3D Food Modeling from Images: Advancements in Physically-Aware Reconstruction

Abstract

The growing focus on computer vision for applications in nutritional monitoring and dietary tracking has spurred the creation of sophisticated 3D reconstruction methods for various food items. A lack of high-quality data, combined with insufficient collaboration between academic research and industry applications, has hindered advancements in this area. This paper outlines a comprehensive workshop and challenge centered on physically informed 3D food reconstruction, leveraging recent progress in 3D reconstruction technologies. The central objective of this challenge is to create volume-accurate 3D models of food using 2D images, with a visible checkerboard serving as a critical size reference. Participants were assigned the task of building 3D models for 20 distinct food items, each presenting varying degrees of difficulty: easy, medium, and hard. The easy category offers 200 images, the medium provides 30, and the hard level includes only a single image to facilitate the reconstruction process. During the final evaluation stage, 16 teams presented their results. The methodologies developed during this challenge have yielded encouraging outcomes in 3D food reconstruction, demonstrating considerable potential for enhancing portion estimation in dietary evaluations and nutritional tracking.

1 Introduction

The merging of computer vision with the culinary domain has unveiled new possibilities in dietary oversight and nutritional evaluation. The 3D Food Modeling Workshop Challenge signifies a notable advancement in this domain, responding to the escalating demand for precise and adaptable techniques for estimating food portions and monitoring nutritional consumption. These technological solutions are essential for encouraging beneficial eating patterns and addressing health issues related to diet.

This initiative aims to close the divide between current methodologies and practical needs by concentrating on the development of accurate 3D models of food items from multi-view and single-view image data. The challenge promotes the creation of novel methods capable of managing the intricacies of food forms, textures, and variations in lighting, all while adhering to the practical limitations inherent in real-world dietary assessment situations.

Conventional methods for diet assessment, like 24-Hour Recall or Food Frequency Questionnaires (FFQ), frequently depend on manual data entry, which can be imprecise and difficult to manage. Additionally, the lack of 3D data in 2D RGB food images poses significant hurdles for methods that rely on regression to estimate food portions directly from images of eating occasions. By making progress in 3D reconstruction techniques for food, the aim is to provide tools for nutritional assessment that are more accurate and easier to use. This technology holds the potential to enhance the way food experiences are shared and could significantly influence areas such as nutritional science and public health initiatives.

Participants were tasked with creating 3D models of 20 different food items from 2D images, simulating a scenario where a smartphone equipped with a depth-sensing camera is employed for dietary recording and nutritional oversight. The challenge was divided into three levels of complexity:

.

The easy level provided approximately 200 frames uniformly sampled from a video, the medium level offered about 30 images, and the hard level presented participants with just one monocular top-view image. This arrangement was intended to assess the resilience and adaptability of the suggested solutions under various real-world conditions. One of the main aspects of the challenge involves the use of a visible checkerboard as a tangible benchmark, coupled with the inclusion of depth images for each frame of the video, thereby ensuring the generated 3D models retain precise real-world measurements for estimating portion sizes.

2 Related Work

Estimating food portions is a crucial part of image-based dietary assessment, with the objective of determining the volume, energy content, or macronutrient breakdown directly from images of meals. Unlike the extensively researched area of food recognition, determining food portions presents a distinct difficulty because of the lack of 3D data and physical benchmarks, which are necessary for precisely deducing the actual sizes of food portions. Specifically, accurately estimating portion sizes requires an understanding of the volume and density of the food, aspects that cannot be easily determined from a two-dimensional image, which highlights the need for advanced methodologies and technologies to address this issue. Current methods for estimating food portions are classified into four primary categories.

Stereo-Based Approaches. These techniques depend on multiple frames to deduce the 3D configuration of food items. For instance, some methods calculate food volume through multi-view stereo reconstruction based on epipolar geometry, while others use a two-view dense reconstruction approach. Another technique, Simultaneous Localization and Mapping (SLAM), is employed for continuous, real-time estimation of food volume. However, the need for multiple images limits the practicality of these methods in real-world situations.

Model-Based Approach. This approach uses predefined shapes and templates to estimate the target volume. Some methods assign specific templates to foods from a reference set and make adjustments based on physical cues to gauge the size and position of the food. A similar approach that matches templates is employed to estimate food volume from just one image. However, these methods struggle to accommodate foods with shapes that do not conform to the established templates.

Depth Camera-Based Approach. This method utilizes depth cameras to create maps that indicate the distance from the camera to the food in the picture. The depth map is then used to create a voxel representation of the image, which aids in estimating the food's volume. The primary drawbacks are the need for high-quality depth maps and the additional processing steps required for depth sensors used by consumers.

Deep Learning Approach. Techniques based on neural networks use the vast amount of image data available to train sophisticated networks for estimating food portions. Some use regression networks to estimate the caloric value of food from a single image or from an "Energy Distribution Map" that correlates the input image with the energy distribution of the foods shown. Others use regression networks trained on images and depth maps to deduce the energy, mass, and macronutrients of the food in the image. These methods require extensive data for training and are generally not transparent. Their performance can significantly decline if the input test image deviates substantially from the training data.

Despite the progress these methods have made in estimating food portions, they each have limitations that restrict their broad use and precision in practical scenarios. Methods based on stereo are not suitable for single-image inputs, those based on models have difficulty with a variety of food shapes, approaches using depth cameras necessitate specialized equipment, and deep learning methods are not easily interpretable and have difficulty with samples that are different from those they were trained on. To tackle these issues, 3D reconstruction provides a viable solution by offering thorough spatial data, accommodating different food shapes, possibly functioning with just one image, presenting results that are visually understandable, and facilitating a uniform method for estimating food portions. These benefits were the driving force behind the organization of the 3D Food Reconstruction challenge, which seeks to surmount the current limitations and create techniques for food portion estimation that are more accurate, user-friendly, and broadly applicable, thereby making a significant impact on nutritional assessment and dietary monitoring.

3 Datasets and Evaluation Pipeline

3.1 Dataset Description

The dataset for the 3D Food Modeling Challenge includes 20 carefully chosen food items, each having been scanned with a 3D scanner and also captured on video. To ensure the reconstructed 3D models accurately represent size, each food item was captured alongside a checkerboard and pattern mat, which provide a physical reference for scaling. The challenge is segmented into three levels of difficulty, based on the number of 2D images provided for reconstruction:

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

Table 1: 3D Food Modeling 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 is divided into two stages, focusing on the accuracy of the reconstructed 3D models in terms of their form (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 as the metric to evaluate the accuracy of portion size. The calculation for MAPE is as follows:

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

where A_i represents the actual volume (in milliliters) of the i-th food item, as determined from the scanned 3D 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 provide full 3D mesh files for each food item. This phase includes multiple steps to guarantee both accuracy and fairness:

1. **Model Verification**: Submitted models are checked against the final submissions from Phase-I to ensure they are consistent. Visual inspections are also conducted to prevent any violations of the rules, such as submitting basic shapes (like spheres) rather than detailed reconstructions.
2. **Model Alignment**: Participants are given the true 3D models and the script used for calculating the final Chamfer distance. They must align their models with these true models and create a transformation matrix for each item submitted. The ultimate Chamfer distance score is then calculated using the submitted models and their corresponding transformation matrices.
3. **Chamfer Distance Calculation**: The accuracy of the shape is assessed using the Chamfer distance. For two sets of points, X and Y, the Chamfer distance is computed as follows:

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

This metric offers a thorough assessment of how closely the reconstructed 3D models match the actual models. The ultimate ranking is determined by merging the scores from both Phase-I (accuracy of volume) and Phase-II (accuracy of shape). It should be noted that after evaluating Phase-I, some issues with the data quality for object 12 (steak) and object 15 (chicken nugget) were found. To maintain the competition's quality and fairness, these two items have been removed 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 accurately determine 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, supported by perceptual hashing and blur detection, ensures data quality. The estimation of camera poses and object segmentation establishes the basis for neural surface reconstruction, resulting in detailed meshes for volume estimation. Refinement processes, such as removing isolated parts and adjusting the scaling factor, improve accuracy.

4.1.2 The Team's Proposal: VolETA

The team starts their process by obtaining input data, specifically RGBD images and their corresponding food object masks. These RGBD images are denoted as $I^D = \{I_i^D\}_{i=1}^n$, where n is the total number of frames, providing the necessary depth information alongside the RGB images. The food object masks, denoted as $\{M_i^F\}_{i=1}^n$, help identify the regions of interest within these images.

Next, the team proceeds with keyframe selection. From the set $\{I_i^D\}_{i=1}^n$, keyframes $\{I_j^K\}_{j=1}^k \subseteq \{I_i^D\}_{i=1}^n$ are selected. The team implements a method to detect and remove duplicates and blurry images to ensure high-quality frames. This involves applying the Gaussian blurring kernel followed by the fast Fourier transform method. Near-Image Similarity employs a perceptual hashing and hamming distance thresholding to detect similar images and keep overlapping. The duplicates and blurry images are excluded from the selection process to maintain data integrity and accuracy.

Using the selected keyframes $\{I_j^K\}_{j=1}^k$, the team estimates the camera poses through a Structure from Motion approach (i.e., extracting features using a feature detection method, matching them

using a matching algorithm, and refining them). The outputs are the set of camera poses $\{C_j\}_{j=1}^{k}$, which are crucial for spatial understanding of the scene.

In parallel, the team utilizes a segmentation algorithm for reference object segmentation. This algorithm segments the reference object with a user-provided segmentation prompt (i.e., user click), producing a reference object mask M^R for each keyframe. This mask is a foundation for tracking the reference object across all frames. The team then applies a memory tracking method, which extends the reference object mask M^R to all frames, resulting in a comprehensive set of reference object masks $\{M_i^R\}_{i=1}^{n}$. This ensures consistency in reference object identification throughout the dataset.

To create RGBA images, the team combines the 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 the various data sources into a unified format suitable 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 the accurate reconstruction of the scene.

The modeled data D_m is then input into a neural surface reconstruction algorithm for mesh reconstruction. This algorithm generates colorful meshes $\{R_f, R_r\}$ for the reference and food objects, providing detailed 3D representations of the scene components. The team applies the "Remove Isolated Pieces" technique to refine the reconstructed meshes. Given that the scenes contain only one food item, the team sets the diameter threshold to 5% of the mesh size. This method deletes isolated connected components whose diameter is less than or equal to this 5% threshold, resulting in a cleaned mesh $\{RC_f, RC_r\}$. This step ensures that only significant and relevant parts of the mesh are retained.

The team manually identifies an initial scaling factor S using the reference mesh via a mesh processing tool for scaling factor identification. This factor is then fine-tuned 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 generate unitless meshes (i.e., no physical scale) by default. To overcome this limitation, the team manually identifies the scaling factor by measuring the distance for each block for the reference object mesh. Next, the team takes the average of all blocks lengths l_{avg}, while the actual real-world length is constant $l_{real} = 0.012$ in meter. Furthermore, the team applies the scaling factor $S = l_{real}/l_{avg}$ on the clean food mesh RC_f, producing the final scaled food mesh RF_f in meter.

The team leverages depth information alongside food and reference object masks to validate the scaling factors. The team's method for assessing food size entails utilizing overhead RGB images for each scene. Initially, the team determines the pixel-per-unit (PPU) ratio (in meters) using the reference object. Subsequently, the team extracts the food width (fw) and length (fl) employing a food object mask. To ascertain the food height (fh), the team follows a two-step process. Firstly, the team conducts binary image segmentation using the overhead depth and reference images, yielding a segmented depth image for the reference object. The team then calculates the average depth utilizing the segmented reference object depth (dr). Similarly, employing binary image segmentation with an overhead food object mask and depth image, the team computes the average depth for the segmented food depth image (df). Finally, the estimated food height fh is computed as the absolute difference between dr and df. Furthermore, to assess the accuracy of the scaling factor S, the team computes the food bounding box volume ((fw × fl × fh) × PPU). The team evaluates if the scaling factor S generates a food volume close to this potential volume, resulting in S_{fine}.

For one-shot 3D reconstruction, the team leverages a single view reconstruction method for reconstructing a 3D from a single RGBA view input after applying binary image segmentation on both food RGB and mask. Next, the team removes isolated pieces from the generated mesh. After that, the team reuses the scaling factor S, which is closer to the potential volume of the clean mesh.

4.2 Experimental Results

4.2.1 Implementation settings

The experiments were conducted using two GPUs: a GeForce GTX 1080 Ti with 12GB of memory and an RTX 3060 with 6GB of memory. For near-image similarity detection, the Hamming distance was set to 12. To identify blurry images, even numbers within the range of [0...30] were used as the Gaussian kernel radius. In the process of removing isolated pieces, a diameter threshold of 5% was applied. Neural surface reconstruction involved 15,000 iterations, with a mesh resolution of 512x512. The unit cube parameters were set with an "aabb scale" of 1, "scale" at 0.15, and "offset" at [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. More Briefly, the team leverages their approach for each food scene separately. A one-shot food volume estimation approach is applied if the number of keyframes k equals 1. Otherwise, a few-shot food volume estimation is applied. The team's keyframe selection process chooses 34.8% of total frames for the rest of the pipeline, where it shows the minimum frames with the highest information.

Table 2: List of Extracted Information Using RGBD and Masks

Level	Id	Label	S_f	PPU	$R_w \times R_l$	$f_w \times f_l \times f_h$	Volume (cm^3)
Easy	1	strawberry	0.08955	0.01786	320 × 360	238 × 257 × 2.353	45.91
	2	cinnamon bun	0.10435	0.02347	236 × 274	363 × 419 × 2.353	197.07
	3	pork rib	0.10435	0.02381	246 × 270	435 × 778 × 1.176	225.79
	4	corn	0.08824	0.01897	291 × 339	262 × 976 × 2.353	216.45
	5	french toast	0.10345	0.02202	266 × 292	530 × 581 × 2.53	377.66
	6	sandwich	0.12766	0.02426	230 × 265	294 × 431 × 2.353	175.52
	7	burger	0.10435	0.02435	208 × 264	378 × 400 × 2.353	211.03
	8	cake	0.12766	0.02143	256 × 300	298 × 310 × 4.706	199.69
Medium	9	blueberry muffin	0.08759	0.01801	291 × 357	441 × 443 × 2.353	149.12
	10	banana	0.08759	0.01705	315 × 377	446 × 857 × 1.176	130.80
	11	salmon	0.10435	0.02390	242 × 269	201 × 303 × 1.176	40.94
	13	burrito	0.10345	0.02372	244 × 271	251 × 917 × 2.353	304.87
	14	frankfurt sandwich	0.10345	0.02115	266 × 304	400 × 1022 × 2.353	430.29
Hard	16	everything bagel	0.08759	0.01747	306 × 368	458 × 484 × 1.176	79.61
	17	croissant	0.12766	0.01751	319 × 367	395 × 695 × 2.176	183.39
	18	shrimp	0.08759	0.02021	249 × 318	186 × 195 × 0.987	14.64
	19	waffle	0.01034	0.01902	294 × 338	465 × 537 × 0.8	72.29
	20	pizza	0.01034	0.01913	292 × 336	442 × 651 × 1.176	123.97

After generating the scaled meshes, the team calculates the volumes and Chamfer distance with and without transformation metrics. The team registered their meshes and ground truth meshes to obtain the transformation metrics using ICP.

5 Second Place Team - ININ-VIAUN

5.1 Methodology

This section provides a detailed explanation of the proposed network, demonstrating how to progress from the original images to the final mesh models step by step.

5.1.1 Scale factor estimation

The pipeline for coordinate-level scale factor estimation is described as follows. The team follows a corner projection matching method. Specifically, using a dense reconstruction model, the team

Table 3: Quantitative Comparison of Team's Approach with Ground Truth

L	Id	Team's Vol.	GT Vol.	Ch. w/ t.m	Ch. w/o t.m
E	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
	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
M	9	224.08	232.74	3.91	160.07
	10	153.76	163.09	2.67	138.45
	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
H	16	163.13	262.15	18.06	28.33
	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: Overall Method Performance

MAPE	Ch. sum w/tm	mean	Ch. w/o tm	mean
10.973	0.130	0.007	1.715	0.095

obtains the pose of each image as well as dense point cloud information. For any image img_k and its extrinsic parameters $[R|t]_k$, the team first performs a threshold-based corner detection with the threshold set to 240. This 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:

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

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 the following equation, where 0.012 represents the known length of each square (1.2 cm):

$$\text{scale} = \frac{0.012}{\text{med}(d^k)}$$

5.1.2 3D Reconstruction

Considering the differences in input viewpoints, the team utilizes two pipelines to process the first fifteen objects and the last five single view objects.

For the first fifteen objects, the team uses a Structure from Motion algorithm to estimate the poses and segment the food using the provided segment masks in the dataset. Then, they apply advanced multi-view 3D reconstruction methods to reconstruct the segmented food. In practice, the team employs three different reconstruction methods. They select the best reconstruction results from these methods and extract the mesh from the reconstructed model. Next, they scale the extracted mesh using the estimated scale factor. Finally, they apply some optimization techniques to obtain a refined mesh.

For the last five single-view objects, the team experiments with several single-view reconstruction methods. They choose a specific method to obtain a 3D food model consistent with the distribution of the input image. In practice, they use the intrinsic camera parameters from the fifteenth object and employ an optimization method based on reprojection error to refine the extrinsic parameters of the single camera. However, due to the limitations of single-view reconstruction, the team needs to incorporate depth information from the dataset and the checkerboard in the monocular image to determine the size of the extracted mesh. Finally, they apply optimization techniques to obtain a refined mesh.

5.1.3 Mesh refinement

In the 3D Reconstruction phase, the team observes that the model's results often suffer from low quality due to the presence of holes on the object surface and substantial noise.

To address the holes, the team employs an optimization method based on computational geometry. For surface noise, they utilize Laplacian Smoothing for mesh smoothing operations. The Laplacian Smoothing method works by adjusting the position of each vertex to the average of its neighboring vertices:

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

In their implementation, the team sets the smoothing factor λ to 0.2 and performs 10 iterations.

5.2 Experimental Results

5.2.1 Estimated scale factor

The scale factors estimated using the method described earlier 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.

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

5.2.2 Reconstructed meshes

The refined meshes obtained using the methods described earlier are shown in Figure 12. The predicted model vol- umes, ground truth model volumes, and the percentage errors between them are shown in Table 6. The unit is cubic millimeters.

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

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 team calculates the central points of both the predicted model and the ground truth model, and moves the predicted model to align the central point of the ground truth model. Next, they perform ICP registration for further alignment, significantly reducing the Chamfer distance. Finally, they use gradient descent for additional fine-tuning, and obtain the final transformation matrix. The total Chamfer distance between all 18 predicted models and the ground truths is 0.069441169.

6 Best 3D Mesh Reconstruction Team - FoodRiddle

6.1 Methodology

To achieve high-quality food mesh reconstruction, the team designed two pipeline processes. For simple and medium cases, they employed a structure-from-motion approach to determine the pose of each image, followed by mesh reconstruction. Subsequently, a series of post-processing steps were implemented to recalibrate scale and enhance mesh quality. For cases with only a single image, the team utilized image generation methods to aid in model generation.

6.1.1 Multi-View Reconstruction

For Structure from Motion (SfM), the team extended the state-of-the-art method by incorporating methodologies. This significantly mitigated the issue of sparse keypoints in weakly textured scenes. For mesh reconstruction, the team's method is based on a differentiable renderer and incorporates regularization terms for depth distortion and normal consistency. The Truncated Signed Distance Function (TSDF) results are used to generate a dense point cloud. In the post-processing stage, the team applied filtering and outlier removal techniques, identified the contour of the supporting surface, and projected the lower mesh vertices onto the supporting surface. They used the reconstructed checkerboard to rectify the scale of the model and used Poisson reconstruction to generate a watertight, complete mesh of the subject.

6.1.2 Single-View Reconstruction

For 3D reconstruction from a single image, the team employed state-of-the-art methods to generate an initial prior mesh. This prior mesh was then jointly corrected with depth structure information.

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

In this report, we provide a summary and analysis of the methodologies and findings from the 3D Food Reconstruction challenge. The primary goal of this challenge was to push the envelope in 3D reconstruction technologies, with an emphasis on the unique challenges presented by food items, such as their varied textures, reflective surfaces, and complex geometries. The competition featured 20 diverse food items, captured under various conditions and with varying numbers of input images, specifically designed to challenge participants in developing 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 made it to the final submission, showcasing a range of innovative solutions. Team VolETA won first place with the overall best performance on both Phase-I and Phase-II, followed by team ININ-VIAUN who won second place. In addition, FoodRiddle team demonstrated superior performance in Phase-II, indicating a competitive and high-caliber field of entries for 3D mesh reconstruction. The challenge has successfully pushed the boundaries of 3D food reconstruction, demonstrating the potential for accurate volume estimation and shape reconstruction in nutritional analysis and food presentation applications. The innovative approaches developed by the participating teams provide a solid foundation for future research in this field, potentially leading to more accurate and user-friendly methods for dietary assessment and monitoring.