https://doi.org/10.2352/EI.2023.35.7.IMAGE-268

Conditional Synthetic Food Image Generation

Wenjin Fu[1], Yue Han[2], Jiangpeng He[2], Sriram Baireddy[2], Mridul Gupta[2], Fengqing Zhu[2]

[1] School of Computer Science and Engineering, The Ohio State University, Columbus, OH, United States
[2]Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, United States

Abstract

Generative Adversarial Networks (GAN) have been widely investigated for image synthesis based on their powerful representation learning ability. In this work, we explore the StyleGAN and its application of synthetic food image generation. Despite the impressive performance of GAN for natural image generation, food images suffer from high intra-class diversity and inter-class similarity, resulting in overfitting and visual artifacts for synthetic images. Therefore, we aim to explore the capability and improve the performance of GAN methods for food image generation. Specifically, we first choose StyleGAN3 as the baseline method to generate synthetic food images and analyze the performance. Then, we identify two issues that can cause performance degradation on food images during the training phase: (1) inter-class feature entanglement during multi-food classes training and (2) loss of high-resolution detail during image downsampling. To address both issues, we propose to train one food category at a time to avoid feature entanglement and leverage image patches cropped from high-resolution datasets to retain fine details. We evaluate our method on the Food-101 dataset and show improved quality of generated synthetic food images compared with the baseline. Finally, we demonstrate the great potential of improving the performance of downstream tasks, such as food image classification by including high-quality synthetic training samples in the data augmentation.

Introduction

Healthy diet is one of the key factors for human wellness and disease prevention. There is a growing trend for people to track their dietary intake to adhere to or maintain a healthy diet. Traditional dietary assessment methods [1, 2] rely on manual self-reporting, which can be tedious and time-consuming. Image-based dietary assessment [3, 4] aims to develop automated methods to analyze consumed food types[5, 6], portion size [7, 8, 9] directly from captured eating occasion images. One of the major challenges of image-based dietary assessment is the lack of enough food images in existing datasets [10, 11] to train a robust deep learning model for food analysis. For example, the food recognition performance on less commonly seen food categories could drop significantly [12, 13] due to the few available training data. Many efforts have been made to solve the problem of lacking enough food images, such as for long-tailed classification [14] to address severe class-imbalance issue, continual learning [15, 16, 17] to learn from new data incrementally, and other food analysis scenarios [18, 19] that focus on real world food data distribution.

Generative network is widely applied as an effective data augmentation method to help address the issue of insufficient training data. Over the years, generative networks have been revolutionized from a basic autoencoder for reconstructing input data to a learning feature representation for creating non-existent objects. In recent years, the paradigms of the state-of-the-art generative models focus on three structures: Variational Autoencoders (VAEs) [20] (VDVAE [21] offers high image diversity), Diffusion models [22] (DDPM2 [23] offers advanced image quality and variety, but low sampling speed), and Generative Adversarial Networks (GANs) [24] (StyleGAN [25] offers good image quality and sampling speed). In general, GANs have been demonstrated to generate high-fidelity synthetic images efficiently.

Food image synthesis using GAN has been widely investigated such as CookGAN [26], built on a cycle-consistent network [27], RamenGAN [28], built on a standard conditional network [29], and multi-ingredients pizza generator [30], built on StyleGAN2 [31] have shown a decent performance on food image generation. However, the food images generated by these methods either do not provide sufficient details or contain many artifacts. Among existing GAN methods, StyleGAN3 [32] shows an ability to generate highly realistic images. In this work, we explore StyleGAN3 with its capability of generating food images corresponding to their labels.

Despite several improvements had been made in StyleGAN3, we discovered two issues could be addressed when generating synthetic food images: (1) inter-class feature entanglement (the generated image for a specific class contains features from other image classes) and (2) loss of high-resolution details during data normalization (e.g., image size rescaling and downsampling). Then, we propose two training strategies to address these issues, including single-class training to avoid features being correlated between different classes, and image-patches training on any-resolution data to avoid image normalization. We evaluate our proposed method on the Food-101 dataset [33] with the Frechet Inception Distance Metric (FID) [34] and a subjective survey to demonstrate the effectiveness in improving the visual resolution and fidelity of our generated food images. Finally, we use our synthetic food images as additional training images for training a food image classifier to explore the impact of data augmentation using synthetic images.

Preliminaries

The main idea of Generative Adversarial Networks[23] is to train a generator network (G) that maps the noise vectors to real training data distribution to create realistic image instances. Meanwhile, the discriminator model (D) attempts to distinguish the real data from generated samples via an estimated probability.

The two networks are optimized simultaneously during training.

Following the idea of GAN, StyleGAN3 [32] also uses a generator to generate synthetic images and a discriminator to distinguish real from synthetic samples. However, the generator network is made more complex. Instead of directly feeding the noise vector to the generator, StyleGAN3 goes through a mapping network to reduce correlation among different features during training. With different combinations of style (feature) information learned from the network, StyleGAN3 has a synthesis framework, which composes 14 layers to collect and generate coarse and fine styles sequentially to generate high-quality synthetic images. The improvements of StyleGAN3 also include solving the feature adhesion under coarse layers and making the generation process invariant to image translation and rotation.

In order to generate synthetic food images corresponding to their class label, we investigate conditional image generation where the image class labels are supervised during training. The StyleGAN3 network controls the generation of image class from two basic parts: the mapping network in the generator and the discriminator network. The mapping network conditions the latent code with a one-hot label vector which defines a set of specific characteristics from a certain class for the generator to study, while the discriminator is trained to classify real and generated data conditioned on their class labels. Therefore, with a feature vector to control the image's underlying content spatial structure, the generator can generate synthetic images for specific classes.

Proposed Methods

The proposed methods aim to improve the training of StyleGAN3 for generating realistic synthetic food images. The first method involves training with a single-class food dataset to avoid feature entanglement, while the second method involves training with any-resolution data to capture fine-grain details in high-resolution images. These approaches have the potential to address specific challenges in training and enhance the performance of StyleGAN3.

Training StyleGAN3 with a single-class at a time. According to the results of training StyleGAN3 on low-resolution multi-class food datasets in the Experiment section, we find that even though the conditional StyleGAN3 model is trained to stabilize and converged based on the FID metric evaluation on generated synthetic images, the results of generated synthetic food images still look unnatural and the reason of artificial-looking and distorted synthetic images are caused by inter-class feature entanglement (e.g., Figure 4 shows that synthetic hamburger images include features from spring roll). Either features in different classes are not well-distributed in the mapping network, or the discriminator could not classify the real and synthetic images into their perspective classes due to complex and similar features learned in different classes. To avoid features being correlated and affecting each other, we trained StyleGAN3 with a single-class food dataset one at a time to avoid feature entanglement.

Training StyleGAN3 with any-resolution data. After analyzing the results from StyleGAN3, we found that this baseline method has a few drawbacks. To train a network, the input images have to be fixed at certain resolutions, such as 256×256, 512×512, or 1024×1024. This requirement could lead to image warping and loss of image details when downsampling the input images from high-resolution to the required low-resolutions. To avoid losing fine-grain details in high-resolution images during downsampling, we adopt the method from Anyres GAN [35] to project and capture previously discarded high-resolution image details.

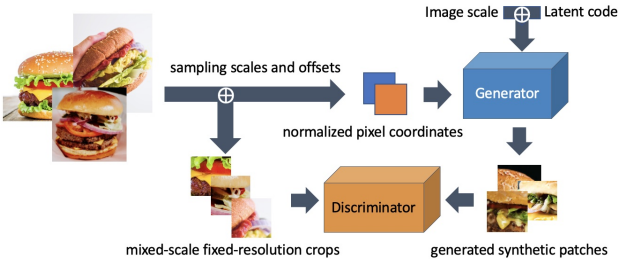

Figure 1: **Training Architecture of Anyres GAN.**

More specifically, Anyres GAN [35] includes two stages of generator training: global fixed-resolution pretraining and mixed-resolution patch-based training. In the first stage, the network follows the standard training procedure of StyleGAN3, which is trained on 256×256 resolution of single food class images (e.g., images from hamburger class) to capture the global structure of the given training images. This model is then used as a teacher model during the second stage of training. To better learn the fine-grain details for synthetic images, we randomly crop square-shaped patches from the same class of high-resolution images with any resolutions in the second stage of training. These images include synthetic food images generated by our pretrained StyleGAN3 and high-resolution images scrapped from Google at various random scales and locations. The generator takes three inputs: normalized pixel coordinates of our sampled patches, the original image resolution which the patches are created from, and the latent code z representing the underlying features of the original image for the generator to produce synthetic square patches.

The discriminator compares the synthetic patches with real patches to help the generator for obtaining fine details in generated image patches. The generated patches are then adjusted to match the teacher's global fixed-resolution output after proper downsampling and alignment. In the end, the patch features are projected into global fixed low-resolution images to obtain fine details in those high-resolution images. The training architecture of the Anyres GAN is illustrated in Figure 1.

Experiments

In this section, we evaluate our proposed method by conducting different experiments on low-resolution multi-class datasets, low-resolution single-class datasets, and any-resolution single-class food datasets. In addition, we perform both objective and subjective tests to show the perceptual visual realism of our generated food images. Finally, we demonstrate the effectiveness of using synthetic data as data augmentation to improve the performance of food image classification.

Datasets

Low-resolution Multi-class and Single-class Food Dataset We construct a dataset with ten random food classes selected from the Food-101 dataset [33] for evaluating the baseline conditional StyleGAN3. These ten food classes include cannoli, cupcake, donut, hamburger, pancake, strawberry, shortcake, pizza, spring

roll, panna cotta, and waffle images. Each class contains 1,000 images, and each image has a maximum resolution of 512 pixels and a minimum resolution of 384 pixels. We pre-process the images to a dimension of 256×256 to meet the input image resolution requirement of StyleGAN3 and refer to this dataset as the low-resolution food dataset (LR). In contrast, the low-resolution single-class Food Dataset consists of only hamburger food images from Food-101 downsampled to 256×256.

Any-resolution Dataset for Anyres Training. We use 600 high-resolution hamburger images scraped from Google as part of our selected dataset for training the second stage of any-resolution GAN. In this dataset, the minimum side length is 512 pixels, the maximum side length is 5,472 pixels, the mean side length is 1,250.06 pixels, and the median side length is 1,000 pixels. We combine the images from the Food-101 dataset and the high-resolution hamburger images from Google to form the selected Any-resolution dataset where all images have a resolution greater than 256. During image-patches training, image patches with 256 resolution are cropped from both Any-resolution and LR datasets.

Evaluation Metrics

The Frechet Inception Distance (FID) [34] is a commonly used metric to evaluate the similarity between the distribution of real and synthetic images. The lower the FID scores, the more realistic of generated images are. The FID metric is been shown to be computationally efficient and consistent with human assessment of synthetic image discrimination [36]. In our experiment, we compute the FID for every five training epochs based on the saved model and sample results. In addition, we also conducted a subjective study to qualitatively evaluate our model, where we ask 82 adult participants to assess the perceptual realism of our synthetically generated food images.

Results on Low-resolution Multi-class Food Datasets

During conditional StyleGAN3 training, we calculate the FID score to evaluate the network's performance and report the lowest FID metrics. As shown in Figure 2, we select the synthetic hamburger images as a representative class to show that the FID metric effectively evaluates the efficiency of the StyleGAN3 network and the visual quality of the generated image. We record the FID score for every 20 iteration of training until the network converged at the score of 17.348.

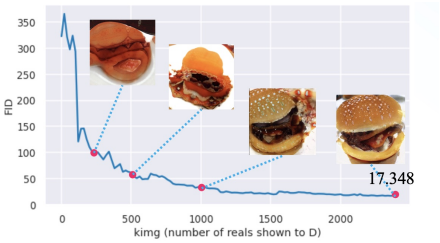

Figure 2: **Evaluation of Synthetic Multi-Class Food Images on conditional StyleGAN3**

Figure 3 shows some example of synthetic food image resulted from conditional StyleGAN3. However, the generated images do not look realistic and contain obvious visual artifacts.

The most obvious artifact we notice in those generated images is the inter-class feature entanglement. For example, in Figure 4, the small disc-shape feature appears in the synthetic hamburger images, but it should only appear in the spring roll food images. This issue can be resolved when we only train one class at a time.

Figure 3: **Example of conditional synthetic images results from global fixed resolution**

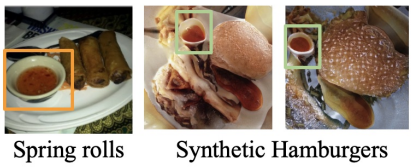

Figure 4: **An Illustration of Inter-class Feature Entanglement Issue in Synthetic Hamburger Images**

Results on Low-resolution Single-class Food Datasets

Figure 5 shows the comparison results of synthetic hamburger images between multi-class and single-class trained on StyleGAN3. Without inter-class feature entanglement, our generator only captures in-class features and the synthetic hamburger image results are much more realistic compared to the baseline of training with multiple food classes.

Figure 5: **Sample Synthetic Hamburger images from Baseline and Improved Methods**

Similar to the multi-class training, we train the network for about 2,000 iterations for the network to converge at 17.295. With the trained StyleGAN3 model on hamburger image samples, we use it as our pretrained model for our next phase of any-resolution training. Results are shown in Figure 6. Although the FID score is similar to training on multi-class food images, the visual artifacts are significantly reduced.

Results on Any-resolution Single-class Food Datasets

To avoid image warping and loss of high-resolution details during image normalization using StyleGAN3's fixed resolution, we train square-shaped image patches cropped from any-resolution datasets. Following the two-phase training of any-resolution dataset, we train the StyleGAN3 model with 256×256 low-resolution hamburger images from Food-101 for the first stage of StyleGAN3 pretraining and our HR dataset for second stage image-patches training. Figure 7 shows the comparison of

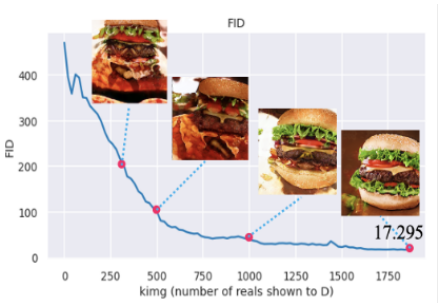

Figure 6: **Evaluation of Synthetic Hamburger Images on StyleGAN3**

the synthetic hamburger images between the first improvement on StyleGAN3 with single-hamburger class training and the second improvement with our any-resolution training. The visual quality of hamburgers is further improved, and details are better preserved.

Quantitative Results

As shown in Table 1, we calculate the standard FID metric for the first and second improvements made to the StyleGAN3 training strategy. The standard FID metrics between training using single-class and image-patches are similar, despite the obvious visual improvement from image-patches training. This is because the standard FID metric assumes all the training images are of 256×256 resolution, which ignores the fine-grain details in the training dataset. Thus, the standard FID scores is not suitable for evaluating our image-patches training results. Instead, we adopt the patch-FID (pFID) metric, which extracts 50K image patches cropped from our Any-resolution dataset at various scales and locations. To avoid downsampling the training images, it computes the FID score on the generated patches and real patches with corresponding scales and locations. The pFID score in Table 1 confirms our observation that with image-patches training, the food images contain details and are visually more realistic.

Table 1: FID and patch-FID Metric Evaluation on Two methods at 256 Image Resolution

Improvement Methods	FID	pFID
Train with Single-class Dataset	17.871	90.113
Train with any-resolution Dataset	17.723	**30.863**

Subjective Study

We conduct a subjective study to assess the perceptual realism of synthetically generated food images to qualitatively evaluate our conditional synthetic food image generation model. This subjective measure is an important complement to the Frechet Inception Distance (FID). The synthetic food image should look realistic so that it can be used for downstream tasks such as food image classification as training examples. In the survey, 82 adult participants were asked to evaluate 88 food images which contain 51 synthetic food images and 37 real food images. The synthetic images are evenly distributed among three classes — hamburger, pizza, and spring roll. For real food images, we select 12 images of hamburgers, 12 images of pizzas, and 13 images of spring rolls. Participants were asked to select images that looked real to them (*i.e.*, did not look synthetic) and were asked to look at each image for no more than 3 seconds. We also set a scoring system to evaluate our model performance — 51 is the full score since they are 51 synthetic images, and participants received one point for selecting the synthetic image.

Figure 7: **Comparison between two improved methods on StyleGAN3**

On average, participants scored 33.02 out of 51, which means that they mistook 64.75% of the generated synthetic food images as real images. Every synthetic food image has at least twenty-five participants who thought it is real. Among the 17 synthetic pizza images, on average, participants selected 45.65% of them as real images. Among the 17 synthetic hamburger images, on average, participants selected 38.29% of them as real images. Among the 17 synthetic spring roll images, on average, participants selected 52.17% of them as real images. From our survey results, more than half of our generated synthetic images are realistic enough to make participants select them as real images. We can conclude from the subjective study that our proposed methods can effectively learn realistic features from the real samples and may be good enough to be used as representative training examples for downstream tasks, such as food image recognition when there is a lack of training images.

Impact of Using Synthetic Images on Food Classification

Most deep learning-based methods require a large number of training data which can be challenging for many applications. Synthetic images that closely resemble real ones could be a potential resolution to address this problem. In order to assess the effect of using synthetic images as part of training data, we design experiments to explore the impact of synthetic images as data augmentation for the food image classification task.

Our experiment aims to classify food images from three different classes (hamburger, pizza, and spring roll). The images we used for training are either randomly picked from the Food-101 dataset (LR dataset) or high-resolution food images from the Any-resolution dataset. We consider 3 experimental setups as described below.

1. We train the ResNet-50 with only 200 real food images (100 from LR and 100 from the Any-resolution dataset).
2. We train the same model with 200 real images from the first experiment and an extra 200 of our generated synthetic images.
3. We train the same model with the same 200 real images from experiment one and an extra 200 real images (100 from LR and 100 from the Any-resolution dataset). (Upper bound)

All the three experiments use the same testing set containing 100 images (50 from LR and 50 from the Any-resolution dataset). We apply ResNet-50 as the backbone and keep the training settings the same with a batch size of 64, and training epoch around 100 for all three experiments to ensure a fair comparison.

Figure 8 shows the comparison of the best training, validation, and testing accuracy results from three different experiments. As expected from those plotted figures, Experiment 1 has the overall lowest accuracy, *i.e.* 62.33%, since it uses the least amount

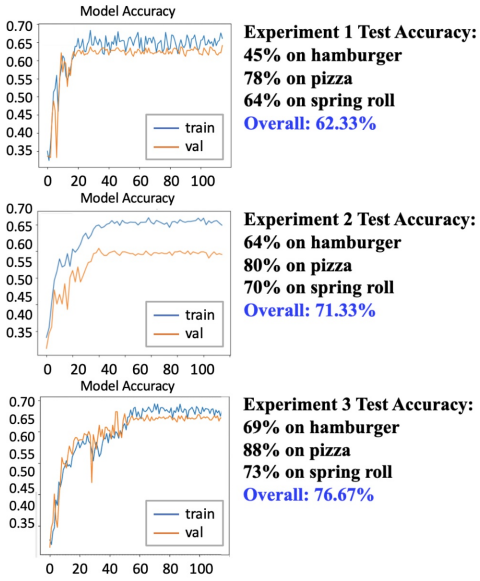

Figure 8: **ResNet-50 Training (Blue) and Validation (Orange) Accuracy for Three Different Experiments**

of training data. Experiment 2 achieves 71.33% accuracy and greatly improves the model by almost 10% by training with the additional synthetic images. Finally, Experiment 3 (upper bound) has the highest testing accuracy, 76.67%. As observed from the plotted figure of Experiment 2, the validation accuracy is lower than the training accuracy, which is caused by different data distribution between training and testing datasets since our generated synthetic images still contain unnatural artifacts. Nonetheless, our preliminary experiments demonstrate that using synthetic images to augment datasets is effective in improving the model's performance on the food image classification task. Furthermore, compared to pizza, hamburgers and spring rolls have lower accuracy due to their more complex and dynamic features, which is still difficult for our food image generation model to produce good feature representation. This is consistent with our visual observation that the synthetic images of hamburgers and spring rolls are less realistic than pizzas. Overall, we show that the generated synthetic images are realistic enough to be used as training samples when real data is scarce and can greatly improve the performance of a deep learning model (in our case food classification).

Conclusion

In this paper, we propose an improved conditional synthetic food image generation based on the StyleGAN3 baseline method. The first improvement uses single-class training instead of multi-class to avoid the inter-class feature entanglement. Next, we leverage square-shaped image patches training to retain high-resolution details in our generated images as opposed to a fixed resolution input. With our improved methods, our synthetic food image generation results are more realistic and contain more details compared to the baseline method. In addition to the quantitative evaluation of our proposed method, we conduct a subject study to qualitatively assess the perceptual realism of generated synthetic images. On average, participants mistaken 64.75% of the generated synthetic food images as real images. To show the impact of synthetic images for downstream tasks, we conducted a set of experiments where synthetic images were used to augment training data for the food image classification task using a ResNet-50 model. Results show significant improvement in classification accuracy. Our future work will focus on developing a multi-label training strategy to generate multiple food classes in a single image and apply it to other vision tasks such as food image localization and volume estimation.

References

[1] M Barbara E Livingstone, PJ Robson, and JMW Wallace, "Issues in dietary intake assessment of children and adolescents," *British journal of nutrition*, vol. 92, no. S2, pp. S213–S222, 2004.

[2] Kamila Poslusna, Jiri Ruprich, Jeanne HM de Vries, Marie Jakubikova, and Pieter van't Veer, "Misreporting of energy and micronutrient intake estimated by food records and 24 hour recalls, control and adjustment methods in practice," *British Journal of Nutrition*, vol. 101, no. S2, pp. S73–S85, 2009.

[3] Zeman Shao, Yue Han, Jiangpeng He, Runyu Mao, Janine Wright, Deborah Kerr, Carol Jo Boushey, and Fengqing Zhu, "An integrated system for mobile image-based dietary assessment," *Proceedings of the 3rd Workshop on AIxFood*, p. 19–23, 2021.

[4] Fengqing Zhu, Marc Bosch, Insoo Woo, SungYe Kim, Carol J Boushey, David S Ebert, and Edward J Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *IEEE journal of selected topics in signal processing*, vol. 4, no. 4, pp. 756–766, 2010.

[5] Runyu Mao, Jiangpeng He, Zeman Shao, Sri Kalyan Yarlagadda, and Fengqing Zhu, "Visual aware hierarchy based food recognition," *arXiv preprint arXiv:2012.03368*, 2020.

[6] Runyu Mao, Jiangpeng He, Luotao Lin, Zeman Shao, Heather A. Eicher-Miller, and Fengqing Zhu, "Improving dietary assessment via integrated hierarchy food classification," *2021 IEEE 23rd International Workshop on Multimedia Signal Processing*, pp. 1–6, 2021.

[7] Zeman Shao, Shaobo Fang, Runyu Mao, Jiangpeng He, Janine Wright, Deborah Kerr, Carol Jo Boushey, and Fengqing Zhu, "Towards learning food portion from monocular images with cross-domain feature adaptation," *arXiv preprint arXiv:2103.07562*, 2021.

[8] Jiangpeng He, Zeman Shao, Janine Wright, Deborah Kerr, Carol Boushey, and Fengqing Zhu, "Multi-task image-based dietary assessment for food recognition and portion size estimation," *2020 IEEE Conference on Multimedia Information Processing and Retrieval*, pp. 49–54, 2020.

[9] Jiangpeng He, Runyu Mao, Zeman Shao, Janine L Wright, Deborah A Kerr, Carol J Boushey, and Fengqing Zhu, "An end-to-end food image analysis system," *Electronic Imaging*, vol. 2021, no. 8, pp. 285–1, 2021.

[10] Zeman Shao, Jiangpeng He, Ya-Yuan Yu, Luotao Lin, Alexandra Cowan, Heather Eicher-Miller, and Fengqing Zhu, "Towards the creation of a nutrition and food group based image database," *arXiv preprint arXiv:2206.02086*, 2022.

[11] Yue Han, Sri Kalyan Yarlagadda, Tonmoy Ghosh, Fengqing Zhu, Edward Sazonov, and Edward J Delp, "Improving food detection for images from a wearable egocentric camera," *Electronic Imaging*, vol. 33, pp. 1–7, 2021.

[12] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng, "Deep long-tailed learning: A survey," *arXiv preprint arXiv:2110.04596*, 2021.

[13] Cheng Zhang, Tai-Yu Pan, Tianle Chen, Jike Zhong, Wenjin Fu, and Wei-Lun Chao, "Learning with free object segments for long-tailed instance segmentation," pp. 655–672, 2022.

[14] Jiangpeng He, Luotao Lin, Heather Eicher-Miller, and Fengqing Zhu, "Long-tailed food classification," *arXiv preprint arXiv:2210.14748*, 2022.

[15] Jiangpeng He, Runyu Mao, Zeman Shao, and Fengqing Zhu, "Incremental learning in online scenario," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13926–13935, 2020.

[16] Jiangpeng He and Fengqing Zhu, "Online continual learning for visual food classification," *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 2337–2346, October 2021.

[17] Siddeshwar Raghavan, Jiangpeng He, and Fengqing Zhu, "Online class-incremental learning for real-world food classification," *arXiv preprint arXiv:2301.05246*, 2023.

[18] Shuqiang Jiang, Weiqing Min, Yongqiang Lyu, and Linhu Liu, "Few-shot food recognition via multi-view representation learning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 3, pp. 1–20, 2020.

[19] Xinyue Pan, Jiangpeng He, Andrew Peng, and Fengqing Zhu, "Simulating personal food consumption patterns using a modified markov chain," *Proceedings of 7th International Workshop on Multimedia Assisted Dietary Management*, p. 61–69, 2022.

[20] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[21] Rewon Child, "Very deep vaes generalize autoregressive models and can outperform them on images," *arXiv:2011.10650*, 2020.

[22] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," pp. 2256–2265, 2015.

[23] Alexander Quinn Nichol and Prafulla Dhariwal, "Improved denoising diffusion probabilistic models," pp. 8162–8171, 2021.

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[25] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," pp. 4401–4410, 2019.

[26] Fangda Han, Ricardo Guerrero, and Vladimir Pavlovic, "Cookgan: Meal image synthesis from ingredients," 2020.

[27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," pp. 2223–2232, 2017.

[28] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," pp. 2223–2232, 2017.

[29] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[30] Fangda Han, Guoyao Hao, Ricardo Guerrero, and Vladimir Pavlovic, "Mpg: A multi-ingredient pizza image generator with conditional stylegans," *arXiv preprint arXiv:2012.02821*, 2020.

[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," pp. 8110–8119, 2020.

[32] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 852–863, 2021.

[33] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool, "Food-101 – mining discriminative components with random forests," 2014.

[34] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[35] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang, "Any-resolution training for high-resolution image synthesis," *arXiv preprint arXiv:2204.07156*, 2022.

[36] Ali Borji, "Pros and cons of gan evaluation measures," *Computer Vision and Image Understanding*, vol. 179, pp. 41–65, 2019.

Author Biography

Wenjin Fu received her B.S. degree with Cum Laude honor from The Ohio State University in December 2022. She is currently studying in Electrical and Computer Engineer MS program at Carnegie Mellon University. Her research interests are in Artificial Intelligence applications specifically in computer vision, software development, and autonomy for mobile.

Yue Han received his B.S degree with distinction from Purdue University in 2019. He is currently pursuing a Ph.D. degree at Purdue University and working as a research assistant in the Video and Image Processing (VIPER) Laboratory at Purdue University. His research interests include image processing, computer vision, and deep learning.

Jiangpeng He received his Ph.D. degree in Electrical and Electronic Engineering from Purdue University in August 2022. He is currently a postdoc research assistant at the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. His research interests include image processing, computer vision, image-based dietary assessment and deep learning.

Sriram Baireddy is a Ph.D. candidate in Electrical Engineering at Purdue University. He earned his B.S. and M.S. degrees in Electrical Engineering at Purdue in 2018 and 2021, respectively, with minors in economics, math, and physics. He currently investigates the application of machine learning techniques to signals, images, and videos for forensic and agricultural research.

Mridul Gupta received his B.Tech degree in Civil Engineering from Indian Institute of Technology Roorkee in 2017. He is currently a Ph.D. candidate in Video and Image Processing Lab (VIPER) advised by Prof. Edward J. Delp at Purdue University. His research focuses on applying deep learning and machine learning tools to problems in computer vision and image processing.

***Fengqing Zhu** is an Associate Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, Indiana. She received the B.S.E.E. (with highest distinction), M.S. and Ph.D. degrees in Electrical and Computer Engineering from Purdue University. She is the recipient of an NSF CISE Research Initiation Initiative (CRII) award in 2017, a Google Faculty Research Award in 2019, and an ESI and trainee poster award for the NIH Precision Nutrition workshop in 2021. She is a senior member of the IEEE.*