Ramen as You Like : Sketch-based Food Image Generation and Editing

Jaehyeong Cho Wataru Shimoda Keiji Yanai
The University of Electro-Communications, Tokyo
{cho,shimoda-k,yanai}@mm.inf.uec.ac.jp

ABSTRACT

In recent years, a large number of images are being posted on SNS. The users often synthesize or modify their photos before uploading them. However, the task of synthesizing and modifying photos requires a lot of time and skill. In this demo, we demonstrate easy and fast image synthesis and modification through "sketch-based food image generation". The proposed system uses pix2pix to generate realistic food images based on sketched images, and DeepLab V3+ to obtain sketch masks from real photos. A user can create a realistic food image easily and fast by sketching a mask image consisting of food elements. In addition, a user can also edit a mask image automatically generated from a real photo food photo, and generate a modified food image. For training, we have created a new ramen image dataset consisting of 555 images with 15 kinds of pixel-wise labels.

CCS CONCEPTS

• **Information systems → Multimedia content creation**.

KEYWORDS

Image-to-Image Transformation, Food Image Generation, Food Image Segmentation, Food Image Dataset

1 INTRODUCTION

In recent years, a large number of images have been posted daily on blog or social media such as Twitter and Instagram. In general, users want to upload more attractive pictures to SNS. For this reason, users often try to synthesize or modify photos to make them more attractive. However, synthesizing and modifying photos is a difficult task that requires a lot of time and skill.

Our objective is to make a system that quickly and easily generates realistic images based on images sketched or modified by the user. To do that, we propose an image generation system using image-to-image translation network based on Generative Adversarial Networks (GAN) [1] and semantic segmentation network.

In the most of the works of the GAN-based method, the dataset on human face images, numeric character images, landscape images, and cityscapes images are widely used. On the other hand, there is very little work to generate or transform food images using GANs. In addition, "ramen" is the most popular food in Japan, and is also the popular Japanese food over the world including Asia, America, and Europe. For that reason, we created a new dataset of ramen photos annotated with pixel-wise labels for image generation and segmentation. In this demo, we propose a sketch-based interactive food image generation/editing system using an image-to-image translation network called "pix2pix" [3] and semantic segmentation network called "DeepLab V3+" [2]. At the conference site, we will demonstrate an interactive food image generation/editing system, "Ramen as You Like", based on sketched images and segmentation images in a web browser. Note that the system can treat with any kinds of the datasets containing pixel-wise labels such as the MS-COCO dataset, although we focus on ramen images in this demo.

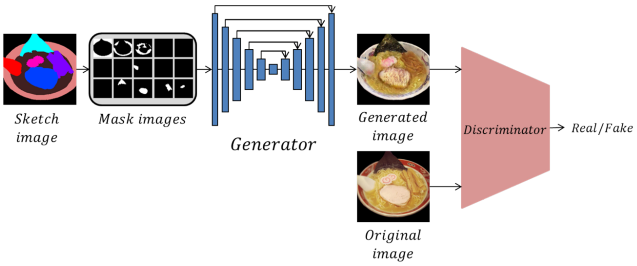

Figure 1: The architecture of a generator network based on U-Net [5]. In the network, the blue network represent a generator, and the red network represent a discriminator.

2 METHOD

2.1 Image Generation

Image-to-image translation networks called "pix2pix" [3] which is an extension of conditional GAN [4]. This pix2pix architecture consists of U-Net [5] which has an encoder-decoder with skip connections.

Our generation network is based on "pix2pix" with U-Net. The input of the network is the channel-concatenated feature map of 15-channel binary masks (Figure 1).

The objective function used Eq.1 as adversarial loss and Eq.2 as L1 loss. Eq.3 shows the overall loss function.

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[\log(1 - D(x,G(x,z)))], \quad (1)$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1], \quad (2)$$

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G), \quad (3)$$

where x and y present input data of set of mask image and real image, z present random noise vector.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10.
https://doi.org/10.1145/3343031.3350604

2.2 Image Segmentation

DeepLab V3+ [2] is a semantic segmentation model with an encoder-decoder structure consisting of a powerful encoder module and a simple yet effective decoder module. We use DeepLab V3+ as a segmentation network to generate a semantic mask from a real photo for sketch-based image editing. We train the model with the ramen dataset.

3 EXPERIMENTS

3.1 Ramen Image Dataset

We have created a new dataset which consists of ramen images and corresponding segmentation mask images (Figure 2). The mask images contain 15 classes of pixel-level semantic labels, which represent background, bowl, soup, spoon, chopsticks and toppings including cut egg, seaweed, sliced roasted pork and so on. We prepared 555 pairs of original ramen images and mask images, of which 500 pairs images were used to train a generation network and a segmentation network. The remaining images were used to test.

3.2 Results of Ramen Image Generation and Editing

Figure 3 shows the results of ramen images generated from sketch images drawn by users. Users can freely draw the size of the bowl and the amount of soup, select the soup type, add/remove spoon, chopsticks and toppings in the mask images, and convert them into realistic images instantly (less than 1 second). This shows that users can draw ramen images as they like interactively with the proposed system. In addition, distorted bowl can be generated as shown in the two rightmost examples of Figure 3 , while all bowls were in circle shapes in the training image set.

Figure 4 shows the work flow of semantic image editing. The original photos are converted into mask images by the segmentation network, and users modified them as they like such as adding/removing toppings and changing soup taste. Finally, we obtained the modified ramen images Users can semantically edit the generated masks by adding, removing, changing each element, which is easier than editing raw photos directly.

Note that additional results can be see at https://mm.cs.uec.ac.jp/RamenAsYouLike/.

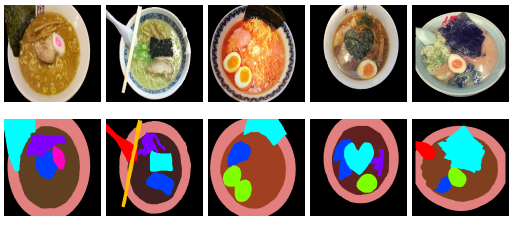

Figure 2: Some images and mask images in the ramen image dataset. Top: Original ramen images, Bottom: Corresponding mask images (Each of the class labels was drawn in the different colors.)

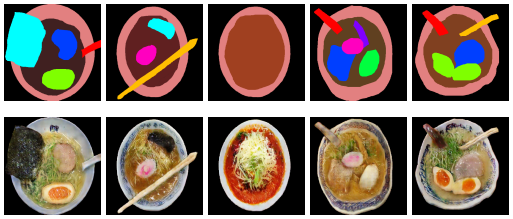

Figure 3: The results of ramen image generation from sketch image drawn by the user. Top : Sketched ramen images drawn by the user, Bottom : Generated ramen images

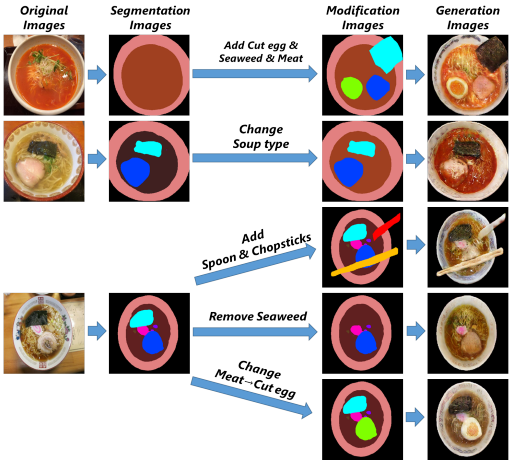

Figure 4: This figure shows the work flow of modifying and regenerating the segment result image of the actual image.

4 CONCLUSION

We have presented an application which can generate and modify food images from mask images sketched interactively. It employs an image-to-image translation network and semantic segmentation network.

For future work, we plan to extend this system to other domains than ramen. Regarding the food domain, we plan to extend it for the healthy purpose such as reducing the amount of over-calorie meals and changing unhealthy meals to healthy ones by replacing high-fat foods with vegetables.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 17J10261, 15H05915, 17H01745, 17H06100 and 19H04929.

REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS* .

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*.

[4] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI* .