VolE^{++}: A Text-Guided Point-Cloud Framework for Food 3D Reconstruction and Volume Estimation

Umair Haroon[1(✉)], Ahmad AlMughrabi[1], Ricardo Marques[2], and Petia Radeva[1,3]

[1] Universitat de Barcelona, Barcelona, Spain
umairharoon@ub.edu
[2] Universitat Pompeu Fabra, Grup de Tecnologies Interactives (GTI), Barcelona, Spain
[3] Institut de Neurociències, Universitat de Barcelona, Barcelona, Spain

Abstract. Accurate food volume estimation is crucial for health monitoring, medical nutrition management, and food intake applications. Current 3D food volume estimation methods are too generic, missing the context of the estimated objects, and thus their performance is suboptimal. We present VolE^{++}, a framework designed to achieve food objects' 3D reconstruction and volume estimation. This approach enables users to specify a target food item through text input, allowing for precise segmentation of specific food objects in a real-world environment. Once segmented, the object is reconstructed using the VolE 3D reconstruction framework. This process uses Multi-View Stereo techniques to transform a point cloud into a refined mesh, ensuring high spatial fidelity for accurate 3D volume estimation. Extensive evaluations of the FoodKit and MetaFood3D datasets demonstrate the effectiveness of our method in isolating and reconstructing food items, with improvements across multiple datasets achieving a 0.2% MAPE, highlighting its superior performance in food volume estimation.

Keywords: Text-Guided Food Selection · 3D Reconstruction · Food Volume Estimation · Real-World Scale · Training-Free

1 Introduction

Accurate dietary assessment is crucial for understanding health trends and individual nutritional needs. While significant progress has been made in food segmentation and recognition through advancements in machine learning and computer vision, accurately estimating food volume remains a difficult challenge. This difficulty arises primarily due to the inherent difficulty in recovering accurate 3D data from 2D images, which is a fundamental requirement for reliable volume estimation. Additionally, in real-world scenarios, food items are often arranged in complex ways, making it challenging to isolate and analyse specific

M. Castrillón-Santana et al. (Eds.): CAIP 2025, LNCS 15621, pp. 386–397, 2026.
https://doi.org/10.1007/978-3-032-04968-1_33

objects. This situation emphasises the need for methods that enable users to guide the 3D reconstruction and volume estimation processes.

Traditional dietary assessment methods are often costly and prone to human error, leading to a growing interest in image-based solutions for automating food segmentation, recognition, and volume estimation [20]. Perhaps recent view synthesis methods like Neural Radiance Fields (NeRF) [16] and Gaussian splatting [14] allow recovering the 3D geometric information of a real-world scene from 2D images. Still, they require accurate camera poses from Structure from Motion (SfM) [19] tools, and cannot determine real-world scale without prior information [13]. Further difficulties arise from low-quality images and varied backgrounds, impacting reconstruction accuracy. Existing volume estimation techniques often rely on specialized and expensive hardware [8] or predefined models [24], struggling to accommodate diverse food shapes. While supervised learning can yield high accuracy, it often requires extensive annotated data and fixed environments, limiting practicality [9]. Many existing methods also lack user control for focusing on specific objects in complex scenes, highlighting the need for a framework that allows for guided 3D reconstruction and accurate volume estimation. To address these limitations, we introduce VolE^{++}, an advanced VolE framework designed for accurate volume estimation in challenging free-motion and multi-object environments, which also allows for text-guided control. Building on the original VolE pipeline for 3D reconstruction [12], VolE^{++} produces dense 3D point clouds using images and camera locations from the FoodKit dataset, captured with AR-enabled mobile devices. A key advancement is the integration of text-guided, decoupled video segmentation, which improves user control over identifying and analyzing individual food items. By utilizing Decoupled Video Segmentation (DEVA) [7], VolE^{++} enhances object tracking and refines segmentations. This user-controlled framework overcomes existing limitations, enabling text-guided 3D reconstruction for more consistent outcomes. Our key contributions include (Fig. 1):

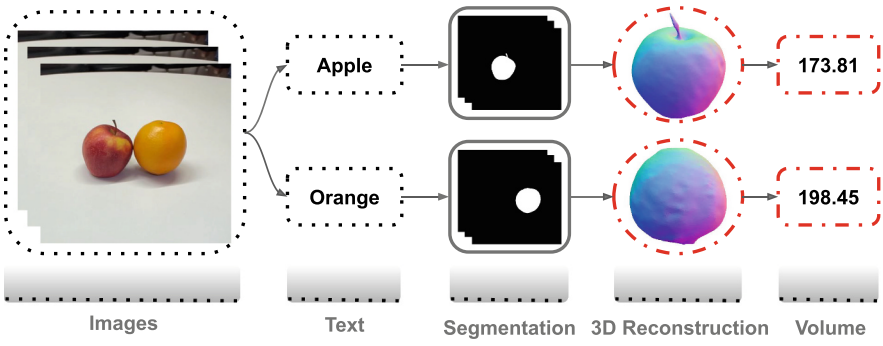

Fig. 1. Visualization of the processing pipeline: Given 2D input images and a text prompt (e.g., "apple"/"orange"), our framework performs text-guided segmentation to isolate food items and reconstruct them in 3D for volume estimation.

- We propose an innovative approach that allows users to accurately identify and isolate specific food items within a scene using simple text prompts. This represents the first exploration of user-prompted guidance in food volume estimation, enabling text-guided segmentation of the desired food object for subsequent accurate volume measurement.
- Building upon the VolE pipeline, we leverage its advanced capability for generating scaled 3D reconstruction of semantically selected objects and estimating the selected object volume, ensuring spatial fidelity and consistency.
- We conduct extensive experiments using the challenging MTF dataset [13] and our FoodKit dataset. Our evaluations thoroughly assess the framework's performance regarding segmentation quality, 3D reconstruction accuracy, and overall volume estimation precision in real-world food scenarios.

The remainder of this paper is structured as follows: Sect. 2 offers a review of related work, while Sect. 3 describes the proposed method in detail. In Sect. 4, we present the experimental results, and Sect. 5 summarizes our contributions and outlines future research directions.

2 Related Work

Recent advancements in accurate food volume estimation have been driven by improvements in 3D reconstruction using visual data. Techniques like SfM and COLMAP [19] have laid the groundwork for this progress by reconstructing 3D scene geometry from multiple images. Innovations like NeRF [16] and related implicit neural representations have further transformed 3D vision, creating dense volumetric representations from sparse 2D images and often relying on SfM for camera pose estimations. Despite enhancements in speed and quality from methods like InstantNGP [17] and NeuS/2 [22,23], challenges like scale ambiguity and diverse food shapes and textures still hinder accurate volume estimation.

Advanced 3D reconstruction techniques for estimating food volume have become more popular, particularly within initiatives like the MetaFood CVPR challenge [13]. A leading approach, VolETA [2], combines SfM with neural surface reconstruction methods, such as NeuS/2 [23], to produce detailed food meshes and accurate volume measurements. Other significant contributions include ININ-VIAUN [13], which integrates deep learning with Multi-View Stereo for depth information, and FoodRiddle [13], which employs 3D Gaussian splatting to address challenges related to data scarcity and complex food shapes. However, many methods, including VolETA, rely on reference objects, limiting their effectiveness in diverse real-world scenarios. While some datasets provide precomputed object masks, practical applications require on-the-fly mask generation. Models like FoodMem [1] enable near real-time food segmentation in videos, but integrating this segmentation into 3D reconstruction for text-guided control remains an active research area. Additionally, some methods are constrained by the need for specialized hardware [8] or fixed environments [21], limiting their broader applicability.

Recognizing the limitations of existing methods, the original VolE [12] framework was introduced as a novel solution for improving 3D reconstruction and volume estimation in food scenes without needing a reference object or depth sensors. Unlike previous approaches that often relied on reference objects [2], depth sensors, or specialized hardware [13], VolE [12] leverages standard mobile device capabilities through technologies like ARCore [10] and ARKit [3]. This allows for capturing real-world measurements using device location and IMU data during video recording. This innovative approach effectively addresses scale ambiguity and enhances adaptability to various food shapes by integrating advanced segmentation techniques with robust 3D reconstruction methods.

Building on the solid foundation of VolE [12], VolE^{++} addresses a significant gap in existing reconstruction pipelines: the lack of explicit user control over the segmentation process. While earlier methods often rely on automated segmentation, real-world food volume estimation frequently involves scenes with multiple food items, which necessitates selective segmentation for accurate volume calculations. To address this issue, VolE^{++} integrates DEVA [7], which enhances segmentation accuracy through a bi-directional propagation mechanism that refines masks over time, reducing inter-frame inconsistencies. This integration enables precise user-specified food segmentation using textual prompts, enhancing volume estimation accuracy. Thus, VolE^{++} offers improved control and accuracy for dietary assessment tools.

3 Our Proposal: VolE^{++}

Our VolE^{++} framework introduces a robust, text-guided approach to estimating food volumes by combining advanced 3D reconstruction with text-driven video segmentation. This section outlines our pipeline, focusing on parameter extraction, text-guided segmentation, point cloud masking, 3D mesh reconstruction, mesh refinement, and volume estimation.

3.1 Overview

Our VolE^{++} framework consists of three phases: (a) Parameters Extraction, (b) 3D Mesh Reconstruction, and (c) Volume Estimation. Initially, we process input image sequences and their associated camera 3D Coordinates to refine camera poses and, crucially, employ a text-guided segmentation module to isolate the desired food object. The segmented data and refined poses feed into the 3D Mesh Reconstruction phase, generating a detailed mesh of the selected food item. In the final stage of Volume Estimation, we calculate the volume of the reconstructed 3D mesh.

3.2 Parameters Extraction

Accurate 3D reconstruction fundamentally relies on precise camera parameters and robust object segmentation. Given a sequence of input images, $\mathcal{I} =$

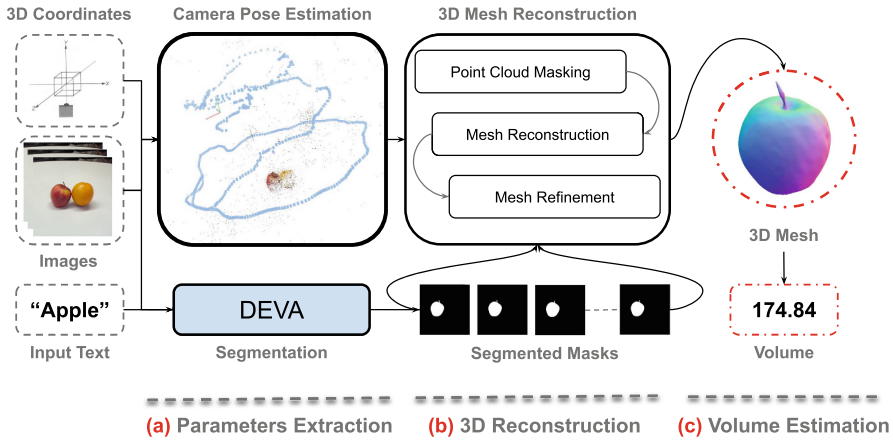

Fig. 2. Overview of the VolE^{++} Framework: (a) Parameter Extraction: Camera poses are refined with 3D coordinates via COLMAP, while DEVA generates segmentation masks for food objects based on prompts (e.g., "Apple"/"Orange"). (b) 3D Reconstruction: Refined poses and segmented images create a dense 3D point cloud, which is then converted into a refined mesh. (c) Volume Estimation: The volume of the food mesh is computed through tetrahedral decomposition.

$\{I_i | i = 1 \dots N_I\}$, we first refine the camera intrinsics and extrinsics, denoted as $\mathcal{C} = \{C_i | i = 1 \dots N_I\}$. Unlike previous approaches relying solely on image data, VolE^{++} leverages AR captured 3D coordinates data and COLMAP [19], a robust SfM pipeline, to enhance these initial camera poses. COLMAP refines poses by performing feature extraction (e.g., SIFT [15]), feature matching across views, and geometric verification to ensure spatial consistency. In parallel to pose estimation, the semantic isolation of specific food items is a key feature for an accurate volume estimation. Here, VolE^{++} leverages DEVA [7], a robust framework that excels in video object segmentation. DEVA's architecture combines an image segmentation model (trained for task-specific hypotheses at the frame level) with a universal temporal propagation model (developed with class-agnostic mask propagation datasets). This decoupled design allows DEVA to generalize effectively, even in scenarios with limited labeled training data. Crucially for VolE^{++}, we adapt DEVA to enable text-guided object segmentation via text prompts. The user provides a textual label, such as "apple" or "orange", to specify the target food object. This input guides DEVA to produce a set of precise segmentation masks $\mathcal{S} = S_1, S_2, \dots, S_T$, where T is the total number of frames in the sequence. These masks are then applied to the input images, isolating the desired objects from the background. This crucial step ensures that only the relevant regions corresponding to the preselected objects contribute to the subsequent 3D reconstruction process, as illustrated in Fig. 2(a).

3.3 3D Mesh Reconstruction

We reconstruct the 3D food mesh using Point Cloud Masking with segmented images and refined camera poses. We utilize the point cloud from COLMAP and masks from DEVA, projecting each 3D point P_i onto the 2D masks. The camera poses are represented as $mathcalC$, and the point cloud as P. We compute the image coordinates with the intrinsic matrix K and retain 3D points within the masks, defining valid points as $M_j = \{P_i \mid p_{ij} \in \mathcal{S}_j\}$. M_j contains valid points P_i projected into the j^{th} segmented image area $\mathcal{S}_j$. The final segmented point cloud, $\mathcal{P}$, is obtained by the intersection of the valid points across all images: $\mathcal{P} = \bigcap_{j=1}^{N_I} M_j$, focusing on the object of interest discarding background noise.

After point cloud masking, we perform Mesh Reconstruction using a Multi-View Stereo (MVS) approach [11]. The filtered point cloud $\mathcal{P}$ is transformed into an initial tetrahedral mesh $\mathcal{T}$ through Delaunay triangulation, represented as $\mathcal{T} = f_D(\mathcal{P})$ [6]. A graph-cut optimization then labels each tetrahedron as inside or outside the object, denoted as $\mathcal{L} = f_G(\mathcal{T})$. Finally, the marching cubes algorithm extracts the mesh surface, yielding an accurate representation of the object's geometry, $\mathcal{M} = \mathbf{M}(\mathcal{T}, \mathcal{L})$. The mesh refinement process enhances the quality of the reconstructed mesh, improving accuracy and surface representation while eliminating artifacts. The process involves several steps: mesh simplification reduces vertices for efficiency, followed by mesh smoothing using techniques like Laplacian or bilateral filtering to create a more uniform surface. Additional denoising removes remaining noise through filtering, and finally, mesh optimization, which includes vertex relaxation and edge flipping, refines triangle quality. This produces a clean and precise food mesh, denoted as $\hat{\mathcal{M}}$, ensuring optimal results as shown in Fig. 2(b).

3.4 Volume Estimation

After refining the 3D mesh of the food item, the final step is accurately determining its volume, which relies on AR-generated 3D coordinates scaled to real-world dimensions. To calculate the volume of our closed triangular mesh $\hat{\mathcal{M}}$ with N faces, we employ the divergence theorem [18], connecting volume and surface integrals. We compute the internal volume by summing the signed volumes of tiny tetrahedra. Each tetrahedron is formed by the three vertices of a mesh triangle v_1^k, v_2^k, v_3^k, and a common origin point as its fourth vertex. The formula used is: $V = \frac{1}{6} \sum_{k=1}^{N} (v_1^k \cdot (v_2^k \times v_3^k))$, where the scalar triple product $(v_2^k \times v_3^k)$ gives the signed volume of the parallelepiped formed by vectors, with a factor of 1/6 representing the volume of the tetrahedron [5]. By summing the signed tetrahedron volumes for all triangle faces, we can efficiently calculate the total volume of the 3D food object in a single pass.

4 Experimental Results

This section evaluates VolE^{++} performance in object volume estimation and 3D reconstruction. We conducted experiments on two datasets to assess its accuracy,

robustness, and text-guided control. VolE^{++} is compared with SOA methods, including those evaluated on the MTF [13] and FoodKit [12] dataset. Our analysis includes quantitative metrics and qualitative visual assessments to provide a comprehensive evaluation of VolE^{++} effectiveness in real-world scenarios.

4.1 Implementation Settings

All VolE^{++} experiments were conducted on a system with an NVIDIA GeForce RTX 3090 GPU (24 GB). We configured our 3D reconstruction pipeline for the FoodKit and MTF datasets with a point cloud masking "max-resolution" of 512. Mesh reconstruction used a "close-holes" setting of 50 and a "smooth" factor of 5 to achieve accurate surface regularization.

4.2 Evaluation Protocol

To evaluate the accuracy of our framework for volume estimation, we use the Mean Average Percentage Error (MAPE) metric. MAPE is calculated as follows: $\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{V_{est,i}-V_{gt,i}}{V_{gt,i}}\right| \times 100\%$, where $V_{est,i}$ represents the estimated volume, $V_{gt,i}$ denotes the ground truth volume, and n is the total number of objects. For 3D reconstruction evaluation, we use the Chamfer distance [4]. This metric measures the average closest distance between points in the reconstructed model and the real ground truth model, and vice versa. This 2-way assessment provides a reliable assessment for assessing 3D reconstruction quality.

4.3 Datasets

We evaluate VolE^{++} on 2 datasets: the FoodKit [12] and the MTF dataset [13].

Foodkit Dataset. [12] is essential for evaluating food volume estimation frameworks in real-world conditions. It overcomes the limitations of existing datasets by providing diverse food items in free-motion scenarios, allowing accurate volume estimation with standard smartphone cameras. The dataset includes 21 food items with ground truth measurements for volume and mass, serving as a benchmark for estimation techniques. Data was collected using a combination of augmented reality (AR) and traditional 3D reconstruction, ensuring precise real-world alignment. Ground truth validation used the water displacement method (± 5 mL error margin) and digital scale mass measurements. FoodKit includes video scenes, image sets, ARKit-estimated 3D coordinates, image masks, and associated metadata, making it a valuable resource for advancing food volume estimation techniques.

MTF Dataset. [13] is a key benchmark for food volume estimation, consisting of 20 food scenes categorized into three difficulty tiers: "easy" with eight scenes (around 200 images each), "medium" with 7 scenes (approximately 30 images

each), and "hard" with single-image scenes. Each image includes food masks and depth. For our evaluation with VolE^{++}, we focused on the easy and medium scenes, as our framework needs multiple input images for 3D reconstruction. We used the reference board from the MTF dataset to ensure accurate scaling of the reconstructed scenes to their original physical dimensions.

4.4 Comparative Analysis

We evaluated VolE^{++} against various state-of-the-art methods for volume estimation and 3D reconstruction. This comparative analysis focuses on relevant, recent, and reproducible techniques, enabling a clear assessment of VolE^{++}'s strengths and weaknesses.

Qualitative Results. Visual comparison of 3D reconstructions shows that VolE^{++} consistently outperforms other methods in detail and geometric accuracy. Compared to VolE [12] on the FoodKit dataset, we observe comparable or slightly improved performance. The framework effectively handles variations in shape, size, and surface texture, demonstrating adaptability to real-world food characteristics, as shown in Fig. 3. A comparison with VolETA [2] and VolE on the MTF dataset further highlights VolE^{++}'s superior ability to capture delicate geometries for more accurate 3D representations, as shown in Fig. 4.

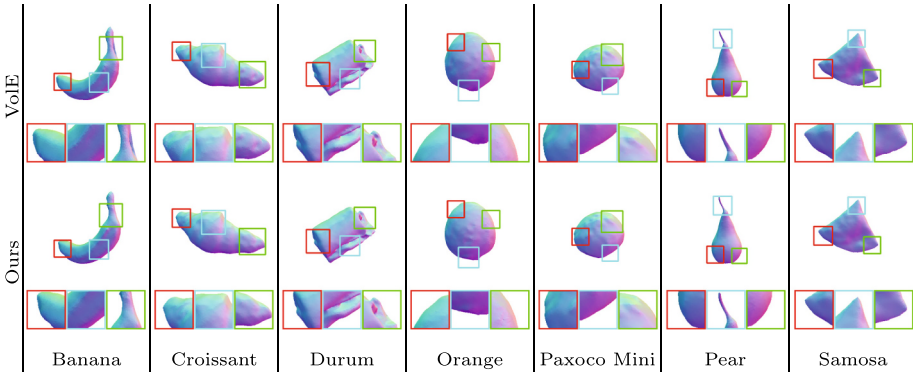

Fig. 3. Qualitative comparison to VolE [12] framework on the FoodKit.

Quantitative Results. We evaluated volume estimation using MAPE and 3D reconstruction accuracy with CD. The results indicate that VolE^{++} performs competitively, achieving impressive MAPE values on both FoodKit and MTF datasets, often surpassing existing methods. An analysis of the FoodKit dataset demonstrates its high accuracy in estimating food volumes compared to the VolE

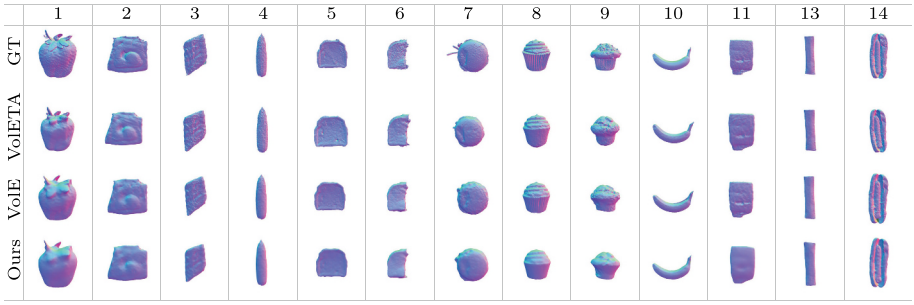

Fig. 4. Our framework 3D reconstruction Visual Results on MTF dataset in comparison with VolETA [2] and ground truth (GT).

[12] framework, as shown in Table 1. VolE^{++} also shows strong performance against VolETA [2], ININ [13], FoodR. [13], and VolE [12] consistently yielding superior or comparable volume estimation (lower MAPE) and 3D reconstruction quality (lower CD) as shown in Table 2. Overall, VolE^{++} is a robust contender in food volume estimation and 3D reconstruction.

4.5 Discussions

The experimental results demonstrate that VolE^{++} effectively estimates object volumes in complex scenarios, handling intricate geometries and diverse textures even in unbounded scenes. Its innovative use of AR-capable devices for initial spatial understanding, particularly showcased with the FoodKit dataset, enhances accuracy. Coupled with a novel text-guided segmentation method using DEVA, VolE^{++} allows for precise identification of target food items, streamlining reconstruction and improving reliability. It significantly outperforms SOA methods in volume estimation accuracy, as indicated by lower MAPE and CD. These results were confirmed across multiple datasets, highlighting VolE^{++}'s versatility in accurate volume estimation, especially in dietary assessments.

4.6 Limitations

The VolE^{++} framework demonstrates promising results; however, it still faces several limitations. One significant issue is the high computational demand associated with camera pose estimation and Multi-View stereo reconstruction, which can hinder real-time processing speeds. While our framework represents a substantial advancement toward practical applications, achieving true real-time performance, particularly with more complex datasets, remains a challenge.

Table 1. Comparison of volume estimation metrics between VoIE [12] and ours for various food items, including estimated volumes, absolute errors, and accuracies.

Items	Images	GT (±5)	Estimated Volume (5x)				Absolute Error (5x)				Accuracy ↑	
			Mean		Std Dev. ↓		Mean ↓		Std Dev. ↓			
			VoIE	Ours	VoIE	Ours	VoIE	Ours	VoIE	Ours	VoIE	Ours
Apple	1005	175	176.68	173.81	0.93	**0.76**	0.96	**0.68**	0.53	**0.43**	99.04	**99.32**
Orange	1001	200	201.06	198.45	3.45	**0.99**	1.35	**0.78**	1.02	**0.49**	98.65	**99.22**
Aguacate	1078	85	83.11	83.31	**1.11**	0.77	2.23	**1.99**	1.30	**0.90**	97.77	**98.01**
Lemon	887	140	134.74	134.14	3.19	**0.98**	1.76	**0.75**	1.33	**0.58**	98.24	**99.25**
Donut	780	245	242.24	242.64	2.57	**1.32**	1.13	**0.96**	1.04	**0.54**	98.87	**99.04**
Durum	1006	200	200.79	198.19	**0.47**	1.20	**0.40**	0.90	**0.24**	0.60	**99.60**	99.10
Pear	849	170	168.13	168.53	**0.71**	0.89	1.10	**0.86**	**0.42**	0.52	98.90	**99.14**
Choc. Cake	781	195	195.38	193.18	1.21	**0.90**	**0.38**	0.93	0.50	**0.46**	**99.62**	99.07
Choc. Croissant	1122	275	274.99	271.99	3.69	**1.26**	**0.95**	1.10	0.82	**0.46**	**99.05**	98.90
Samosa	848	145	144.10	143.10	2.76	**1.11**	1.53	**1.31**	1.10	**0.76**	98.47	**98.69**
Apple Pie	1201	135	135.52	133.32	**1.02**	1.19	**0.51**	1.24	**0.66**	0.88	**99.49**	98.76
Choc. Bomb	1111	200	197.65	197.05	4.39	**1.84**	2.08	**1.47**	1.06	**0.92**	97.92	**98.53**
Empanadilla	926	95	94.86	93.66	**1.12**	1.40	**0.89**	1.73	**0.65**	0.97	**99.11**	98.27
Falafel	929	48	47.58	46.98	2.49	**0.35**	3.96	**2.12**	2.87	**0.72**	96.04	**97.88**
French Bread	1139	163	162.49	161.49	1.54	**1.03**	**0.78**	0.93	**0.50**	0.63	**99.22**	99.07
Paxoco Mini	911	150	148.08	148.68	**1.62**	2.05	1.40	**1.26**	**0.89**	0.92	98.60	**98.74**
Napolitanas	1071	233	232.79	231.99	1.28	**0.71**	**0.41**	0.43	0.32	**0.31**	**99.59**	99.57
Capsicum	881	320	318.64	318.44	2.54	**0.74**	0.76	**0.49**	0.35	**0.23**	99.24	**99.51**
Choc. Panettone	1209	293	290.79	290.99	1.74	**1.19**	0.75	**0.69**	0.59	**0.41**	99.25	**99.31**
Banana	1156	150	153.03	152.63	1.74	**0.80**	2.02	**1.76**	1.16	**0.53**	97.98	**98.24**
Yellow Cane	715	350	350.07	349.47	1.45	**0.81**	0.30	**0.23**	0.24	**0.13**	99.70	**99.77**
Mean		**188.90**	188.23	**187.24**	1.95	**1.06**	1.22	**1.08**	0.84	**0.59**	98.78	**98.92**

Table 2. Comparison of volume estimation and 3D reconstruction methods on the MTF dataset, showing predicted volumes, percentage errors, and Chamfer Distance for VolETA [2], ININ [13], FoodR. [13], VolE [12], and ours method.

ID	Scene Name	Predicted Volume					GT	Error Percentage ↓					Chamfer Distance ↓				
		VolETA	ININ	FoodR.	VolE	Our		VolETA	ININ	FoodR.	VolE	Our	VolETA	ININ	FoodR.	VolE	Our
1	Strawberry	40.06	37.65	44.51	37.47	37.57	38.53	3.97	**2.28**	15.52	2.74	2.49	0.0016	0.0020	**0.0011**	0.0028	0.0021
2	Cinnamon bun	216.90	325.44	321.26	275.38	276.57	280.36	22.64	16.08	14.59	1.78	**1.35**	0.0071	0.0036	0.0031	**0.0022**	0.0023
3	Pork rib	278.86	473.40	336.11	268.93	264.84	249.65	11.70	89.63	34.63	7.72	**6.08**	0.0137	**0.0049**	0.0053	0.0068	0.0063
4	Corn	279.02	294.32	347.54	277.56	276.66	295.13	5.46	**0.27**	17.76	5.95	6.26	0.0020	0.0038	**0.0015**	0.0046	0.0043
5	French toast	395.76	353.66	389.28	394.04	390.54	392.58	0.81	9.91	0.84	**0.37**	0.52	0.0137	**0.0020**	0.0040	0.0021	0.0028
6	Sandwich	205.17	237.88	197.82	215.21	216.13	218.31	6.02	8.96	9.39	1.42	**1.06**	0.0067	0.0038	**0.0025**	0.0039	0.0040
7	Burger	372.93	361.49	412.52	370.69	366.80	368.77	1.13	1.97	11.86	**0.52**	0.54	0.0047	0.0048	**0.0025**	0.0036	0.0039
8	Cake	186.62	172.32	181.21	176.43	171.56	173.13	7.79	**0.47**	4.67	1.91	0.91	0.0030	0.0019	**0.0010**	0.0012	0.0017
9	Blueberry muffin	224.08	253.01	233.79	233.95	230.98	232.74	3.72	8.71	**0.45**	0.52	0.75	0.0039	0.0029	0.0033	0.0029	**0.0027**
10	Banana	153.76	157.58	160.06	159.20	155.60	163.23	5.80	3.46	**1.94**	2.47	4.59	0.0027	0.0034	**0.0019**	0.0118	0.0081
11	Salmon	80.40	76.46	86.00	82.75	83.98	85.18	5.61	10.24	**0.96**	2.85	1.41	0.0034	**0.0015**	0.0015	0.0021	0.0022
12	Burrito	363.99	246.60	334.70	297.09	298.79	308.28	18.07	20.01	8.57	3.63	**3.08**	0.0052	**0.0026**	0.0041	0.0055	0.0052
14	Hotdog	535.44	495.10	517.75	541.58	544.68	589.82	9.22	16.06	12.22	8.18	**7.66**	**0.0043**	0.0044	0.0046	0.0082	0.0090
MAPE ↓		7.84	14.47	10.26	3.08	**2.82**	**S.D. ↓**	6.36	23.47	9.48	2.63	**2.61**					
CD (Sum) ↓													0.0720	0.0416	**0.0364**	0.0576	0.0547
CD (Mean) ↓													0.0055	0.0032	**0.0028**	0.0044	0.0042

5 Conclusions

This paper introduces an advanced framework designed to enhance food volume estimation through a text-guided 3D reconstruction pipeline. Users can isolate target food items using text prompts, leveraging DEVA for accurate segmentation even in complex scenes. The framework also includes camera pose refinement to generate real-world scaled 3D meshes, eliminating the need for physical reference objects. Validation using the MTF and our FoodKit datasets demonstrates that VolE++ outperforms existing methods, effectively handling a diverse range of food shapes and textures. However, it does encounter challenges related to computational intensity, which affects real-time performance on edge devices, as well as occasional segmentation errors. Future work will focus on optimising the computational pipeline for mobile use and enhancing the robustness of segmentation, further establishing VolE++ as a valuable tool for dietary assessment.

Acknowledgment. This work was partially funded by the EU project MUSAE (No. 01070421), 2021-SGR-01094 (AGAUR), Icrea Academia'2022 (Generalitat de Catalunya), Robo STEAM (2022-1-BG01-KA220-VET000089434, Erasmus+ EU), DeepSense (ACE053/22/000029, ACCIÓ), CERCA Programme/Generalitat de Catalunya, and Grants PID2022141566NB-I00 (IDEATE), PDC2022-133642-I00 (DeepFoodVol), and CNS2022-135480 (A-BMC) funded by MICIU/AEI/10.13039/501100011033, by FEDER (UE), and by European Union NextGenerationEU/PRTR. A. AlMughrabi acknowledges the support of FPI Becas, MICINN, Spain. U. Haroon acknowledges the support of FI-SDUR Becas, MICINN, Spain.

References

1. AlMughrabi, A., Galán, A., Marques, R., Radeva, P.: Foodmem: near real-time and precise food video segmentation. arXiv preprint arXiv:2407.12121 (2024)
2. AlMughrabi, A., Haroon, U., Marques, R., Radeva, P.: Voleta: one-and few-shot food volume estimation. arXiv preprint arXiv:2407.01717 (2024)
3. Apple Developer: Arkit (2024). https://developer.apple.com/augmented-reality/arkit/. Accessed 10 Dec 2024
4. Barrow, H., Tenenbaum, J., Bolles, R., Wolf, H.: Parametric correspondence and chamfer matching: two new techniques for image matching. In: Proceedings of Image Understanding Workshop, pp. 21–27. Science Applications (1977)
5. Botsch, M., Kobbelt, L., Pauly, M., Alliez, P., Lévy, B.: Polygon Mesh Processing. CRC Press (2010)
6. Cernea, D.: Openmvs: open multiple view stereovision (2015). https://github.com/cdcseacave/openMVS/
7. Cheng, H.K., Oh, S.W., Price, B., Schwing, A., Lee, J.Y.: Tracking anything with decoupled video segmentation. In: CVPR, pp. 1316–1326 (2023)
8. Dehais, J., Anthimopoulos, M., Shevchik, S., Mougiakakou, S.: Two-view 3D reconstruction for food volume estimation. IEEE Trans. Multimedia **19**(5) (2016)
9. Ferdinand Christ, P., Schlecht, S., Ettlinger, e.: Diabetes60-inferring bread units from food images using fully convolutional neural networks. In: CVPR (2017)

10. Google Developers: Arcore overview (2024). https://developers.google.com/ar/develop/. Accessed 10 Dec 2024
11. Haroon, U., AlMughrabi, A., Marques, R., Radeva, P.: Mvsboost: an efficient point cloud-based 3D reconstruction. arXiv preprint arXiv:2406.13515 (2024)
12. Haroon, U., AlMughrabi, A., Zoumpekas, T., Marques, R., Radeva, P.: Vole: a point-cloud framework for food 3D reconstruction and volume estimation. arXiv:2505.10205 (2025)
13. He, J., et al.: Metafood CVPR 2024 challenge on physically informed 3D food reconstruction: methods and results. arXiv:2407.09285 (2024)
14. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139-1 (2023)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**, 91–110 (2004)
16. Mildenhall, B., Srinivasan, P.P., et al.: Nerf: representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (2021)
17. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4) (2022)
18. O'Rourke, J.: Computational Geometry in C. Cambridge University Press (1998)
19. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR, pp. 4104–4113 (2016)
20. Tahir, G.A., Loo, C.K.: A comprehensive survey of image-based food recognition and volume estimation methods for dietary assessment. In: Healthcare, vol. 9 (2021)
21. Thames, Q., Karpur, A., Norris, W., Xia, F., et.al.: Nutrition5k: towards automatic nutritional understanding of generic food. In: CVPR, pp. 8903–8911 (2021)
22. Wang, P., Liu, L., Liu, Y., et al.: Neus: learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv:2106.10689 (2021)
23. Wang, Y., Han, Q., et al.: Neus2: fast learning of neural implicit surfaces for multi-view reconstruction. In: CVPR, pp. 3295–3306 (2023)
24. Xu, C., He, Y., Khannan, N., Parra, A., Boushey, C., Delp, E.: Image-based food volume estimation. In: Proceedings of MADIMA, pp. 75–80 (2013)