region segmentation dataset

ACM Reference Format:
Yuma Honbu Keiji Yanai. 2022. [Demo] SetMealAsYouLike: Sketch-based Set Meal Image Synthesis with Plate Annotations. In *Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management (MADiMa '22), October 10, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3552484.3555749

1 INTRODUCTION

It is believed that the appearance of food and appetite are closely related. Therefore, the arrangement and style of plates and dishes are important when preparing set meal, and the field of image generation using deep learning has attracted a lot of attention as various image generation models have been published one after another, including GAN (Generative Adversarial Networks) published in 2014. In particular, by generating food images from segmentation mask images, it is possible to generate food images with the shape and food category specified by the user. However, in food image generation from mask images, it is common that the food segmentation dataset is not annotated with plate regions, which often causes problems such as distortion of the generated plate regions or merging of plates. In this study, we address this problem by using the Few-shot segmentation method to extract the plate region using a small number of plate annotations. Furthermore, we create a new food annotated dataset by adding estimated plate regions to the existing food annotated dataset. This plate annotation allows the style and shape of the plate and food regions to be generated independently of each other, resulting in high-quality generation. In our demonstration system, we show how SEAN [18], a typical GAN for mask-based image generation, can be used to generate high-quality set meal images based on free-sketch images and edit various styles.

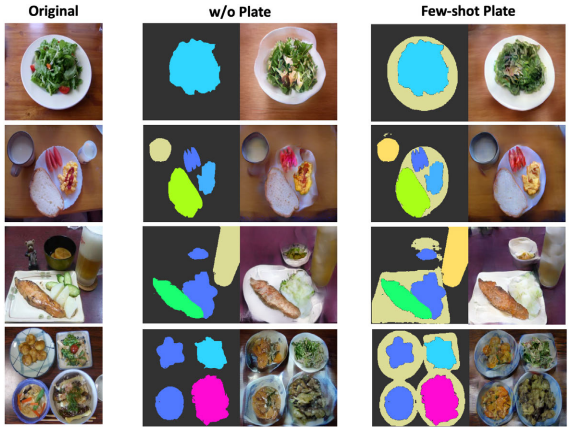

Figure 1: Example results of food image generation with/without plate masks.

2 RELATED WORKS

2.1 Food image generation

With the advent of GAN, published in 2014, various image generation and transformation studies have been conducted, resulting in significant improvements in the quality of generated images. In the study by Ito et al. [6] focused on the distortion of the generated dish region, which has been a problem in food image generation, and proposed a method that extends cGAN. In the proposed method, a discriminator that identifies whether dishes are round or distorted

MADiMa '22, October 10, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9502-1/22/10…$15.00
https://doi.org/10.1145/3552484.3555749

is added to cGAN. This was found to reduce the percentage of images with distorted plates after generation and improve the quality of the generated images.

Horita et al. [5] proposed a method called conditional CycleGAN (cCycleGAN), which transformed a food image into multiple food categories by providing a class-conditional vector to CycleGAN. cCycleGAN enables high-quality food image transformation while maintaining shape among 10 different food categories.

In addition, Horita et al. [4] have proposed a conditional StyleGAN to control the stochastic fluctuations of the separated style features of StyleGAN. The conditional style generator allows the user to manipulate the style of an arbitrary region by providing a conditional vector. By applying this conditional StyleGAN to a food image region, it is possible to generate food images that contain conditions specifying multiple classes. PizzaGAN [11] reflects the step-by-step pizza creation procedure in the model and achieves pizza ingredient editing by learning the presence or absence of each ingredient using CycleGAN.

SPADE [12] used semantic masks in the normalization layer to enable normalization in the spatial direction, resulting in realistic image generation.Ketchup GAN [1] demonstrated that ketchup letters can be drawn on omelets by training SPADE on a dataset of omelet images collected from Twitter. Nyamdorj's [8] study used SPADE and suggested that the presence or absence of a plate region mask affects the quality of food image generation. Shimoda et al. [16] found that plate distortion and plate-to-plate merging problems were mitigated by using a plate region mask generated with the

In SEAN [18], in addition to the SPADE layer, category feature of each segmentation mask region is convoluted to make it easier to transfer the features of each par. This network architecture can be created that controls the style of each semantic domain separately. At the same time as SEAN, RamenStyleAsYouLike [2] was proposed to add styles to SPADE using segmentation masks. The difference between the two methods is whether the mean and variance of the normalization layer is divided between the style feature part and the segmentation image part. RamenStyleAsYouLike is implemented as a web-based system, creating an application that generates the specified ramen when the user draws a mask of favorite ramen and selects a style.

2.2 Few-shot segmentation

Few-shot Segmentation aims to learn category and object identification knowledge for unknown classes using a small number of identical domain data and knowledge previously learned for existing classes when there is little teacher data for unknown categories that are not present in the training data.

In this task, there are no commonalities between the categories at the time of training and at the time of validation. Therefore, the input for validation consists of a query image of an unknown category, several images of the same category and its GT mask image as a support set. The query image is segmented using the support set as the minority data. There is also a task called Zero-shot Segmentation, which uses word-embedded features and similarity between pre-trained features to segment unknown categories into regions.

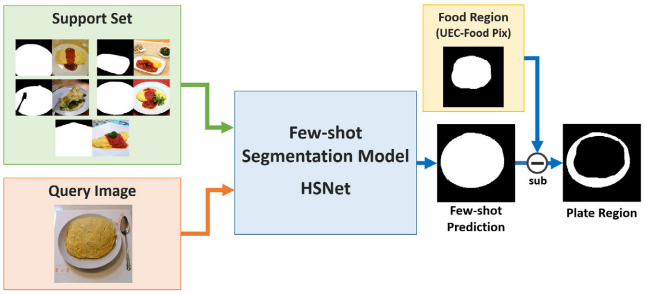

Figure 2: Network architecture of the plate region extraction method using HSNet

Many Few-shot Segmentation methods use deep metric learning techniques. In general, a prototype vector is extracted by Masked Average Pooling (MAP), which masks and averages features in the target region using a support set of masks. In many cases, the similarity between the prototype vector and the query features is calculated, or the common parts with the query are estimated by concatenating and convolving the prototype vectors.

HSNet [7], which achieved the best performance in Few-shot Segmentation, proposes a method to extract features for query and support images using a backbone pre-trained by ImageNet, masking them to support features, and then convolving the 4D tensor by computing pixel-level cosine similarity between all block features. Since the features used in this model deal with support and query similarity, it addresses the problems of Few-shot, such as domain differences and biased inference toward the training class. This allows inference to be independent of class and domain, making it possible to estimate segmentation masks with high accuracy.

In our demonstration system, we used the plate regions extracted by HSNet in the training data and added to the UEC-FoodPix Complete [9] dataset to generate set meal image based on user-drawn sketches using SEAN.

2.3 Plate annotation

Shimoda et al. [16] proposed a weakly supervised food and plate region detection method using CAM [17] to segment plate regions by learning a plate region model using the difference between the visualized regions of the food/non-food image classifier and those of the food classifier as pseudo plate regions. The proposed method uses the SSDD [15] module, which takes two segmentation masks as input and outputs a single integrated mask.

3 METHOD

The procedure of builiding the demo system is summarized as follows:

(1) Annotate the plate regions of the five images for each of the 100 category in the food segmentation dataset, UEC-FoodPix Complete [9].
(2) Apply the few-shot segmentation method, HSNet [7] with 5-shot annotated images to estimate plate regions of the other images than the five annotated image for each food category.

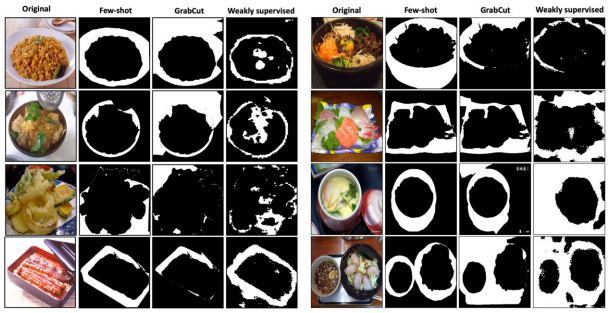

Figure 3: Estimated plate region masks (shown by the white regions in the figures) by three methods. left:HSNet [7], center:GrabCut [13],right:Weakly [16]

Figure 4: Plate area style conversion results.

(3) Train the SPADE [12] and SEAN [18] models with the UEC-FoodPix Complete augmented with the plate region annotation.
(4) Implement an Web-based interactive system for sketch-based meal image synthesis.

3.1 Plate region extraction by HSNet

In the research of Nyamdorj [8], they found that the plate region played a significant role in the quality of mask-based food image generation, since they obtained the synthesized food images with distorted plates using the model trained with a food region dataset with no plate annotation. Therefore we realized we need to add plate region masks to the existing food image dataset, although pixel-wise annotation is very costly. To make annotation cost the smallest, we use a few-shot segmentation method, HS-Net [7], to add plate region masks to the UEC-FoodPix Complete dataset. The network diagram of HS-Net is shown in Figure 2, where HSNet uses the model trained on the Pascal-5^i dataset [14] and the average of each region estimated by the model trained on the four splits. We also inferred plate regions in a 5-shot setting with 5 support sets. The GT mask images included in the support set used here were manually annotated with five images of each class for the plate and food regions. Since the food and plate regions are inferred during inference, the plate region was added to the dataset as the region from which the food region was removed from the estimated plate region mask image.

3.2 Implementation of mask-based image generator

The SEAN model was trained by adding an FC layer that generates variance and mean with a normalization layer for the number of categories in UEC-FoodPix Complete with plate regions. For comparison, we also trained a plate region generated using bounding boxes and GrabCut [13], and SEAN [18] without a plate region, and compared the quality of the generated images with Fréchet Inception Distance (FID) [3]. FID is a metric to measure the feature distance between an image and the generated image.

4 EXPERIMENTS

4.1 Dataset

The UEC-FoodPix Complete [10] created by Okamoto et al. was used as a food segmentation dataset, which had pixel-wise food region masks for 102 food classes. The number of the total images was about 10,000. To estimate plate masks using HSNet, five images of each class of the UEC-FoodPix Complete were manually annotated with plate region masks. We used them as a support set.

For the training of SEAN [18] and SPADE [12], we used the 100 food categories plus the others and the beverage categories, 9,000 images for training and 1,000 images for validation and testing. In addition, we prepared 200 manually annotated plate region masks for quantitative evaluation of estimated plate region masks.

4.2 Evaluation of the generated plate region masks

For evaluation of plate mask generation, we have prepared two baselines, GrabCut [13] and the weakly-supervised plate estimation method [16] in addition to the few-shot method. With GrabCut we estimated the plate masks based on the bounding boxes attached to UEC-FoodPix, while with the weakly-supervised method we estimated the plate masks without any additional annotation. In all the methods except the weakly method, to obtain plate masks, we subtracted the meal regions given in the UEC-FoodPix from the estimated dish regions which contain both food and plate regions.

Table 1 shows the quantitative evaluation of estimated plate region masks with mean Intersection over Union (mIoU), while Figure 3 shows some estimated plate masks by the three methods. These shows the plate masks estimates by the few-shot methods are much superior to the baselines.

Regarding the plate region extraction method of Shimoda et al. [16], the resolution of the mask is low because it is estimated using CAM, and the reliability of the details of the plate region tends to be low. In the method using GrabCut, a bounding box is used to extract the food region and the plate region, and only the food region is extracted using the original data set. Therefore, if the estimation is incorrect, extra regions contained in the bounding box is also sometimes extracted, resulting in an unnatural shape for the plate region. On the other hand, in the plate region extraction method using HSNet, similar regions are extracted by simply

Table 1: Quantitative evaluation of plate area extraction methods by mIoU.

	Few-shot [7]	GrabCut [13]	Weakly [16]
mIoU	78.5	66.1	50.2

Table 2: Quantitative evaluation of image generation by different plate region methods (FID ⇓).

	Few-shot [7]	GrabCut [13]	No Plate
SPADE	68.44	77.33	71.54
SEAN	50.86	51.59	51.10

specifying several food and plate regions in the support set, and as a result, the plate regions are extracted in a natural way.

4.3 Evaluation of generated images

Table 2 shows the FID scores which evaluate the quality of the food images generated from the GT food masks and the estimated plate masks. We used two mask-based GAN methods, SPADE and SEAN. For SEAN, the style feature for each region class is needed. We extracted style features based on the masks of the testing images. In addition to estimated plate maskes, we made experiments with the case with no plate masks and the other case with the plate masks estimated by GrabCut [13] based on the bounding boxes attached to UEC-FoodPix.

In generating the test images with SEAN and SPADE, the experimental results show that the generation quality was better when plate masks were used. Furthermore, we found that the models trained using Few-shot plate regions had better generation quality. Thus, it was found that using a high-quality plate region dataset for training data improves the quality of food image generation, indicating that plate regions have an impact on food image generation quality. It was also found that using SEAN rather than SPADE resulted in better quality of generated images and smaller FID.

On the other hand, the plate region generated by GrabCut resulted in a lower quality of generation than the model without the plate. In the plate region generated by GrabCut, the bounding box may cause a large outlier from the actual plate region. Thus, it was found that learning with incorrect plate regions degrades the quality of food image generation.

Figure 1 in the first paper shows generated results in case of the SEAN model trained with/without plate masks. In case of no plate masks, although the plate regions are generated even without plate masks in the given sketches, the generated plates tends to be distorted along the boundary of food regions, and no circle-shaped or square-shaped plates are generated.

5 DEMO SYSTEM

We have implemented an Web-based interactive demo system the name of which is "SetMealAsYouLike". The Web user interface of the system is shown in Figure 5. It is designed for the usage on both tables and PCs. The system can be operated via GUI, allowing users to draw masks interactively using a apple pencil or a PC mouse.

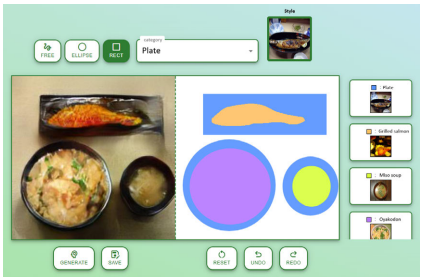

Figure 5: The user interface of "SetMealAsYouLike."

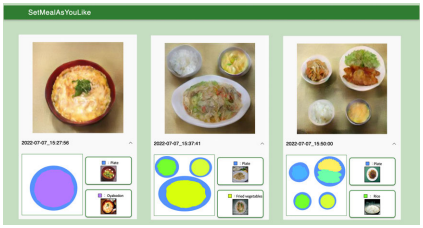

Figure 6: The collection page to show generated set meals in "SetMealAsYouLike".

To operate the system, users select the category they wish to draw among the food categories of the UEC-FoodPix Complete. Then, they select a pen and draw plate and food masks on the right side of the canvas, and a meal image is generated instantly. The three buttons in the lower right corner of Figure 5 are, from left to right, reset, undo, and redo.

The demo system allows us to transfer various styles to the drawn regions. Figure 4 shows the results of changing the style of plates with various styles of the given reference images. The impression of the plates changes greatly depending on the style of the plates. When favorite meal images for a user are generated, it can be saved and kept in the system. As shown in Figure 6, all the saved results can be seen on the UI, and the masks of the generated images and the style images used in the time of image synthesis can be viewed.

6 CONCLUSIONS

In this paper, we have created a new dataset by extracting plate regions using the the few-shot segmentation method and adding them to the UEC-FoodPix Complete dataset. Furthermore, to address the problem of distorted plate regions in food image generation, we improved the quality of food image generation using SEAN by using the UEC-FoodPix Complete augmented with plate region masks. and demonstrated the effectiveness of plate regions in food image generation. Furthermore, by freely selecting the shape, style, and arrangement of the plate region, the ideal set meal can be generated. We hope that users use this application to generate their favorite set meals. Note that we release the new dateset of pixel-wise plate annotation for the UEC-FoodPix Complete dataset at https://mm.cs.uec.ac.jp/uecfoodpix/.

Acknowledgment This work was supported by JSPS KAKENHI Grant Numbers, 21H05812, 22H00540, 22H00548, and 22K19808.

REFERENCES

[1] G. Benitez-Garc and K. Yanai. 2021. Ketchup GAN: A New Dataset for Realistic Synthesis of Letters on Food. In *Proc. of ICMR WS on Multimedia Artworks Analysis and Attractiveness Computing (MMArt).*

[2] J. Cho, W. Shimoda, and A. K. Yanai. 2021. Mask-Based Style-Controlled Image Synthesis Using a Mask Style Encoder. In *Proc. of IAPR International Conference on Pattern Recognition (ICPR).*

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. In *Advances in Neural Information Processing Systems.*

[4] D. Horita, W. Shimoda, and K. Yanai. 2019. Unseen Food Creation by Mixing Existing Food Images with Conditional StyleGAN. In *Proc. of ACMMM Workshop on Multimedia Assisted Dietary Management (MADIMA).*

[5] D. Horita, R. Tanno, W. Shimoda, and K. Yanai. 2018. Food Category Transfer with Conditional Cycle GAN and a Large-scale Food Image Dataset. In *Proc. of International Workshop on Multimedia Assisted Dietary Management (MADIMA).*

[6] Y. Ito, W. Shimoda, and K. Yanai. 2018. Food Image Generation using A Large Amount of Food Images with Conditional GAN: RamenGAN and RecipeGAN. In *Proc. of International Workshop on Multimedia Assisted Dietary Management (MADIMA).*

[7] Juhong Min, Dahyun Kang, and Minsu Cho. 2021. Hypercorrelation Squeeze for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*

[8] E. Nyamdorj. 2020. Generation of set meal images based on sketch images. In *Master Thesis at the University of Electro-Communications, Tokyo.*

[9] K. Okamoto and K Yanai. 2021. UEC-FoodPix Complete: A Large-scale Food Image Segmentation Dataset. In *Proc.of the ICPR Workshop on Multimedia Assisted Dietary Management (MADIMA).*

[10] K. Okamoto and K Yanai. 2021. UEC-FoodPix Complete: A Large-scale Food Image Segmentation Dataset. In *Proc.of the ICPR Workshop on Multimedia Assisted Dietary Management (MADIMA).*

[11] Dim P. Papadopoulos, Youssef Tamaazousti, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2019. How to make a pizza: Learning a compositional layer-based GAN model. In *Proc.of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[12] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proc.of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[13] C. Rother, V. Kolmogorov, and A. Blake. 2004. "GrabCut": interactive foreground extraction using iterated graph cuts. *Proc.of ACM Trans. Graph.* 23, 3 (2004), 309–314.

[14] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots. 2017. One-shotlearning for semantic segmentation.. In *Proc.of British Machine Vision Conference (BMVC).*

[15] W. Shimoda and K. Yanai. 2019. Self-supervised Difference Detection for Weakly-supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*

[16] W. Shimoda and K. Yanai. 2020. Predicting Plate Regions for Weakly-supervised Food Image Segmentation. In *Proc. of IEEE International Conference on Multimedia and Expo (ICME).*

[17] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *Proc.of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[18] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. SEAN: Image Synthesis with Semantic Region-Adaptive Normalization. In *Proc.of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*