3D Mesh Reconstruction of Foods from a Single Image

Shu Naritomi Keiji Yanai
The University of Electro-Communications, Tokyo
Chofu, Tokyo, Japan
{naritomi-s,yanai}@mm.inf.uec.ac.jp

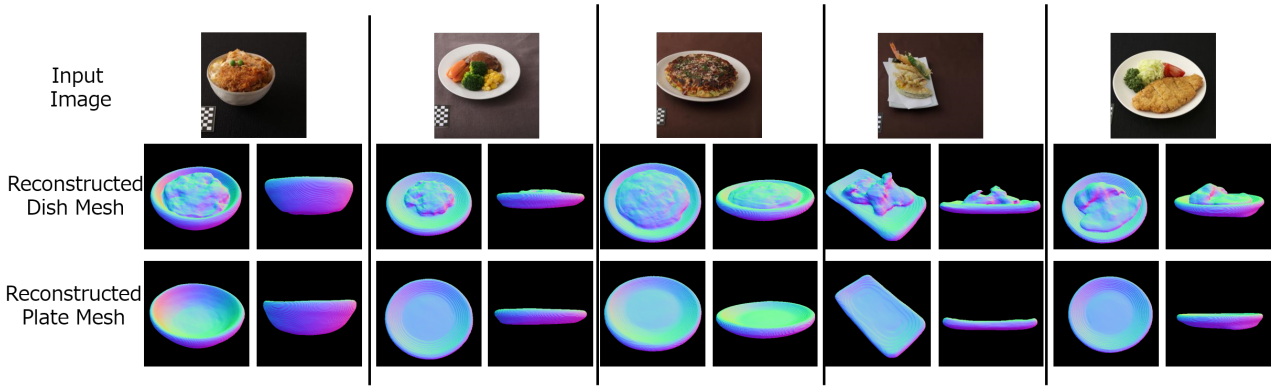

Figure 1: 3D reconstruction results from real food photos with ResNet18, $\lambda_3 = 20$ (w/ plate consistency loss) and backgrounds.

ABSTRACT

Dietary calorie management has been an important topic in recent years, and various methods and applications on image-based food calorie estimation have been published in the multimedia community. Most of the existing methods of estimating food calorie amounts use 2D-based image recognition. On the other hand, in this extended abstract, we would like to introduce our work on 3D food volume estimation employing a recent DNN-based 3D mesh reconstruction technique. We performed 3D mesh reconstruction of a dish (food and plate) and a plate (without foods) from a single image. We succeeded in restoring the 3D shape with high accuracy while maintaining the consistency between a plate part of an estimated 3D dish and an estimated 3D plate. To achieve this, the following contributions were made in our recent work [18]. (1) Proposal of "Hungry Networks," a new network that generates two kinds of 3D volumes from a single image. (2) Introduction of plate consistency loss that matches the shapes of the plate parts of the two reconstructed models. (3) Creating a new dataset of 3D food models that are 3D scanned of actual foods and plates. We also conducted an experiment to infer the volume of only the food region from the difference of the two reconstructed volumes. As a result, it was shown that the introduced new loss function not only matches the 3D shape of the plate, but also contributes to obtaining the volume with higher accuracy. Although there are some existing studies that consider 3D shapes of foods, this is the first study to generate a 3D mesh volume from a single dish image. In addition, we have implemented a web-based 3D dish reconstruction system, "Pop'n Food" [19], which enables reconstruction of 3D shapes from a single dish image in a real-time way. The demo video of the system is available at https://youtu.be/YyIu8bL65EE.

AIxFood '21, October 20, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8673-9/21/10…$15.00
https://doi.org/10.1145/3475725.3483626

CCS CONCEPTS

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

KEYWORDS

food volume estimation, 3D reconstruction from a single image, food image recognition

ACM Reference Format:
Shu Naritomi Keiji Yanai. 2021. 3D Mesh Reconstruction of Foods from a Single Image. In *Proceedings of the 3rd Workshop on AIxFood (AIxFood '21), October 20, 2021, Virtual Event, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3475725.3483626

1 INTRODUCTION

It is necessary to consider the amount of food for accurate estimation of the amounts of food calories for dietary management. Various methods and applications on image-based food calorie estimation have been published in the multimedia community. Most of the existing methods of estimating food calorie amounts use 2D-based image recognition. Some methods infer calorie amounts directly with regression [5, 6], while the others estimate calorie amounts based on 2D area sizes using detection and segmentation methods [4, 7]. However, most of the image-based methods cannot estimate the actual size of foods. Then, size-known reference objects were commonly used for accurate food calorie estimation. Recently, some works use AR/MR devices to estimate accurate actual food size without a reference object [17, 25]. However, the accuracy of the calorie estimation by 2D-based methods is limited due to the 3D nature of real foods. For this reason, there are methods that use depth estimation [2, 14, 16] or depth cameras [1], but all of them assume that the meal is on a flat plate. Therefore, in this work, we propose "Hungry Networks," which is a network for simultaneous 3D reconstruction of both a dish and a plate from a single 2D image. By using the difference between the estimated volumes of a dish and a plate, we can obtain only the food volume, which is difficult to obtain in general. To estimate the difference of two volumes, we

introduce plate consistency loss, which is a new loss function for matching the plate parts of the two output models. Note that we do not estimate 3D food-only volumes directly, since it is difficult to create a dataset containing 3D food-only volumes.

Although some existing dietary datasets contain depth images, none contain a complete 3D shape of foods. Therefore, in our work, we captured dishes and plates with a 3D scanner, and created a 3D mesh food data set. The corresponding dish image was created by rendering a scanned 3D model. We also experimented with whether the model learned from the rendered image can be reconstructed from the actual dish image. The contributions in our recent paper [18] are as follows:

- Proposing "Hungry Networks," a new network that generates two models from a single image.
- Introduction of plate consistency loss that matches the shape of the plate part of the two reconstructed 3D models.
- Creating a new dataset of 3D models with 3D scans of real food and plate.

2 RELATED WORK

2.1 3D reconstruction from a single image

There are several methods for reconstructing a 3D shape from a single image regarding 3D representation: voxel-based [3, 27**?**], point-cloud-based [8], and mesh-based [10, 11, 20, 23, 28]. Recently, implicit function-based methods [15, 21, 24] have been proposed, and they are attracting a lot of attention because of their high representation power and computational efficiency. These methods eventually apply a marching cube algorithm to extract the mesh.

2.2 Food recognition considering 3D shapes

In this section, we review some works on diets that consider 3D shape or volume. The ultimate goal of these studies is to estimate the amounts of calories and ingredients. In Chen et al. [2], a depth sensor is used to take a depth image to estimate the amount of calories in a food. Some methods such as Puri et al. [22] and DietCam [12] obtained a 3D shape by estimating a classical camera matrix from multiple viewpoints. In recent years, CNN-based has been acitively explored. Lu et al. [14] generated a depth image using a neural network and tried to infer the amount of food from the generated depth image. Im2calories[16] was trying to estimate the calorific value by estimating the 3D shape in voxel representation from a color image. Ando et al. [1] used depth-camera built in a smartphone to obtain a RGB-D food image for calorie estimation. Recently, Nutrition5k [26] containing 5,000 RGB-D food images annotated with nutrition information has been released, which is the largest open food image dataset having nutrition information.

3 HUNGRY NETWORKS

The Hungry Network is a deep neural network that reconstructs two 3D shapes of a dish (food and plate) and a plate (without foods) from a single food image. The network consists of one encoder and two decoders as shown in Figure 2. The encoder extracts features of a dish image, which consists of a pre-trained backbone network such as ResNet. Dish image features and 3D points, $p \in \mathbb{R}^3$, are used as decoder inputs. The decoders output the occupancy for a dish (containing a food part and a plate part) and a plate, respectively. The occupancy represents if each of 3D point is inside the mesh or outside the mesh with 1/0 binary values. Finally, by applying the marching cubes algorithm [13] to the occupancy field obtained by inference with multiple times using decoders, the isosurface is extracted as a mesh.

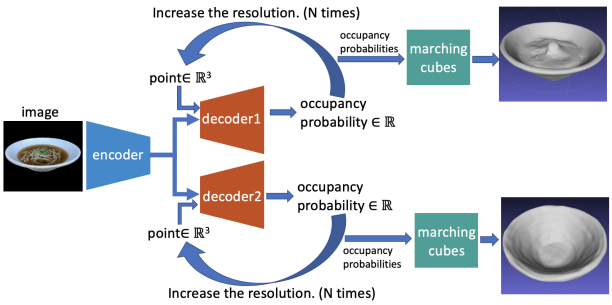

Figure 2: The overview of "Hungry Networks."

Table 1: occupancy table

dish occupancy ($f_{d1}(p)$)	plate occupancy ($f_{d2}(p)$)	$f_{d2}(p) - f_{d1}(p)$
0	0	0
1	0	-1
0	1	1
1	1	0

3.1 Training

We explain how to train the network. $p \in \mathbb{R}^3$ is the input point, x is the feature vector of the input image, and the decoder network for the dish and the plate are represented as $f_{d1}(x, p)$ and $f_{d2}(x, p)$, respectively. In addition, the occupancy of training data is represented by $o(p) \in \mathbb{R}$ corresponding to the point p. Training of occupancy is equivalent to the binary classification problem of whether the point is inside or outside the mesh surface. Then, the loss function for learning the occupancy is represented in Eq.1. Binary cross entropy loss is used for the loss function because it results in binary classification.

$$\mathcal{L}_O(f_d(x, p), o(p)) = \mathcal{L}_{bce}(f_d(x, p), o(p)) \tag{1}$$

Next, we introduce a plate consistency loss to match the plate parts of both the output mesh models to each other. First, the possible patterns of the combination of occupancy of the corresponding points on two mesh models are shown in Table 1. When the occupancy of both models at the corresponding point is the same, it is in the desirable condition. In addition, the condition where the occupancy of the dish is 1 and the occupancy of the plate is 0 is no problem, since such a point corresponds to a part of the food part of the dish model. On the other hand, the condition where the occupancy of the dish is 0 and the occupancy of the plate is 1 is problematic, since this means that inconsistency happens between the dish model and the plate model, which should be resolved. Penalties were applied during training only if the dish occupancy is 0 and the plate occupancy is 1, which corresponds to the condition where $f_{d2}(p) - f_{d1}(p)$ equals 1 as shown in Table 1. So, $\max(f_{d2}(p) - f_{d1}(p), 0)$ is used as a loss function to be minimize. We will call this "plate consistency loss".

$$\mathcal{L}_C(f_{d1}(p), f_{d2}(p)) = \max(f_{d2}(p) - f_{d1}(p), 0) \tag{2}$$

The above two formulas (Eq.1, Eq.2) are put together to determine the loss $\mathcal{L}_{\mathcal{B}}$ for each mini-batch of the entire learning. Here, $\mathcal{B}$ is the sampled mini-batch, I_i is the i-th image of the batch, and K points are sampled from the i-th batch, and $p_{i,j}$ represents the sampled j-th point of the i-th image. It is assumed that f_e is the encoder that output image features, and f_{d1} and f_{d2} are decoder outputs that output food and plate occupancy rates, respectively.

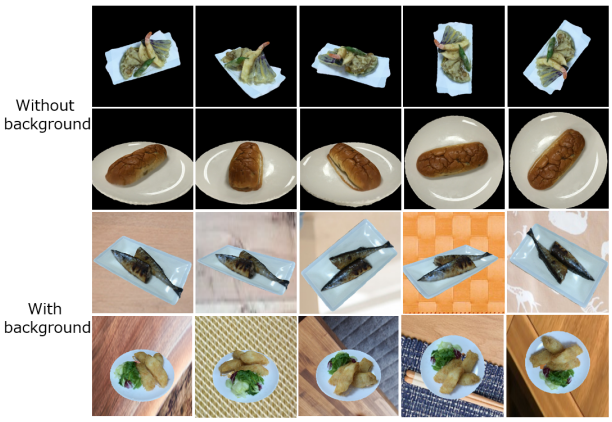

Figure 3: Images rendered for training without/with backgrounds.

$o1, o2$ represents ground truth occupancy value.

$$x_i = f_e(I_i) \tag{3}$$
$$y1_{i,j} = f_{d1}(x_i, p_{i,j}) \tag{4}$$
$$y2_{i,j} = f_{d2}(x_i, p_{i,j}) \tag{5}$$

$$\mathcal{L}_{\mathcal{B}} = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{K}\Big(\lambda_1\mathcal{L}_O(y1_{i,j}, o1_i(p_{i,j})) + \lambda_2\mathcal{L}_O(y2_{i,j}, o2_i(p_{i,j})) + \lambda_3\mathcal{L}_C(y1_{i,j}, y2_{i,j})\Big) \tag{6}$$

4 DATASET CONSTRUCTION

Some existing dietary datasets contain color and depth images [9]. However, no dietary dataset contains 3D Mesh models of foods. Therefore, for this work, we had to create a new 3D dietary dataset. We created the dataset consisting of 240 3D models of foods and 38 models of plates. To create the models, we used a commercially available 3D sensor called "Structure Sensor" and a dedicated 3D scanning application. Since the same plate was used for different dishes, the number of plate models is smaller than that of dish models.

4.1 Generating input images by rendering

In this work, we rendered images for learning using software called blender. 25 images were rendered for each model, taken from various angles. In this work, we collected textures of various types of tables and tablecloths from the Web as the background of the rendered dish images, and created composite images. Figure 3 shows the image created by rendering. The top two lines are just rendered, and the bottom two lines are a composite of the background.

5 EXPERIMENTS

We made experiments with the proposed model, "Hungry Networks", on the following conditions: (1) we set three values as the plate consistency loss weight (λ_3 in Eq.6), (2) we train the model with rendered dish images with/without backgrounds. To train the proposed network, we used 216 models for training and 24 models for evaluation among 240 models in the constructed dataset. The hyperparameters, $\lambda 1, \lambda 2$, were fixed at 1, and only λ_3 was changed in the experiments. We used Adam as an optimizer.

5.1 Metrics

For quantitative evaluation, we use Volumetric IoU, Chamber L1 distance, plate consistency, and volume error. The plate consistency and volume error are proprietary metrics.

The plate consistency is the mean distance from points on the generated plate mesh to the nearest neighbor points on the generated food mesh. This value indicates how different the plate part of the dish volume is from the plate volume.

Volume error is the mean distance from the inferred volume of the food region to the ground-truth food volume. The food volume is obtained by subtracting the plate volume from the dish volume.

On IoU, the higher value is better, while on the other metrics, the Chamfer L1 distance, plate consistency and the volume error, the lower values are better.

5.2 Quantitative evaluation

First, we investigated the effect of λ_3 on evaluation. The encoder was based on ResNet34, and training images without backgrounds were used for training and evaluation. We made experiments with 0, 20 and 50 for λ_3. Note that 0 means we did not used the plate consistency loss. The results are shown in Table 2. As a result, it indicates that the volume error is greatly reduced when plate consistency loss is used. On the other hand, the 3D meshes of dishes and plates were estimated the most accurately without the plate consistency loss. However, as shown in Figure 5, in case of no plate consistency loss invisible parts of the dish volume and the plate volume were differently reconstructed. Note that the mesh captured by the 3D sensor contains a lot of noise, unlike the handcrafted 3D models included in ShapeNet [29]. Since both the dish decoder and the plate decoder were optimized independently using only independent occupancy loss functions to each other, individual evaluation tends to become better and integrated evaluation such as volume error tends to become worse.

In the next experiments, we evaluated how much accuracy was affected by backgrounds of training images. We used $\lambda_3 = 20$ with ResNet18 and ResNet50 as backbones. With backgrounds in the training images, we achieved the best results regarding the volume error and the plate consistency.

5.3 Qualitative evaluation

Figure 4 shows the estimated 3D meshes of both dishes and plates with $\lambda_3 = 20$, ResNet18 and training images without background. The 3D meshes of both the dishes and the plates were correctly estimated for the corresponding images. In addition, we can see that most of the plate parts of the dish meshes were identical to the plate meshes.

Figure 1 shows the results of using a real food photo as an input image for a network trained on ResNet18, $\lambda_3 = 20$, with a background image. Although real images of actual foods were not used for network training, the trained model was able to reconstruct 3D volumes of the dishes and the plates. Various kinds of the plates such as square flat plates, rectangular plates, round flat plates, and bowls were successfully reconstructed, although the shape and the height of each of the plates were different greatly.

6 POP'N FOOD

We have created "Pop'n food" as an application to easily view 3D models inferred by the network. In the works on 3D reconstruction, we often see videos created from images of models rendered from different angles and displayed in a browser. This application, on the other hand, uses WebGL via a JavaScript library called three.js,

Table 2: The evaluation results with three kinds of λ_3 using ResNet34 and non-background images.

λ_3	IoU (dish)	IoU (plate)	Chamfer L1 (dish)	Chamfer L1 (plate)	**plate consistency**	**volume error**
0	**0.624**	**0.621**	**0.0189**	0.0186	0.0256	0.0252
20	0.550	0.607	0.0262	**0.0182**	0.0168	**0.0155**
50	0.542	0.610	0.0260	0.0209	**0.0152**	0.0161

Table 3: The evaluation results with training images with/without backgrounds with $\lambda_3 = 20$ and ResNet18/50 backbones.

encoder	background	IoU (dish)	IoU (plate)	Chamfer L1 (dish)	Chamfer L1 (plate)	Plate consistency score	Volume error
ResNet 18	none	0.560	0.634	0.0265	0.0193	**0.0146**	0.0150
ResNet 50	none	0.564	0.617	**0.0251**	0.0186	0.0148	0.0147
ResNet 18	yes	**0.565**	**0.645**	0.0254	**0.0173**	**0.0146**	**0.0146**
ResNet 50	yes	0.558	0.628	0.0252	**0.0173**	0.0157	0.0157

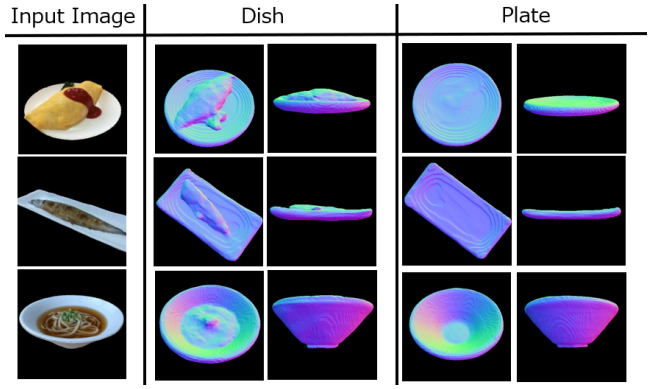

Figure 4: The estimated volumes of both dishes and plates with the model trained from non-background images with a ResNet18-based encoder and $\lambda_3 = 20$.

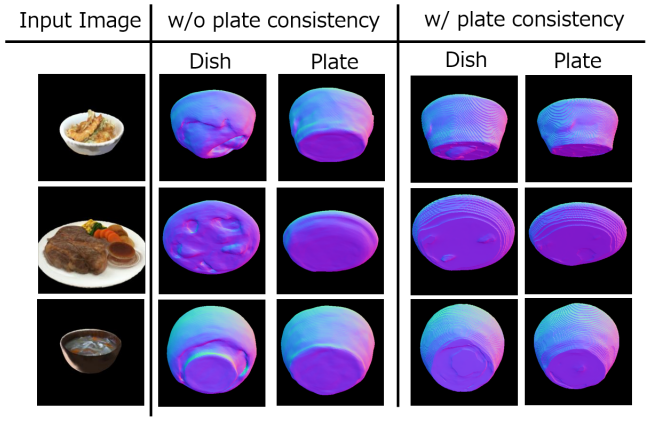

Figure 5: Comparative results with/without the plate consistency loss. Note that training condition is the same as Fig.4.

which allows users to interactively view the 3D model from all angles.

The user interface of the application is shown in Figure 6. There is a list of images on the left and two canvases on the right. The 3D model of the dish is displayed on the upper side of the canvas, and the 3D model of the plate is displayed on the lower side. By clicking the button next to the image, the user can see the 3D model reconstructed from the clicked image.

By loading Hungry Networks when the web server starts and deploying it on the GPU, the response to the request is within about 2 seconds.

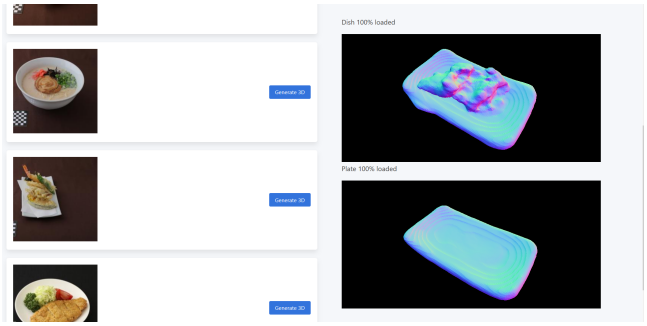

Figure 6: Web Interface. There is a list of images on the left and two canvases on the right. The 3D model of the dish is displayed on the upper side of the canvas, and the 3D model of the plate is displayed on the lower side. By clicking the button next to the image, the user can see the 3D model reconstructed from the clicked dish image. The 3D model is also rendered in real-time, so you can interact with it (not a video).

7 CONCLUSIONS

In this work, we proposed "Hungry Networks" that enabled 3D shape reconstruction of dishes and plates from a single food image. For training, we introduced a new loss, plate consistency loss, in order to maintain the consistency between the plate part of the dish and the plate. In addition, for experiments, we created a dataset consisting of 3D mesh models of dishes. By the experiments, it was shown that 3D shapes could be reconstructed with high accuracy by using rendered images of dishes and composite rendered images of backgrounds for training. In addition, by introducing plate consistency loss, we succeeded in maintaining and restoring the consistency of the plate parts of the two meshes, which contributed to the estimation of the volume of the dietary area. It was shown that the network learned from the dish images obtained by synthesizing the background image can be correctly reconstructed even if the real dish image is input as well.

As a future task, the current 3D shape restoration is performed in a normalized space, and the actual size cannot be taken into consideration. In order to estimate the amount of calories, it is necessary to be able to consider the actual size. Therefore, we would like to use the environment recognition function of the AR device, RGB-D depth images, reference objects and so on to perform 3D shape restoration considering the actual size, which will lead to accurate estimation of the amounts of food calorie intake.

REFERENCES

[1] Y. Ando, T. Ege, J. Cho, and K. Yanai. 2019. DepthCalorieCam: A Mobile Application for Volume-Based FoodCalorie Estimation using Depth Cameras. In *Proc. of the 5th International Workshop on Multimedia Assisted Dietary Management*. 76–81.

[2] M. Y. Chen, Y. H. Yang, C. J. Ho, S. H. Wang, S. M. Liu, E. Chang, C. H. Yeh, and M. Ouhyoung. 2012. Automatic chinese food identification and quantity estimation. In *Proc. of SIGGRAPH Asia 2012 Technical Briefs*. 1–4.

[3] C. B. Choy, Danfei. Xu, J. Gwak, K. Chen, and S. Savarese. 2016. 3D-R2N2: A unified approach for single and multi-view 3d object reconstruction. In *Proc. of European Conference on Computer Vision*. 628–644.

[4] T. Ege and K. Yanai. 2017. Estimating Food Calories for Multiple-dish Food Photos. In *Proc. of Asian Conference on Pattern Recognition*.

[5] T. Ege and K. Yanai. 2017. Imag-Based Food Calorie Estimation Using Knowledge on Food Categories, Ingredients and Cooking Directions. In *Proc. of ACM Multimedia Thematic Workshop*.

[6] T. Ege and K Yanai. 2018. Image-Based Food Calorie Estimation Using Recipe Information. *IEICE Transactions on Information and Systems* E101-D, 5 (2018), 1333–1341.

[7] T. Ege and K Yanai. 2018. Multi-task Learning of Dish Detection and Calorie Estimation. In *Proc. of IJCAI and ECAI Workshop on Multimedia Assisted Dietary Management*.

[8] H. Fan, H. Su, and L. J. Guibas. 2017. A Point Set Generation Network for 3D Object Reconstruction from a Single Image. In *Proc. of IEEE Computer Vision and Pattern Recognition*. 605–613.

[9] C. P. Ferdinand, S. Schlecht, F. Ettlinger, F. Grun, C. Heinle, S. Tatavatry, S. A. Ahmadi, K. Diepold, and B. H. Menze. 2017. Diabetes60-Inferring Bread Units From Food Images Using Fully Convolutional Neural Networks. In *Proc. of the IEEE International Conference on Computer Vision Workshops*. 1526–1535.

[10] G. Gkioxari, J. Malik, and J. Johnson. 2019. Mesh R-CNN. In *Proc. of IEEE International Conference on Computer Vision*. 9785–9795.

[11] A. Kanazawa, S. Tulsiani, A. A. Efros, and J. Malik. 2018. Learning category-specific mesh reconstruction from image collections. In *Proc. of European Conference on Computer Vision*. 371–386.

[12] F. Kong and J. Tan. 2011. DietCam: Regular Shape Food Recognition with a Camera Phone. In *2011 International Conference on Body Sensor Networks*. 127–132.

[13] W. E. Lorensen and H. E. Cline. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *ACM siggraph computer graphics* 21, 4 (1987), 163–169.

[14] Y. Lu, D. Allegra, M. Anthimopoulos, F. Stanco, G. M. Farinella, and S. Mougiakakou. 2018. A multi-task learning approach for meal assessment. In *Proc. of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management*. 46–52.

[15] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. 2019. Occupancy Networks: Learning 3d reconstruction in function space. In *Proc. of IEEE Computer Vision and Pattern Recognition*. 4460–4470.

[16] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy. 2015. Im2Calories: towards an automated mobile vision food diary. In *Proc. of the IEEE International Conference on Computer Vision*. 1233–1241.

[17] S. Naritomi and K. Yanai. 2020. CalorieCaptorGlass: Food Calorie Estimation Based on Actual Size using HoloLens and Deep Learning. In *Proc. of IEEE Conference on Virtual Reality and 3D User Interfaces*.

[18] S. Naritomi and K Yanai. 2020. Hungry Networks: 3D Mesh Reconstruction of a Dish and a Plate from a Single Dish Image for Estimating Food Volume. In *Proc. of ACM Multimedia Asia*.

[19] S. Naritomi and K Yanai. 2021. Pop'n Food: 3D Food Model Estimation System from a Single Image. In *Proc. of IEEE 4th International Conference on Multimedia Information Processing and Retrieval*.

[20] Albert P., Jordi S., Gary P. T. C., Alberto S., and Francesc M. 2019. 3DPeople: Modeling the Geometry of Dressed Humans. In *Proc. of IEEE International Conference on Computer Vision*.

[21] Jeong J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *Proc. of IEEE Computer Vision and Pattern Recognition*.

[22] M. Puri, Zhiwei Zhu, Q. Yu, A. Divakaran, and H. Sawhney. 2009. Recognition and volume estimation of food intake using a mobile device. In *2009 Workshop on Applications of Computer Vision (WACV)*. 1–8.

[23] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black. 2018. Generating 3D faces using convolutional mesh autoencoders. In *Proc. of European Conference on Computer Vision*. 704–720.

[24] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. 2019. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *Proc. of IEEE International Conference on Computer Vision*.

[25] R. Tanno, T. Ege, and K. Yanai. 2018. AR DeepCalorieCam V2: food calorie estimation with CNN and AR-based actual size estimation. In *Proc. of the 24th ACM Symposium on Virtual Reality Software and Technology*. 1–2.

[26] Q. Thames, A. Karpur, W. Norris, F. Xia, L. Panait, T. Weyand, and J. Sim. 2021. Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food. In *Proc. of IEEE Computer Vision and Pattern Recognition*.

[27] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. 2017. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proc. of IEEE Computer Vision and Pattern Recognition*. 2626–2634.

[28] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y. G. Jiang. 2018. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proc. of European Conference on Computer Vision*. 52–67.

[29] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 2015. 3D ShapeNets: A deep representation for volumetric shapes. In *Proc. of IEEE Computer Vision and Pattern Recognition*. 1912–1920.