Physically Informed 3D Food Reconstruction: Methods and Results

Jiangpeng He[1†], Yuhao Chen[2†], Gautham Vinod[1], Xiaoyan Zhang[3], Talha Ibn Mahmud[1], Ahmad AlMughrabi[5], Umair Haroon[5], Ricardo Marques[10], Petia Radeva[5,11], Jiadong Tang[6], Dianyi Yang[6], Yu Gao[6], Zhaoxiang Liang[6], Yawei Jueluo[7], Chengyu Shi[8], Pengyu Wang[9], Pengcheng Xi[4], Alexander Wong[2], Edward Delp[1], Fengqing Zhu[1*]

[1*]Purdue University.
[2]University of Waterloo.
[3]Anhui University.
[4]National Research Council Canada.
[5]Universitat de Barcelona.
[6]Beijing Institute of Technology.
[7]Baidu Inc..
[8]XPeng Motors.
[9]Beijing University of Posts and Telecommunications.
[10]Universitat Pompeu Fabra.
[11]Institut de Neurociències.

*Corresponding author(s). E-mail(s): zhu0@purdue.edu;
†These authors contributed equally to this work.

Abstract

Accurate food portion size estimation is a critical challenge in nutrition analysis and dietary assessment. Recent 3D reconstruction methods primarily focus on surface geometry, often neglecting the volumetric accuracy necessary for precise portion size estimation. The MetaFood Workshop hosted a challenge on Physically Informed 3D Food Reconstruction, with the goal of reconstructing volume-accurate 3D models of food items from 2D images, using a visible checkerboard as a size reference. The challenge included three difficulty levels—easy, medium, and hard—where participants reconstructed 3D models of 20 selected

food items. A total of 16 teams submitted methods that demonstrated the potential to improve 3D food reconstruction and enhance food portion estimation accuracy. These solutions address real-world complexities such as varying camera positions, lighting conditions, and food shape diversity, contributing to more robust and scalable tools for nutritional monitoring and dietary assessment. The specific information of the MetaFood Workshop can be found in this link. This paper discusses the workshop challenge, its outcomes, and future directions for 3D food reconstruction in nutritional monitoring and dietary assessment.

Keywords: 3D food reconstruction, Physical reference, Image-based dietary assessment, Computer vision, Deep learning

1 Introduction

The intersection of computer vision and culinary arts has opened new frontiers in dietary monitoring and nutritional analysis. The CVPR 2024 MetaFood Workshop Challenge represents a significant step in this direction, addressing the growing need for accurate, scalable methods of food portion estimation and nutritional intake tracking. These technologies are crucial for promoting healthy eating habits and managing diet-related health conditions.

This challenge aims to bridge the gap between existing methods and real-world requirements by focusing on the reconstruction of accurate 3D models of food items from both multi-view and single-view inputs. The challenge encourages the development of innovative techniques capable of handling the complexities of food shapes, textures, and lighting conditions. While also addressing practical constraints in real-world dietary assessment scenarios, including dynamic camera movement, irregular capturing topologies, and varying distances between the camera and the food. These techniques must be sufficiently robust to process images taken from multiple angles and orientations, thus overcoming the limitations of current approaches that often rely on controlled environments with fixed camera positions. By bringing together researchers and practitioners in computer vision, machine learning, and nutrition science, this challenge seeks to catalyze advancements in 3D food reconstruction that could significantly improve the accuracy and applicability of food portion estimation in various contexts, ranging from personal health monitoring to large-scale nutritional studies.

Traditional diet assessment methods [1], such as the 24-Hour Recall and Food Frequency Questionnaire (FFQ), often rely on manual input, which can be both inaccurate and cumbersome. Furthermore, the absence of 3D information in 2D RGB food images poses significant challenges for regression-based methods [2, 3] that estimate food portions directly from eating occasion images. By advancing 3D reconstruction techniques for food items, we aim to provide more precise and intuitive tools for nutritional assessment. This technology has the potential to enhance the sharing of food experiences and significantly impact fields such as nutrition science and public health.

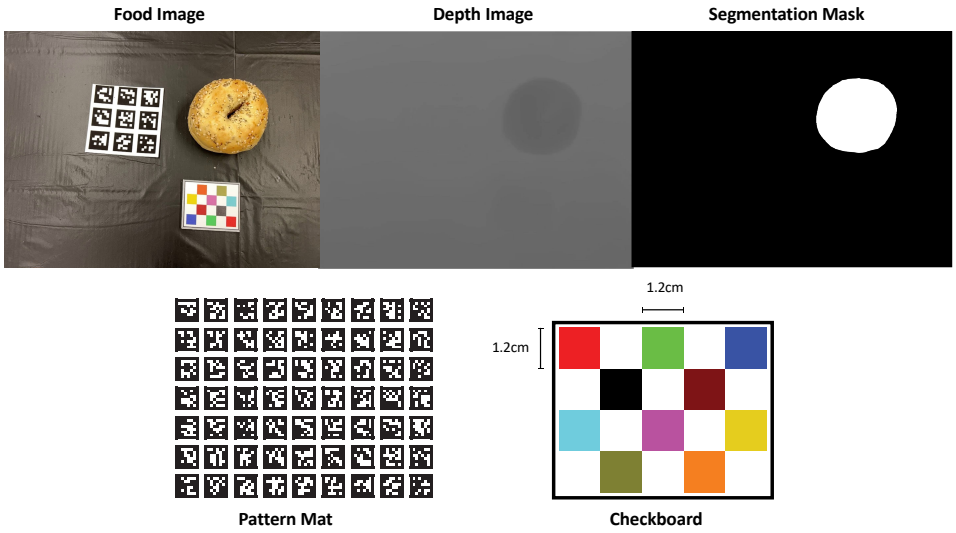


Fig. 1: Sample challenge data for "Everything bagel".

The challenge tasked participants with reconstructing 3D models of 20 selected food items from 2D images, simulating a scenario in which a smartphone equipped with a depth camera is used for food logging and nutritional monitoring. The challenge was structured into three difficulty levels. The easy-level food objects provided approximately 200 frames uniformly sampled from videos, the medium level offered about 30 images, and the hard level presented participants with a single monocular top-view image. This structure was designed to test the robustness and versatility of the proposed solutions across various real-world scenarios. A sample of a hard food object is shown in Figure 1. The key features of the challenge include the use of a visible checkerboard as a physical reference, as well as the availability of the depth image for each video frame, ensuring that the reconstructed 3D models maintain accurate real-world scaling for portion size estimation.

This challenge not only pushes the boundaries of 3D reconstruction technology but also paves the way for more accurate, robust, and user-friendly applications in real-world scenarios, such as image-based dietary assessment. The solutions developed in this challenge have the potential to significantly impact how we monitor and understand our nutritional intake, contributing to broader goals in health and wellness. As we continue advancing in this field, we anticipate the emergence of innovative applications that could revolutionize personal health management, nutritional research, and the food industry. The structure of this technical report is organized as follows: In Section 2, we review existing related work for food portion size estimation. Section 3 introduces the dataset utilized in this challenge and the detailed evaluation pipelines. Next, we summarize the methodologies proposed by the three winning teams (*VolETA*, *ININ-VIAUN*, *FoodRiddle*) in Section 4.1, Section 4.2, and Section 4.3, respectively. Finally, we present the shared results across the three teams in Section 5.1 along with

the intermediate results and specific analysis results for each team in Section 5.2.1, Section 5.2.2, and Section 5.2.3.

The main contributions of the challenge are summarized as follows:

- The challenge showcased novel 3D reconstruction approaches that overcome the limitations of traditional food portion estimation methods, such as the need for multiple images in stereo-based methods, lack of flexibility in model-based approaches, and reliance on specialized hardware in depth camera-based techniques.
- It highlighted the advancement of 3D mesh reconstruction techniques capable of reconstructing food items accurately from limited inputs, including single-view images. This development is crucial for making food portion estimation more feasible in real-world settings, where capturing multiple views is often impractical.
- The challenge focused on real-world complexities, such as varying camera positions, lighting, and food shape diversity, which traditional methods often fail to handle effectively. By incorporating these aspects, the challenge contributed towards more robust and accurate 3D reconstruction models suitable for practical dietary assessment applications.

2 Related Work

Food portion estimation is an important component of image-based dietary assessment [4, 5] with the goal of estimating the volume, energy or macronutrients directly from the input eating occasion images. Compared to the widely studied food recognition task [6–12], food portion estimation presents a unique challenge due to the absence of 3D information and physical references, which are essential for accurately inferring the real-world size 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. Existing food portion estimation methods can be categorized into four main groups [13].

Stereo-Based Approach. These methods rely on multiple frames to reconstruct the 3D structure of the food. In [14], food volume is estimated using multi-view stereo reconstruction based on epipolar geometry. Similarly, [15] performs two-view dense reconstruction. Simultaneous Localization and Mapping (SLAM) is utilized in [16] for continuous and real-time food volume estimation. The primary limitation of these methods is the requirement for multiple images, which is impractical for real-world deployment.

Model-Based Approach. Predefined shapes and templates are leveraged to estimate the target volume. For instance, [17] assigns certain templates to foods from a library and applies transformations based on physical references to estimate the size and location of the food. A similar template matching approach is used in [18] to estimate food volume from a single image. However, these methods cannot accommodate variations in food shapes that deviate from predefined templates. The most recent work [19] leveraged the 3D food mesh as the template to align both camera pose and object pose for portion size estimation.

Depth Camera-Based Approach. The depth camera is utilized to produce a depth map that captures the distance from the camera to the food in the image. In [20, 21], the depth map is used to form a voxel representation of the image, which is then used to estimate the food volume. The main limitation is the requirement for high-quality depth maps and additional post-processing needed for consumer depth sensors.
Deep Learning Approach. Neural network-based methods leverage the abundance of image data to train complex networks for food portion estimation. Regression networks are used in [2, 22] to estimate the energy value of food from a single image input and from an "Energy Distribution Map," which maps the input image to the energy distribution of the foods in the image. In [23], regression networks trained on input images and depth maps produce energy, mass, and macronutrient information for the food(s) in the image. Deep learning-based methods require large amounts of data for training and are generally not explainable. Their performance often degrades when the input test image differs significantly from the training data.

Although these approaches have made significant strides in food portion estimation, they all face limitations that hinder their widespread adoption and accuracy in real-world scenarios. Stereo-based methods are impractical for single-image input, model-based approaches struggle with diverse food shapes, depth camera-based methods require specialized hardware, and deep learning approaches lack explainability and struggle with out-of-distribution samples. To address these challenges, 3D reconstruction offers a promising solution by providing comprehensive spatial information, adapting to various food shapes, potentially working with single images, offering visually interpretable results, and enabling a standardized approach to food portion estimation. These advantages motivated the organization of the 3D Food Reconstruction challenge, with the aim of overcoming existing limitations and developing more accurate, user-friendly, and widely applicable food portion estimation techniques that can significantly impact nutritional assessment and dietary monitoring.

There is a notable shortage of food datasets that include both 2D images and 3D models. For instance, SimpleFood45 [24] comprises 12 food types, each containing 3-4 food items. The dataset provides enough 2D images, but food items within the same category share a single 3D model. SimpleFood45 also includes detailed food portion annotations, such as weight, energy, and volume, making it a valuable resource for training food portion estimation models. However, the challenge mainly focused on the precise estimation of volume and the intricate reconstruction of 3D meshes from 2D images. In contrast, the new dataset MetaFood3D [25], encompasses a wide variety of food types, and each food item is represented by its 3D ground truth mesh scanned by a 3D scanner, along with 2D images. Additionally, it contains explicit nutritional annotations, including volume. Therefore, We chose MetaFood3D as the challenge dataset for evaluating volume-accurate 3D reconstruction methods.

3 Datasets and Evaluation Pipeline

In this section, we introduce the detailed composition of the challenge dataset and establish the two-step rule for the evaluation method.

3.1 Dataset Description

The MetaFood Challenge dataset comprises 20 carefully selected food items from MetaFood3D dataset [25] [1], each scanned with the 3D scanner (Revopoint POP 2[2]) and accompanied by corresponding video captures recorded using an iPhone 13 Pro with the Record3D app[3]. To ensure accurate size representation in the reconstructed 3D models, each item was captured alongside a checkerboard and pattern matrix, serving as physical references for scaling [26]. The challenge is structured into three difficulty levels, determined by the number of 2D images available for reconstruction:

- Easy: Approximately 200 images sampled from video; Medium: 30 images; Hard: A single monocular top-view image.

Table 1 provides detailed information about the food items 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 consists of two phases, focusing on the precision of the reconstructed 3D models in terms of their shape (3D structure) and portion size (volume).

[1] MetaFood3D - the dataset can be accessed at this link.
[2] https://www.revopoint3d.com/pages/face-3d-scanner-pop2
[3] https://record3d.app/

3.2.1 Phase-I: Volume Accuracy

In the first phase, we employ Mean Absolute Percentage Error (MAPE) as the metric to assess the accuracy of portion size. The MAPE is calculated as follows:

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

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

3.2.2 Phase-II: Shape Accuracy

The top-ranking teams from Phase-I are invited to submit complete 3D mesh files for each food item. This phase involves several steps to ensure accuracy and fairness:

- Model Verification: We verify the submitted models against the final Phase-I submissions to ensure consistency. Additionally, we conduct visual inspections to prevent rule violations, such as submitting primitive objects (e.g., spheres) instead of detailed reconstructions.
- Model Alignment: We will provide participants with the ground truth 3D models, along with the script used to compute the final Chamfer distance. Participants are required to align their models with the ground truth and prepare a transform matrix for each submitted object. The final Chamfer distance score will be computed using these submitted models and transformation matrices.
- Chamfer Distance Calculation: The shape accuracy is evaluated using the Chamfer distance metric. Given two point sets X and Y, the Chamfer distance is defined as in the following.

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

This metric provides a comprehensive measure of the similarity between the reconstructed 3D models and the ground truth. The final ranking will be determined by combining the scores from both Phase-I (volume accuracy) and Phase-II (shape accuracy). Note that after the Phase-I evaluation, we observed some quality issues with the provided data for object 12 (steak) and object 15 (chicken nugget). To ensure the quality and fairness of the competition, we have decided to exclude these two items from the final overall evaluation process.

4 Methodology

In this section, we summarize the methods proposed by three winning teams, VolETA, ININ-VIAUN, and FoodRiddle, in Sections 4.1, 4.2, and 4.3, respectively.

4.1 First Place Team - VolETA

The team's study utilizes multi-view reconstruction to create intricate food meshes and calculate precise food volumes. The paper can be accessed at [4].

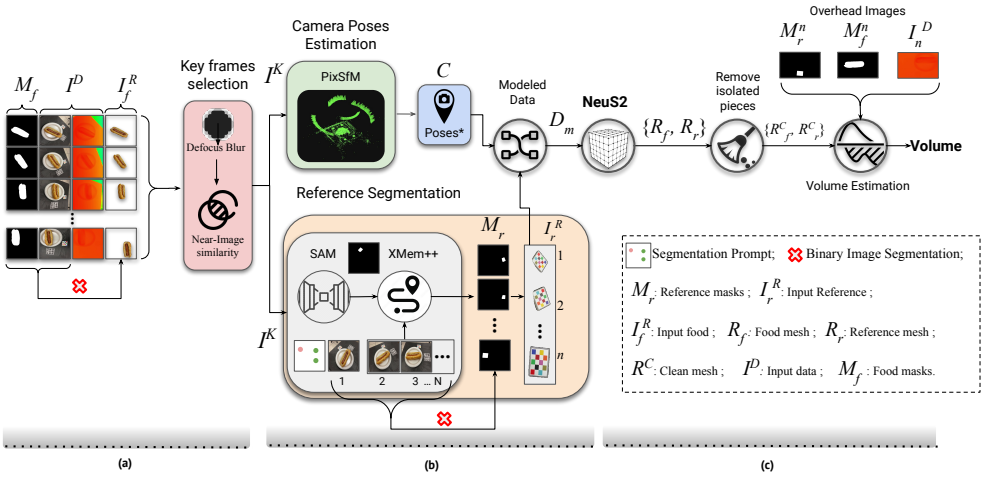


Fig. 2: The team's few-shot approach for estimating food volume in (a) a few shots involves taking ($\mathcal{I}^D$) and food object masks as input. The team starts by selecting keyframes based on the RGB images, removing blurry and highly overlapped images resulting (I^K). Then, (b) the team uses PixSfM to estimate camera poses (C). Simultaneously, the team segments the reference object using SAM with a segmentation prompt provided by a user. The team then uses the XMem++ method for memory-tracking to produce reference object masks for all frames, using the reference object mask and RGB images. After that, the team applies a binary image segmentation method to RGB images (I^K), reference object masks (M_r), and food object masks (M_f), resulting in RGBA images (I_r^R). In contrast, the team transforms the RGBA images and poses to generate meaningful metadata and create modeled data (D_m). Next, (c) the team inputs the modeled data into NeuS2 to reconstruct colorful meshes for reference (R_r) and food objects (R_f). To ensure accuracy, the team uses "Remove Isolated Pieces" with diameter thresholding to clean up the mesh and remove small isolated pieces that do not belong to the reference or food mesh resulting ($\{R_r^C, R_f^C\}$). Finally, the team manually identifies the scaling factor using the reference mesh via MeshLab (S). The team fine-tunes the scaling factor using depth information and the food masks and then applies the fine-tuned scaling factor (S_f) to the cleaned food mesh to generate a scaled food mesh (R_f^F) in meter unit.

[4] https://arxiv.org/abs/2407.01717

4.1.1 Overview

The team's approach combines computer vision and deep learning techniques to estimate food volume from RGBD images and masks accurately. Keyframe selection ensures data quality, aided by perceptual hashing and blur detection. Camera pose estimation and object segmentation lay the groundwork for neural surface reconstruction, producing detailed meshes for volume estimation. Refinement steps enhance accuracy, including isolated piece removal and scaling factor adjustment. The team's approach offers a comprehensive solution for precise food volume assessment with potential applications in nutrition analysis.

4.1.2 The Team's Proposal: VolETA

The team begins their approach by acquiring input data, specifically RGBD images and corresponding food object masks. These RGBD images, denoted as $\mathcal{I}^D = \{I_i^D\}_{i=1}^n$, where n is the total number of frames, provide the necessary depth information alongside the RGB images. The food object masks, denoted as $\{M_f^i\}_{i=1}^n$, aid in identifying the regions of interest within these images.

Next, the team proceeds with keyframe selection. From the set $\{I_i^D\}_{i=1}^n$, keyframes $\{I_i^K\}_{j=1}^k \subseteq \{I_i^D\}_{i=1}^n$ are selected. The team implements a method to detect and remove duplicates [27] and blurry images [28] to ensure high-quality frames. This involves applying the Gaussian blurring kernel followed by the fast Fourier transform method. Near-Image Similarity [27] 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, as shown in Fig. 2(a).

Using the selected keyframes $\{I_j^K\}_{j=1}^k$, the team estimates the camera poses through PixSfM [29] (i.e., extracting features using SuperPoint [30], matching them using SuperGlue [31], 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 the SAM [32] for reference object segmentation. SAM 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 the XMem++ [33] method for memory tracking, 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, as shown in Fig. 2(b).

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 NeuS2 [34] for mesh reconstruction. NeuS2 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 $\{R_f^C, R_r^C\}$. This step ensures that only significant and relevant parts of the mesh are retained.

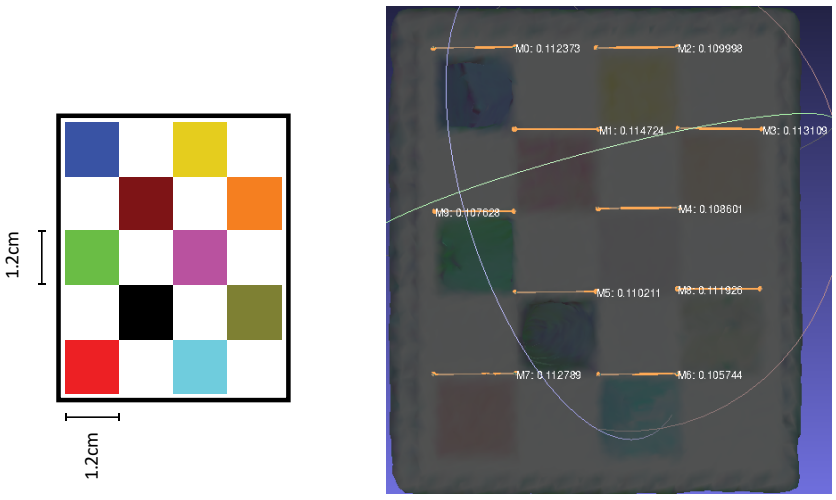
The team manually identifies an initial scaling factor S using the reference mesh via MeshLab [35] for scaling factor identification. This factor is then fine-tuned S_f using depth information and food and reference masks, ensuring accurate scaling relative to real-world dimensions. Finally, the fine-tuned scaling factor S_f is applied to the cleaned food mesh R_f^C, producing the final scaled food mesh R_f^F. This step culminates in an accurately scaled 3D representation of the food object, enabling precise volume estimation, as shown in Fig. 2(c).

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, as shown in Fig. 3b. Next, the team takes the average of all blocks lengths l_{avg}, while the actual real-world length (as shown in Fig. 3a) 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 R_f^C, producing the final scaled food mesh R_f^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 (f_w) and length (f_l) employing a food object mask. To ascertain the food height (f_h), 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 (d_r). 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 (d_f). Finally, the estimated food height f_h is computed as the absolute difference between d_r and d_f. Furthermore, to assess the accuracy of the scaling factor S, the team computes the food bounding box volume ($(f_w \times f_l \times f_h) \times PPU$). The team evaluates if the scaling factor S generates a food volume close to this potential volume, resulting in S_{fine}. Detailed results can be found in Appendix A.

For one-shot 3D reconstruction, the team leverages One-2-3-45 [36] 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 to ensure the volume of the new mesh is closer to the potential volume of the clean mesh, as illustrated in Fig. 4.



(a) (b)

Fig. 3: The team manually measures the scaling factor using MeshLab's Measuring tool. The team measures multiple blocks in the reference object mesh; then, the team takes the average of blocks lengths l_{avg}.

4.2 Second Place Team - ININ-VIAUN

This team provides a detailed explanation of their proposed network, demonstrating how to progress from the original images to the final mesh models step by step. The code is available at [5].

4.2.1 Scaling factor estimation

The pipeline for coordinate-level scaling factor estimation is shown in Figure 5. The team follows a corner projection matching method. Specifically, using the COLMAP[37] dense model, the team 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:

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

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

[5] https://github.com/BITyia/cvpr-metafood

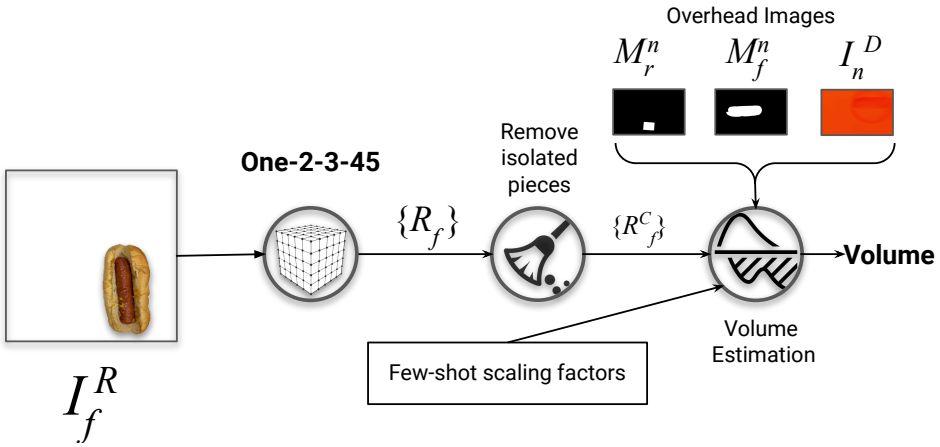


Fig. 4: The team's one-shot food volume estimation approach. The team begins with a food-segmented image (I_f^R), and then uses the One-2-3-45 model to generate a mesh (R_f). Next, the team cleans up the isolated pieces that are less than 5% of the (R_f) size, resulting in a cleaned food mesh R_f^C. Furthermore, the team chooses a scaling factor based on the depth information S_f. Finally, the team applies the chosen scaling factor on R_f^C to have a scaled mesh (R_f^F) where the team extracts the volume.

formula is given by Equation 4, where 0.012 represents the known length of each square (1.2 cm):

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

4.2.2 3D Reconstruction

The pipeline for 3D reconstruction is shown in Figure 6. 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 COLMAP[37] 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: COLMAP[37], DiffusioNeRF[38], and NeRF2Mesh[39]. 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 scaling factor. Finally, they apply some optimization techniques to obtain a refined mesh as illustrated in Section 4.2.3.

For the last five single-view objects, the team experiments with several single-view reconstruction methods, such as Zero123[40], Zero123++[41], One2345[36], ZeroNVS[42], and DreamGaussian[43]. They choose ZeroNVS[42] 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

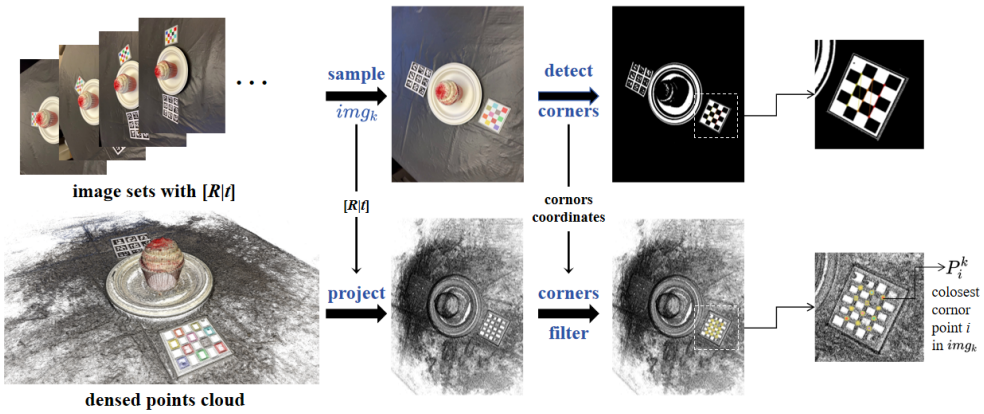


Fig. 5: The pipeline of scaling factor estimation. The process involves corner detection in images, projection of dense point clouds onto image planes, calculation of distances between corner points, and determination of checkerboard square lengths. The final scale is computed using the median of minimum distances and the known square length (1.2 cm)

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 as illustrated in Section 4.2.3.

4.2.3 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, as illustrated in Figure 7.

To address the holes, the team employs MeshFix[44], 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:

$$v_i^{(new)} = v_i^{(old)} + \lambda \left(\frac{1}{N} \sum_{j \in \mathcal{N}(i)} v_j^{(old)} - v_i^{(old)} \right) \tag{3}$$

where $v_i^{(old)}$ is the point on the noisy surface, and $v_i^{(new)}$ is the adjusted new position of the old point. $\mathcal{N}(i)$ denotes the area of vertices in the neighborhood of point $v_i^{(old)}$, $v_j^{(old)}$ is the neighboring point of $v_i^{(old)}$, and N is the total number of neighboring points. In their implementation, the team sets the smoothing factor λ to 0.2 and performs 10 iterations.

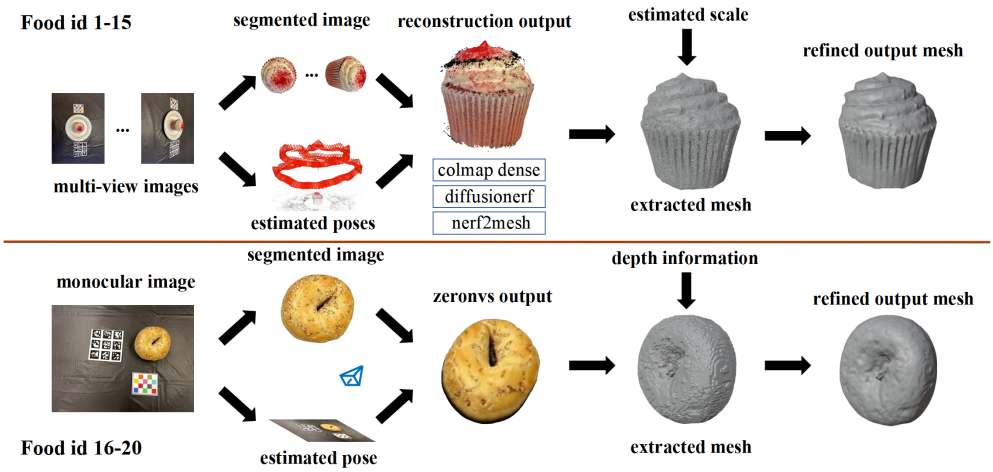


Fig. 6: The pipeline of 3d reconstruction. For multi-view objects, the process involves pose estimation, food segmentation, and application of various 3D reconstruction methods (COLMAP, DiffusioNeRF, NeRF2Mesh), followed by mesh extraction, scaling, and refinement. For single-view objects, ZeroNVS is used with camera parameter optimization and depth information incorporation. Mesh refinement includes hole filling with MeshFix and noise reduction using Laplacian Smoothing.

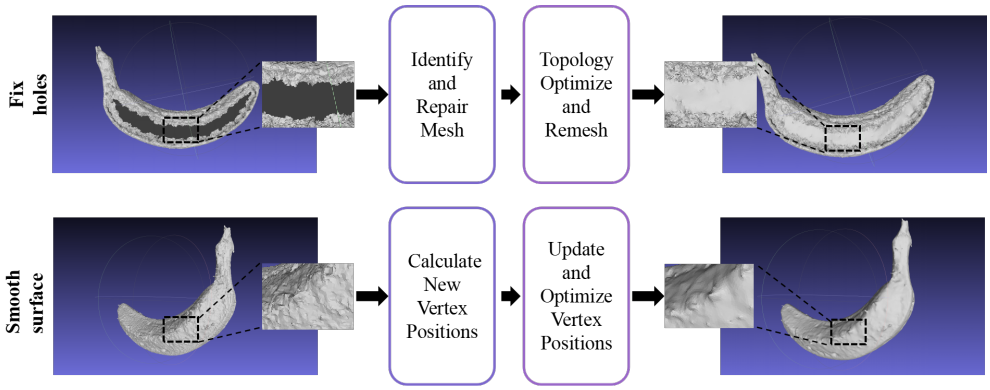


Fig. 7: The overview of Mesh refinement pipeline to fix holes and smooth surface.

4.3 Best 3D Mesh Reconstruction Team - FoodRiddle

To achieve high-quality food mesh reconstruction, the team designs two pipeline processes as shown in Figure 16. For simple and medium cases, they employ a structure-from-motion approach to determine the pose of each image, followed by mesh reconstruction. Subsequently, a series of post-processing steps are implemented to recalibrate scale and enhance mesh quality. For cases with only a single image, the

Fig. 8: The top left image showcases the vanilla COLMAP sfm points. The light-colored areas on the bread have fewer feature points, leading to the incomplete mesh in the top-right image. However, by integrating SuperPoint and SuperGlue into COLMAP, more interest points are obtained, resulting in an excellent final mesh, as shown in the bottom-right image.

team utilizes image generation methods to aid in model generation. The source code is available at [6].

4.3.1 Multi-View Reconstruction

For Structure from Motion (SfM), the team utilizes Pixel-SfM [29], which extends the state-of-the-art method COLMAP [45] by incorporating SuperPoint [46] and SuperGlue [31] methodologies. This significantly mitigates the issue of sparse key points in weakly textured scenes, as shown in Figure 8.

For mesh reconstruction, the team's method is based on 2D Gaussian Splatting[47], which utilizes a differentiable 2D Gaussian 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 applies filtering and outlier removal techniques, identifies the contour of the supporting surface, and projects the lower mesh

[6] https://github.com/jlyw1017/FoodRiddle-MetaFood-CVPR2024

vertices onto the supporting surface. They use the reconstructed checkerboard to rectify the scale of the model and use Poisson reconstruction to generate a watertight, complete mesh of the subject.

4.3.2 Single-View Reconstruction

For 3D reconstruction from a single image, the team employs state-of-the-art methods such as LGM[48], Instant Mesh[49], and One-2-3-45[50] to generate an initial prior mesh. This prior mesh is then jointly corrected with depth structure information.

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

5 Experimental Results

5.1 Shared Results

Three teams, VolETA, ININ-VIAUN, and FoodRiddle, extensively validated their approach on the 3D food dataset from the MetaFood CVPR workshop, as described in Section 3. They compared their results against ground truth meshes using two key metrics: Mean Absolute Percentage Error (MAPE) and Chamfer distance.

To accurately reconstruct food meshes from limited 2D images and provide precise volume estimates, the teams applied different mesh reconstruction strategies for multi-image and single-image food types, aided by refining procedures. With the scaling factor, each mesh was adjusted to its absolute size. Participants were tasked with calculating the volume and evaluating the shape accuracy of the reconstructed meshes, which demonstrated the effectiveness of the 3D food reconstruction methods.

5.1.1 Scaling Factors Calculation

The scaling factors estimated by Team VolETA and Team ININ-VIAUN using the previously described methods are presented in Table 2. Each food item and its corresponding reconstructed 3D model yield a scaling factor, and the table summarizes the average scaling factor for the first fourteen objects.

5.1.2 Volume Estimation

Table 3 presents the quantitative comparison of the predicted volumes from refined meshes and the corresponding percentage errors for three teams, compared to the ground truth model volumes. In addition, We summarize the mean absolute percentage error (MAPE) and the standard deviation (S.D.) for various food items, categorized into three groups based on the number of input images.

Team VolETA achieved the best overall performance, with an average percentage error of 10.98%. They excelled particularly in the Multi-view conditions, recording an impressive score of 7.84%, which is significantly lower than that of the other competing teams. Furthermore, Team VolETA showed excellent consistency across different input

Table 2: Estimated scaling factors from Team VolETA and Team ININ-VIAUN

Object Index	Food Item	VolETA	ININ-VIAUN
1	Strawberry	0.089552	0.060058
2	Cinnamon bun	0.104348	0.081829
3	Pork rib	0.104348	0.073861
4	Corn	0.088235	0.083594
5	French toast	0.103448	0.078632
6	Sandwich	0.127660	0.088368
7	Burger	0.104348	0.103124
8	Cake	0.127660	0.068496
9	Blueberry muffin	0.087591	0.059292
10	Banana	0.087591	0.058236
11	Salmon	0.104348	0.083821
13	Burrito	0.103448	0.069663
14	Hotdog	0.103448	0.073766

conditions, as indicated by the lowest S.D. in all three scenarios. Team FoodRiddle demonstrated strong proficiency in Single-view scenarios. They achieved the lowest error in this category with a percentage error of 15.56%. Their overall performance was also remarkable and placed them as the second-best team across all categories.

Table 4 summarizes the absolute differences between the predicted volumes and the GT volume for each food type among three participant teams. The average and standard deviation of these absolute values are also presented.

For the Multi-view scenarios, Team VolETA achieved the smallest errors with the lowest average absolute difference of 21.33cm^3 and demonstrated smaller fluctuations compared to the other teams. In the Single-view scenarios, Team FoodRiddle outperformed the others, achieving an average absolute difference of 21.06cm^3, which is more than 10cm^3 lower than that of Team VolETA. Considering all food types, Team VolETA still showed the best results, with a mean difference higher by approximately 3cm^3 and a slightly higher S.D. by about 2.5.

Table 3 and Table 4 highlight Team VolETA's overall consistency and accuracy, while also underscoring Team FoodRiddle's capability to effectively handle Single-view predictions.

5.1.3 Shape Accuracy Evaluation

Apart from accurate volume estimation, the precise reproduction of 3D structures is equally important. An accurate 3D mesh not only demonstrates the models' reconstruction capabilities but also promotes the development of food science and dietary health. Participants evaluated their reconstructed meshes using the Chamfer distance metric, where a smaller value indicates a more accurate reproduction of the food item. Table 5 presents the Chamfer distance of the refined predicted models from the three winning teams.

FoodRiddle outperformed any other team in 3D reconstruction for both multi-view and single-view approaches. After the post-processing procedures, the discrete points were removed, models were scaled to the estimated size according to the checkboard, and meshes were refined to the watertight forms. Upon completion of these processes, the average Chamfer distance across the final reconstructions of the 20

Table 3: Comparison of Predicted Volumes and Error Statistics Among Teams. The table presents the predicted volumes (cm^3) of three teams (VolETA, ININ-VIAUN, FoodRiddle) for various food items, along with the ground truth (GT) volume. Additionally, it shows the error percentages (%) for each team's predicted volume compared to the GT. The table also includes the mean and standard deviation (S.D.) of the error percentages for Multi-view, Single-view, and All data categories in order to highlight consistency and accuracy. Team VolETA achieved the lowest average error in both Multi-view and All data categories, while Team FoodRiddle showcased superior performance in processing Single-view data. Team VolETA also demonstrated the most stable performance with the lowest S.D. across all scenarios.

Type	L	ID	**Predicted Volume**			GT Volume	**Error percentage ↓**		
			VolETA	ININ-VIAUN	FoodRiddle		VolETA	ININ-VIAUN	FoodRiddle
Multi-view	E	1	40.06	37.65	44.51	38.53	3.97	2.28	15.52
		2	216.90	325.44	321.26	280.36	22.64	16.08	14.59
		3	278.86	473.40	336.11	249.65	11.70	89.63	34.63
		4	279.02	294.32	347.54	295.13	5.46	0.27	17.76
		5	395.76	353.66	389.28	392.58	0.81	9.91	0.84
		6	205.17	237.88	197.82	218.31	6.02	8.96	9.39
		7	372.93	361.49	412.52	368.77	1.13	1.97	11.86
		8	186.62	172.32	181.21	173.13	7.79	0.47	4.67
	M	9	224.08	253.01	233.79	232.74	3.72	8.71	0.45
		10	153.76	157.58	160.06	163.23	5.80	3.46	1.94
		11	80.40	76.46	86.00	85.18	5.61	10.24	0.96
		13	363.99	246.60	334.70	308.28	18.07	20.01	8.57
		14	535.44	495.10	517.75	589.82	9.22	16.06	12.22
	Mean ↓		**7.84**	14.47	10.26	**S.D. ↓**	**6.108**	22.550	9.108
Single-view	H	16	163.13	89.71	176.24	262.15	37.77	65.78	32.77
		17	224.08	181.37	180.68	181.32	23.58	0.03	0.35
		18	25.40	20.00	13.58	20.58	23.42	2.82	34.01
		19	110.05	130.85	117.72	108.35	1.57	20.77	8.65
		20	130.96	100.79	117.43	119.83	9.29	15.89	2.00
	Mean ↓		19.13	21.06	**15.56**	**S.D. ↓**	**12.576**	23.671	14.829
All	**Mean ↓**		**10.98**	16.30	11.73	**S.D. ↓**	**9.820**	23.057	11.253

objects amounted to 0.0031 meters as shown in Table 5, significantly smaller than the results of other teams.

5.2 Team Specific Results

In this section, we present the immediate results of the mesh generation procedure and analyze the effectiveness of key components within three methods in Sections 5.2.1, 5.2.2, and 5.2.3.

5.2.1 Experiments of VolETA

Results and Analysis

The team leverages their approach for each food scene separately. A Single-view food volume estimation approach is applied if the number of keyframes k equals 1. Otherwise, a few-shot food volume estimation is applied. Notably, Fig. 9 shows that 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 4: Comparison of Absolute Differences Between Predicted Volumes and Ground Truth. The table presents the absolute differences (cm^3) between the predicted volumes of the three teams and the ground truth (GT) volumes for each food item. It also includes the mean and standard deviation (S.D.) of the absolute differences across all food items for both Multi-view and Single-view settings, as well as overall. Team VolETA achieved the best performance overall and in the Multi-view setting, while Team FoodRiddle demonstrated superior performance in Single-view volume prediction, highlighting its strong reconstruction capability.

Type	L	ID	**Predicted Volume**		
			VolETA	ININ-VIAUN	FoodRiddle
Multi-view	E	1	1.53	0.88	5.98
		2	63.46	45.08	40.9
		3	29.21	223.75	86.46
		4	16.11	0.81	52.41
		5	3.18	38.92	3.3
		6	13.14	19.57	20.49
		7	4.16	7.28	43.75
		8	13.49	0.81	8.08
	M	9	8.66	20.27	1.05
		10	9.47	5.65	3.17
		11	4.78	8.72	0.82
		13	55.71	61.68	26.42
		14	54.38	94.72	72.07
	Mean ↓		**21.33**	40.63	28.07
	S.D. ↓		**21.229**	59.412	27.738
Single-view	H	16	99.02	172.44	85.91
		17	42.76	0.05	0.64
		18	4.82	0.58	7.00
		19	1.70	22.5	9.37
		20	11.13	19.04	2.4
	Mean ↓		31.89	42.92	**21.06**
	S.D. ↓		36.605	65.411	**32.573**
All	**Mean ↓**		**24.26**	41.26	26.12
	S.D. ↓		**26.834**	61.147	29.330

After finding the keyframes, PixSfM [29] estimates the poses and point cloud (see Fig. 10).

After calculating the scaling factors S, the team determined the potential volume using PPU, along with the food width f_w, food length F_l, and food height f_h, to refine the scaling and generate S_f. The values for PPU, 2D Reference object dimensions, and 3D food object dimensions are presented in Table A1.

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 [51] (see Fig. 11).

Figure 12 shows the comparison of the team's Chamfer distance with and without the estimated transformation metrics from ICP.

Table 5: Quantitative comparison of three approaches with ground truth in terms of shape accuracy on the challenge dataset. These teams evaluate their approaches using Chamfer distance in meters (*textm*) for each type of food. The mean and sum of the 18 scenes are also presented.

L	ID	Chamfer distance		
		VolETA	ININ-VIAUN	FoodRiddle
E	1	0.0016	0.0020	0.0011
	2	0.0071	0.0036	0.0031
	3	0.0137	0.0049	0.0053
	4	0.0020	0.0038	0.0015
	5	0.0137	0.0020	0.0040
	6	0.0067	0.0038	0.0025
	7	0.0047	0.0048	0.0025
	8	0.0030	0.0019	0.0010
M	9	0.0039	0.0029	0.0033
	10	0.0027	0.0034	0.0019
	11	0.0034	0.0015	0.0015
	13	0.0052	0.0026	0.0041
	14	0.0043	0.0044	0.0046
H	16	0.0181	0.0083	0.0033
	17	0.0094	0.0070	0.0045
	18	0.0043	0.0041	0.0033
	19	0.0113	0.0047	0.0047
	20	0.0156	0.0037	0.0034
Mean		0.0073	0.0039	**0.0031**
Sum		0.1306	0.0694	**0.0556**

Additionally, Fig. 13 shows the qualitative results on the one and few-shot 3D reconstruction from the challenge dataset. The figures show that the team's model excels in texture details, artifact correction, missing data handling, and color adjustment across different scene parts.

5.2.2 Experiments of ININ-VIAUN

Mesh Refinement After applying the scaling factors to the initial meshes, several refinement steps, such as filling holes and removing isolated points, are performed. The refined meshes obtained are shown in Figure 14.

3D Mesh Registration The team designs a multi-stage alignment method for evaluating reconstruction quality. Figure 15 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. Finally, they perform ICP registration for further alignment, significantly reducing the Chamfer distance.

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

5.2.3 Experiments of FoodRiddle

In Phase-I of volume prediction, the team achieved an overall MAE error of 11.73 across 18 food objects. In Phase-II, through nonlinear optimization and a rescaling

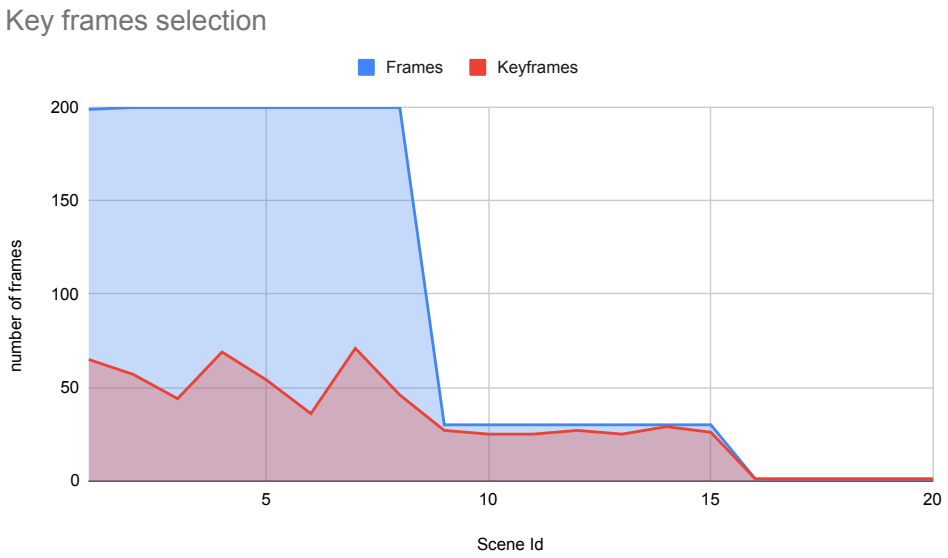


Fig. 9: A quantitative results to the number of frames before and after the keyframe selection phase. The team's approach is only using 34.8% of the data.

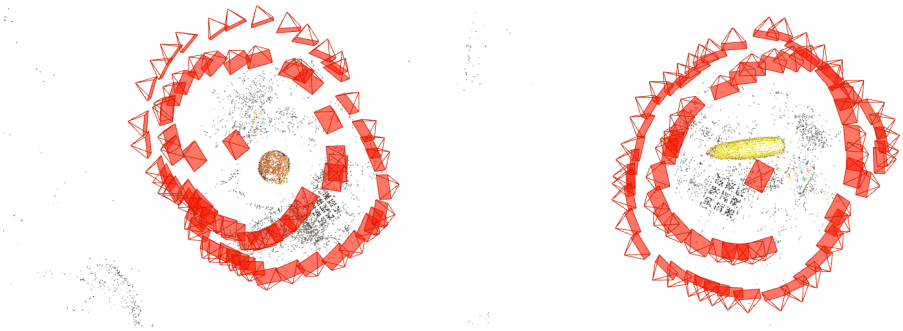
process, the team aimed to refine the mesh models, resulting in a more accurate representation of the food items. The shape accuracy of the final results exceeded that of all other teams, with an average Chamfer distance of 0.0031 meters and a total of 0.0556 meters. Moreover, as illustrated in Table 6, Team FoodRiddle excelled in both multi-view and single-view reconstructions, surpassing the performance of other teams in the competition.

Table 6: Total Chamfer Distance for Different Teams on Multi-view and Single-view Data. Team FoodRiddle has the best score.

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

6 Limitations

Despite the promising results demonstrated by these teams' methods, Many limitations need to be addressed in future work.



(a) burger (7) (b) corn (4)

Fig. 10: PixSfm results after applying keyframes selection. PixSfM excels in estimating and refining camera poses by providing a rich point cloud using Superpoint feature extractors.

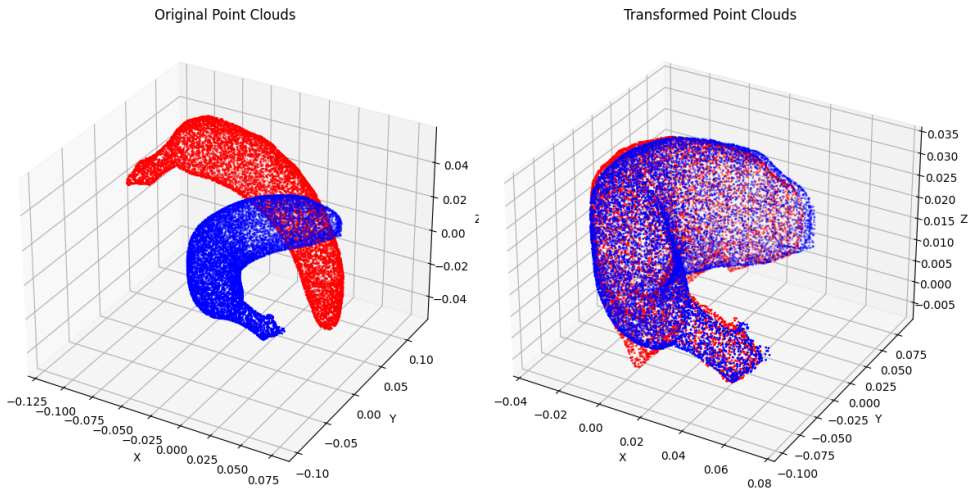


Fig. 11: Performed ICP mesh registration between the team's generated and ground truth meshes for banana_2 scene. Unregistered meshes are on the left, while registered meshes are on the right. The team's Point cloud is red, and the ground truth is blue.

6.1 Combined Limitations Across All Methods

- **Manual processes:** All pipelines include manual steps. VolETA provided a segmentation prompt and identified scaling factors, ININ-VIAUN selected the best mesh reconstruction results, set the corner detection threshold, incorporated depth information using the checkerboard reference, and optimized extrinsic camera parameters

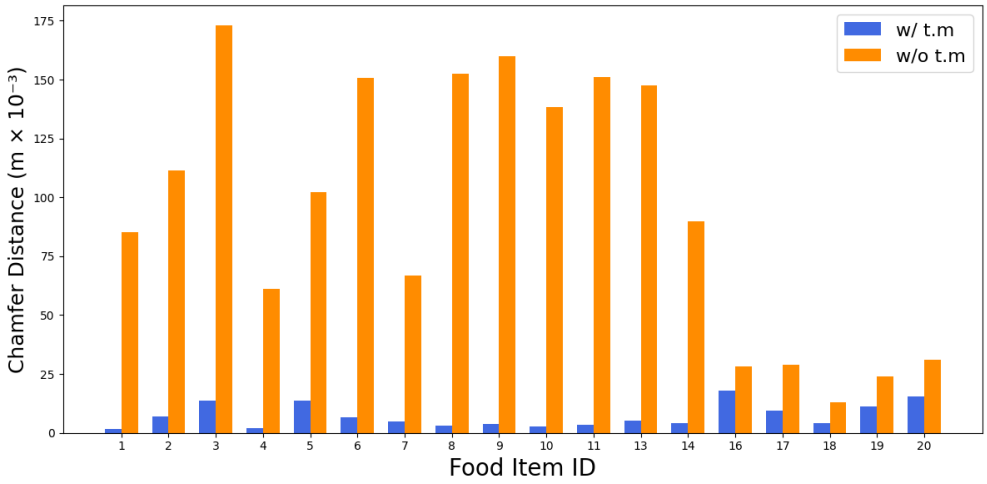


Fig. 12: Comparison of the VolETA approach using challenge dataset. The team evaluates their approach using Chamfer distance in $\times 10^{-3}$ with and without transformation metrics, which are presented in blue and orange bars respectively.

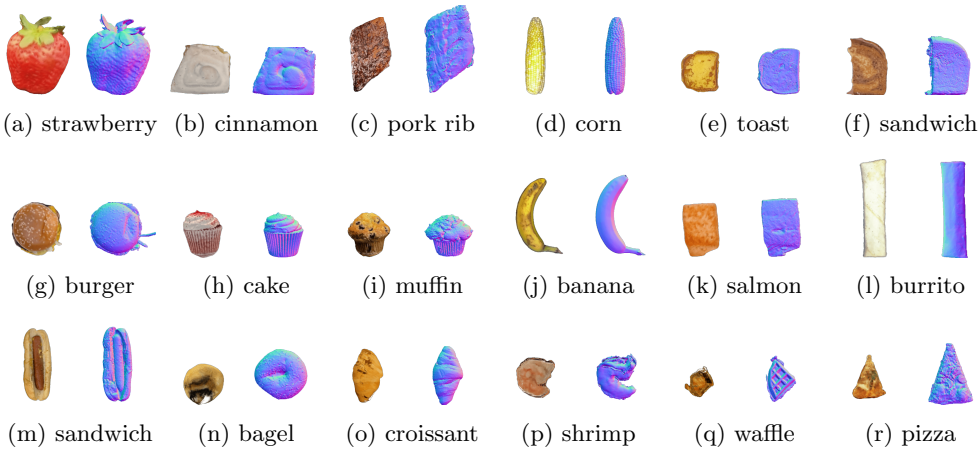


Fig. 13: Comparisons to the VolETA's results and ground truth using the challenge dataset. Each scene shows the team's reconstruction (left) and ground truth (right).

in single-view reconstruction, and FoodRiddle needed to adjust the scaling factors. These steps should be automated to enhance efficiency and reduce human intervention.

- **Complex Backgrounds and Food Items:** None of the teams have tested their methods on datasets with highly complex backgrounds or intricate food items. Applying their approach to datasets with more complex food items, such as the

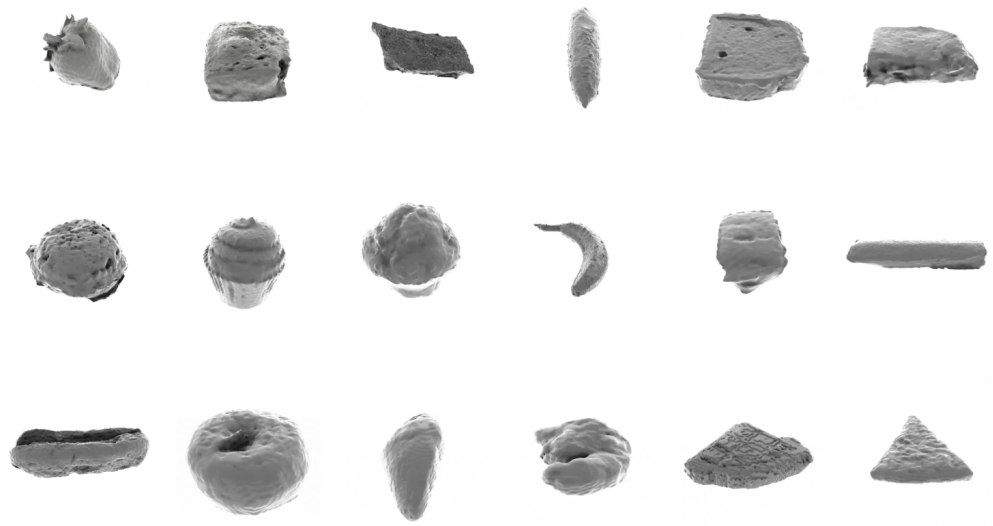


Fig. 14: The visualization of ININ-VIAUN's refined meshes achieved by fixing holes and smoothing the surface.

Nutrition5k [23] dataset, would be challenging and could help identify corner cases that need to be addressed.

- **Capturing complexities:** The proposed methods have not been evaluated under diverse capturing conditions, such as different camera distances, varying speeds, and other scenarios as defined in the Fruits and Vegetables [52] dataset. The robustness and performance of these approaches remains unproven under such circumstances.
- **High Input Requirements:** These methods require multiple inputs, such as depth data, food masks, or checkerboard references. Streamlining these input requirements would simplify the methods and potentially increase its applicability in diverse and less controlled environments.

6.2 Team Specific Limitations

6.2.1 Limitations of VolETA

- **Pipeline complexity:** For Single-view neural rendering, the team currently uses the One-2-3-45 [36] method. However, they aim to use only the 2D diffusion model, Zero123 [40], in their pipeline to reduce complexity and improve the efficiency of their approach.

6.2.2 Limitations of ININ-VIAUN

- **High Dependence on Mesh Refinement:** The need for significant mesh refinement using MeshFix [44] and Laplacian Smoothing [53] suggests that the initial reconstructions may contain substantial noise and holes, which indicates a potential weakness in the reconstruction pipeline. Heavy reliance on post-processing could affect the efficiency of the final mesh output.

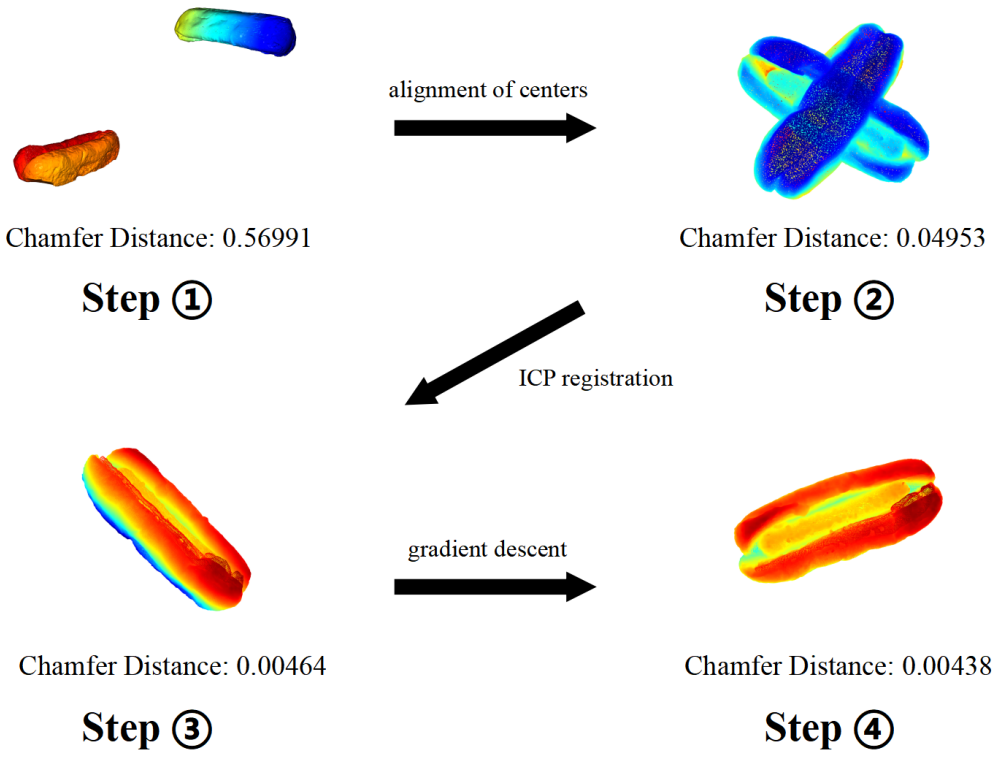


Fig. 15: The process of aligning object 14.

- **Single-view Limitations:** The single-view reconstruction approach relies heavily on depth information and checkerboard references for scale estimation. Moreover, the team not only explored multiple reconstruction methods, but also required extrinsic parameter optimization and depth data to achieve satisfactory results. This dependence highlights the current limitations of their single-view reconstruction capabilities and reduces its practicality in scenarios where depth information or physical references are not available.

6.2.3 Limitations of FoodRiddle

- **Complex Post-Processing for Multi-View Reconstruction:** The post-processing stage for multi-view reconstruction involves several steps, such as filtering, outlier removal, contour identification, and Poisson reconstruction. This extensive post-processing increases the computational complexity and limits the pipeline's efficiency, particularly for large-scale or real-time applications.
- **Single-View Limitations and Assumptions:** In the single-view reconstruction, the method assumes that the object and the reference checkerboard are on the same plane, which may be the same situation in real-world scenarios when there are differences in height or uneven surfaces.

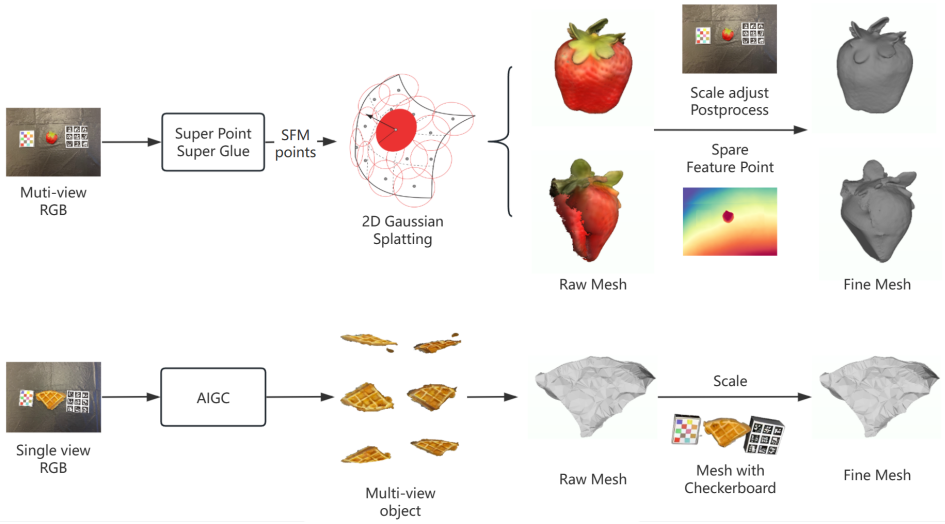


Fig. 16: For multi-view image inputs, COLMAP integrated with SuperPoint and SuperGlue generates SFM points, which are used to create the initial Gaussian. Using 2D Gaussian Splatting, we obtain the mesh of the observed object. Subsequently, we adjust the mesh size and fit the unobserved underside of the object using the RGB checkerboard and depth maps. Finally, a complete and realistic food mesh is produced. For single-view input, we use the AIGC method to generate multi-view images consistent with the input image in 3D using a multi-view diffusion model. Then, using the Sparse-view Large Reconstruction Model, we directly predict the mesh. Lastly, using the simultaneously reconstructed checkboard, adjust the size of the food.

7 Conclusion

This paper summarizes the MetaFood CVPR Workshop Challenge on 3D Food Reconstruction, which aimed to advance reconstruction techniques for food items, addressing challenges posed by varying textures, reflective surfaces, and complex geometries. The competition utilized 20 diverse food items captured under different conditions and challenge participants to develop robust models for volume estimation and shape reconstruction. The evaluation was based on two phases: portion size accuracy (MAPE) and shape accuracy (Chamfer distance). Three teams advanced to the final submission, with Team VolETA securing first place, followed by Team FoodRiddle and ININ-VIAUN. FoodRiddle showed 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, which potentially leads to more accurate and more user-friendly methods for

dietary assessment and monitoring. Future work could expand datasets to include more complex scenarios, such as varying lighting, occlusions, more intricate food items and different food-item backgrounds. Additionally, challenges could focus on single-view reconstruction or predicting nutritional content directly from the reconstructed models, which links geometric information to food content estimation. These directions will push the boundaries of 3D reconstruction techniques and intelligent nutrition science, and encourage the development of more robust, adaptable, and practical solutions for dietary assessment and nutritional monitoring.

Declarations

Conflict of interest The authors declare that they have no relevant financial interests or competing interests to declare that are relevant to the content of this article.

References

[1] Thompson, F.E., Subar, A.F.: Dietary assessment methodology. Nutrition in the Prevention and Treatment of Disease, 5–48 (2017)

[2] Shao, Z., Fang, S., Mao, R., He, J., Wright, J.L., Kerr, D.A., Boushey, C.J., Zhu, F.: Towards learning food portion from monocular images with cross-domain feature adaptation. Proceedings of 2021 IEEE 23rd International Workshop on Multimedia Signal Processing, 1–6 (2021) https://doi.org/10.1109/MMSP53017.2021.9733557

[3] He, J., Shao, Z., Wright, J., Kerr, D., Boushey, C., Zhu, F.: Multi-task image-based dietary assessment for food recognition and portion size estimation. 2020 IEEE Conference on Multimedia Information Processing and Retrieval, 49–54 (2020) https://doi.org/10.1109/MIPR49039.2020.00018

[4] Shao, Z., Han, Y., He, J., Mao, R., Wright, J., Kerr, D., Boushey, C.J., Zhu, F.: An integrated system for mobile image-based dietary assessment. Proceedings of the 3rd Workshop on AIxFood, 19–23 (2021) https://doi.org/10.1145/3475725.3483625

[5] He, J., Mao, R., Shao, Z., Wright, J.L., Kerr, D.A., Boushey, C.J., Zhu, F.: An end-to-end food image analysis system. Electronic Imaging **2021**(8), 285–1 (2021)

[6] Min, W., Wang, Z., Liu, Y., Luo, M., Kang, L., Wei, X., Wei, X., Jiang, S.: Large scale visual food recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 9932–9949 (2023)

[7] Mao, R., He, J., Shao, Z., Yarlagadda, S.K., Zhu, F.: Visual aware hierarchy based food recognition. Proceedings of the International conference on pattern recognition, 571–598 (2021). Springer

[8] Mao, R., He, J., Lin, L., Shao, Z., Eicher-Miller, H.A., Zhu, F.: Improving dietary assessment via integrated hierarchy food classification. 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP), 1–6 (2021) https://doi.org/10.1109/MMSP53017.2021.9733586

[9] He, J., Zhu, F.: Online continual learning for visual food classification. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2337–2346 (2021)

[10] Pan, X., He, J., Zhu, F.: Personalized food image classification: Benchmark datasets and new baseline. In: 2023 57th Asilomar Conference on Signals, Systems, and Computers, pp. 1095–1099 (2023). IEEE

[11] He, J., Lin, L., Eicher-Miller, H.A., Zhu, F.: Long-tailed food classification. Nutrients **15**(12), 2751 (2023)

[12] Huang, Y., A Hassan, M., He, J., Higgins, J., McCrory, M., Eicher-Miller, H., Thomas, J.G., Sazonov, E., Zhu, F.: Automatic recognition of food ingestion environment from the aim-2 wearable sensor. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3685–3694 (2024)

[13] Lo, F.P.W., Sun, Y., Qiu, J., Lo, B.: Image-Based Food Classification and Volume Estimation for Dietary Assessment: A Review. IEEE Journal of Biomedical and Health Informatics **24**(7), 1926–1939 (2020)

[14] Puri, M., Zhu, Z., Yu, Q., Divakaran, A., Sawhney, H.: Recognition and volume estimation of food intake using a mobile device. Proceedings of the 2009 Workshop on Applications of Computer Vision, 1–8 (2009) https://doi.org/10.1109/WACV.2009.5403087

[15] Rahman, M.H., Li, Q., Pickering, M., Frater, M., Kerr, D., Bouchey, C., Delp, E.: Food volume estimation in a mobile phone based dietary assessment system. Proceedings of the 2012 8th International Conference on Signal Image Technology and Internet Based Systems, 988–995 (2012) https://doi.org/10.1109/SITIS.2012.146

[16] Gao, A., Lo, F.P.-W., Lo, B.: Food volume estimation for quantifying dietary intake with a wearable camera. Proceedings of the 2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks, 110–113 (2018) https://doi.org/10.1109/BSN.2018.8329671

[17] Xu, C., He, Y., Khanna, N., Boushey, C.J., Delp, E.J.: Model-based food volume estimation using 3d pose. Proceedings of the 2013 IEEE International Conference on Image Processing, 2534–2538 (2013) https://doi.org/10.1109/ICIP.2013.6738522

[18] Jia, W., Yue, Y., Fernstrom, J.D., Zhang, Z., Yang, Y., Sun, M.: 3d localization

of circular feature in 2d image and application to food volume estimation. Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 4545–4548 (2012). IEEE

[19] Vinod, G., He, J., Shao, Z., Zhu, F.: Food portion estimation via 3d object scaling. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3741–3749 (2024)

[20] Lo, F.P.-W., Sun, Y., Lo, B.: Depth estimation based on a single close-up image with volumetric annotations in the wild: A pilot study. Proceedings of the 2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, 513–518 (2019) https://doi.org/10.1109/AIM.2019.8868629

[21] Shao, Z., Vinod, G., He, J., Zhu, F.: An end-to-end food portion estimation framework based on shape reconstruction from monocular image. Proceedings of 2023 IEEE International Conference on Multimedia and Expo, 942–947 (2023) https://doi.org/10.1109/ICME55011.2023.00166

[22] Vinod, G., Shao, Z., Zhu, F.: Image based food energy estimation with depth domain adaptation. Proceedings of 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval, 262–267 (2022) https://doi.org/10.1109/MIPR54900.2022.00054

[23] Thames, Q., Karpur, A., Norris, W., Xia, F., Panait, L., Weyand, T., Sim, J.: Nutrition5k: Towards automatic nutritional understanding of generic food. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8903–8911 (2021)

[24] Vinod, G., He, J., Shao, Z., Zhu, F.: Food portion estimation via 3d object scaling. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3741–3749 (2024). https://doi.org/10.1109/CVPRW63382.2024.00378

[25] Chen, Y., He, J., Czarnecki, C., Vinod, G., Mahmud, T.I., Raghavan, S., Ma, J., Mao, D., Nair, S., Xi, P., et al.: Metafood3d: Large 3d food object dataset with nutrition values. arXiv preprint arXiv:2409.01966 (2024)

[26] Xu, C., Zhu, F., Khanna, N., Boushey, C.J., Delp, E.J.: Image enhancement and quality measures for dietary assessment using mobile devices. Computational Imaging X **8296**, 153–162 (2012). SPIE

[27] Jain, T., Lennan, C., John, Z., Tran, D.: Imagededup. https://github.com/idealo/imagededup (2019)

[28] De, K., Masilamani, V.: Image sharpness measure for blurred images in frequency domain. Procedia Engineering **64**, 149–158 (2013)

[29] Lindenberger, P., Sarlin, P.-E., Larsson, V., Pollefeys, M.: Pixel-perfect structure-from-motion with featuremetric refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5987–5997 (2021)

[30] DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 224–236 (2018)

[31] Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4938–4947 (2020)

[32] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., *et al.*: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)

[33] Bekuzarov, M., Bermudez, A., Lee, J.-Y., Li, H.: Xmem++: Production-level video segmentation from few annotated frames. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 635–644 (2023)

[34] Wang, Y., Han, Q., Habermann, M., Daniilidis, K., Theobalt, C., Liu, L.: Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3295–3306 (2023)

[35] Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., Ranzuglia, G., *et al.*: Meshlab: an open-source mesh processing tool. In: Eurographics Italian Chapter Conference, vol. 2008, pp. 129–136 (2008). Salerno, Italy

[36] Liu, M., Xu, C., Jin, H., Chen, L., Varma T, M., Xu, Z., Su, H.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. Advances in Neural Information Processing Systems **36** (2024)

[37] Schonberger, J.L., Frahm, J.-M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4104–4113 (2016)

[38] Wynn, J., Turmukhambetov, D.: Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4180–4189 (2023)

[39] Tang, J., Zhou, H., Chen, X., Hu, T., Ding, E., Wang, J., Zeng, G.: Delicate textured mesh recovery from nerf via adaptive surface refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17739–17749 (2023)

[40] Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9298–9309 (2023)

[41] Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)

[42] Sargent, K., Li, Z., Shah, T., Herrmann, C., Yu, H.-X., Zhang, Y., Chan, E.R., Lagun, D., Fei-Fei, L., Sun, D., *et al.*: Zeronvs: Zero-shot 360-degree view synthesis from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9420–9429 (2024)

[43] Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=UyNXMqnN3c

[44] Attene, M.: A lightweight approach to repairing digitized polygon meshes. The Visual Computer (2010) https://doi.org/10.1007/s00371-010-0416-3

[45] Schönberger, J.L., Frahm, J.-M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

[46] DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 337–33712 (2018). https://doi.org/10.1109/CVPRW.2018.00060

[47] Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: SIGGRAPH 2024 Conference Papers. Association for Computing Machinery, ??? (2024). https://doi.org/10.1145/3641519.3657428

[48] Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In: European Conference on Computer Vision, pp. 1–18 (2025). Springer

[49] Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191 (2024)

[50] Liu, M., Xu, C., Jin, H., Chen, L., Varma T, M., Xu, Z., Su, H.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. Advances in Neural Information Processing Systems **36** (2024)

[51] Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: Proceedings Third International Conference on 3-D Digital Imaging and Modeling, pp. 145–152 (2001). IEEE

[52] Steinbrener, J., Dimitrievska, V., Pittino, F., Starmans, F., Waldner, R., Holzbauer, J., Arnold, T.: Learning metric volume estimation of fruits and vegetables from short monocular video sequences. Heliyon **9**(4) (2023)

[53] Aupy, G., Park, J., Raghavan, P.: Locality-aware laplacian mesh smoothing. In: 2016 45th International Conference on Parallel Processing (ICPP), pp. 588–597 (2016). https://doi.org/10.1109/ICPP.2016.74

Appendix A Intermediate Results

Table A1 presents the intermediate results used by Team VolETA to estimate the scaling factors. This table includes data extracted from RGBD images and segmentation masks, providing key information such as the scene ID, Pixel-Per-Unit (PPU) ratio in centimeters, the 2D dimensions of the reference object, and the 3D dimensions of the food object in pixels. These intermediate results help produce more accurate scaling factors by comparing the reference and food object sizes in both 2D and 3D.

Table A1: A list of information extracted using the RGBD and masks, where Team VolETA presents the scene ID, Pixel-Per-Unit (in cm), 2D reference object dimensions ($R_w \times R_l$), and 3D food object dimensions ($f_w \times f_l \times f_h$) in pixels.

L	ID	Food Item	PPU	$R_w \times R_l$		$(f_w \times f_l \times f_h)$		
E	1	Strawberry	0.01786	320	360	238	257	2.353
	2	Cinnamon bun	0.02347	236	274	363	419	2.353
	3	Pork rib	0.02381	246	270	435	778	1.176
	4	Corn	0.01897	291	339	262	976	2.353
	5	French toast	0.02202	266	292	530	581	2.530
	6	Sandwich	0.02426	230	265	294	431	2.353
	7	Burger	0.02435	208	264	378	400	2.353
	8	Cake	0.02143	256	300	298	310	4.706
M	9	Blueberry muffin	0.01801	291	357	441	443	2.353
	10	Banana	0.01705	315	377	446	857	1.176
	11	Salmon	0.02390	242	269	201	303	1.176
	13	Burrito	0.02372	244	271	251	917	2.353
	14	Hotdog	0.02115	266	304	400	1022	2.353
H	16	Everything bagel	0.01747	306	368	458	484	1.176
	17	Croissant	0.01751	319	367	395	695	2.176
	18	Shrimp	0.02021	249	318	186	195	0.987
	19	Waffle	0.01902	294	338	465	537	0.800
	20	Pizza	0.01913	292	336	442	651	1.176