Foodverse: A Dataset of 3D Food Models for Nutritional Intake Estimation

Chi-en Amy Tai — Vision and Image Processing Lab, University of Waterloo
Yuhao Chen — Vision and Image Processing Lab, University of Waterloo
Matthew Keller — Vision and Image Processing Lab, University of Waterloo
Mattie Kerrigan — Vision and Image Processing Lab, University of Waterloo
Saeejith Nair — Vision and Image Processing Lab, University of Waterloo
Pengcheng Xi — National Research Council Canada
Alexander Wong — Vision and Image Processing Lab, University of Waterloo
Email: {amy.tai, yuhao.chen1, m6keller, makerrig, smnair, alexander.wong}@uwaterloo.ca, pengcheng.xi@nrc-cnrc.gc.ca

Abstract

77% of adults over 50 want to age in place today, presenting a major challenge of ensuring adequate nutritional intake. Recent advancements in machine learning and computer vision show promise of automated tracking methods, but require a large high-quality dataset to have accurate performance. Existing datasets comprise of 2D images with discretely sampled camera views, unrepresentative of the different angles and quality taken by older individuals. By leveraging view synthesis for 3D models, an infinite number of 2D images can be generated for any given viewpoint/camera angle. In this paper, we develop a methodology for collecting high-quality 3D models for food items with a particular focus on speed and consistency, and introduce Foodverse, a large-scale high-quality high-resolution multimodal dataset of 52 3D food models, in conjunction with their associated weight, food name, language description, and nutritional value. We also demonstrate 2D view synthesis using these 3D food models.

1 Introduction

The desire to age in place has grown immensely in the past decade with 77% of adults over 50 wanting to stay at home in 2021 [1]. However, one of the main challenges with aging in place is ensuring adequate food nutritional intake. It has been reported that one in four older adults that are 65 years or older are malnourished [2]. Given the direct link between malnutrition and decreased quality of life [3], there have been numerous studies conducted on how to efficiently track food nutritional intake.

While self-reporting methods such as food frequency questionnaires, food diaries, and 24-hour recall [4] are subject to substantial bias with errors of up to 400% for 24-hour recall [5], more technologically enhanced methods such as mobile phone applications [6, 7], digital photography [8], and personal assistants [9] are more time-consuming and may require trained personnel. Promising results have been shown by pairing technological methods with machine learning and computer vision algorithms [10, 11]. However, developing high performing machine learning methods requires a large-scale high-quality dataset.

Unfortunately, existing datasets comprise of 2D images with fixed or randomly selected camera views that are discretely sampled [10, 12–16]. These set views introduce bias in terms of how individuals take images with their camera which would affect the training and accuracy of the model. In addition, the majority of older individuals struggle with taking photos and hence, these discrete views are not representative of the food images that would be taken by aging in place individuals.

In order to create an effective model for nutritional intake tracking for aging in place, an assortment of images of different angles and quality should be obtained to train an automated model. Yet, the manual creation of a large-scale dataset would be time-consuming and it would be difficult to capture all potential angles and photo quality. On the other hand, 3D models allow for view synthesis as these models can be postprocessed to generate an infinite number of 2D images taken from any angle to reduce imbalance or bias towards a certain viewing angle.

In this paper, we develop a methodology for collecting quality 3D models for food items with a particular focus on speed and consistency, and introduce Foodverse, a large-scale high-quality high-resolution multimodal dataset of 52 3D food models, in conjunction with their associated weight, food name, language description, and nutritional value. We also demonstrate 2D view synthesis using these 3D food models.

2 Methodology

The two primary factors considered in the design of the data collection pipeline are speed and consistency. Speed is important to maximize the number of food models that can be collected in a feasible amount of time for a large-scale dataset. Likewise, consistency is also critical to minimize human interaction and likelihood of variation in collecting data so that the number of high-quality food models obtained is optimized.

Though it is now feasible to use automated wearable cameras, these devices have been found to be incredibly intrusive [17] and pose significant ethical ramifications [18]. Given that the main goal is convenient nutritional intake tracking for older individuals, recent advances in mobile phone applications [6, 7] demonstrate that nutritional intake tracking through mobile devices would be more convenient and accepted by older individuals. Subsequently, mobile devices are chosen for collecting images and specifically, the iPhone [19] was chosen as the primary image capturing device due to its popularity and quality camera resolution (though any phone with a suitable camera could be used too).

Fig. 1: Setup for the data collection process for an exemplar sushi piece.

To generate a quality 3D model of a food item, various 3D scanner applications were compared based on their review rating, exporting capabilities, and ease of usage. In addition to having a high review rating and a variety of model export formats, Polycam [20] also has a web interface with a shareable account for easy image input captured from multiple devices [21]. Hence, leveraging the Polycam app, 3D models of food items are generated from 2D images taken by the iPhone. Consequently, three main restrictions are imposed by using the Polycam app. First, at least 70% overlap between the photos is needed to produce quality 3D food models without holes or blurs. Second, a variety of angles of the food need to be captured to render a full model, and third, a maximum limit of

250 images is allowed for each food item.

To address the first restriction, an electric turntable with the default rotation speed of a full rotation in 24 seconds and a custom image taking script implemented using the built-in Shortcuts iPhone application is used to automatically collect consistent images of each food item in a short period of time whilst allowing for at least 70% overlap between the photos.

However, to meet the second limitation, a variety of angles need to be obtained for each food item. To ensure consistency between item captures, the camera angles and food 6D-poses collected for each food item should be the same. In experimenting with the number of camera angles, faster and more consistent performance is obtained using two iPhones set at two different angles compared to only one iPhone. Unfortunately, using two iPhones causes shadow interference in the image captures for each iPhone due to the lighting conditions in the room. In particular, the room has sparse fluorescent ceiling lights that are about 1 meter apart from each other. Therefore, we experimented with a variety of tripod layouts to discern the setup with the least amount of shadows on the turntable. As seen in Fig. 1, the setup for the data collection process has two iPhones on adjacent tripods with very specific tripod distances for each iPhone and low shadow interference on the exemplar sushi piece on the turntable.

In terms of the third main Polycam limitation, coordination between the number of photos taken and the combinations of the food item 6D-pose and the camera angle had to be determined. With a limit of 250 photos, the ideal scenario for data collection is to position the food in four different ways with two different camera angles. As such, the photo limit and the number of combinations leads to roughly 30 photos per food 6D pose-camera angle combination for a total of 240 images. Hence, as seen in the custom Shortcuts app in Fig. 2, the iPhones are configured to automatically take 30 consecutive photos. After taking 60 photos of the food item on one 6D-pose (30 photos per iPhone-tripod), the food item is rotated to another 6D-pose and the custom Shortcuts app is started again. Occasionally, due to the shape of some food items, four different food 6D-poses is infeasible. For example, the egg and cheese bite could not stand on its side without rolling when the turntable rotated. Thus, to ensure consistency between image captures, the number of camera angles is increased to compensate for the lower number of possible food 6D-poses as seen in Table 1.

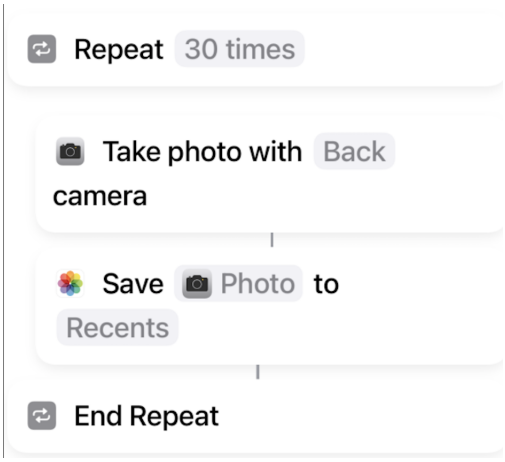


Fig. 2: Custom Shortcuts app used to take photos on the iPhones.

Table 1: Overview of the food 6D-poses and camera settings combinations in data collection to produce a total of 240 images for the first and last row and a total of 180 images for the second row.

Num of Food 6D-poses	Num of Camera Angles
2	4
3	2
4	2

Though the setup led to successful 3D model renderings, these models often had pieces of their background included in the model itself. To address this problem, the object masking feature in the Polycam app is used to remove background from the images and render only the food item. After conducting several experiments using plates with different textures or colours, it was determined that placing the food item on a white plate with low reflectivity and having a black tablecloth on top of the table rendered the most consistent quality of 3D models. Though the turntable has a white-ish colour, the food item is not placed directly on the electric turntable as cleaning the turntable is risky and hence, may result in irreparable damage.

The overall process to generate the 3D models of food items is shown in Fig. 3 with an example of a successful 3D model rendering displayed in Fig. 4. The total weight and protein weight of each food item is weighed using a food scale and the food name along with the language description is recorded for each food item. The nutritional value is obtained from the food packaging or from the Nutrient Value of Some Common Foods set posted on the Government of Canada website [22] for non-packaged food such as apples.

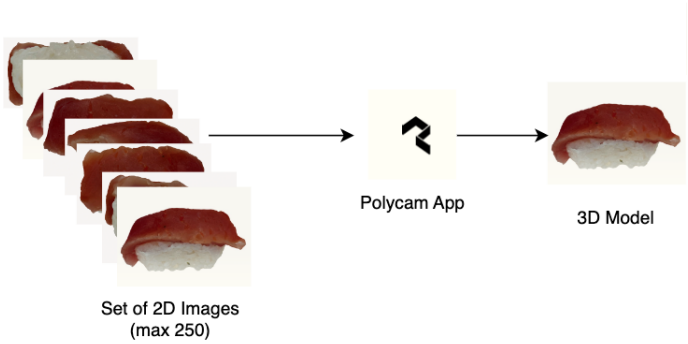
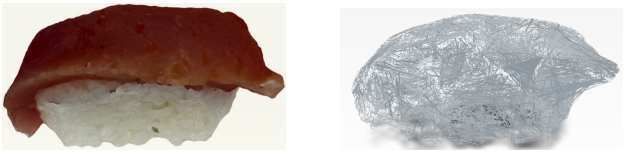


Fig. 3: Overall process to generate 3D models of food items.



(a) 3D Model Image *(b)* 3D Model Mesh

Fig. 4: Example of a successful 3D model Polycam rendering.

2.1 Item-Specific Challenges

In the collection of various food items, we quickly discovered that it is easier to render 3D models of certain types of food compared to others. Specifically, models for textureless food such as cheese, thin foods such as chips, and small items such as grapes often failed to render or rendered in an unrecognizable fashion. On the other hand, it is easier to generate 3D models of larger items with more texture such as chicken strips or a chicken wing. Yet, irrespective of texture or size, items that fall apart (have high fragility) throughout the entire data collection process also led to poor model renderings. Such an instance is the tuna rice ball. Though the 3D model for one tuna rice ball is successfully created, most of the tuna rice balls failed to capture as the tuna would slip or change shape when the sushi is flipped which resulted in a poor 3D model rendering. Thus, extra care had to be taken during data collection for fragile food items to ensure that a high-quality model could be captured. A generalized summary of properties that contribute to the success of a 3D model rendering along with examples is displayed in Table 2.

3 Foodverse Dataset

52 food models comprising of 20 unique types are created successfully using the pipeline proposed in Section 2 and are listed in Table 3. These models are saved in the OBJ and PLY file formats, two of the most widely used file formats for 3D models [23]. Saved along with the models are their associated weight, food name, language description, and nutritional value. Examples of a language description are "a piece of tuna on rice" and "cucumber wrapped with seaweed and rice". The total number of food models per category is

Table 2: Quantifiers and examples for various properties that contribute to a good quality (green) and poor quality (red) model rendering.

Property	Quantifier	Example
Texture	Low	Cheese Block
	High	Granola Bar
Volume	Low	Grape
	High	Apple
Thickness	Low	Potato Chip
	High	Salad Chicken Strip
Fragility	Low	Chicken Wing
	High	Tuna Rice Ball

Table 3: Tabular listing of all 52 collected food items.

Food Item Type	Number of Different Weights
Salad Chicken Strip	7
Salad Beef Strip	6
Nature Valley Granola Bar	3
Apple	2
Carrot	1
Cucumber Piece	2
Chicken Wing	1
Half Bread Loaf	1
Captain Crunch Granola Bar	1
Near Whole Chicken	1
Chicken Breast	2
Chicken Leg	2
Meatloaf	4
Asian Pear	1
Egg and Cheese Bite	1
Salad Sushi Roll	6
Cucumber Sushi Roll	1
Shrimp Sushi Roll	4
California Sushi Roll	5
Tuna Rice Ball	1

shown in Table 4 with mixed protein referring to food items (e.g., tuna rice ball) that contain almost equal amounts of protein and other categories such as carbohydrates. Roughly 4357 2D images of these images are also stored with the 3D models and descriptors.

Table 4: Count of 3D food models in each category.

Food Category	Total Count
Protein	23
Fruits	5
Vegetables	1
Carbohydrates	5
Mixed Protein	18

The main benefit of these 3D food models is that they allow for view synthesis. Examples of leveraging view synthesis with a 3D food model are shown in Fig. 5, Fig. 6, and Fig. 7 for a chicken leg, egg and cheese bite, and apple respectively. View synthesis is utilized in these figures as the postprocessed sample of generated 2D images includes angles of the food that were not captured in the initial data collection process. As a result, similar 2D images obtained by postprocessing 3D food models extend beyond the fixed camera angles used in the data collection process to reduce imbalance or bias towards a certain viewing angle.

4 Conclusion

In this paper, we introduced Foodverse, a large-scale high-quality, high-resolution multimodal dataset of 52 3D food models in conjunction with their associated weight, food name, language description, and nutritional value. The methodology to collect this dataset was also presented along with the encountered challenges to develop the pipeline. Leveraging the 3D models in the dataset, 3D food scenes can be generated and when coupled with automated view synthesis algorithms, an infinite number of 2D images can be obtained from any angle. Such an approach would allow for a more

(a) 3D Model Mesh *(b)* Generated 2D Images

Fig. 5: Postprocessed sample of 2D images obtained from 3D food model of a chicken leg.

(a) 3D Model Mesh *(b)* Generated 2D Images

Fig. 6: Postprocessed sample of 2D images obtained from 3D food model of an egg and cheese bite.

(a) 3D Model Mesh *(b)* Generated 2D Images

Fig. 7: Postprocessed sample of 2D images obtained from 3D food model of an apple.

representative and unbiased image dataset that can be used to develop an effective model for nutritional intake tracking for older adults.

5 Future Work

Further studies can be conducted using Foodverse to generate an assortment of 3D food scenes and an automated collection of a variety of 2D images from different angles, quality, and lighting condition. A major challenge with creating a food dataset is accounting for numerous dish combinations and layouts. Having 3D models of individual food items permits efficient swapping and the automated assembly of a variety of dish combinations. As an example, by having the individual pieces of a salad, one could assemble various types of salad without actually having to obtain and image each salad type. Furthermore, substitution of sides in a dish would be as easy as swapping the 3D food models. Such substitution could be easily automated using food categories and adding constraints to ensure realistic food renderings.

Acknowledgments

The authors thank National Research Council Canada and the Aging in Place (AiP) Challenge Program. The authors also thank their

partners in the Kinesiology and Health Sciences department Dr. Heather Keller, Dr. Sharon Kirkpatrick, and Meagan Jackson.

References

[1] M. R. Davis. (2021) Despite pandemic, percentage of older adults who want to age in place stays steady. [Online]. Available: https://www.aarp.org/home-family/your-home/info-2021/home-and-community-preferences-survey.html

[2] M. J. Kaiser, J. M. Bauer, C. Rämsch, W. Uter, Y. Guigoz, T. Cederholm, D. R. Thomas, P. S. Anthony, K. E. Charlton, M. Maggio, A. C. Tsai, B. Vellas, C. C. Sieber, and for the Mini Nutritional Assessment International Group, "Frequency of malnutrition in older adults: A multinational perspective using the mini nutritional assessment," *Journal of the American Geriatrics Society*, vol. 58, no. 9, pp. 1734–1738, 2010. [Online]. Available: https://agsjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.2010.03016.x

[3] H. H. Keller, T. Østbye, and G. Richard, "Nutritional risk predicts quality of life in elderly community-living canadians," *The Journals of Gerontology: Series A*, vol. 59, no. 1, p. M68–M74, 2004. [Online]. Available: https://academic.oup.com/biomedgerontology/article/59/1/M68/533583

[4] A. F. Subar, S. I. Kirkpatrick, B. Mittl, T. P. Zimmerman, F. E. Thompson, C. Bingley, G. Willis, N. G. Islam, T. Baranowski, S. McNutt, and N. Potischman, "The automated self-administered 24-hour dietary recall (asa24): A resource for researchers, clinicians, and educators from the national cancer institute," *Journal of the Academy of Nutrition and Dietetics*, vol. 112, no. 8, pp. 1134–1137, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212267212005898

[5] S. A. Bingham, "Limitations of the various methods for collecting dietary intake data," *Annals of Nutrition and Metabolism*, vol. 35, no. 3, p. 117–127, 1991. [Online]. Available: https://www.karger.com/Article/Abstract/177635

[6] "A mobile phone app intervention targeting fruit and vegetable consumption: The efficacy of textual and auditory tailored health information tested in a randomized controlled trial," *Journal of Medical Internet Research*, vol. 18, no. 6, p. e147, 2016. [Online]. Available: https://www.jmir.org/2016/6/e147

[7] W. Zhang, Q. Yu, B. Siddiquie, A. Divakaran, and H. Sawhney, ""snap-n-eat": Food recognition and nutrition estimation on a smartphone," *Journal of Diabetes Science and Technology*, vol. 9, no. 3, pp. 525–533, 2015, pMID: 25901024. [Online]. Available: https://doi.org/10.1177/1932296815582222

[8] D. A. Williamson, H. R. Allen, P. D. Martin, A. J. Alfonso, B. Gerald, and A. Hunt, "Comparison of digital photography to weighed and visual estimation of portion sizes," *Journal of the American Dietetic Association*, vol. 103, no. 9, p. 1139–1145, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000282230300974X

[9] A. Rusu, M. Randriambelonoro, C. Perrin, C. Valk, B. Álvarez, and A.-K. Schwarze, "Aspects influencing food intake and approaches towards personalising nutrition in the elderly," *Journal of Population Ageing*, vol. 13, p. 239–256, 2020. [Online]. Available: https://link.springer.com/article/10.1007/s12062-019-09259-1

[10] G. Ciocca, P. Napoletano, and R. Schettini, "Food recognition: A new dataset, experiments, and results," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 588–598, 2017.

[11] Y. Ando, T. Ege, J. Cho, and K. Yanai, "Depthcaloriecam: A mobile application for volume-based foodcalorie estimation using depth cameras," in *Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management*, ser. MADiMa '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 76–81. [Online]. Available: https://doi.org/10.1145/3347448.3357172

[12] P. Kaur, K. Sikka, W. Wang, S. J. Belongie, and A. Divakaran, "Foodx-251: A dataset for fine-grained food classification," *CoRR*, vol. abs/1907.06167, 2019. [Online]. Available: http://arxiv.org/abs/1907.06167

[13] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, 2012.

[14] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, and S. Jiang, "Large scale visual food recognition," *CoRR*, vol. abs/2103.16107, 2021. [Online]. Available: https://arxiv.org/abs/2103.16107

[15] X. Chen, H. Zhou, and L. Diao, "Chinesefoodnet: A large-scale image dataset for chinese food recognition," *CoRR*, vol. abs/1705.02743, 2017. [Online]. Available: http://arxiv.org/abs/1705.02743

[16] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision*, 2014.

[17] P. Kelly, S. J. Marshall, H. Badland, J. Kerr, M. Oliver, A. R. Doherty, and C. Foster, "An ethical framework for automated, wearable cameras in health behavior research," *American Journal of Preventive Medicine*, vol. 44, no. 3, pp. 314–319, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0749379712008628

[18] T. M. Mok, F. Cornish, and J. Tarr, "Too much information: Visual research ethics in the age of wearable cameras," *Integrative Psychological and Behavioral Science*, vol. 49, pp. 309–322, 2015. [Online]. Available: https://link.springer.com/article/10.1007/s12124-014-9289-8

[19] (2022) Apple. [Online]. Available: https://www.apple.com/ca/iphone/

[20] (2022) Polycam - lidar & 3d scanner for iphone & android. [Online]. Available: https://poly.cam/

[21] J. Chambers, T. Hullette, and P. Gharge. (2022, Sep) The best 3d scanner apps of 2022 (iphone & android). [Online]. Available: https://all3dp.com/2/best-3d-scanner-app-iphone-android-photogrammetry/

[22] (2008) Nutrient value of some common foods. [Online]. Available: https://www.canada.ca/en/health-canada/services/food-nutrition/healthy-eating/nutrient-data/nutrient-value-some-common-foods-2008.html

[23] K. McHenry and P. Bajcsy, "An overview of 3d data content, file formats and viewers," *National Center for Supercomputing Applications*, vol. 1205, p. 22, 2008.