Advancements in 3D Food Modeling: A Review of the MetaFood Challenge Techniques and Outcomes

Abstract

The growing focus on leveraging computer vision for dietary oversight and nutrition tracking has spurred the creation of sophisticated 3D reconstruction methods for food. The lack of comprehensive, high-fidelity data, coupled with limited collaborative efforts between academic and industrial sectors, has significantly hindered advancements in this domain. This study addresses these obstacles by introducing the MetaFood Challenge, aimed at generating precise, volumetrically accurate 3D food models from 2D images, utilizing a checkerboard for size calibration. The challenge was structured around 20 food items across three levels of complexity: easy (200 images), medium (30 images), and hard (1 image). A total of 16 teams participated in the final assessment phase. The methodologies developed during this challenge have yielded highly encouraging outcomes in 3D food reconstruction, showing great promise for refining portion estimation in dietary evaluations and nutritional tracking. Further information on this workshop challenge and the dataset is accessible via the provided URL.

1 Introduction

The convergence of computer vision technologies with culinary practices has pioneered innovative approaches to dietary monitoring and nutritional assessment. The MetaFood Workshop Challenge represents a landmark initiative in this emerging field, responding to the pressing demand for precise and scalable techniques for estimating food portions and monitoring nutritional consumption. Such technologies are vital for fostering healthier eating behaviors and addressing health issues linked to diet.

By concentrating on the development of accurate 3D models of food derived from various visual inputs, including multiple views and single perspectives, this challenge endeavors to bridge the disparity between current methodologies and practical needs. It promotes the creation of unique solutions capable of managing the intricacies of food morphology, texture, and illumination, while also meeting the real-world demands of dietary evaluation. This initiative gathers experts from computer vision, machine learning, and nutrition science to propel 3D food reconstruction technologies forward. These advancements have the potential to substantially enhance the precision and utility of food portion estimation across diverse applications, from individual health tracking to extensive nutritional investigations.

Conventional methods for assessing diet, like 24-Hour Recall or Food Frequency Questionnaires (FFQs), are frequently reliant on manual data entry, which is prone to inaccuracies and can be burdensome. The lack of 3D data in 2D RGB food images further complicates the use of regression-based methods for estimating food portions directly from images of eating occasions. By enhancing 3D reconstruction for food, the aim is to provide more accurate and intuitive nutritional assessment tools. This technology could revolutionize the sharing of culinary experiences and significantly impact nutrition science and public health.

Participants were tasked with creating 3D models of 20 distinct food items from 2D images, mimicking scenarios where mobile devices equipped with depth-sensing cameras are used for dietary

.

recording and nutritional tracking. The challenge was segmented into three tiers of difficulty based on the number of images provided: approximately 200 images for easy, 30 for medium, and a single top-view image for hard. This design aimed to rigorously test the adaptability and resilience of proposed solutions under various realistic conditions. A notable feature of this challenge was the use of a visible checkerboard for physical referencing and the provision of depth images for each frame, ensuring the 3D models maintained accurate real-world measurements for portion size estimation.

This initiative not only expands the frontiers of 3D reconstruction technology but also sets the stage for more reliable and user-friendly real-world applications, including image-based dietary assessment. The resulting solutions hold the potential to profoundly influence nutritional intake monitoring and comprehension, supporting broader health and wellness objectives. As progress continues, innovative applications are anticipated to transform personal health management, nutritional research, and the wider food industry. The remainder of this report is structured as follows: Section 2 delves into the existing literature on food portion size estimation, Section 3 describes the dataset and evaluation framework used in the challenge, and Sections 4, 5, and 6 discuss the methodologies and findings of the top three teams (VolETA, ININ-VIAUN, and FoodRiddle), respectively.

2 Related Work

Estimating food portions is a crucial part of image-based dietary assessment, aiming to determine the volume, energy content, or macronutrients directly from images of meals. Unlike the well-studied task of food recognition, estimating food portions is particularly challenging due to the lack of 3D information and physical size references necessary for accurately judging the actual size of food portions. Accurate portion size estimation requires understanding the volume and density of food, elements that are hard to deduce from a 2D image, underscoring the need for sophisticated techniques to tackle this problem. Current methods for estimating food portions are grouped into four categories.

Stereo-Based Approaches use multiple images to reconstruct the 3D structure of food. Some methods estimate food volume using multi-view stereo reconstruction based on epipolar geometry, while others perform two-view dense reconstruction. Simultaneous Localization and Mapping (SLAM) has also been used for continuous, real-time food volume estimation. However, these methods are limited by their need for multiple images, which is not always practical.

Model-Based Approaches use predefined shapes and templates to estimate volume. For instance, certain templates are assigned to foods from a library and transformed based on physical references to estimate the size and location of the food. Template matching approaches estimate food volume from a single image, but they struggle with variations in food shapes that differ from predefined templates. Recent work has used 3D food meshes as templates to align camera and object poses for portion size estimation.

Depth Camera-Based Approaches use depth cameras to create depth maps, capturing the distance from the camera to the food. These depth maps form a voxel representation used for volume estimation. The main drawback is the need for high-quality depth maps and the extra processing required for consumer-grade depth sensors.

Deep Learning Approaches utilize neural networks trained on large image datasets for portion estimation. Regression networks estimate the energy value of food from single images or from an "Energy Distribution Map" that maps input images to energy distributions. Some networks use both images and depth maps to estimate energy, mass, and macronutrient content. However, deep learning methods require extensive data for training and are not always interpretable, with performance degrading when test images significantly differ from training data.

While these methods have advanced food portion estimation, they face limitations that hinder their widespread use and accuracy. Stereo-based methods are impractical for single images, model-based approaches struggle with diverse food shapes, depth camera methods need specialized hardware, and deep learning approaches lack interpretability and struggle with out-of-distribution samples. 3D reconstruction offers a promising solution by providing comprehensive spatial information, adapting to various shapes, potentially working with single images, offering visually interpretable results, and enabling a standardized approach to food portion estimation. These benefits motivated the organization of the 3D Food Reconstruction challenge, aiming to overcome existing limitations and develop more accurate, user-friendly, and widely applicable food portion estimation techniques, impacting nutritional assessment and dietary monitoring.

3 Datasets and Evaluation Pipeline

3.1 Dataset Description

The dataset for the MetaFood Challenge features 20 carefully chosen food items from the MetaFood3D dataset, each scanned in 3D and accompanied by video recordings. To ensure precise size accuracy in the reconstructed 3D models, each food item was captured alongside a checkerboard and pattern mat, serving as physical scaling references. The challenge is divided into three levels of difficulty, determined by the quantity of 2D images provided for reconstruction:

- Easy: Around 200 images taken from video.
- Medium: 30 images.
- Hard: A single image from a top-down perspective.

Table 1 details the food items included in the dataset.

Table 1: MetaFood Challenge Data Details

Object Index	Food Item	Difficulty Level	Number of Frames
1	Strawberry	Easy	199
2	Cinnamon bun	Easy	200
3	Pork rib	Easy	200
4	Corn	Easy	200
5	French toast	Easy	200
6	Sandwich	Easy	200
7	Burger	Easy	200
8	Cake	Easy	200
9	Blueberry muffin	Medium	30
10	Banana	Medium	30
11	Salmon	Medium	30
12	Steak	Medium	30
13	Burrito	Medium	30
14	Hotdog	Medium	30
15	Chicken nugget	Medium	30
16	Everything bagel	Hard	1
17	Croissant	Hard	1
18	Shrimp	Hard	1
19	Waffle	Hard	1
20	Pizza	Hard	1

3.2 Evaluation Pipeline

The evaluation process is split into two phases, focusing on the accuracy of the reconstructed 3D models in terms of shape (3D structure) and portion size (volume).

3.2.1 Phase-I: Volume Accuracy

In the first phase, the Mean Absolute Percentage Error (MAPE) is used to evaluate portion size accuracy, calculated as follows:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - F_i}{A_i} \right| \times 100\% \tag{1}$$

where A_i is the actual volume (in ml) of the i-th food item obtained from the scanned 3D food mesh, and F_i is the volume calculated from the reconstructed 3D mesh.

3.2.2 Phase-II: Shape Accuracy

Teams that perform well in Phase-I are asked to submit complete 3D mesh files for each food item. This phase involves several steps to ensure precision and fairness:

- Model Verification: Submitted models are checked against the final Phase-I submissions for consistency, and visual inspections are conducted to prevent rule violations.
- Model Alignment: Participants receive ground truth 3D models and a script to compute the final Chamfer distance. They must align their models with the ground truth and prepare a transformation matrix for each submitted object. The final Chamfer distance is calculated using these models and matrices.
- Chamfer Distance Calculation: Shape accuracy is assessed using the Chamfer distance metric. Given two point sets X and Y, the Chamfer distance is defined as:

$$d_{CD}(X,Y) = \frac{1}{|X|}\sum_{x\in X}\min_{y\in Y}\|x-y\|_2^2 + \frac{1}{|Y|}\sum_{y\in Y}\min_{x\in X}\|x-y\|_2^2 \tag{2}$$

This metric offers a comprehensive measure of similarity between the reconstructed 3D models and the ground truth. The final ranking is determined by combining scores from both Phase-I (volume accuracy) and Phase-II (shape accuracy). Note that after the Phase-I evaluation, quality issues were found with the data for object 12 (steak) and object 15 (chicken nugget), so these items were excluded from the final overall evaluation.

4 First Place Team - VolETA

4.1 Methodology

The team's research employs multi-view reconstruction to generate detailed food meshes and calculate precise food volumes.

4.1.1 Overview

The team's method integrates computer vision and deep learning to accurately estimate food volume from RGBD images and masks. Keyframe selection ensures data quality, supported by perceptual hashing and blur detection. Camera pose estimation and object segmentation pave the way for neural surface reconstruction, creating detailed meshes for volume estimation. Refinement steps, including isolated piece removal and scaling factor adjustments, enhance accuracy. This approach provides a thorough solution for accurate food volume assessment, with potential uses in nutrition analysis.

4.1.2 The Team's Proposal: VolETA

The team starts by acquiring input data, specifically RGBD images and corresponding food object masks. The RGBD images, denoted as $I_D = \{I_{Di}\}_{i=1}^{n}$, where n is the total number of frames, provide depth information alongside RGB images. The food object masks, $\{M_i^f\}_{i=1}^{n}$, help identify regions of interest within these images.

Next, the team selects keyframes. From the set $\{I_{Di}\}_{i=1}^{n}$, keyframes $\{I_j^K\}_{j=1}^{k} \subseteq \{I_{Di}\}_{i=1}^{n}$ are chosen. A method is implemented to detect and remove duplicate and blurry images, ensuring high-quality frames. This involves applying a Gaussian blurring kernel followed by the fast Fourier transform method. Near-Image Similarity uses perceptual hashing and Hamming distance thresholding to detect similar images and retain overlapping ones. Duplicates and blurry images are excluded to maintain data integrity and accuracy.

Using the selected keyframes $\{I_j^K\}_{j=1}^{k}$, the team estimates camera poses through a method called PixSfM, which involves extracting features using SuperPoint, matching them with SuperGlue, and refining them. The outputs are the camera poses $\{C_j\}_{j=1}^{k}$, crucial for understanding the scene's spatial layout.

In parallel, the team uses a tool called SAM for reference object segmentation. SAM segments the reference object with a user-provided prompt, producing a reference object mask M^R for each keyframe. This mask helps track the reference object across all frames. The XMem++ method extends the reference object mask M^R to all frames, creating a comprehensive set of reference object masks $\{M_i^R\}_{i=1}^n$. This ensures consistent reference object identification throughout the dataset.

To create RGBA images, the team combines RGB images, reference object masks $\{M_i^R\}_{i=1}^n$, and food object masks $\{M_i^F\}_{i=1}^n$. This step, denoted as $\{I_i^R\}_{i=1}^n$, integrates various data sources into a unified format for further processing.

The team converts the RGBA images $\{I_i^R\}_{i=1}^n$ and camera poses $\{C_j\}_{j=1}^k$ into meaningful metadata and modeled data D_m. This transformation facilitates accurate scene reconstruction.

The modeled data D_m is input into NeuS2 for mesh reconstruction. NeuS2 generates colorful meshes $\{R^f, R^r\}$ for the reference and food objects, providing detailed 3D representations. The team uses the "Remove Isolated Pieces" technique to refine the meshes. Given that the scenes contain only one food item, the diameter threshold is set to 5% of the mesh size. This method deletes isolated connected components with diameters less than or equal to 5%, resulting in a cleaned mesh $\{RC^f, RC^r\}$. This step ensures that only significant parts of the mesh are retained.

The team manually identifies an initial scaling factor S using the reference mesh via MeshLab. This factor is fine-tuned to S_f using depth information and food and reference masks, ensuring accurate scaling relative to real-world dimensions. Finally, the fine-tuned scaling factor S_f is applied to the cleaned food mesh RC^f, producing the final scaled food mesh RF^f. This step culminates in an accurately scaled 3D representation of the food object, enabling precise volume estimation.

4.1.3 Detecting the scaling factor

Generally, 3D reconstruction methods produce unitless meshes by default. To address this, the team manually determines the scaling factor by measuring the distance for each block of the reference object mesh. The average of all block lengths l_{avg} is calculated, while the actual real-world length is constant at $l_{real} = 0.012$ meters. The scaling factor $S = l_{real}/l_{avg}$ is applied to the clean food mesh RC^f, resulting in the final scaled food mesh RF^f in meters.

The team uses depth information along with food and reference object masks to validate the scaling factors. The method for assessing food size involves using overhead RGB images for each scene. Initially, the pixel-per-unit (PPU) ratio (in meters) is determined using the reference object. Subsequently, the food width (f_w) and length (f_l) are extracted using a food object mask. To determine the food height (f_h), a two-step process is followed. First, binary image segmentation is performed using the overhead depth and reference images, yielding a segmented depth image for the reference object. The average depth is then calculated using the segmented reference object depth (d_r). Similarly, employing binary image segmentation with an overhead food object mask and depth image, the average depth for the segmented food depth image (d_f) is computed. The estimated food height f_h is the absolute difference between d_r and d_f. To assess the accuracy of the scaling factor S, the food bounding box volume $(f_w \times f_l \times f_h) \times PPU$ is computed. The team evaluates if the scaling factor S generates a food volume close to this potential volume, resulting in S_{fine}. Table 2 lists the scaling factors, PPU, 2D reference object dimensions, 3D food object dimensions, and potential volume.

For one-shot 3D reconstruction, the team uses One-2-3-45 to reconstruct a 3D model from a single RGBA view input after applying binary image segmentation to both food RGB and mask images. Isolated pieces are removed from the generated mesh, and the scaling factor S, which is closer to the potential volume of the clean mesh, is reused.

4.2 Experimental Results

4.2.1 Implementation settings

Experiments were conducted using two GPUs: GeForce GTX 1080 Ti/12G and RTX 3060/6G. The Hamming distance for near image similarity was set to 12. For Gaussian kernel radius, even numbers in the range [0...30] were used for detecting blurry images. The diameter for removing isolated pieces was set to 5%. NeuS2 was run for 15,000 iterations with a mesh resolution of 512x512, a unit cube "aabb scale" of 1, "scale" of 0.15, and "offset" of [0.5, 0.5, 0.5] for each food scene.

4.2.2 VolETA Results

The team extensively validated their approach on the challenge dataset and compared their results with ground truth meshes using MAPE and Chamfer distance metrics. The team's approach was applied separately to each food scene. A one-shot food volume estimation approach was used if the number of keyframes k equaled 1; otherwise, a few-shot food volume estimation was applied. Notably, the keyframe selection process chose 34.8% of the total frames for the rest of the pipeline, showing the minimum frames with the highest information.

Table 2: List of Extracted Information Using RGBD and Masks

Level	Id	Label	S_f	PPU	$R_w \times R_l$	$(f_w \times f_l \times f_h)$
	1	Strawberry	0.08955223881	0.01786	320×360	$(238 \times 257 \times 2.353)$
	2	Cinnamon bun	0.1043478261	0.02347	236×274	$(363 \times 419 \times 2.353)$
	3	Pork rib	0.1043478261	0.02381	246×270	$(435 \times 778 \times 1.176)$
Easy	4	Corn	0.08823529412	0.01897	291×339	$(262 \times 976 \times 2.353)$
	5	French toast	0.1034482759	0.02202	266×292	$(530 \times 581 \times 2.53)$
	6	Sandwich	0.1276595745	0.02426	230×265	$(294 \times 431 \times 2.353)$
	7	Burger	0.1043478261	0.02435	208×264	$(378 \times 400 \times 2.353)$
	8	Cake	0.1276595745	0.02143	256×300	$(298 \times 310 \times 4.706)$
	9	Blueberry muffin	0.08759124088	0.01801	291×357	$(441 \times 443 \times 2.353)$
	10	Banana	0.08759124088	0.01705	315×377	$(446 \times 857 \times 1.176)$
Medium	11	Salmon	0.1043478261	0.02390	242×269	$(201 \times 303 \times 1.176)$
	13	Burrito	0.1034482759	0.02372	244×271	$(251 \times 917 \times 2.353)$
	14	Frankfurt sandwich	0.1034482759	0.02115	266×304	$(400 \times 1022 \times 2.353)$
	16	Everything bagel	0.08759124088	0.01747	306×368	$(458 \times 134 \times 1.176)$
Hard	17	Croissant	0.1276595745	0.01751	319×367	$(395 \times 695 \times 2.176)$
	18	Shrimp	0.08759124088	0.02021	249×318	$(186 \times 95 \times 0.987)$
	19	Waffle	0.01034482759	0.01902	294×338	$(465 \times 537 \times 0.8)$
	20	Pizza	0.01034482759	0.01913	292×336	$(442 \times 651 \times 1.176)$

After finding keyframes, PixSfM estimated the poses and point cloud. After generating scaled meshes, the team calculated volumes and Chamfer distance with and without transformation metrics. Meshes were registered with ground truth meshes using ICP to obtain transformation metrics.

Table 3 presents quantitative comparisons of the team's volumes and Chamfer distance with and without estimated transformation metrics from ICP. For overall method performance, Table 4 shows the MAPE and Chamfer distance with and without transformation metrics.

Additionally, qualitative results on one- and few-shot 3D reconstruction from the challenge dataset are shown. The model excels in texture details, artifact correction, missing data handling, and color adjustment across different scene parts.

Limitations: Despite promising results, several limitations need to be addressed in future work:

- Manual processes: The current pipeline includes manual steps like providing segmentation prompts and identifying scaling factors, which should be automated to enhance efficiency.
- Input requirements: The method requires extensive input information, including food masks and depth data. Streamlining these inputs would simplify the process and increase applicability.
- Complex backgrounds and objects: The method has not been tested in environments with complex backgrounds or highly intricate food objects.
- Capturing complexities: The method has not been evaluated under different capturing complexities, such as varying distances and camera speeds.
- Pipeline complexity: For one-shot neural rendering, the team currently uses One-2-3-45. They aim to use only the 2D diffusion model, Zero123, to reduce complexity and improve efficiency.

Table 3: Quantitative Comparison with Ground Truth Using Chamfer Distance

L	Id	Team's Vol.	GT Vol.	Ch. w/ t.m	Ch. w/o t.m
	1	40.06	38.53	1.63	85.40
	2	216.9	280.36	7.12	111.47
	3	278.86	249.67	13.69	172.88
E	4	279.02	295.13	2.03	61.30
	5	395.76	392.58	13.67	102.14
	6	205.17	218.44	6.68	150.78
	7	372.93	368.77	4.70	66.91
	8	186.62	173.13	2.98	152.34
	9	224.08	232.74	3.91	160.07
	10	153.76	163.09	2.67	138.45
M	11	80.4	85.18	3.37	151.14
	13	363.99	308.28	5.18	147.53
	14	535.44	589.83	4.31	89.66
	16	163.13	262.15	18.06	28.33
H	17	224.08	181.36	9.44	28.94
	18	25.4	20.58	4.28	12.84
	19	110.05	108.35	11.34	23.98
	20	130.96	119.83	15.59	31.05

Table 4: Quantitative Comparison with Ground Truth Using MAPE and Chamfer Distance

MAPE	Ch. w/ t.m		Ch. w/o t.m	
(%)	sum	mean	sum	mean
10.973	0.130	0.007	1.715	0.095

5 Second Place Team - ININ-VIAUN

5.1 Methodology

This section details the team's proposed network, illustrating the step-by-step process from original images to final mesh models.

5.1.1 Scale factor estimation

The procedure for estimating the scale factor at the coordinate level is illustrated in Figure 9. The team adheres to a method involving corner projection matching. Specifically, utilizing the COLMAP dense model, the team acquires the pose of each image along with dense point cloud data. For any given image img_k and its extrinsic parameters $[R|t]_k$, the team initially performs threshold-based corner detection, setting the threshold at 240. This step allows them to obtain the pixel coordinates of all detected corners. Subsequently, using the intrinsic parameters k and the extrinsic parameters $[R|t]_k$, the point cloud is projected onto the image plane. Based on the pixel coordinates of the corners, the team can identify the closest point coordinates P_i^k for each corner, where i represents the index of the corner. Thus, they can calculate the distance between any two corners as follows:

$$D_{ij}^k = (P_i^k - P_j^k)^2 \quad \forall i \neq j \tag{3}$$

To determine the final computed length of each checkerboard square in image k, the team takes the minimum value of each row of the matrix D^k (excluding the diagonal) to form the vector d^k. The median of this vector is then used. The final scale calculation formula is given by Equation 4, where 0.012 represents the known length of each square (1.2 cm):

$$\text{scale} = \frac{0.012}{\sum_{i=1}^{n} med(d^k)} \tag{4}$$

5.1.2 3D Reconstruction

The 3D reconstruction process, depicted in Figure 10, involves two different pipelines to accommodate variations in input viewpoints. The first fifteen objects are processed using one pipeline, while the last five single-view objects are processed using another.

For the initial fifteen objects, the team uses COLMAP to estimate poses and segment the food using the provided segment masks. Advanced multi-view 3D reconstruction methods are then applied to reconstruct the segmented food. The team employs three different reconstruction methods: COLMAP, DiffusioNeRF, and NeRF2Mesh. They select the best reconstruction results from these methods and extract the mesh. The extracted mesh is scaled using the estimated scale factor, and optimization techniques are applied to obtain a refined mesh.

For the last five single-view objects, the team experiments with several single-view reconstruction methods, including Zero123, Zero123++, One2345, ZeroNVS, and DreamGaussian. They choose ZeroNVS to obtain a 3D food model consistent with the distribution of the input image. The intrinsic camera parameters from the fifteenth object are used, and an optimization method based on reprojection error refines the extrinsic parameters of the single camera. Due to limitations in single-view reconstruction, depth information from the dataset and the checkerboard in the monocular image are used to determine the size of the extracted mesh. Finally, optimization techniques are applied to obtain a refined mesh.

5.1.3 Mesh refinement

During the 3D Reconstruction phase, it was observed that the model's results often suffered from low quality due to holes on the object's surface and substantial noise, as shown in Figure 11.

To address the holes, MeshFix, an optimization method based on computational geometry, is employed. For surface noise, Laplacian Smoothing is used for mesh smoothing operations. The Laplacian Smoothing method adjusts the position of each vertex to the average of its neighboring vertices:

$$V_i^{(\text{new})} = V_i^{(\text{old})} + \lambda \left(\frac{1}{|N(i)|} \sum_{j \in N(i)} V_j^{(\text{old})} - V_i^{(\text{old})} \right) \tag{5}$$

In their implementation, the smoothing factor λ is set to 0.2, and 10 iterations are performed.

5.2 Experimental Results

5.2.1 Estimated scale factor

The scale factors estimated using the described method are shown in Table 5. Each image and the corresponding reconstructed 3D model yield a scale factor, and the table presents the average scale factor for each object.

5.2.2 Reconstructed meshes

The refined meshes obtained using the described methods are shown in Figure 12. The predicted model volumes, ground truth model volumes, and the percentage errors between them are presented in Table 6.

5.2.3 Alignment

The team designs a multi-stage alignment method for evaluating reconstruction quality. Figure 13 illustrates the alignment process for Object 14. First, the central points of both the predicted and ground truth models are calculated, and the predicted model is moved to align with the central point of the ground truth model. Next, ICP registration is performed for further alignment, significantly reducing the Chamfer distance. Finally, gradient descent is used for additional fine-tuning to obtain the final transformation matrix.

The total Chamfer distance between all 18 predicted models and the ground truths is 0.069441169.

Table 5: Estimated Scale Factors

Object Index	Food Item	Scale Factor
1	Strawberry	0.060058
2	Cinnamon bun	0.081829
3	Pork rib	0.073861
4	Corn	0.083594
5	French toast	0.078632
6	Sandwich	0.088368
7	Burger	0.103124
8	Cake	0.068496
9	Blueberry muffin	0.059292
10	Banana	0.058236
11	Salmon	0.083821
13	Burrito	0.069663
14	Hotdog	0.073766

Table 6: Metric of Volume

Object Index	Predicted Volume	Ground Truth	Error Percentage
1	44.51	38.53	15.52
2	321.26	280.36	14.59
3	336.11	249.67	34.62
4	347.54	295.13	17.76
5	389.28	392.58	0.84
6	197.82	218.44	9.44
7	412.52	368.77	11.86
8	181.21	173.13	4.67
9	233.79	232.74	0.45
10	160.06	163.09	1.86
11	86.0	85.18	0.96
13	334.7	308.28	8.57
14	517.75	589.83	12.22
16	176.24	262.15	32.77
17	180.68	181.36	0.37
18	13.58	20.58	34.01
19	117.72	108.35	8.64
20	117.43	119.83	20.03

6 Best 3D Mesh Reconstruction Team - FoodRiddle

6.1 Methodology

To achieve high-fidelity food mesh reconstruction, the team developed two procedural pipelines as depicted in Figure 14. For simple and medium complexity cases, they employed a structure-from-motion strategy to ascertain the pose of each image, followed by mesh reconstruction. Subsequently, a sequence of post-processing steps was implemented to recalibrate the scale and improve mesh quality. For cases involving only a single image, the team utilized image generation techniques to facilitate model generation.

6.1.1 Multi-View Reconstruction

For Structure from Motion (SfM), the team enhanced the advanced COLMAP method by integrating SuperPoint and SuperGlue techniques. This integration significantly addressed the issue of limited keypoints in scenes with minimal texture, as illustrated in Figure 15.

In the mesh reconstruction phase, the team's approach builds upon 2D Gaussian Splatting, which employs a differentiable 2D Gaussian renderer and includes regularization terms for depth distortion and normal consistency. The Truncated Signed Distance Function (TSDF) results are utilized to produce a dense point cloud.

During post-processing, the team applied filtering and outlier removal methods, identified the outline of the supporting surface, and projected the lower mesh vertices onto this surface. They utilized the reconstructed checkerboard to correct the model's scale and employed Poisson reconstruction to create a complete, watertight mesh of the subject.

6.1.2 Single-View Reconstruction

For 3D reconstruction from a single image, the team utilized advanced methods such as LGM, Instant Mesh, and One-2-3-45 to generate an initial mesh. This initial mesh was then refined in conjunction with depth structure information.

To adjust the scale, the team estimated the object's length using the checkerboard as a reference, assuming that the object and the checkerboard are on the same plane. They then projected the 3D object back onto the original 2D image to obtain a more precise scale for the object.

6.2 Experimental Results

Through a process of nonlinear optimization, the team sought to identify a transformation that minimizes the Chamfer distance between their mesh and the ground truth mesh. This optimization aimed to align the two meshes as closely as possible in three-dimensional space. Upon completion of this process, the average Chamfer dis- tance across the final reconstructions of the 20 objects amounted to 0.0032175 meters. As shown in Table 7, Team FoodRiddle achieved the best scores for both multi- view and single-view reconstructions, outperforming other teams in the competition.

Table 7: Total Errors for Different Teams on Multi-view and Single-view Data

Team	Multi-view (1-14)	Single-view (16-20)
FoodRiddle	0.036362	0.019232
ININ-VIAUN	0.041552	0.027889
VolETA	0.071921	0.058726

7 Conclusion

This report examines and compiles the techniques and findings from the MetaFood Workshop challenge on 3D Food Reconstruction. The challenge sought to enhance 3D reconstruction methods by concentrating on food items, tackling the distinct difficulties presented by varied textures, reflective surfaces, and intricate geometries common in culinary subjects.

The competition involved 20 diverse food items, captured under various conditions and with differing numbers of input images, specifically designed to challenge participants in creating robust reconstruction models. The evaluation was based on a two-phase process, assessing both portion size accuracy through Mean Absolute Percentage Error (MAPE) and shape accuracy using the Chamfer distance metric.

Of all participating teams, three reached the final submission stage, presenting a range of innovative solutions. Team VolETA secured first place with the best overall performance in both Phase-I and Phase-II, followed by team ININ-VIAUN in second place. Additionally, the FoodRiddle team exhibited superior performance in Phase-II, highlighting a competitive and high-caliber field of entries for 3D mesh reconstruction. The challenge has successfully advanced the field of 3D food reconstruction, demonstrating the potential for accurate volume estimation and shape reconstruction in nutritional analysis and food presentation applications. The novel methods developed by the participating teams establish a strong foundation for future research in this area, potentially leading to more precise and user-friendly approaches for dietary assessment and monitoring.