Ketchup GAN: A New Dataset for Realistic Synthesis of Letters on Food

Gibran Benitez-Garcia, Keiji Yanai
gibran@ieee.org,yanai@cs.uec.ac.jp
Department of Informatics, The University of Electro-Communications
Chofu-shi, Tokyo, Japan

(a) Plain omelette set (b) Ketchup letter set (c) Segmentation set (d) Synthetic ketchup-free set

Figure 1: Samples of the four sets included in our Ketchup GAN dataset. Each set is subdivided into 1000 images for training and 150 for testing. Note that segmentation and ketchup-free sets were automatically generated by using weakly supervised segmentation [1] and unpaired image-to-image translation [15] methods, respectively.

ABSTRACT

This paper introduces a new dataset for the realistic synthesis of letters on food. Specifically, the "Ketchup GAN" dataset consists of real-world images of egg omelettes decorated with ketchup letters. Our dataset contains sufficient size and variety to train and evaluate deep learning-based generative models. In addition, we generate a synthetic ketchup-free set, which enables us to train paired-based generative adversarial networks (GAN). The ketchup GAN dataset comprises more than two thousand images of omelette dishes collected from Twitter. Automatically generated segmentation masks of egg and ketchup are also provided as part of the dataset. Thus, we can evaluate generative models based on segmentation inputs as well. With our dataset, two state-of-the-art GAN models (Pix2Pix and SPADE) are reviewed on photorealistic ketchup letter synthesis. We finally present an automatic application of omelette decoration with ketchup text input from users. The dataset and more details are publicly available at https://mm.cs.uec.ac.jp/omrice/.

CCS CONCEPTS

• **Applied computing** → *Media arts.*

MMArt-ACM '21, August 21, 2021, Taipei, Taiwan

ACM ISBN 978-1-4503-8531-2/21/08…$15.00
https://doi.org/10.1145/3463946.3469241

KEYWORDS

Letters on Food, Food Image Synthesis, Food Image Segmentation, Food Image Dataset

ACM Reference Format:
Gibran Benitez-Garcia, Keiji Yanai. 2021. Ketchup GAN: A New Dataset for Realistic Synthesis of Letters on Food. In *Proceedings of the International Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia 2021 (MMArt-ACM '21), August 21, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3463946.3469241

1 INTRODUCTION

Recently, food image generation and translation have been attracted particular attention due to the outstanding performance of generative adversarial networks (GAN) [13]. GAN-based models can generate food images that have not been cooked or translate from one dish to another [4, 5, 9]. Nonetheless, these models need a significantly large amount of images for learning. Indeed, it has been proved that the generated image quality becomes better as the number of training images increases [5]. Thus, reliable and substantial food image datasets are crucial for GAN-based approaches. On the other hand, as far as we know, there is no GAN-based model proposed for food image decoration with handwriting patterns, which can be suitable for SNS image enhancement, or even for helping chefs to visualize their dishes before cooking them. Therefore, in this paper we present a new dataset for the realistic synthesis of letters on food.

Our "Ketchup GAN" dataset consists of real-world images of omelettes decorated with ketchup letters. Omelette rice (omurice) is a popular dish in Japan (within the top-5 of food photos daily uploaded on Twitter [14]), often decorated with ketchup patterns. Hence, we gathered omelette images with and without ketchup

letters from Twitter. We also generate egg and ketchup masks, as well as synthetic ketchup-free images. Accordingly, Ketchup GAN consists of four sets with more than four thousand images in total, as shown in Figure 1. As additional contributions, we evaluate two state-of-the-art (SOTA) GAN models (Pix2Pix [6] and SPADE [10]) for synthetic ketchup image generation, and present an automatic application of omelette decoration with handwriting ketchup patterns based on the SPADE model and free-input from users.

2 RELATED WORK

GAN-based models that learn to synthesize and translate food images usually consist of a generator and a discriminator. The goal of the generator is to produce realistic images so that the discriminator cannot distinguish them from the real ones. Previous works for food image generation and translation have been proposed their own datasets. For instance, Horita et al. introduced "Bowl10" [5] and "Food13" [4] datasets, including 10 and 13 food categories derived from the 100 classes of the "UECFOOD-100" [7] dataset. Papadopoulos et al. introduced the "PizzaGAN" [9] dataset comprised of cooked and uncooked pizza images with binary levels of 12 topping ingredients. Although the number of images included in these datasets is big enough for training GAN models (230K, 227K, and 9K photos, respectively), they do not have food decorated with letters nor segmentation masks of the decorations. On the other hand, a few food datasets include semantic segmentation masks, such as "UEC-FoodPix Complete" [8] and "Ramen555" [2]. Note that these datasets do not have decorated food, and their size is relatively small (10K images from 100 classes and 555 from 15 ingredients, respectively).

3 THE KETCHUP GAN DATASET

We composed Ketchup GAN based on the gathered images from the large-scale food image dataset created by mining food images from the Twitter stream for more than eight years [14]. We focused only on the "omelette rice" category defined in the UECFOOD-100 dataset [7], which includes 100 classes from common Japanese foods. Based on the analysis of Twitter photos in Japan [14], omelette rice is ranked in fourth place of popularity behind ramen, curry, and sushi. Thus, we had access to a vast group of real-world images (about 100K photos) within this category. Note that, in this paper, we focus on omelet dishes because, in Japan, it is common to decorate these with ketchup patterns. On the other hand, there are other dishes and foods that follow the same process. For example, cakes decorated with whipped cream and pancakes decorated with jam or chocolate. We believe that using GAN-based methods could complement the food design usually achieved by the 3D-printing technology [3].

To choose meaningful omelette dishes decorated with ketchup, we trained a CNN classifier with five classes, as shown in Figure 2. We sorted all images using the confidence scores obtained from the classifier and choose 2000 per class. With this dataset, we train a weakly supervised segmentation algorithm to obtain coarse egg and ketchup masks. We use the PSA algorithm [1], which learns Pixel-level Semantic Affinity from Class Activation Maps (CAMs). Hence, the training only relies on image-level class labels. We further use

Figure 2: Original five classes derived from the "omelette rice" category of the large-scale food image dataset [14]. From top to down rows: plain omelette, ketchup letters, ketchup drawings, ketchup patterns, and sauce.

dense CRF (Conditional Random Field) to obtain smoother regions from roughly estimated PSA masks potentials. Subsequently, we train SOTA real-time semantic segmentation method with the CRF results of the 2000 images per class. In this paper, we use FASSD-Net [11], a U-shape encoder-decoder network designed to keep a low computational complexity and exploit high-level features' contextual information by using asymmetric convolutions. We train FASSD-Net with 1800 and 200 images per class for training and testing, respectively. Note that we normalize all images at 360×360 pixel resolution. Then, we generate segmentation masks of all images from plain omelette and ketchup letters to finally choose manually the best 1150 segmentation masks related to each class, as shown in Figure 1.

In order to be able to train paired-based GAN models, we seek to generate a ketchup-free counterpart of the ketchup letter set. Thus, we use CycleGAN [15], an unpaired image-to-image translation method that can converts an image from one specific domain into another domain with outstanding results. CycleGAN introduces the cycle consistency loss that enables generating images with their original structure without a need for paired images for training. We train CycleGAN with the 2000 images per class at 360×360 pixel resolution resolution described in Figure 2. Then, we choose the counterpart images of the letter set, as shown in Figure 1. Note that we follow the official open-source implementation of PSA [1], FASSD-Net [11], and CycleGAN [15] for training the three models. In resume, Ketchup GAN comprises 2300 real-world photos of plain omelette and ketchup letters, as well as 1150 semantic segmentation masks of egg and ketchup sauce, and 1150 ketchup-free images generated from the ketchup letter set.

4 PHOTOREALISTIC KETCHUP LETTER SYNTHESIS

In this section, we describe the experiments based on two SOTA GAN-based models: Pix2Pix [6] and SPADE [10].

(a) Inputs (b) Pix2Pix (d) SPADE

Figure 3: Samples of the test set showing the results of the GAN-based models for photorealistic ketchup synthesis.

(a) Inputs (b) Pix2Pix (d) SPADE

Figure 4: Results of the GAN-based models for photorealistic ketchup synthesis with customized ketchup letters.

4.1 Pix2Pix

Pix2Pix [6] proposed by Isola et al. is a historic milestone in image-to-image translation. Before this method, encoder-decoder-based image translation models were trained using the L2 mean square loss only, making them unsuitable for image conversion between different domains. Pix2Pix firstly introduced an adversarial loss in addition to the L1 loss for achieving domain translation from paired image datasets. Due to the paired requirement of the training dataset, the results of Pix2Pix tend to be more realistic than those of CycleGAN [15]. Therefore we use the former as a baseline for photorealistic ketchup letter synthesis.

In the experiment, we use the network architecture and the official implementation by the authors [6]. We train the model with 1000 images per class of size 256×256. Note that to control the ketchup decoration generated by the model, we concatenate the segmentation masks to the ketchup-free RGB image. So that, the input to the generator is a tensor of size $5 \times 256 \times 256$ corresponding to the RGB channels plus the binary masks of egg and ketchup, respectively. Figure 3 (a) and (b) show some samples of the test set depicting the inputs and outputs of the trained Pix2Pix model. The results show a correct ketchup letter generation. However, the consistency and light reflection of the sauce is not entirely realistic.

We have also evaluated our Pix2Pix model to a customized ketchup pattern as input. In this experiment, we use a modified ketchup mask with a user-input pattern drawn within the omelette region, as shown in Figure 4(a). The results of this experiment are shown in Figure 4(b), where the trend of the previous results is more evident. Although the dish reconstruction is almost identical and the ketchup message is readable, Pix2Pix is not able to synthesize the characteristics of the sauce properly.

4.2 SPADE

SPADE [10], defined as "SPatially-ADaptivE normalization," is a GAN-based approach aimed to learn a mapping function that can convert an input segmentation mask to a photorealistic image. Its main contribution is a conditional normalization layer similar to batch normalization, in which activations are normalized in a channel-wise manner and then modulated with affine parameters learned from the segmentation mask. SPADE based its architecture on the decoder of the HD version of Pix2Pix [12], excluding the encoder part, and adopting SPADE normalization on all layers of the generator. The affine parameters of all normalization layers are learned using the segmentation mask resized accordingly to the corresponding feature maps. The generator is trained using the same multi-scale discriminator and similar loss functions as Pix2pixHD [12]. Thus, we use SPADE as a feasible alternative to synthesizing photorealistic images from egg and ketchup segmentation masks.

In the experiments, we follow the official implementation by the authors [10] and use the vanilla architecture, where the input to the generator is a random noise and the segmentation mask is inputted on each SPADE normalization layer. We choose the vanilla architecture due to its simplicity (it does not include any encoder) and because we are focus on the photorealistic synthesis of ketchup letters. We train SPADE with 1000 images per class of size 256×256, as suggested by the authors. Note that the input of the network is only the segmentation mask. Hence the ketchup-free set of our dataset is not needed. The results of the test set and the customized ketchup patterns are shown in Figure 3 and Figure 4, respectively. The second column of (a) shows the mask inputs while (c) presents the synthesized images. As expected, the egg and ketchup regions look more realistic than the results of Pix2Pix, particularly in the test with customized ketchup letters. Still, the images show a low-quality aspect due to the plate and background regions are not included in the segmentation mask input.

5 OMELETTE DECORATION

We build up an omelette decoration application using the SPADE model trained with our dataset and handwriting ketchup patterns

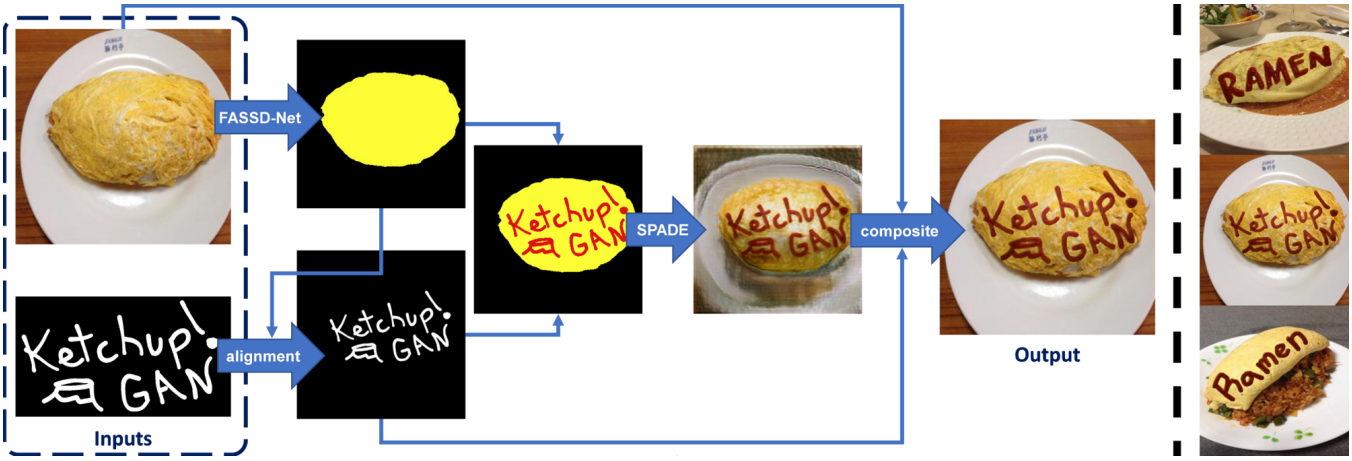

Figure 5: Omelette decoration process (left), and the results obtained with inputs shown in Figure 4(a) (right).

Figure 6: Comparison between the results of our method (left) and an image generated by a photo editor (right).

input by users. Figure 5 (left) shows the general process of the proposed omelette decoration, which is based on four steps described as follows.

i *Segmentation.* The FASSD-Net model is used to generate the segmentation mask of the omelette in real-time.
ii *Text alignment.* The handwriting text is resized based on the egg region. Then, we adjust the text position using the orientation of the omelette obtained by applying PCA to the egg shape. The final segmentation mask combines the aligned text and the omelette mask.
iii *SPADE.* We use the SPADE model to synthesize the ketchup letters from the segmentation mask.
iv *Composition.* We compose the final output I as $I = I_S M + I_O$, where I_S is the image generated by SPADE, M represents the segmentation mask, and I_O is the original omelette image.

Composed images using our omelette decoration are shown in Figure 5 (right). From these generated images, we prove that it is possible to achieve photorealistic results with free-handwriting ketchup patterns. Although the method of *step iv* generates letters with rough shapes, the results of the ketchup sauce look more realistic than a manually edited image by superposing red letters, as shown in Figure 6. In this example, we use a handwriting-looking font with a ketchup color to edit the plain omelette. It is clear that the results are better when using our proposal than the edited photo with the same font and message.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we introduced a new dataset for the realistic synthesis of ketchup letters on omelettes. Our dataset includes real-world images from omelettes with and without ketchup patterns, as well as synthetic paired ketchup-free images and segmentation masks of egg and ketchup. We evaluated two SOTA GAN-based methods for photorealistic ketchup synthesis using our dataset, and proposed a simple yet effective automatic method for omelette decoration with handwriting patterns. Our results show that the proposed dataset is suitable for training GAN-based methods, and SPADE can be used to generate ketchup letters for omelette decoration.

As future work, we plan to improve the decoration method by proposing a GAN-based model capable of aligning the input texts and synthesizing omelettes with ketchup patterns simultaneously.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Jiwoon Ahn and Suha Kwak. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*. IEEE, 4981–4990.
[2] Jaehyeong Cho, Wataru Shimoda, and Keiji Yanai. 2019. Ramen as You Like: Sketch-based Food Image Generation and Editing. In *ACM International Conference on Multimedia*.
[3] Fernanda C Godoi, Sangeeta Prakash, and Bhesh R Bhandari. 2016. 3d printing technologies applied for food design: Status and prospects. *Journal of Food Engineering* 179 (2016), 44–54.
[4] Daichi Horita, Wataru Shimoda, and Keiji Yanai. 2019. Unseen food creation by mixing existing food images with conditional stylegan. In *5th International Workshop on Multimedia Assisted Dietary Management MADiMa*. 19–24.
[5] Daichi Horita, Ryosuke Tanno, Wataru Shimoda, and Keiji Yanai. 2018. Food category transfer with conditional cyclegan and a large-scale food image dataset. In *Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management*. 67–70.
[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*. 1125–1134.
[7] Yuji Matsuda and Keiji Yanai. 2012. Multiple-food recognition considering co-occurrence employing manifold ranking. In *21st International Conference on Pattern Recognition ICPR*. IEEE, 2017–2020.
[8] Kaimu Okamoto and Keiji Yanai. 2021. UEC-FoodPix Complete: A Large-Scale Food Image Segmentation Dataset. In *6th International Workshop on Multimedia*

Assisted Dietary Management MADiMa. 647–659.

[9] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*. 8002–8011.

[10] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR*. 2337–2346.

[11] Leonel Rosas-Arias, Gibran Benitez-Garcia, Jose Portillo-Portillo, Gabriel Sanchez-Perez, and Keiji Yanai. 2021. Fast and Accurate Real-Time Semantic Segmentation with Dilated Asymmetric Convolutions. In *25th International Conference on Pattern Recognition ICPR*. IEEE, 1–8. https://doi.org/10.23919/MVA.2019.8757973

[12] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*. 8798–8807.

[13] Keiji Yanai, Daichi Horita, and Jaehyeong Cho. 2020. Food Image Generation and Translation and Its Application to Augmented Reality. In *IEEE Conference on Multimedia Information Processing and Retrieval MIPR*. IEEE, 181–186.

[14] Keiji Yanai, Kaimu Okamoto, Tetsuya Nagano, and Daichi Horita. 2019. Large-Scale Twitter Food Photo Mining and Its Applications. In *IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, 77–85.

[15] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision ICCV*. IEEE, 2223–2232.