Unseen Food Creation by Mixing Existing Food Images with Conditional StyleGAN

Daichi Horita Wataru Shimoda Keiji Yanai
The University of Electro-Communications, Tokyo
{horita-d,shimoda-k,yanai}@mm.inf.uec.ac.jp

ABSTRACT

In recent years, thanks to the development of generative adversarial networks (GAN), it has become possible to generate food images. However, the quality is still low and it is difficult to generate appetizing and delicious-looking images.

In the latest GAN study, StyleGAN [12] enabled high-level feature separation and stochastic variation of generated images by unsupervised learning. However, to manipulate any style, it is necessary to understand the representation of the latent space and to input reference images.

In this paper, we propose a conditional version of StyleGAN to control probabilistic fluctuations of disentangled features. The conditional style-based generator can manipulate the style of any domain by providing conditional vectors. By applying the conditional StyleGAN to the food image domain, we successfully have generated higher quality food images than before. In addition, introducing conditional vectors enabled us to control the food categories of generated images and to generate unseen foods by mixing multiple kinds of foods.

In the experiment, to show the result of the proposed method explicitly, Food13 dataset is constructed and evaluated by both qualitative evaluation and quantitative evaluation.

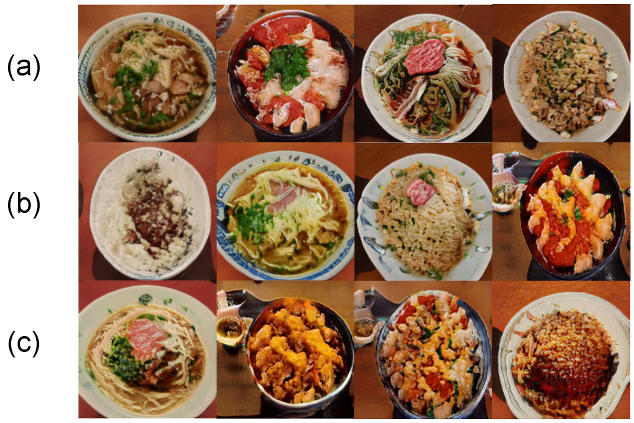

Figure 1: From top to bottom: example images of generated from inputs containing condition information specifying (a) one, (b) two, (c) three domains, respectively. All images are generated with the same style noise. Our model can generate images that include class information from an input that includes condition information that specifies multiple classes.

CCS CONCEPTS

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

KEYWORDS

GAN, CNN, generative adversarial model, food image generation

ACM Reference Format:
Daichi Horita Wataru Shimoda 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), October 21, 2019, Nice, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3347448.3357166

1 INTRODUCTION

Recently, in the field of computer vision, there are many studies for generative methods, variational auto-encoders (VAEs) [14] and generative adversarial nets (GANs) [3]. In particular, GANs show remarkable results in various computer vision tasks such as image generation [1, 11], image conversion [9, 31], super-resolution [15] and text-to-image synthesis [29]. GANs have already been applied to the food image domain. Ito et al. [10] proposed RecipeGAN and RamenGAN both of which are based on conditional GAN [16]. By using conditional vectors, we can control the categories or attributes of generate images. RecipeGAN generates food images with the condition on food ingredients, and RamenGAN generates ramen images with the condition on ramen fine-grained categories. However, both the GANs failed to generate realistic food images which look delicious. To solve this problem, we introduce the StyleGAN [12] which is one of the state-of-the-art GANs regarding image quality into food image generation. To control food categories of generated images, we propose a conditional StyleGAN which is the StyleGAN with conditional inputs.

Regarding StyleGAN, Karras et al. [12] re-designed the generator architecture to control the image synthesis process, motivated by style transfer literature [7]. The generator starts from the input of learned constants and adjusts the style of the image at each convolutional layer based on the latent code. Thus, the generator directly controls the strength of image features at different scales. The generator separates high dimensional attributes from probabilistic fluctuations in unsupervised learning. However, the change of noise from high dimensional latent space cannot be controlled

MADiMa '19, October 21, 2019, Nice, France

ACM ISBN 978-1-4503-6916-9/19/10...$15.00
https://doi.org/10.1145/3347448.3357166

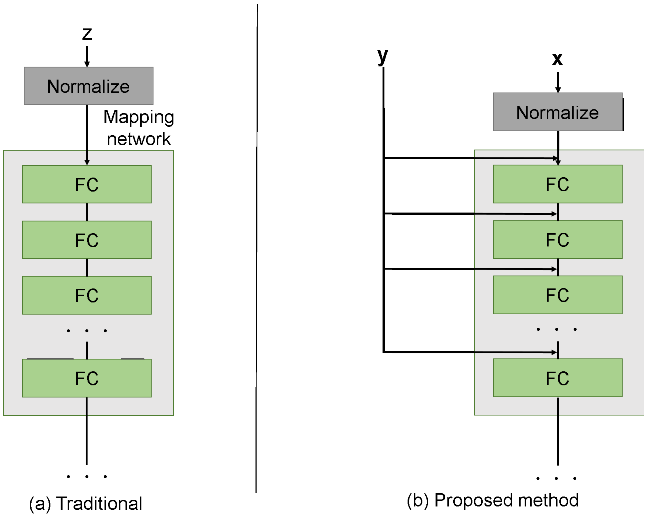

Figure 2: We add conditional information y after each FC layer (Right) to the styled-based generator (Left) [12] . Conditional information y makes it possible to specify any style. Please refer to Section 4.4 for details about how to input conditions many times.

to be the style of a particular domain. To solve this problem, we extend the generator to manipulate the latent space in a specific domain by combining the conditional vector with the noise injected directly into the network.

Mirza et al. [16] propose that the generative adversarial nets can be extended to a conditional model if both the generator and the discriminator are conditioned based on some additional information y. y is any kind of auxiliary information, such as class labels and data from other modalities. In this paper, we propose a the conditional style-based generator that can control the latent variables representing the style of the images for more separation and understanding of the potential space. A conditional vector that specifies an arbitrary style associated with a latent variable representing a style controls the latent space and allows the generation of images of any domain.

Through the experiments, we found interesting characteristics of conditional StyleGAN on food domain. By modifying the random noise, our proposed model can generate images with various variations within the same food categories. In addition, our model can generate unknown dishes by mixing multiple conditions. For example, as shown in Figure 2, our model generate images representing each feature by providing multiple food class condition vectors as inputs. As a result, we confirmed that the proposed model had the potential of creating new foods.

Major contributions of this paper can be summarized as follows:

(1) We propose the style-based conditional generator to control the latent space.
(2) Our proposed generator can generate images that reflect each feature from condition information that specifies multiple classes.
(3) In the experiments with a food dataset, we reveal that the noises in the style-based generator indicate additional elements such as topping information, and unseen types of food images can be generated by mixing multiple kinds of foods with mixing conditional weights.

2 RELATED WORK

2.1 Generative Adversarial Nets

GANs [1, 3] have achieved impressive results in image generation [11, 23], image-to-image translation [9, 13, 31], text-to-image translation [24, 29], and super-resolution [15]. A GAN consists of a generator and a discriminator. The discriminator learns to distinguish the produced fake samples from the real ones, while the generator learns to generate fake samples which are not distinguishable from the real ones.

To the best of our knowledge, cGAN [16] is the first GAN model which leverages conditional information. The input of cGAN's generator is the combination of the latent vector noise and label vector, and the input to the discriminator is the generated fake or real sample and a corresponding label. One of the cGANs methods is AC-GANs [21], which directly manipulates the loss objectives. AC-GANs use a classifier that shares part of its structure with the discriminator and augments the original discriminator's objective function using the generated fake samples and the likelihood score of real samples. Label information is included in the objective function. Another method using of conditional information is the method proposed by [20]. Miyato et al. [20] propose the projection based way to incorporate the conditional information into the discriminator of GANs. Therefore, the model takes into account the structure of the assumed conditional probabilistic models underlined by the structure of the discriminator. We propose a condition version of StyleGAN [12] using the method proposed by Miyato et al [20].

2.2 Feature Disentanglement

There are some studies for disentangled representation learning. For instance, there are some studies [2, 26, 28] to learn disentangled representations without supervision. Some other works [8, 18] focus on disentangling content and style by encoding from an image. In [8], "content" is defined as the underlying spatial structure and "style" is defined as the rendering of the structure. However, Karras et el. [12] propose a method to disentangle by learning various style information by supplying a part of the latent code to multiple generator layers. This method enables the generator to learn distributions in which latent codes represent conditional information. However, it is difficult to sample a feature that represents an arbitrary style from the latent code without reference images because no condition that restricts the distribution of the latent code. In this paper, we propose to specify a latent code representing an arbitrary style by learning the distribution of conditional information by combining it with the latent code.

2.3 Food Image Generation

GANs have been applied for generating food images from ingredients [10] and recipes [25, 30], and translating [5, 6, 27]. Horita

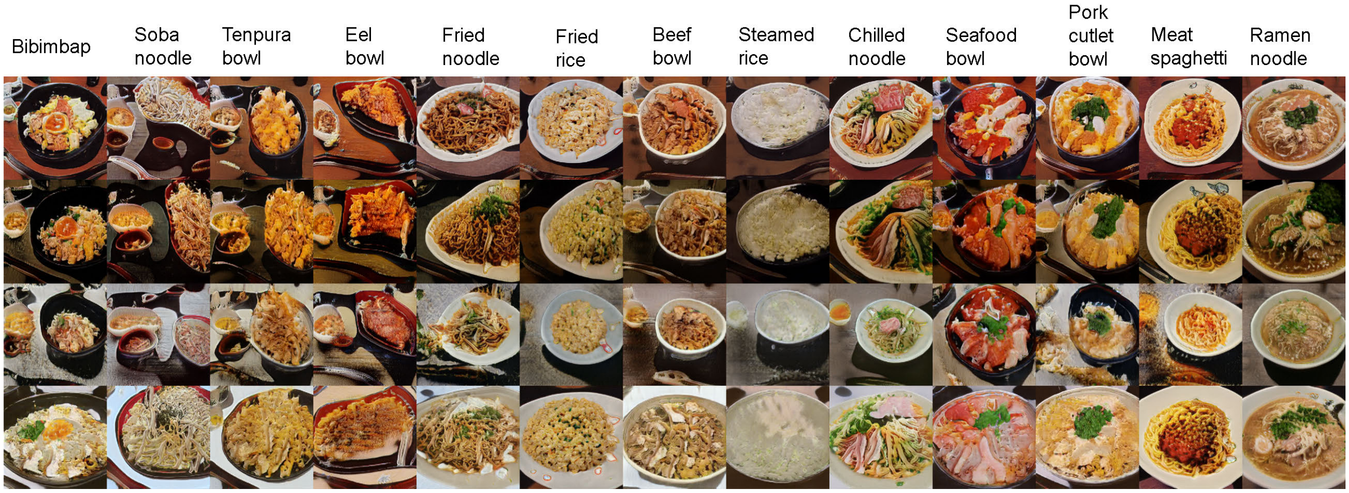

Figure 3: Given the fixed noise, style, and conditional vector, our proposed model generates the image corresponding to the food class of each row. Eel bowl and soba noodle have distorted food images because they have few round samples.

et al. [6] propose to extend CycleGAN [31] to convert food images among 10 categories. However, this work converts only the style between the two types of food without changing the content of the food. On the other hand, in this paper, we manipulate the structure by fluctuating transformation and style noise including two or more styles by setting multiple domains as conditions.

3 OUR APPROACH

3.1 Conditional style-based generator

Our goal is to learn a conditional version of the style-based generator to control the latent space. To achieve this, we propose a method to input conditional information to the mapping function f configured in the MLP layer, as shown in Figure. 1. This architecture differs from the traditional way of providing conditional information to the first layer input of the generator. Given latent code z and the conditional information y, The mapping function f converts to conditional style information$(y_{s,c}, y_{b,c})$ to control an adaptive instance normalization layer (AdaIN) [7] by performing the nonlinear conversion. The AdaIN operation is defined as

$$AdaIN(x_i, y_c) \quad = \quad y_{s,i,c}\frac{x-\mu(x_i)}{\sigma(x_i)} + y_{b,i,c}, \qquad (1)$$

where x_i is normalized with AdaIN operation using the conditional styles. As a result, AdaIN operation can be normalized so that mapping networks show any style. The dimension of the conditional style is twice that of the feature map z.

For the objective functions, we use the hinge version of the standard adversarial loss [17], which is different from StyleGAN [12], defined as

$$\begin{aligned} \mathcal{L}(\hat{G}, D) &= E_{q(y)}\left[E_{q(x|y)}\left[max(0, 1 - D(x,y))\right]\right] \\ &+ E_{q(y)}\left[E_{p(z)}\left[max(0, 1 + D(\hat{G}(z,y),y))\right]\right], \quad (2) \end{aligned}$$

$$\mathcal{L}(G, \hat{D}) = -E_{q(y)}\left[E_{p(z)}\left[\hat{D}(G(z,y),y)\right]\right], \qquad (3)$$

where $p(z)$ is the standard Gaussian distribution. Our conditional style mapping network f consists of multiple fully-connected (FC) layers. We analyze the effect of the number of FC layers on the image quality in Section 4.4.

3.2 Conditional discriminator

We adopt the projection discriminator proposed by Miyato et al. [20]. The projection discriminator incorporates conditional vectors differently from the method of concatenating the embedded conditional vectors into feature vectors.

Table 1: Food13 dataset. The number of samples in each class.

Category	# number of images
bibimbap	9433
fried rice	28406
beef bowl	9720
steamed rice	6387
ramen noodle	80000
eel bowl	5100
fried noodle	25000
pork cutlet bowl	10000
chilled noodle	13600
seafood bowl	10000
tempura bowl	10000
meat spaghetti	7000
soba noodle	3300
total	227946

4 EXPERIMENTS

4.1 Dataset

We have created a relatively large food image dataset, which is "Food13". As shown in Table 1, Food13 dataset consists of a total of 13 categories and of about 230k of food images that contains

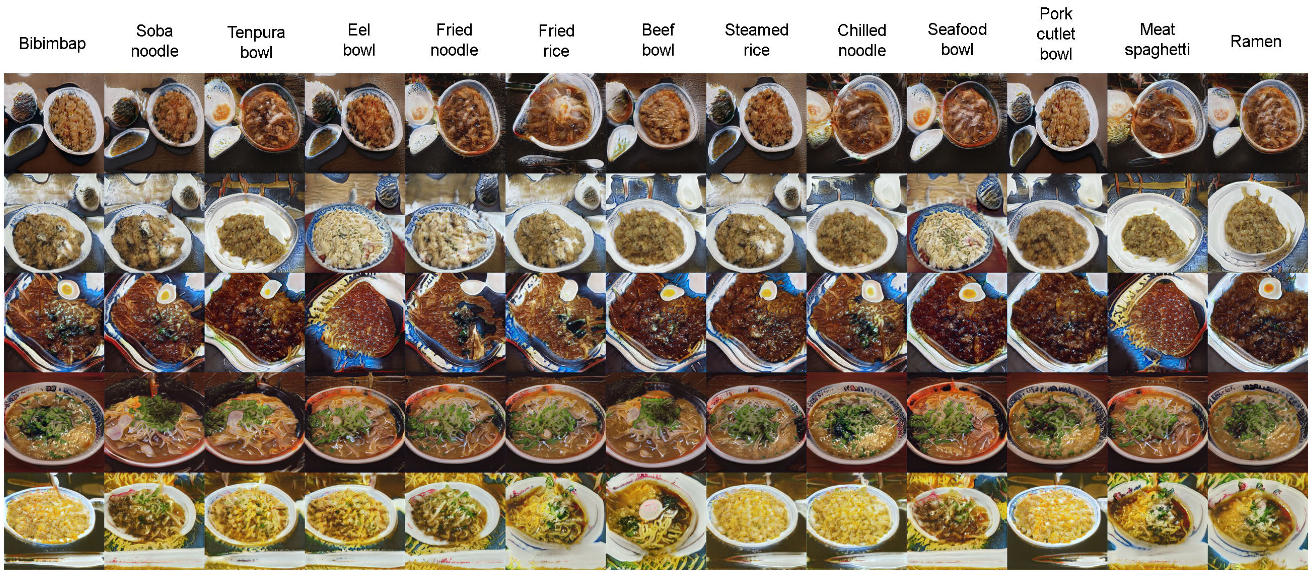

Figure 4: Images generated by the model learned without inputting condition information other than the first one. That is, the model is given the same input as in Figure 3. Each column corresponds to a conditional vector of each class.

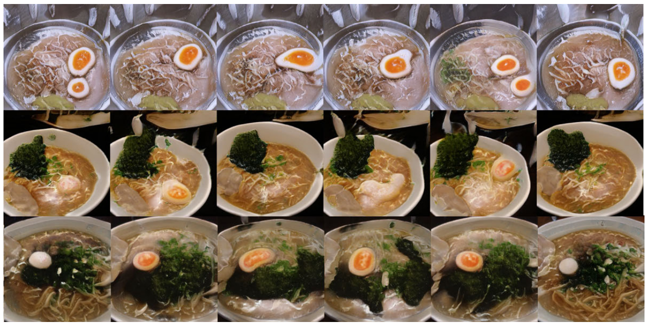

Figure 5: Images generated from the fixed styles, the condition vectors, and randomly sampled noises. Each row shows the difference in randomly sampled noise. Random noise plays a role in expressing differences such as food topping.

the dataset proposed by [6, 19] and the additional food images collected from the Twitter stream. We divide the set of the dataset into a training set and a test set. The test set consists of 1000 images of each category. Note that most of the images are not high-resolution, and they are not suitable for high-resolution image generation. We normalize the color values to $[-1, 1]$ before given the image into the model.

4.2 Implementation Details

We implement our method using PyTorch [22]. We also start to generate resolutions from 8×8 , as we do with Karras et al. [12]. The models are optimized by stochastic gradient descent. For all experiments, we use the same learning rates as Karras et al. [11]. We update the generator once for each discriminator update.

4.3 Manipulation of latent space

Figure 3 shows that images are generated from fixed the input noises styles, and the condition vectors representing each class by our model. We can see that our model can generate an arbitrary class of images thanks to the condition vectors. However, although the noises and styles are fixed, eel bowl and soba noodle have distorted food images because they have few round samples. Figure 5 shows the images generated from a fixed style, a condition vector, and randomly sampled noise. Each image is generated by our model given different random noises. We can see that random noise plays a role in expressing differences such as food topping. Figure 6 shows that images generated by our model simultaneously with condition vectors representing two or more classes. We can see that our model can generate each feature even when given the multiple condition vectors.

4.4 Comparison of the number of FC layers

We calculate the FIDs using samples randomly divided into 1000 for evaluation from each class of the Food13 dataset. To extract features, we use Resnet50 [4] which has been pre-learned using UEC-food100 [19]. Table 2 shows how FIDs are affected by the mapping network. We can see that the average value is the highest when the mapping network consists of eight FC layers.

Figure 3 and Table 3 shows results and FIDs in case of the model learned without inputting condition information other than the first one to evaluate the role of the condition information input before each FC layer. We can see that FIDs are lower when we enter the condition information after each FC layer. Note that lower FID indicates better quality.

Figure 4 shows the generation result of the condition vector that controls the class in the model learned without inputting condition information other than the first one. That is, the model is given the

Category	MLP 4	MLP 8	MLP 12	MLP 16
bibimbap	13.61	10.91	20.15	11.89
fried rice	16.30	8.50	14.15	8.65
beef bowl	22.41	11.50	13.96	17.39
steamed rice	15.59	7.72	8.28	6.26
ramen noodle	29.27	25.31	26.31	24.16
eel bowl	39.26	40.33	32.53	31.82
fried noodle	10.56	8.51	13.58	16.93
pork cutlet bowl	13.47	10.22	15.91	11.93
chilled noodle	13.43	11.19	15.22	11.14
seafood bowl	12.76	15.68	15.49	11.12
tempura bowl	15.25	11.37	18.59	13.73
meat spaghetti	14.68	14.20	20.15	12.32
soba noodle	21.15	13.73	14.29	18.0
average	18.28	14.55	17.58	15.02

Table 2: FIDs by the difference in the number of FC layers in the mapping network f.

Category	Input 4	Input 8	Input 12	Input 16
bibimbap	23.11	19.34	21.63	26.13
fried rice	19.35	23.45	19.85	32.23
beef bowl	23.28	21.89	19.26	29.51
steamed rice	29.75	28.87	22.61	43.24
ramen noodle	34.18	30.32	27.15	45.09
eel bowl	51.23	47.13	43.34	45.57
fried noodle	18.41	19.65	20.27	27.16
pork cutlet bowl	22.46	20.77	20.56	22.47
chilled noodle	21.10	20.45	19.91	29.99
seafood bowl	27.21	24.30	21.78	34.13
tempura bowl	21.63	19.06	18.56	20.38
meat spaghetti	21.66	21.37	23.44	28.27
soba noodle	29.79	27.93	23.16	37.65
average	26.39	24.96	23.19	32.44

Table 3: FIDs due to the difference in the number of FC layers when the conditional information y is first given together with the latent code z.

same input as in Figure 1. We can see that the mapping function f can not control the class without multiple condition vector inputs.

5 CONCLUSION

We propose a conditional version of StyleGAN to control probabilistic fluctuations of disentangled features. The style-based generator can manipulate the style of any domain by providing conditional vectors. Experiments show that the proposed model can generate various food images with conditional vectors.

As future work, we plan to generate food images from recipes. This method allows the generator to generate images from various ingredients. For example, we experiment what kind of image can be generated by changing tofu, which is a material contained in mapo tofu, to another ingredient like eggplants. In addition, we plan to learn an image transformation network that changes the calorie appearance of the food, taking into account the noise effect that changes the topping that we clarified in this paper. Thus, our proposed conditional style-based generator has great potential to solve problems that are still unresolved concerning to food images. Finally, we hope that our proposed model will contribute to further food image research.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning(ICML)*.

[2] Chris Donahue, Akshay Balsubramani, Julian McAuley, and Zachary C. Lipton. [n. d.]. Semantically Decomposing the Latent Spaces of Generative Adversarial Networks. In *Proc. of the International Conference on Learning Representation(ICLR)*.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proc.of Neural Information Processing Systems*.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc.of IEEE Computer Vision and Pattern Recognition(CVPR)*.

[5] Daichi Horita, Jaehyeong Cho, Takumi Ege, and Keiji Yanai. 2018. Ramen Spoon Eraser: CNN-based photo transformation for improving attractiveness of ramen photos. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology(VRST)*.

[6] Daichi Horita, Ryosuke Tanno, Wataru Shimoda, and Keiji Yanai. [n. d.]. Food category transfer with conditional cycleGAN and a large-scale food image dataset. In *Proc.of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management(MADiMa)*.

[7] Xun Huang and Serge J. Belongie. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. *Proc.of IEEE International Conference on Computer Vision(ICCV)* (2017).

[8] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. 2018. Multimodal Unsupervised Image-to-Image Translation. (2018).

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Proc.of IEEE Computer Vision and Pattern Recognition(CVPR)*.

[10] Yoshifumi Ito, Wataru Shimoda, and Keiji Yanai. [n. d.]. Food image generation using a large amount of food images with conditional gan: ramengan and recipegan. In *Proc.of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management(MADiMa)*.

[11] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *"Proc. of the International Conference on Learning Representation(ICLR)"*.

[12] Tero Karras, Samuli Laine, and Timo Aila. [n. d.]. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proc.of IEEE Computer Vision and Pattern Recognition(CVPR)*.

[13] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In *International Conference on Machine Learning(ICML)*.

[14] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *Proc. of the International Conference on Learning Representation(ICLR)*.

[15] C. Ledig, L. Theis, F. HuszÃ¡r, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. 2016. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network, In Proc.of IEEE Computer Vision and Pattern Recognition(CVPR). *arXiv preprint arXiv:1609.04802*.

[16] Jae Hyun Lim and Jong Chul Ye. 2014. Conditional Generative Adversarial Nets. (2014). arXiv:1411.1784

[17] Jae Hyun Lim and Jong Chul Ye. 2017. Geometric GAN. (2017). arXiv:1705.02894

[18] Alexander H. Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. [n. d.]. A Unified Feature Disentangler for Multi-Domain Image Translation and Manipulation. ([n. d.]).

[19] Y. Matsuda, H. Hoashi, and K. Yanai. 2012. Recognition of Multiple-Food Images by Detecting Candidate Regions. In *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*.

[20] Takeru Miyato and Masanori Koyama. [n. d.]. cGANs with Projection Discriminator. ([n. d.]).

[21] Augustus Odena, Christopher Olah, and Jonathon Shlens. [n. d.]. Conditional Image Synthesis with Auxiliary Classifier. In *International Conference on Machine Learning(ICML)*.

[22] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.

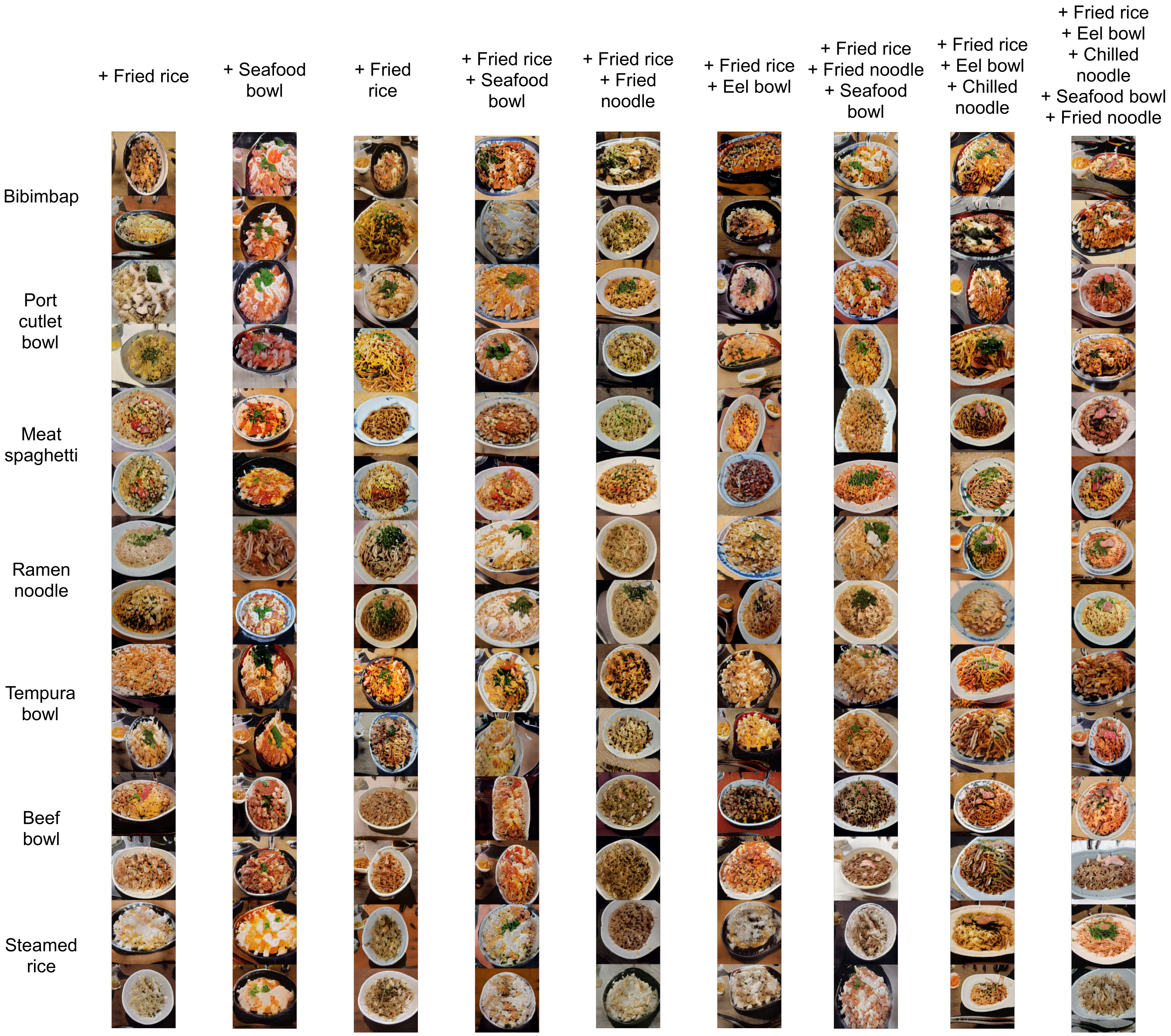
+ Fried rice
+ Seafood bowl
+ Fried rice
+ Fried rice + Seafood bowl
+ Fried rice + Fried noodle
+ Fried rice + Eel bowl
+ Fried rice + Fried noodle + Seafood bowl
+ Fried rice + Eel bowl + Chilled noodle
+ Fried rice + Eel bowl + Chilled noodle + Seafood bowl + Fried noodle
Bibimbap
Port cutlet bowl
Meat spaghetti
Ramen noodle
Tempura bowl
Beef bowl
Steamed rice

Figure 6: Images generated by mixing two or more condition information by our proposed model. The condition vector consists of each food class in the left column plus the food class at the top of each column.

[23] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *Proc. of the International Conference on Learning Representation(ICLR)*.

[24] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative Adversarial Text to Image Synthesis, In International Conference on Machine Learning(ICML). *arXiv preprint arXiv:1605.05396*.

[25] Amaia Salvador, Michal Drozdzal, Xavier Giro-i Nieto, and Adriana Romero. 2019. Inverse Cooking: Recipe Generation From Food Images. In *Proc.of IEEE Computer Vision and Pattern Recognition(CVPR)*.

[26] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. [n. d.]. Style Transfer from Non-Parallel Text by Cross-Alignment. In *Proc.of Neural Information Processing Systems*.

[27] Ryosuke Tanno, Daichi Horita, Wataru Shimoda, and Keiji Yanai. 2018. Magical Rice Bowl: A Real-time Food Category Changer.. In *Proc.of ACM International Conference Multimedia(ACMM)*.

[28] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2018. MoCoGAN: Decomposing Motion and Content for Video Generation. In *Proc.of IEEE Computer Vision and Pattern Recognition(CVPR)*.

[29] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. 2017. Stackgan: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In *Proc.of IEEE International Conference on Computer Vision(ICCV)*.

[30] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. 2019. R2GAN: Cross-Modal Recipe Retrieval With Generative Adversarial Network. In *Proc.of IEEE Computer Vision and Pattern Recognition(CVPR)*.

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networkss. In *Proc.of IEEE International Conference on Computer Vision(ICCV)*.