Shape-Preserving Generation of Food Images for Automatic Dietary Assessment

Guangzong Chen Zhi-Hong Mao Mingui Sun Kangni Liu Wenyan Jia
University of Pittsburgh
{guangzong, zhm4, drsun, connie.liu, wej6}@pitt.edu

Abstract

Traditional dietary assessment methods heavily rely on self-reporting, which is time-consuming and prone to bias. Recent advancements in Artificial Intelligence (AI) have revealed new possibilities for dietary assessment, particularly through analysis of food images. Recognizing foods and estimating food volumes from images are known as the key procedures for automatic dietary assessment. However, both procedures required large amounts of training images labeled with food names and volumes, which are currently unavailable. Alternatively, recent studies have indicated that training images can be artificially generated using Generative Adversarial Networks (GANs). Nonetheless, convenient generation of large amounts of food images with known volumes remain a challenge with the existing techniques. In this work, we present a simple GAN-based neural network architecture for conditional food image generation. The shapes of the food and container in the generated images closely resemble those in the reference input image. Our experiments demonstrate the realism of the generated images and shape-preserving capabilities of the proposed framework.

1. Introduction

Nutrition plays a pivotal role in maintaining health, influencing both our daily well-being and long-term health status. A balanced diet can foster overall wellness, while unhealthy eating habits can lead to a range of health problems, such as diabetes, heart disease, obesity, stroke, and certain types of cancers [9, 21, 52, 68]. Therefore, accurate dietary assessment is a critical component in keeping healthy and treating chronic diseases [35, 64].

Traditional self-reported dietary assessment methods include 24-hour dietary recall (24HR), dietary records, and food frequency questionnaires (FFQ) [4, 22, 48, 54, 64]. All these methods necessitate individuals to report their food consumption, detailing the type and/or volume of food consumed. However, such a process can be time-consuming, cumbersome, and biased, since it relies heavily on self-reporting. Individuals tend to report healthier food choices while neglecting unhealthy items. The reliance on self-reporting introduces potential inaccuracies in capturing a comprehensive and precise picture of an individual's dietary habits [45, 55].

Food images can be conveniently acquired by wearable devices or smartphones, and thus image-assisted dietary assessment has attracted research interest and been extensively investigated. The integration of artificial intelligence (AI), especially deep learning networks, in analyzing food images has markedly advanced the automation of dietary assessment [17, 34, 58]. Developing AI algorithms for dietary assessment requires a substantial collection of labeled images covering a wide range of food types and volumes for effective training. Manual labeling becomes necessary to fulfill this demand, which is a laborious and time-intensive task. While several datasets with large amounts of food images are currently available, there is still a need for training images to recognize foods in specific countries and regions [6, 13, 33, 63]. Additionally, many existing food image datasets lack annotations for food volume, making them unsuitable for dietary assessment. To address these challenges, generative models have been proposed to synthesize images, thereby augmenting training image datasets. Generative Adversarial Networks (GANs) have demonstrated a powerful ability in several areas of image generation, such as super-resolution image generation, image inpainting, and image semantic editing. Although training images can be artificially generated using GANs, it is difficult for existing techniques to maintain both food shapes and image quality simultaneously, significantly affecting the accuracy of dietary assessment. In response, we propose a simple GAN-based network architecture. Our experiments indicate that the new form of GAN can not only generate realistic food images but also preserve the food shape in the reference image. The proposed approach can significantly enhance the performance of AI-based dietary assessment systems by generating training images for both food recognition and volume estimation. It will thus offer an effective, efficient, and scalable solution to overcome the current limitations in automatic dietary assessment.

The major contributions of this paper are twofold. First, we present a straightforward GAN architecture for realistic food image translation. Second, we demonstrate that, in the generated images, it is convenient to control food categories and preserve food shapes using style and category variables.

2. Related Works

2.1. Automatic Dietary Assessment

In the field of automated image-based dietary assessment, identifying and quantifying food nutrition, particularly recognizing their types and estimating volume, poses considerable challenges due to the complicated visual characteristics of various foods and the absence of reference scales in images [1, 34, 36, 51]. Traditional food recognition relies on the extraction of image features, such as the scale-invariant feature transform (SIFT) and the histogram of oriented gradients (HOG), followed by classification using a classifier like the support vector machine (SVM) [6, 14, 27, 34, 41]. However, the classification accuracy is low, and the algorithm is difficult to develop [6, 14, 27]. Recently, with the rapid development of deep learning, deep networks and strategies such as fine-tuning and transfer learning have been effectively employed in the analysis of food images, leading to unprecedented levels of accuracy of food recognition [2, 26, 34, 40, 43, 50, 61, 70, 73]. For training and evaluating the deep network, food image datasets, such as Food-101, and UEC-Food, have been constructed using online sources (e.g., Google Images, Flickr) or collected images for different types of cuisines or specifically controlled environments [5, 6, 16, 18, 32, 34, 44, 47, 69]. Currently, the application of state-of-the-art methodologies to these datasets has achieved an impressive accuracy rate [61]. For example, the EfficientNet-B7 network achieves 93% accuracy in the Food-101 dataset ([61]) and the ensemble method averaging the predictions of ResNeXt and DenseNet models reaches 90.02% in the UEC-Food100 dataset [34].

The challenge in calculating the volume of food from a single image is primarily attributed to the absence of three-dimensional (3D) information inherent in a two-dimensional (2D) image [34, 36, 60, 62]. Previous studies mostly rely on model-based techniques [10, 12, 19, 30, 56]. After a calibration procedure using a reference object with a known size (e.g., a checkerboard, credit card) to determine the camera's location and orientation, a pre-defined shape model is chosen for each food item to match the contour of the food and estimate its volume. However, this procedure is labor-demanding in most cases since manual operations are required, and estimating the volume of irregularly shaped food can be challenging [19, 30, 56]. Recently deep neural networks are expected to automatically learn the scale information of a 2D image from the global cures in the image and use it for volume estimation. Yang et al. propose a novel human-mimetic AI system to virtually gauge the volume of food using a set of internal reference volumes, mimicking the thinking of dietitians who mentally use a standard measuring tool (e.g., cup) as a reference [71]. Several studies employed convolutional neural networks (CNNs) to estimate a depth map or 3D shape (represented by voxels) corresponding to the input food image and obtain volumetric information [15, 20, 46, 53]. In most of these studies, the training images were created by the research group themselves, either manually labeled [20, 71] or captured with a depth sensor [15, 46]. However, obtaining training images with labeled food volume/calorie or depth map is a tedious task. Thus, large-scale food image databases with known volume/nutrient information have not yet been developed.

2.2. Food Image Generation

It is well known that the quantity and quality of images in the training set play a critical role in the performance and generalization ability of deep networks. Therefore, data augmentation techniques (such as random crop, rotation, translation, flip, and rescaling) have been proposed to expand training datasets. To further increase the diversity of images, GANs have proven to be invaluable tools [23, 37, 66]. GANs introduce a novel approach to image generation by training a generator network to produce realistic images that are indistinguishable from real ones, while a discriminator network learns to differentiate between real and generated images.

Several GAN-based structures have been proposed to generate images from a list of ingredients/recipes or reference images [29, 49, 57, 74]. The Multi-ingredient Pizza Generator (MPG) is a conditional GAN framework based on StyleGAN2 designed to generate pizza images with desired ingredients[24]. CookGAN combines an attention-based recipe association model and StackGAN to generate meal images from ingredients [74]. ChefGAN, RDE-GAN, and other related works integrate an image-recipe embedding module into GANs structure to synthesize dish images [49, 57, 65].

RamenGAN uses a conditional GAN to generate ramen images after training with a ramen image dataset [29]. arCycleGAN introduces the mechanism of attribute registration into CycleGAN to transfer the freshness styles from the style-offering images to the input images [11]. DuDGAN improves class-conditional GANs to control the output image using an additional classifier trained with a diffusion-based noise injection process [72]. TransferI2I explores several novel techniques to implement image-to-image translation with limited data labeled data for two-class and multi-class translation tasks [67]. TUNIT is a truly unsupervised image-to-image translation model that simultaneously learns to separate image domains and translates input images into the estimated domains [3]. Besides the GAN structure, diffusion models have also been introduced recently to generate food images [25, 39].

Although promising results have been demonstrated in these studies, currently, recipe-image pairs are only available in the Recipe1M+ dataset [38], and the volumes of the foods in the images generated from the recipe are unknown since they cannot be controlled. Thus in this work, we focus on image-to-image translation approaches designed to produce food images with the volumetric information. We aim to estimate the food volume from a single image, which is a projection of food in 3D. Therefore, the volume of the food is preserved, if the shape and depth map of the projection are unchanged. In our case, we assume that a small set of training images with known volumes exists but its size is insufficient for training. Our goal is to increase the size of this small training set by including new image samples which are created by replacing the foods in the existing images with numerous other foods. As a result, the combined set of images, which may be very large, can then be used to train deep neural networks for volume estimation.

In addition to preserving contours, maintaining the shapes of food containers is equally important since containers serve as references for estimating food volumes. In doing so, the realism of the generated images is enhanced. While CycleGAN, among various GAN structures, can maintain shapes in the generated images, retraining is necessary for each new class of images and this procedure is inefficient. It requires extra training to solve container distortion by introducing a discriminator to identify whether a dish plate observed exhibits a correct round shape, and the results are often not satisfactory [29]. A mask-based image synthesis network has been proposed to ensure a reasonable plate shape in generated images, but images with segmented plate regions are necessary [28]. We propose a simple network to generate diverse, high-quality images while preserving the shapes of both the food and the container of a given dish in the reference image.

3. Methods

We develop a neural network architecture for image generation with specific object constraints. Our goal is to generate an image that retains the same object shape as the given reference image, while the textures are determined by a latent variable. This variable enables the creation of diverse food images with identical shapes. The "shape" in this context includes both the shapes of the food and food container. We also use a category label as a conditional variable to control the object category of the generated image. The architecture of our network is illustrated in Fig. 1.

The network includes three parts. The first is an encoder, which compresses the input image into features. The second part is a generator, which takes the features and a latent variable as inputs, generating an image. The third part is a discriminator, which is used to distinguish between the real and generated images.

Compared with regular GANs, our proposed network architecture includes a shape encoder. The encoder is necessary for shape learning. Our model is remarkably compact, comprising only a single generator and a discriminator without the need for additional components. Given one shape image, multiple food images can be generated from our model. The generated images can be utilized to estimate the food volume in future research. The type of food can be specified through conditions, allowing the images to be used for training a food recognition network. As is widely acknowledged, defining image attributes, especially shapes, is challenging. Shape information cannot be accurately represented by just a few feature variables, making it impractical to use a classifier for defining shape features. Alternatively, we employ the encoder to extract shape information directly from images.

Two datasets are used for training. The first image dataset, $\mathcal{I}^{\mathrm{s}}$, is used as food shape references, where $\mathcal{I}^{\mathrm{s}}$ equals $\{I_i^{\mathrm{s}} \,|\, i = 1, \dots, N\}$, $I_i^{\mathrm{s}} \in \mathbb{R}^{H \times W \times 3}$, H and W are the height and width of the images, 3 is the number of channels of an RGB image, and N is the total number of images. The second image dataset, $\mathcal{I}_{\mathrm{t}}$, is used to provide the food texture information, where $\mathcal{I}^{\mathrm{t}}$ equals $\{I_i^{\mathrm{t}} \,|\, i = 1, \dots, M\}$, $I_i^{\mathrm{t}} \in \mathbb{R}^{H \times W \times 3}$, and M the size of the second dataset. The "textures" mainly encompass various aspects such as the color of the material, grain size, condensed state, and other detailed characteristics of the food. This dataset is provided to the discriminator, D, to train the network. We want to apply $\mathcal{I}^{\mathrm{s}}$ to facilitate the network to generate images with the same shapes as the images in $\mathcal{I}^{\mathrm{s}}$ while maintaining the textures from $\mathcal{I}^{\mathrm{t}}$.

3.1. Functions of Network Components

Encoder. The input images, I^{s}, are compressed by the encoder to extract essential features. These features mainly contain the topological information of an image. The encoder also helps to reduce the resolution of the shape images, leading to a more compact network structure.

The encoder structure is shown in Fig. 2. It consists of downsampling layers and convolutional layers. As the shape features of an image often encode global and topological information, which is of "low-frequency" nature, we apply two downsampling layers to reduce the resolution. This is followed by a stack of three convolutional network blocks, each containing a convolutional layer, a ReLU activation layer, and a downsampling layer.

We use E to represent the encoder model. The encoder takes an image I^{s} from the shape dataset as input and outputs a feature vector, which is set of feature maps $f = E(I^{\mathrm{s}}) \in \mathbb{R}^{H' \times W' \times C'}$, where H' and W' are the height

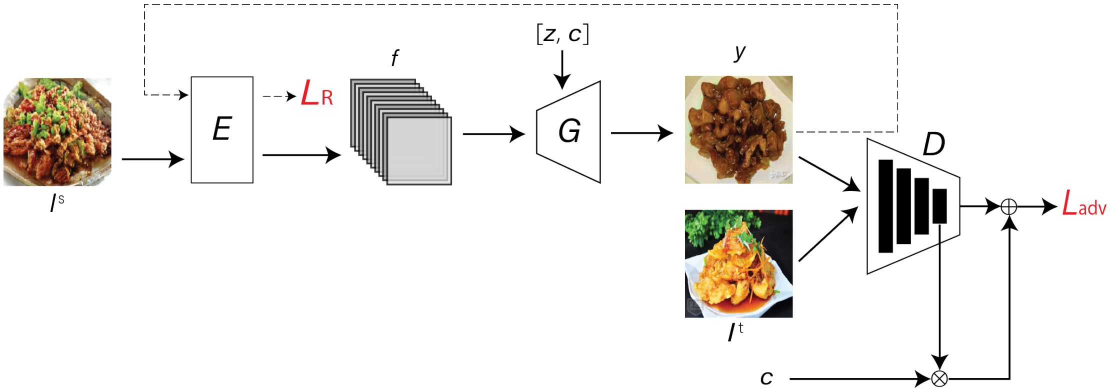


Figure 1. Our network architecture includes three major components, encoder E, generator G, and discriminator D. The encoder produces shape-related features f from the image I^{s}. The generator takes features f, latent variable z, and category label c as conditional inputs and create output image y. The discriminator is used to evaluate the realism of the output image. Loss functions L_{adv} and L_{R} are used for training the network.

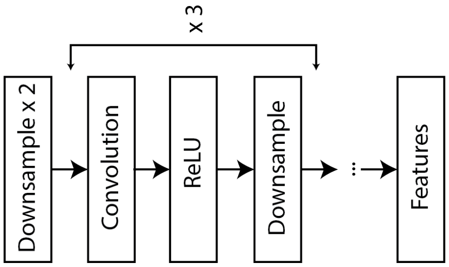


Figure 2. The network structure of the encoder.

and width of a feature map and C' is the number of feature maps. We set $H' = W' = 16$ and $C' = 128$ in the experiment.

In the subsequent stages, the feature vector f provides constraints for generated images, playing a pivotal role in defining the overall shapes of the objects within the generated images.

Generator. The generator's primary role is to create images that adhere to the constraints derived from the shape features. The generator is designed with three inputs: the shape feature vector f, extracted by the encoder, the latent variable z, and the category label c. The shape feature vector f primarily determines the shape of the object. The latent variable z, which is sampled from a Gaussian distribution, influences the texture of the generated image. The category label c determines the image's class. The separated input ports of f, z, and c are essential for isolating the shape, texture, and category of the generated image.

Let G denote the generator model of the proposed network. The output image $y = G(f, z, c)$ is determined by the shape feature vector $f = E(I^{\mathrm{s}})$ (output of the encoder), z (latent variable), and c the category. The dimension of y is the same as that of I^{s}.

Discriminator. For the discriminator D, a conditional discriminator with class embedding is used, which is similar to BigGAN [7]. The discriminator takes the generated image y and an image sample I^{t} from the texture dataset $\mathcal{I}^{\mathrm{t}}$ as inputs and provides metrics for realism evaluation, which are $D(y, c_y) \in \mathbb{R}^1$, and $D(I^{\mathrm{t}}, c_{I^{\mathrm{t}}}) \in \mathbb{R}^1$. The learning process for the discriminator is to detect the differences between the generated images and real images.

3.2. Network Training

Training is performed in two alternating stages. One stage is to train the encoder and generator, and the other is to train the discriminator. Different loss functions are applied in different stages. We apply a reconstruction loss and GAN loss when training the encoder and generator. The reconstruction loss ensures the same shape features are shared by the input and generated images. The GAN loss function and R_1 [42] regularization are applied to train the discriminator. We adopt the GAN loss from [23] given by

$$\begin{aligned} L_{\mathrm{adv}} =& \mathbb{E}_{I^{\mathrm{t}}}[\log(D(I^{\mathrm{t}}, c_{I^{\mathrm{t}}}))] \\ &+ \mathbb{E}_{I^{\mathrm{s}},z,c}[1 - \log(D(G(E(I^{\mathrm{s}}), z, c), c))] \end{aligned} \tag{1}$$

where $\mathbb{E}$ means expectation. The L_1 loss is used as our reconstruction loss:

$$L_{\mathrm{R}} = \mathbb{E}_{I^{\mathrm{s}},z,c}[\|E(I^{\mathrm{s}}) - E(G(E(I^{\mathrm{s}}), z, c))\|_1]. \tag{2}$$

Our learning problem is to solve

$$\min_{E,\, G} \max_{D} \quad L_{\mathrm{adv}} + \lambda L_{\mathrm{R}} \tag{3}$$

where λ is a hyper-parameter indicating the relative weight of the reconstruction loss with respect to the GAN loss.

4. Experiments

4.1. Realism Evaluation of Generated Food Images

The primary objective of the first experiment is to demonstrate the capability of our method to generate realistic food images. The quality of the generated images was quantitatively evaluated using the FID (Frechet Inception Distance) [8], a widely used metric for assessing the fidelity of the generated images. To validate the effectiveness of our approach, we also conducted a comparative analysis with StyleGAN3 [31].

Datasets. A Chinese food image dataset VireoFood-172 [13] and a Western food image dataset Food-101 [6] were used to evaluate the performance of our approach. The VireoFood-172 dataset encompasses 172 distinct classes of Chinese food, with each class featuring between 300 and 1000 images. In total, the VireoFood-172 dataset comprises 110,241 images. Most images in this dataset contain a food item in a container (e.g., plate, bowl). Whether the shape of the container can be preserved was also studied in this experiment. The Food-101 dataset contains 101,000 images of 101 different food classes. As this dataset has been used by other researchers for food image generation, we also evaluated our approach with the Food-101 dataset for comparison with other approaches. Before training the network, the images in both datasets were resized to 256 × 256 pixels to improve computational efficiency.

Evaluation Metric. We used FID as the metric to evaluate the quality of the generated images. FID measures the discrepancy between the features of the generated and real images. These features are extracted using the Inception network [59]. The computation of FID involves comparing the distributions of these features as derived from the Inception network. A lower FID value signifies higher realism in the generated images. At the extreme, a zero value of FID indicates a perfect match in the distribution of the generated and real data, implying that the generated images are indistinguishable from the real.

Results. Fig. 3 displays some image examples created by our network, which was trained by the VireoFood-172 dataset. We selected five random images as inputs, with the first column showing these inputs and the subsequent columns presenting the outputs generated by our network. These outputs were created by combining the same input image in each row with different style variables, indicated by z. This resulted in a notable change in textures, yielding highly realistic food visuals.

To calculate the FID values of the generated images, we randomly selected $30,000$ images as input images. For each input image, we generated one output image and calculated the FID value based on the training dataset and the $30,000$ generated images. The result is shown in Table 1. The FID value of our method is 4.97. To benchmark against StyleGAN3[31], we ran the StyleGAN3 algorithm using the VireoFood-172 dataset. The model trained 100 epochs ($10,000$K images), utilizing the default hyperparameters. Image examples generated by StyleGAN3 are shown in Fig. 4. We randomly generated $30,000$ images using StyleGAN3 and then calculated the FID value based on these generated images. The FID value of StyleGAN3 for the VireoFood-172 dataset is 9.25 as shown in Table 1.

Method	FID
StyleGAN3 [31]	9.25
Ours	4.97

Table 1. Comparison of FID on the VireoFood-172 Dataset.

Method	FID
StyleGAN3 [25]	39.05
Finetuned Latent Diffusion [25]	30.39
ClusDiff [25]	27.73
Ours	22.82

Table 2. Comparison of FID among various food image generation models on the Food-101 dataset.

From Fig. 3, we can see that the structural integrity of the images generated by our network is consistent across different styles. This consistency proves the ability of our method to generate diverse food images while adhering to fixed shape constraints. On the contrary, it can be observed from Fig. 4(b) that the shapes of the containers generated by StyleGAN3 are quite irregular and unpredictable.

The FID value of the $30,000$ generated images when using the Food-101 dataset as the training set is listed in Table 2. For comparison, the FID values for other models using the same dataset [25] are also included in this Table. It shows that our model achieves the lowest FID value, 22.82.

4.2. Evaluation of Shape Preservation Performance

In this subsection, we evaluated the shape-preservation performance of the proposed GAN architecture (Fig. 1). Here the segmentation images were employed by the training data for the network to learn the texture. The images generated by the network only contain foods. Using segmentation images simplifies the evaluation of the network's performance in maintaining accurate shapes.

Dataset. We used the segmented food images in the UEC-FoodPIX dataset [47] for shape-preservation evaluation. This dataset is particularly well-suited for our study as it includes segmentation information for a variety of food items. This allows us to quantitatively assess how well our network preserves the food shape. The UEC-FoodPIX dataset comprises 120 food classes and a total of $9,000$ im-

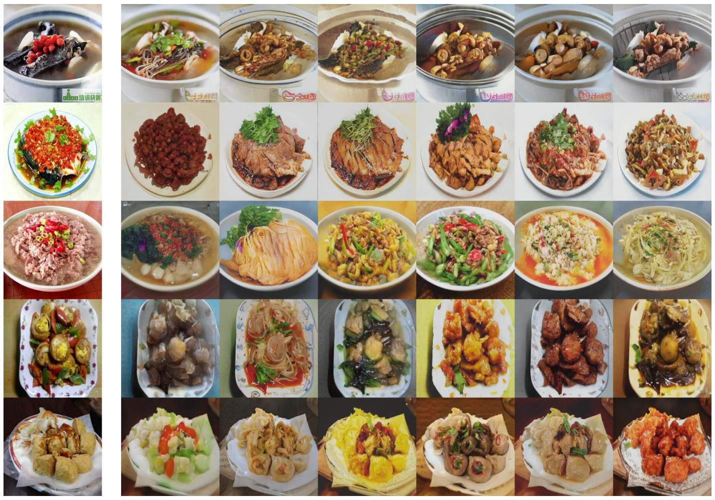
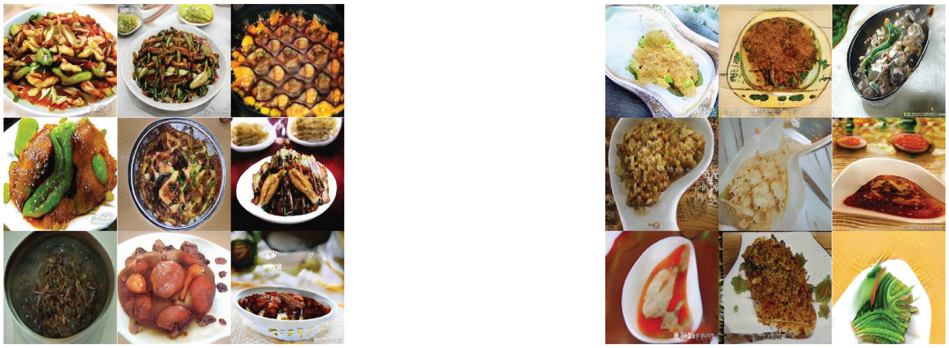


Figure 3. Image examples generated by our network using VireoFood-172 dataset: The first column shows the original input images, and subsequent columns display images created by varying the latent variable z while keeping the corresponding input image from the first column fixed.



(a)

(b)

Figure 4. Image examples generated by StyleGAN3: (a) with round-shaped containers and (b) with irregular-shaped containers.

ages. Each image in the dataset may contain more than one type of food, and the resolutions vary. In pre-processing, we extracted the image of each food item from the original image according to the provided segmentation mask. Then we resized each image to 256×256 pixel resolution for training.

Evaluation Metric. Our network is specifically designed to preserve the shape of food in the input image while changing the food category in the generated images. We use the Intersection Over Union (IoU) metric to evaluate how well food shapes are preserved. IoU is a widely used metric in image processing, specifically object detection, for quantifying the degree of overlap between two areas. To calculate the IoU, we segmented the generated images. A high IoU score signifies effective shape preservation, implying that the network is proficient in replacing the food in the input image with another category of food while maintaining the shape.

Results. Fig. 5 presents some image examples generated by our network. The first column shows the original input images. Subsequent columns display images created by varying the latent variable z while keeping the corresponding input image from the first column fixed.

Fig. 6 displays the IoU scores achieved by our network. We selected five random images as inputs and generated eight output images for each. We calculated the IoU score for each generated image. In Fig. 6, each color corresponds to the IoU scores for the same input image, providing a clear visual representation of the network's performance in terms of shape consistency across multiple outputs. In general, all IoU scores are above 0.8, and the average IoU score of the eight images is 0.91.

4.3. Category Control of the Generated Images

Generating images without conditional constraints (as in Fig. 3) can increase the diversity of the generated images, thus they are perfectly suitable for the purpose of volume estimation. However, they are not suitable for image recog-

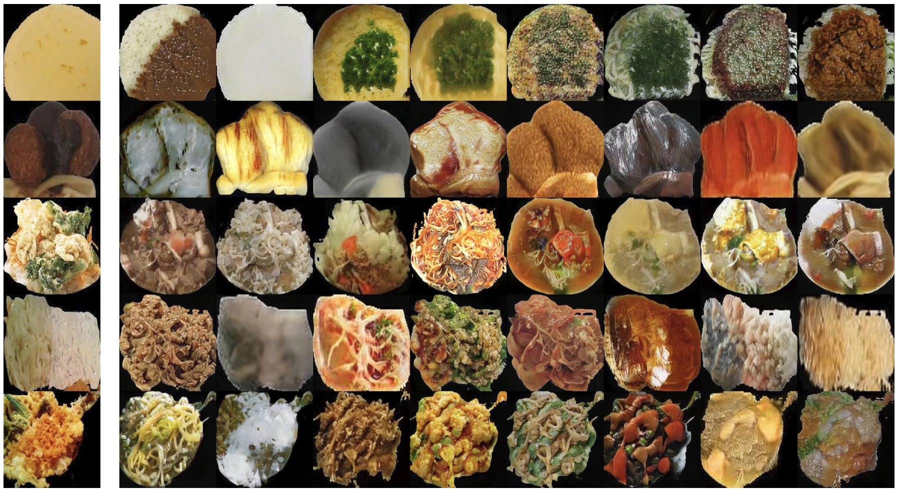


Figure 5. Image examples generated by our model: The first column is the input images, and the rest are generated images.

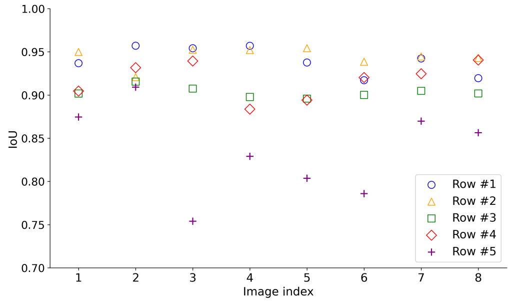


Figure 6. IoU of the generated images shown in Fig. 5.

Category No.	2	10	132	148	150	169
FID	19.57	24.09	50.78	51.12	44.16	39.68

Table 3. FID values of the generated images of six food categories. Image examples of each category are shown in Fig. 7

nition since they do not have labels for food categories. Therefore, to generate images for food recognition, we applied a conditional generator and a conditional discriminator to control the category of the generated images. The output category can be explicitly controlled by variable c.

The VireoFood-172 dataset was applied in the experiment. Fig. 7 shows the results. Three random images in one category were selected as the input, which are shown in the first column. Subsequent images in each column are created by different category labels c with the same input image (in the first row) and latent variable z. The FID values for one thousand generated images in each category are presented in Table 3. The FID values across the whole dataset had also been calculated by generating thirty thousand images with random input images, variable z, and category label c. This value turns out to be 5.18. On the other hand, the FID value of each single category is greater than that of the whole dataset, as shown in Table 3. It may be caused by the small number of images in these categories, the feature distributions may not be accurately estimated from a small set of data.

By manually selecting the desired image categories for generation, the problem of mismatching between the food and container can be avoided. Fig. 8 illustrates the generated images when the containers of the input and the output are inconsistent. The input is a plate of fried vegetables, and the output is a bowl of porridge. It is impossible to keep the shape of a plate when transferring between these two kinds of foods. In addition, the volume of the food in the bowl cannot be assumed to be close to the volume of the food on the plate. These issues may be solved by providing a plate as the input and excluding categories where the food is typically served only in bowls from the generation process.

4.4. Implementation Details

Our model was implemented by PyTorch. The encoder was self-built based on the structure described in Section 3. In our experiments, the dimensions of images in the dataset were different. To accommodate these differences, the number of blocks in the encoder and generator was adjusted while the resolution of features f was fixed. We used the Adam optimizer to train the network, with the learning rate set to 0.0001 for the first 100 epochs and then reduced to 0.00001 for the remaining 150 epochs. The generator and discriminator were trained with a batch size of 64. In Sections 4.1 and 4.2, the category label c is set to "None." In Section 4.3, c is a one-hot vector which encodes the category label. In the experiments, the hyper-parameter λ was set to 50.

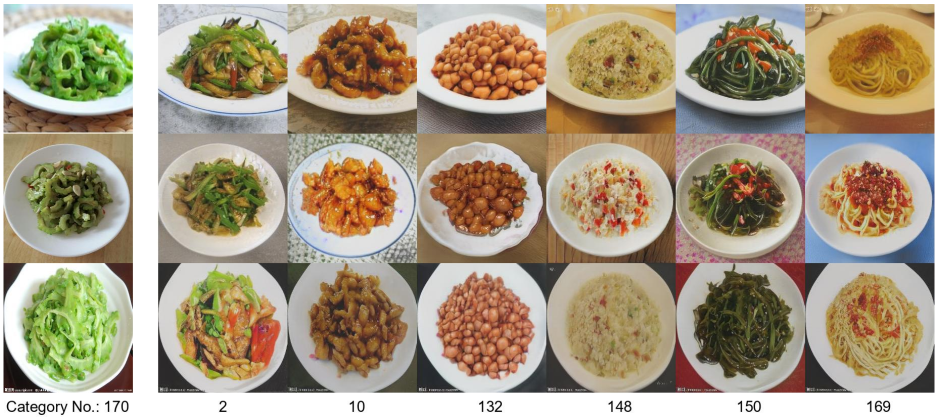


Figure 7. Image examples generated by our model: The first column is the input images, and the rest are generated by our model. Generated images in each column correspond to the same food category, which is controlled by the category label c.

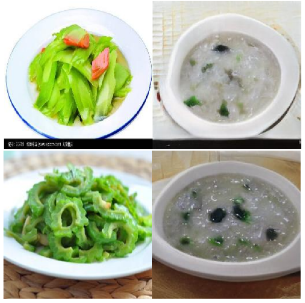
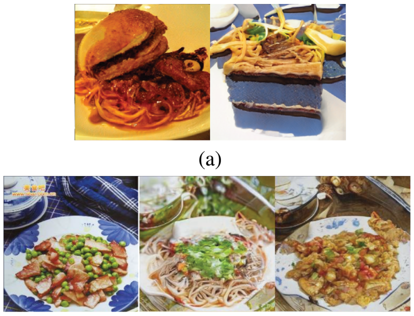


Figure 8. Image examples when the containers of the input and the output are inconsistent. The first column is the input images, and the second column is the output images, where the fried vegetables are substituted with porridge. The containers for the vegetables (i.e., plate) and the porridge (i.e., bowl) are not matched.



(a)

(b)

Figure 9. (a) Image examples of combining incompatible ingredients. (b) Images examples of misinterpreting plate patterns as food.

5. Discussions

In our experiments, the foods in the randomly generated images (i.e., without condition control) sometimes may not correspond to real-world foods. Multiple ingredients were randomly mixed to construct a dish, such as noodles in a burger and chips on a cake, as shown in Fig. 9(a). While these "strange" foods can still be used for volume estimation, it is challenging to assign categories to these foods. Occasionally, the decoration pattern on the plate can be mistakenly recognized as food, causing the food shape in the generated images to extend into the plate area, as illustrated in Fig. 9(b).

These issues might be attributed to inaccurate learning of the network. Expanding the dataset with food images containing various cuisines, appearances, and patterned food containers can enhance training. However, image augmentation may be unnecessary when diverse real-world food images are already available. Currently, a practical approach is deliberately selecting shape reference images and controlling generated image categories, though this may limit the diversity of the generated images. Conducting iterative training sessions with feedback from human annotators to continuously fine-tune the model holds the potential for enhancing the network's performance over time.

6. Conclusion

Our method can generate new images after training with a given image food dataset. The content and volume/shape of the food can be controlled separately. The textures (food category) of the generated image can be controlled by style variable z or category label c, and the shape of the generated image can be constrained by the reference image and the encoder. The generated images are suitable for training deep networks for food recognition and volume estimation, which overcomes the lack of training data in automated dietary assessment.

Acknowledgment. This work was supported in part by the U.S. National Institutes of Health [Grants No. R56 DK113819 and No. R01DK127310] and the Bill Melinda Gates Foundation [OPP1171395].

References

[1] Birdem Amoutzopoulos, Polly Page, Caireen Roberts, Mark Roe, Janet Cade, Toni Steer, Ruby Baker, Tabitha Hawes, Catherine Galloway, Dove Yu, and Eva Almiron-Roig. Portion size estimation in dietary assessment: a systematic review of existing tools, their strengths and limitations. *Nutrition Reviews*, 78(11):885–900, 2020. 2

[2] Berker Arslan, Sefer Memis, Elena Battini Sonmez, and Okan Zafer Batur. Fine-grained food classification methods on the uec food-100 database. *IEEE Transactions on Artificial Intelligence*, 3(2):238–243, 2022. 2

[3] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. In *2021 IEEE/CVF International Conference on Computer Vision*. IEEE, 2021. 3

[4] Tom Baranowski. *24-Hour recall and diet record methods*, pages 49–69. Oxford University Press New York, NY, 2012. 1

[5] Elena Battini Sönmez, Sefer Memiş, Berker Arslan, and Okan Zafer Batur. The segmented UEC food-100 dataset with benchmark experiment on food detection. *Multimedia Systems*, 29(4):2049–2057, 2023. 2

[6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer International Publishing, 2014. 1, 2, 5

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 4

[8] Naresh Babu Bynagari. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Asian Journal of Applied Science and Engineering*, 8(1):25–34, 2019. 5

[9] Michele Cecchini, Franco Sassi, Jeremy A Lauer, Yong Y Lee, Veronica Guajardo-Barron, and Daniel Chisholm. Tackling of unhealthy diets, physical inactivity, and obesity: health effects and cost-effectiveness. *The Lancet*, 376(9754): 1775–1784, 2010. 1

[10] Junghoon Chae, Insoo Woo, SungYe Kim, Ross Maciejewski, Fengqing Zhu, Edward J. Delp, Carol J. Boushey, and David S. Ebert. Volume estimation using food specific shape templates in mobile image-based dietary assessment. In *SPIE Proceedings*. SPIE, 2011. 2

[11] Guangzong Chen, Wenyan Jia, Yifan Zhao, Zhi-Hong Mao, Benny Lo, Alex K. Anderson, Gary Frost, Modou L. Jobarteh, Megan A. McCrory, Edward Sazonov, Matilda Steiner-Asiedu, Richard S. Ansong, Thomas Baranowski, Lora Burke, and Mingui Sun. Food/Non-Food classification of real-life egocentric images in low- and middle-income countries based on image tagging features. *Frontiers in Artificial Intelligence*, 4, 2021. 2

[12] Hsin-Chen Chen, Wenyan Jia, Yaofeng Yue, Zhaoxin Li, Yung-Nien Sun, John D Fernstrom, and Mingui Sun. Model-based measurement of food portion size for image-based dietary assessment using 3D/2D registration. *Measurement Science and Technology*, 24(10):105701, 2013. 2

[13] Jingjing Chen and Chong wah Ngo. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM International Conference on Multimedia*. ACM, 2016. 1, 5

[14] Mei-Yun Chen, Yung-Hsiang Yang, Chia-Ju Ho, Shih-Han Wang, Shane-Ming Liu, Eugene Chang, Che-Hua Yeh, and Ming Ouhyoung. Automatic chinese food identification and quantity estimation. In *Proceedings of SIGGRAPH Asia 2012 Technical Briefs*, New York, NY, USA, 2012. ACM. 2

[15] Patrick Ferdinand Christ, Sebastian Schlecht, Florian Ettlinger, Felix Grün, Christoph Heinle, Sunil Tatavatry, Seyed-Ahmad Ahmadi, Klaus Diepold, and Bjoern H. Menze. Diabetes60 — inferring bread units from food images using fully convolutional neural networks. In *Proceedings of 2017 IEEE International Conference on Computer Vision Workshops*, pages 1526–1535, 2017. 2

[16] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. Food recognition: a new dataset, experiments, and results. *IEEE Journal of Biomedical and Health Informatics*, 21(3):588–598, 2017. 2

[17] Kalliopi V Dalakleidi, Marina Papadelli, Ioannis Kapolos, and Konstantinos Papadimitriou. Applying image-based food-recognition systems on dietary assessment: a systematic review. *Advances in Nutrition*, 13(6):2590–2619, 2022. 1

[18] Takumi Ege, Wataru Shimoda, and Keiji Yanai. A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice. In *Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management*. ACM, 2019. 2

[19] Shaobo Fang, Chang Liu, Fengqing Zhu, Edward J. Delp, and Carol J. Boushey. Single-view food portion estimation based on geometric models. *ISM*, 2015:385–390, 2015. 2

[20] Shaobo Fang, Zeman Shao, Deborah A Kerr, Carol J Boushey, and Fengqing Zhu. An end-to-end image-based automatic food energy estimation technique based on learned energy distribution images: protocol and methodology. *Nutrients*, 11(4):877, 2019. 2

[21] Mohammad H Forouzanfar and Alexander et. al. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the global burden of disease study 2013. *The Lancet*, 386(10010):2287–2323, 2015. 1

[22] Rosalind S Gibson. *Principles of nutritional assessment*. Oxford University PressNew York, NY, 2005. 1

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 2, 4

[24] Fangda Han, Guoyao Hao, Ricardo Guerrero, and Vladimir Pavlovic. MPG: a multi-ingredient pizza image generator with conditional StyleGANs. Technical report, 2020. arXiv:2012.02821 [cs] type: article. 2

[25] Yue Han, Jiangpeng He, Mridul Gupta, Edward J. Delp, and Fengqing Zhu. Diffusion model with clustering-based con-

ditioning for food image generation. In *Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management*. ACM, 2023. 3, 5

[26] Hamid Hassannejad, Guido Matrella, Paolo Ciampolini, Ilaria De Munari, Monica Mordonini, and Stefano Cagnoni. Food image recognition using very deep convolutional networks. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, page 41–49, New York, NY, USA, 2016. ACM. 2

[27] Ye He, Chang Xu, Nitin Khanna, Carol J. Boushey, and Edward J. Delp. Analysis of food images: features and classification. In *Proceedings of 2014 IEEE International Conference on Image Processing*, pages 2744–2748. IEEE, 2014. 2

[28] Yuma Honbu and Keiji Yanai. Setmealasyoulike: sketch-based set meal image synthesis with plate annotations. In *Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management on Multimedia Assisted Dietary Management*. ACM, 2022. 3

[29] Yoshifumi Ito, Wataru Shimoda, and Keiji Yanai. Food image generation using a large amount of food images with conditional GAN: ramenGAN and recipeGAN. In *Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management*. ACM, 2018. 2, 3

[30] Wenyan Jia, Hsin-Chen Chen, Yaofeng Yue, Zhaoxin Li, John Fernstrom, Yicheng Bai, Chengliu Li, and Mingui Sun. Accuracy of food portion size estimation from digital pictures acquired by a chest-worn camera. *Public Health Nutrition*, 17(8):1671–1681, 2014. 2

[31] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *arXiv*, 2021. 5

[32] Yoshiyuki Kawano and Keiji Yanai. *Automatic expansion of a food image dataset leveraging existing categories with domain adaptation*, pages 3–17. Springer International Publishing, 2014. 2

[33] Yoshiyuki Kawano and Keiji Yanai. Offline 1000-class classification on a smartphone. In *Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2014. 1

[34] Fotios S. Konstantakopoulos, Eleni I. Georga, and Dimitrios I. Fotiadis. A review of image-based food recognition and volume estimation artificial intelligence systems. *IEEE Reviews in Biomedical Engineering*, 17:136–152, 2024. 1, 2

[35] Jessica R. L. Lieffers and Rhona M. Hanning. Dietary assessment and self-monitoring: With nutrition applications for mobile devices. *Canadian Journal of Dietetic Practice and Research*, 73(3):e253–e260, 2012. 1

[36] Frank Po Wen Lo, Yingnan Sun, Jianing Qiu, and Benny Lo. Image-based food classification and volume estimation for dietary assessment: a review. *IEEE Journal of Biomedical and Health Informatics*, 24(7):1926–1939, 2020. 2

[37] Sanbi Luo. A survey on multimodal deep learning for image synthesis: applications, methods, datasets, evaluation metrics, and results comparison. In *Proceedings of 2021 the 5th International Conference on Innovation in Artificial Intelligence*. ACM, 2021. 2

[38] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1M+: a dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):187–203, 2021. 3

[39] Olivia Markham, Yuhao Chen, Chi-en Amy Tai, and Alexander Wong. FoodFusion: a latent diffusion model for realistic food image generation. *arXiv preprint arXiv:2312.03540*, 2023. 3

[40] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. Wide-slice residual networks for food recognition. In *Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision*, pages 567–576. IEEE, 2018. 2

[41] Yuji Matsuda, Hajime Hoashi, and Keiji Yanai. Recognition of multiple-food images by detecting candidate regions. In *Proceedings of 2012 IEEE International Conference on Multimedia and Expo*, pages 25–30. IEEE, 2012. 2

[42] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *International Conference on Machine Learning*, pages 3481–3490. PMLR, 2018. 4

[43] Weiqing Min, Linhu Liu, Zhengdong Luo, and Shuqiang Jiang. Ingredient-guided cascaded multi-attention network for food recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, page 1331–1339, New York, NY, USA, 2019. ACM. 2

[44] Weiqing Min, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. ISIA food-500: a dataset for large-scale food recognition via stacked global-local attention network. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 393–401, New York, NY, USA, 2020. ACM. 2

[45] Alanna J Moshfegh, Donna G Rhodes, David J Baer, Theophile Murayi, John C Clemens, William V Rumpler, David R Paul, Rhonda S Sebastian, Kevin J Kuczynski, Linda A Ingwersen, Robert C Staples, and Linda E Cleveland. The US department of agriculture automated multiple-pass method reduces bias in the collection of energy intakes. *The American Journal of Clinical Nutrition*, 88(2):324–332, 2008. 1

[46] Austin Myers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin Murphy. Im2Calories: towards an automated mobile vision food diary. In *Proceedings of 2015 IEEE International Conference on Computer Vision*. IEEE, 2015. 2

[47] Kaimu Okamoto and Keiji Yanai. *UEC-FoodPix complete: a large-scale food image segmentation dataset*, pages 647–659. Springer International Publishing, 2021. 2, 5

[48] Rosa M Ortega, Carmen Pérez-Rodrigo, and Ana M López-Sobaler. Dietary assessment methods: dietary records. *Nutricion Hospitalaria*, 31(3):38–45, 2015. 1

[49] Siyuan Pan, Ling Dai, Xuhong Hou, Huating Li, and Bin Sheng. ChefGAN: food image generation from recipes. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 4244–4252, New York, NY, USA, 2020. Association for Computing Machinery. 2

[50] Arnav A Rajesh, Madhumita Raghu, and J Sangeetha. Fast food image recognition using transfer learning. In *2022 Fourth International Conference on Cognitive Computing and Information Processing*, pages 1–10. IEEE, 2022. 2

[51] Viprav B. Raju and Edward Sazonov. A systematic review of sensor-based methodologies for food portion size estimation. *IEEE Sensors Journal*, 21(11):12882–12899, 2021. 2

[52] Arthur Schatzkin, Amy F. Subar, Steven Moore, Yikyung Park, Nancy Potischman, Frances E. Thompson, Michael Leitzmann, Albert Hollenbeck, Kerry Grace Morrissey, and Victor Kipnis. Observational epidemiologic studies of nutrition and cancer: the next generation (with better observation). *Cancer Epidemiology, Biomarkers & Prevention*, 18 (4):1026–1032, 2009. 1

[53] Zeman Shao, Gautham Vinod, Jiangpeng He, and Fengqing Zhu. An end-to-end food portion estimation framework based on shape reconstruction from monocular image. In *Proceedings of 2023 IEEE International Conference on Multimedia and Expo*, pages 942–947. IEEE, 2023. 2

[54] Jee-Seon Shim, Kyungwon Oh, and Hyeon Chang Kim. Dietary assessment methods in epidemiologic studies. *Epidemiology and Health*, 36:e2014009, 2014. 1

[55] N Slimani, C Casagrande, G Nicolas, H Freisling, I Huybrechts, M C Ocké, E M Niekerk, C van Rossum, M Bellemans, M De Maeyer, L Lafay, C Krems, P Amiano, E Trolle, A Geelen, J H de Vries, and E J de Boer and. The standardized computerized 24-h dietary recall method EPIC - soft adapted for pan-european dietary monitoring. *European Journal of Clinical Nutrition*, 65(S1):S5–S15, 2011. 1

[56] Shamus P. Smith, Marc T. P. Adam, Grace Manning, Tracy Burrows, Clare Collins, and Megan E. Rollo. Food volume estimation by integrating 3D image projection and manual wire mesh transformations. *IEEE Access*, 10:48367–48378, 2022. 2

[57] Yu Sugiyama and Keiji Yanai. Cross-modal recipe embeddings by disentangling recipe contents and dish styles. In *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, 2021. 2

[58] Jamalia Sultana, Benzir Md. Ahmed, Mohammad Mehedy Masud, A. K. Obidul Huq, Mohammed Eunus Ali, and Mahmuda Naznin. A study on food value estimation from images: taxonomies, datasets, and techniques. *IEEE Access*, 11:45910–45935, 2023. 1

[59] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015. 5

[60] Ghalib Ahmed Tahir and Chu Kiong Loo. A comprehensive survey of image-based food recognition and volume estimation methods for dietary assessment. *Healthcare (Basel)*, 9 (12):1676, 2021. 2

[61] Mingxing Tan and Quoc Le. Efficientnet: rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 2

[62] Wesley Tay, Bhupinder Kaur, Rina Quek, Joseph Lim, and Christiani Jeyakumar Henry. Current developments in digital quantitative volume estimation for the optimisation of dietary assessment. *Nutrients*, 12(4):1167, 2020. 2

[63] Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, and Jack Sim. Nutrition5k: towards automatic nutritional understanding of generic food. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8903–8911, 2021. 1

[64] Frances E Thompson and Amy F Subar. *Dietary assessment methodology*, pages 5–48. Elsevier, 2017. 1

[65] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven C. H. Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11564–11573. IEEE, 2019. 2

[66] Lei Wang, Wei Chen, Wenjia Yang, Fangming Bi, and Fei Richard Yu. A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access*, 8: 63514–63537, 2020. 2

[67] Yaxing Wang, Hector Laria, Joost van de Weijer, Laura Lopez-Fuentes, and Bogdan Raducanu. TransferI2I: transfer learning for image-to-image translation from small datasets. In *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, pages 13990–13999. IEEE, 2021. 2

[68] Walter Willett. *Nutritional epidemiology*. Oxford University Press, 2012. 1

[69] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven C.H. Hoi, and Qianru Sun. A large-scale benchmark for food image segmentation. In *Proceedings of the 29th Acm International Conference on Multimedia*, page 506–515, New York, NY, USA, 2021. ACM. 2

[70] Keiji Yanai and Yoshiyuki Kawano. Food image recognition using deep convolutional network with pre-training and fine-tuning. In *Proceedings of 2015 IEEE International Conference on Multimedia & Expo Workshops*, pages 1–6, 2015. 2

[71] Zhengeng Yang, Hongshan Yu, Shunxin Cao, Qi Xu, Ding Yuan, Hong Zhang, Wenyan Jia, Zhi-Hong Mao, and Mingui Sun. Human-mimetic estimation of food volume from a single-view RGB image using an AI system. *Electronics*, 10 (13):1556, 2021. 2

[72] Taesun Yeom, Chanhoe Gu, and Minhyeok Lee. DuDGAN: improving class-conditional GANs via dual-diffusion. *IEEE Access*, pages 1–1, 2024. 2

[73] Heng Zhao, Kim-Hui Yap, Alex Chichung Kot, and Lingyu Duan. JDNet: a joint-learning distilled network for mobile visual food recognition. *IEEE Journal of Selected Topics in Signal Processing*, 14(4):665–675, 2020. 2

[74] Bin Zhu and Chong-Wah Ngo. CookGAN: Causality based text-to-image synthesis. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2020. 2