Diffusion-Guided 3D-Aware Calorie Estimation from a Single Food Image

Mayu Ogishi
Department of Informatics
Univ. of Electro-Communications
Tokyo, Japan
ogishi-m@mm.inf.uec.ac.jp

Hikaru Tanabe
Department of Informatics
Univ. of Electro-Communications
Tokyo, Japan
tanabe-h@mm.inf.uec.ac.jp

Keiji Yanai
Department of Informatics
Univ. of Electro-Communications
Tokyo, Japan
yanai@cs.uec.ac.jp

Abstract

Accurate calorie estimation from food images is a critical yet challenging task for dietary monitoring. Existing image-based approaches predominantly rely on 2D representations, limiting their ability to capture food volume and leading to inaccurate estimations for dishes with complex or occluded geometries. To address this, we propose a novel framework that leverages diffusion-based 3D reconstruction to estimate food volume and calories from a single image. Our framework integrates prompt-based segmentation, inpainting-guided plate completion, and multi-view-consistent 3D reconstruction using a diffusion prior. We also fine-tune a diffusion model to inpaint missing plate regions occluded by food, enabling more accurate plate reconstruction and volume difference estimation. The final calorie estimation is made by fusing volumetric cues, 2D visual features, and 3D geometric features from a point-based transformer within a multi-task learning framework. Experiments on the MetaFood3D dataset demonstrate improved volume estimation, and ablation studies show that incorporating volume information improves calorie estimation accuracy. Our findings highlight the potential of high-quality 3D priors and a volume estimation framework as foundational components for practical and scalable nutrition estimation systems.

CCS Concepts

• **Information systems** → **Multimedia information systems**.

Keywords

3D Reconstruction, Diffusion Models, Calorie Estimation, Volume Estimation, Food Image Recognition

ACM Reference Format:
Mayu Ogishi, Hikaru Tanabe, and Keiji Yanai. 2025. Diffusion-Guided 3D-Aware Calorie Estimation from a Single Food Image. In *Proceedings of the 1st International Workshop on Multi-modal Food Computing (MMFood '25), October 27–28, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3746264.3760487

1 Introduction

Monitoring dietary intake is essential for maintaining long-term health and managing conditions such as diabetes, obesity, and other calorie-related diseases. Traditional methods like food diaries and 24-hour dietary recall are time-consuming, prone to user bias, and often suffer from low adherence. To overcome these challenges, recent research [1, 4, 5, 9, 13, 20, 21] has explored automated image-based food analysis, leveraging the ubiquity of smartphone cameras to facilitate dietary tracking.

Most existing systems rely on 2D representations of food. Although 2D images can capture appearance-based features such as color and contour, they lack depth and volumetric information, which are essential for accurate portion size and calorie estimation. This shortcoming is especially evident for foods placed on plates or in bowls, where occlusions and complex shapes make it difficult to estimate true volume. As a result, 2D-based methods often misestimate caloric content due to the absence of spatial cues.

Recent advances in 3D reconstruction, particularly those based on diffusion models, offer a promising direction. Diffusion models pretrained on large-scale image-text datasets encode strong geometric priors that enable plausible 3D reconstruction from a single image. Motivated by these recent advances, we investigate diffusion-guided 3D reconstruction to improve calorie estimation from a single food image.

We introduce a novel framework that combines open-vocabulary segmentation, diffusion-based inpainting for plate completion, and multi-view-consistent 3D reconstruction. Food volume is estimated by comparing reconstructed meshes of the food-on-plate and the plate alone, with scale inferred from reference objects. We then fuse the estimated volume with 2D and 3D features to jointly predict caloric content and food category using a multi-task learning strategy. On the MetaFood3D dataset, our method outperforms prior approaches in volume estimation (MAE) and achieves strong food classification accuracy. Ablation studies further confirm that incorporating volume information, 3D structure, and 2D appearance improves calorie estimation.

2 Related Work

2.1 3D Reconstruction from a Single Image

3D reconstruction from a single image has evolved from direct 3D dataset–based methods to approaches leveraging diffusion models. Early works like Pixel2Mesh [22] deformed an initial mesh using image features, while implicit methods such as Occupancy Networks [11] and DeepSDF [14] achieved memory-efficient continuous shape representation but relied on costly large-scale datasets.

Diffusion-based approaches using Score Distillation Sampling (SDS), pioneered by DreamFusion [15], exploit pre-trained 2D diffusion models to generate 3D shapes from text or a single image without 3D training data (e.g., RealFusion [10]), though they suffer

MMFood '25, Dublin, Ireland

ACM ISBN 979-8-4007-2046-8/2025/10
https://doi.org/10.1145/3746264.3760487

from high computation and the multi-view inconsistency (Janus problem). Multi-view diffusion models such as MVDream [19] and SyncDreamer [7] address this by correlating features across views. Wonder3D [8], used in our study, extends this to jointly generate color images and normal maps, achieving higher fidelity and consistency.

2.2 Calorie Estimation

Calorie estimation from food images has been approached with 2D methods that infer calories from meal classification and area, sometimes using reference objects (e.g., Okamoto et al. [13], Akpa et al. [1]) or multi-task learning without them (e.g., Ege et al. [5]). However, lacking depth and height information limits their accuracy.

3D-based methods address this by estimating food volume. Multi-view approaches (e.g., Dehais et al. [4]) require multiple images, while single-image methods such as Naritomi et al. [12] reconstruct meal and plate models to compute volume. CalorieVoL [20] integrates 3D estimation with multi-modal LLMs but suffers from reasoning limitations and plate-region overestimation. MFP3D [9] fuses 2D and 3D point cloud features for portion size regression, enabled by datasets like MetaFood3D [3].

In this study, we aim to improve single-image calorie estimation by leveraging a diffusion-based 3D reconstruction method for high-fidelity volume estimation and spatial-aware calorie estimation.

3 Method

Our framework estimates the calorie value and food category from a single RGB image by leveraging multi-modal visual cues. As illustrated in Figure 1, the process begins with segmentation and inpainting to recover the food-occluded region of the plate, followed by 3D reconstruction to estimate the food volume. We then extract 2D and 3D features from the image and point cloud sampled from the reconstructed mesh, and combine them with the estimated volume to perform multi-task estimation of calories and food categories. The pipeline consists of four main stages: (1) object segmentation and inpainting, (2) 3D reconstruction and volume estimation, (3) multi-modal feature extraction, and (4) calorie and category estimation.

Given an RGB food photo $I \in \mathbb{R}^{B\times H\times W\times 3}$, the model outputs the estimated calorie value $\hat{C}$ and food category $\hat{y}$:

$$(\hat{C}, \hat{y}) = \mathcal{F}_\Theta(I), \tag{1}$$

where the *trainable* parameter set is $\Theta = \{\theta_{\text{inp}}, \theta_{\text{est}}\}$; all other modules are kept frozen. The function $\mathcal{F}_\Theta$ represents the multi-task calorie estimation model, composed of inpainting, 3D reconstruction, feature encoding, and calorie estimation stages.

3.1 Mask Extraction and Plate Inpainting

The input image I is first segmented by an open-vocabulary segmentation module $\mathcal{S}$, producing binary masks for the food M_{f}, plate M_{p}, and reference object M_r:

$$(M_{\text{f}}, M_{\text{p}}, M_{\text{r}}) = \mathcal{S}(I, P_{seg}), \tag{2}$$

where the segmentation query P_{seg} consists of terms describing the food, plate, and reference object. The open-vocabulary segmentation model enables flexible region extraction via textual queries,

Figure 1: Overview of the proposed framework

allowing for adjustments when the target food is difficult to express with simple words or when using alternative reference objects.

Furthermore, a combined plate-and-food mask, M_{pf}, is created for 3D reconstruction. However, simply uniting the food and plate masks can leave a gap at their boundary. This gap can degrade the quality of subsequent inpainting and 3D reconstruction. To address this, we dilate the food mask with a structuring element k before uniting it with the plate mask to fill any potential gaps:

$$M_{\text{pf}} = \text{dilate}(M_{\text{f}}, k) \vee M_{\text{p}}, \tag{3}$$

where the value of k represents the dilation rate, set to 8% in this study.

Next, we apply the generated masks to the original image I to obtain images for subsequent modules. The plate-and-food image I_{pf} is created by masking the original image with M_{pf}. The image for inpainting, I_{p}, is created by masking out the food region from the plate image:

$$I_{\text{pf}} = I \odot M_{\text{pf}}, \quad I_{\text{p}} = I \odot (M_{\text{p}} - M_{\text{f}}), \tag{4}$$

Prior to the subsequent 3D reconstruction step, we inpaint the missing region in the plate image I_{p} to generate a complete, artifact-free plate image $\tilde{I}_p$, enabling high-quality 3D mesh reconstruction of the plate.

$$\tilde{I}_{\text{p}} = \mathcal{G}_{\tilde{\theta}_{\text{inp}}}(I_{\text{p}}, M_{\text{f}}, P_{\text{inp}}). \tag{5}$$

where the inpainting query P_{inp} describes the appearance of an empty plate using a concise textual expression. As described later, we train and utilize a diffusion-guided inpainting model $\mathcal{G}_{\tilde{\theta}_{\text{inp}}}$.

3.2 3D Reconstruction and Volume Estimation

To determine the precise geometry and volume of the food, we perform 3D reconstruction using the images constructed in the

previous section. The 3D reconstruction module $\mathcal{R}$ takes the inpainted plate image $\tilde{I}_p$ and the plate-and-food image I_{pf} as input to reconstruct their respective 3D meshes, $\mathcal{M}_{\text{p}}$ and $\mathcal{M}_{\text{pf}}$:

$$\mathcal{M}_{\text{p}} = \mathcal{R}(\tilde{I}_{\text{p}}), \quad \mathcal{M}_{\text{pf}} = \mathcal{R}(I_{\text{pf}}). \tag{6}$$

Since these meshes are not yet scaled to the real world, we estimate the actual scale from the reference object to calculate the volume. In this study, we use the calibration card because it has known physical dimensions and serves as a reliable reference object. Furthermore, since our method uses an open-vocabulary model, other common dining objects like chopsticks or cups could also be used, provided their sizes are known.

The volume of food is estimated by applying a scale factor s, computed from the reference object in the image. The final food volume V_{food} is calculated by scaling the difference between the volumes of the reconstructed meshes:

$$V_{\text{food}} = s\big(\text{vol}(\mathcal{M}_{\text{pf}}) - \text{vol}(\mathcal{M}_{\text{p}})\big) \in \mathbb{R}^{B\times 1}. \tag{7}$$

3.3 Multi-modal Feature Extraction and Calorie Estimation

To enrich the feature representation for the calorie estimation module, we perform both 2D feature extraction to capture visual appearance and 3D feature extraction to capture geometric properties.

First, we apply an image encoder $\mathcal{E}_{\text{2D}}$ to the input image I to extract the 2D appearance feature z_{2D}:

$$z_{\text{2D}} = \mathcal{E}_{\text{2D}}(I) \in \mathbb{R}^{B\times d_{\text{2D}}}. \tag{8}$$

This is expected to improve calorie estimation accuracy by capturing information unavailable from 3D shape features alone, such as the type of food and its surface characteristics (e.g., oiliness or degree of cooking).

Next, we obtain a plate-and-food point cloud $\mathcal{P}_{\text{pf}} = \text{sample}(\mathcal{M}_{\text{pf}})$, and apply a point-cloud encoder $\mathcal{E}_{\text{3D}}$ to extract the 3D shape feature z_{3D}:

$$z_{\text{3D}} = \mathcal{E}_{\text{3D}}(\mathcal{P}_{\text{pf}}) \in \mathbb{R}^{B\times d_{\text{3D}}}. \tag{9}$$

This aims to improve calorie estimation performance by accurately capturing the detailed 3D geometry of the food.

These features are concatenated into a multi-modal feature vector ϕ, which is then input to the calorie estimation module:

$$\phi = \text{concat}(V_{\text{food}}, z_{\text{2D}}, z_{\text{3D}}) \in \mathbb{R}^{B\times(1+d_{\text{2D}}+d_{\text{3D}})}. \tag{10}$$

The calorie estimation module predicts both the calorie value and the food category. This dual-head architecture facilitates effective learning through the multi-task objective described below.

$$(\hat{C}, \hat{y}) = \mathcal{H}_{\theta_{\text{est}}}(\phi). \tag{11}$$

An overview of this process is illustrated in Figure 2.

3.4 Training

We adopt a two-stage training strategy. First, we fine-tune the diffusion-guided inpainting model $\mathcal{G}_{\theta_{\text{inp}}}$ to enhance the performance of the inpainting module. Then, we train the calorie estimation module $\mathcal{H}_{\theta_{\text{est}}}$ based on the extracted multi-modal features.

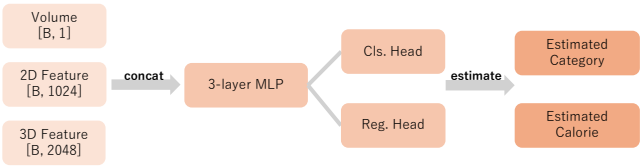

Figure 2: Multi-modal calorie estimator combining volume, 2D and 3D features.

Figure 3: Inpainting result using the default Stable Diffusion v2 model. The model hallucinates food, which motivated our fine-tuning approach.

Figure 4: Examples from our plate dataset used for inpainting model fine-tuning. (Top) real-world plate images collected from public sources. (Bottom) synthetic plate images generated via text-to-image Stable Diffusion v2.

3.4.1 Inpainting Model. For accurate plate reconstruction, we inpaint the occluded plate region using an inpainting model $\mathcal{G}_{\tilde{\theta}_{\text{inp}}}$ with a prompt describing an empty plate. However, as shown in Figure 3, the vanilla model often hallucinates food in the missing region. This behavior is likely due to a bias in the model's training data, which frequently depicts plates containing food rather than empty plates.

To address this issue, we fine-tuned the model on a custom dataset of food-free plates (Figure 4). This dataset was created by collecting 1,000 real-world images from public sources and augmenting them with 2,000 synthetic plate images generated using Stable Diffusion v2. The synthetic images were produced by randomizing plate thickness, color, and background, while applying the negative prompt `"multiple plates, food, text, watermark"` to suppress unwanted artifacts.

This process yielded a balanced dataset of 3,000 images. We fine-tuned Stable Diffusion v2 using LoRA [6] on this dataset. The resulting inpainting model, $\mathcal{G}_{\tilde{\theta}_{\text{inp}}}$, is trained independently and kept frozen for all downstream tasks.

3.4.2 Calorie Estimator. For calorie estimation, the MLP $\mathcal{H}_{\theta_{\text{est}}}$ is trained using the following multi-task objective:

$$\mathcal{L} = \lambda_{\text{reg}} \left| \hat{C} - C_{\text{gt}} \right| + \lambda_{\text{cls}} \ \text{CE}(\hat{y}, y_{\text{gt}}), \tag{12}$$

At this stage, we train the calorie estimator while keeping all modules except θ_{est} frozen.

3.5 Module Details

We use Grounded-SAM [18] as segmentation module $\mathcal{S}$, combining GroundingDINO1.5 [17] and SAM-2 [16]. GroundingDINO1.5 takes the query `"food . plate . chessboard"` as P_{seg} to predict bounding boxes, and SAM-2 generates corresponding masks.

For 3D reconstruction module $\mathcal{R}$, we adopt Wonder3D [8], a multi-view diffusion model producing watertight meshes with high precision and cross-view consistency by leveraging 2D diffusion priors. This enables reconstruction of complex food geometries for accurate volume and calorie estimation.

For 3D shape feature extractor $\mathcal{E}_{\text{3D}}$, we use PointGPT [2] on reconstructed point clouds, employing a ModelNet40-pretrained model with frozen final layer to obtain 2048-dimensional embeddings capturing volumetric and geometric details. We also use CLIP ViT-L/14@336px to extract 1024-dimensional 2D appearance features such as color and texture.

4 Experiments

4.1 Dataset

We used the MetaFood3D dataset [3], which contains 637 food items across 108 categories, including both individual items (e.g., apples, bananas) and plated dishes (e.g., bagels, toast). Each image includes a calibration card for scale reference. The dataset was split into training and test sets with an 80:20 ratio.

4.2 Implementation Details

Table 1 summarizes the implementation parameters used in our experiments. These include thresholds for mask extraction, LoRA inpainting configurations, 3D reconstruction settings, and training hyperparameters for calorie estimation. For inpainting, we use the Stable Diffusion v2 weights as the base model[1]. For the image encoder, we adopt the publicly available pretrained weights[2].

4.3 Results of Volume, Calorie, and Category Estimation

We evaluated volume estimation and calorie prediction using Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). Our method was compared with MFP3D [9], and the results are shown in Table 2.

Our method achieved a lower MAE in volume estimation than MFP3D, indicating higher accuracy. The MAPE was high due to segmentation failures especially on small food items (e.g., nuts, blueberries), leading to large relative errors. Although calorie estimation performance was lower than MFP3D, the results highlight the benefits of incorporating volume and 3D shape features. In category classification, our method achieved high accuracy, demonstrating the effectiveness of the multi-task learning approach.

[1] `stabilityai/stable-diffusion-2-inpainting`.
[2] `vit_large_patch14_clip_336.openai_ft_in12k.in1k`.

Table 1: Implementation details

Category	Parameter	Value
Mask Extraction	Expansion ratio of food area	8%
	`text.threshold`	0.25
	`box.threshold`	0.35
Inpainting	LoRA scale	0.75
	Prompt	`"plain plate"`
3D Reconstruction	Crop size	192
	Camera type	`"perspective"`
	Refinement steps	2
3D Feature Extraction	Input points	1024
	Depth	24
Calorie Estimation	Epochs	50
	Batch size	512
	Training samples	8,357
	Test samples	2,090

Table 2: Evaluation results between MFP3D and the proposed method on MetaFood3D [3].

Method	Volume MAE (ml)	Volume MAPE (%)	Calorie MAE (kcal)	Calorie MAPE (%)	Category Accuracy
MFP3D [9]	62.60	**41.43**	**77.98**	**68.05**	–
Ours	**57.53**	269.31	92.74	212.78	**0.9641**

Table 3: Ablation study on feature contributions to calorie estimation on MetaFood3D. 2D, Vol., and 3D denote 2D appearance, volume, and 3D shape features.

2D	Vol.	3D	Calorie MAE (kcal)	Calorie MAPE (%)
✓			133.27	373.23
✓	✓		101.12	254.09
✓	✓	✓	**92.74**	**212.78**

4.4 Ablation Study on Feature Contributions

We investigated the impact of different feature types on calorie estimation: (1) 2D features only, (2) 2D features + volume, (3) Ours (2D + volume + 3D features).

Results in Table 3 show that incorporating volume information and 3D shape features substantially improves calorie estimation accuracy. Volume features, in particular, yield a significant performance gain, demonstrating the value of our framework's volume estimation.

4.5 Effect of LoRA Scaling Factor

We evaluated the effect of the LoRA scaling factor by varying it across {0.0, 0.25, 0.5, 0.75, 1.0}. Figure 5 illustrates representative inpainting results for each setting. When the scaling factor was too low, the inpainting was barely applied, whereas excessively high values caused blurring even on the target object. From these observations, a value of 0.75 yielded the most visually natural results.

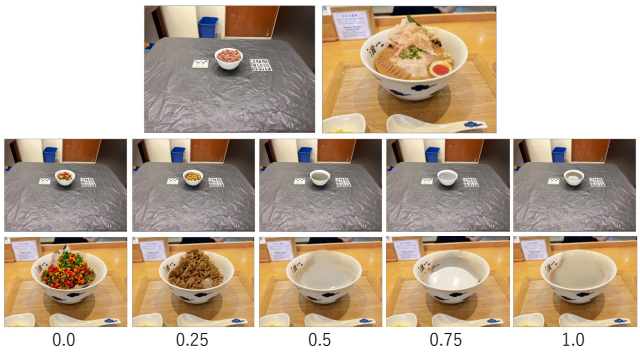

Figure 5: Comparison of inpainting results with different LoRA scales.

4.6 Qualitative Evaluation

Figure 6 illustrates the end-to-end pipeline from a single input image to 3D reconstruction, including mask extraction and plate inpainting.

The predicted masks cleanly separate the relevant regions and accurately preserve fine boundary details. The inpainting step plausibly recovers occluded plate areas, enabling robust reconstruction for complex meals such as bowl-type dishes. Both plate-and-food and plate-only 3D meshes are successfully reconstructed, capturing fine geometry and texture.

4.7 Limitations and Future Work

We found that failures in detecting reference objects, particularly for small or off-plate food items, were the main source of large errors in volume estimation. These cases significantly increased relative errors and worsened the MAPE score. This suggests that the proposed method is effective when segmentation succeeds but remains highly vulnerable to detection failures, especially with small or non-plate items.

To address these issues, future work should aim to improve the accuracy of reference object segmentation, enhance the robustness of volume estimation against segmentation errors, and develop a dataset for more comprehensive evaluation. Additionally, incorporating mechanisms to detect and handle small or off-plate food items could help reduce the sensitivity of the method to such cases and improve overall estimation stability.

The current pipeline takes approximately two minutes per sample, including about one minute for 3D reconstruction when using an A4000 GPU. To enable practical deployment, future work should also explore optimization of both the model and pipeline for faster inference. This includes accelerating the 3D reconstruction process and designing lightweight models suitable for real-time or edge applications.

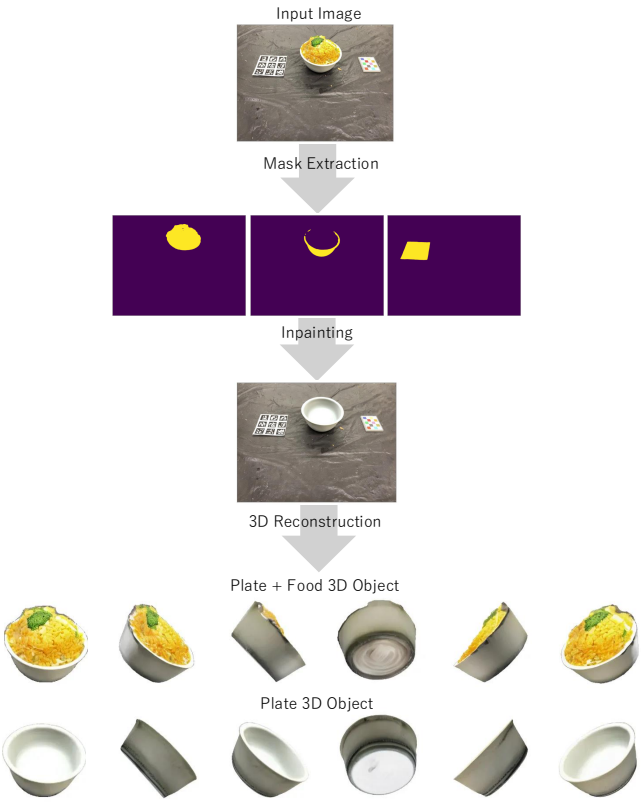

Figure 6: Example outputs from the proposed method, including mask extraction, inpainting, and 3D reconstruction.

5 Conclusion

We introduced a novel framework for calorie estimation from a single food image, leveraging diffusion-guided 3D reconstruction. By integrating open-vocabulary segmentation, inpainting for plate completion, and multi-view-consistent 3D reconstruction, our approach enables high-fidelity volume estimation. These volumetric cues, combined with 2D and 3D features in a multi-task learning setting, enhance calorie estimation performance, as confirmed through ablation studies. The modular design offers a flexible path to integrate stronger 3D priors and alternative reference objects.

To support practical deployment, future work will focus on improving segmentation robustness, handling challenging inputs more effectively, and accelerating the pipeline through optimized reconstruction and model lightweighting. These advancements will bring us closer to realizing accurate, efficient, and scalable dietary monitoring systems.

Acknowledgments

This work was supported by JSPS KAKENHI 22H00548 and JST CRONOS JPMJCS24K4.

References

[1] Elder Akpro Hippocrate Akpa, Hirohiko Suwa, Yutaka Arakawa, and Keiichi Yasumoto. 2017. Smartphone-Based Food Weight and Calorie Estimation Method for Effective Food Journaling. *SICE Journal of Control, Measurement, and System Integration* (2017).

[2] Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. 2024. PointGPT: Auto-regressively generative pre-training from point clouds. *Advances in Neural Information Processing Systems* (2024).

[3] Yuhao Chen, Jiangpeng He, Chris Czarnecki, Gautham Vinod, Talha Ibn Mahmud, Siddeshwar Raghavan, Jinge Ma, Dayou Mao, Saeejith Nair, Pengcheng Xi, Alexander Wong, Edward Delp, and Fengqing Zhu. 2024. MetaFood3D: Large 3D Food Object Dataset with Nutrition Values. In *Proc. of International Conference on Learning Representations.*

[4] Joachim Dehais, Marios Anthimopoulos, Sergey Shevchik, and Stavroula Mougiakakou. 2017. Two-view 3d reconstruction for food volume estimation. *IEEE Transactions on Multimedia* 19, 5 (2017), 1090–1099.

[5] Takumi Ege and Keiji Yanai. 2017. Simultaneous estimation of food categories and calories with multi-task CNN. In *2017 IAPR International Conference on Machine Vision Applications (MVA).* 198–201.

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proc. of International Conference on Learning Representations.*

[7] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2024. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. In *International Conference on Learning Representations (ICLR).*

[8] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. 2024. Wonder3D: Single Image to 3D using Cross-Domain Diffusion. In *Proc. of IEEE Computer Vision and Pattern Recognition.* 9970–9980.

[9] Jinge Ma, Xiaoyan Zhang, Gautham Vinod, Siddeshwar Raghavan, Jiangpeng He, and Fengqing Zhu. 2024. MFP3D: Monocular Food Portion Estimation Leveraging 3D Point Clouds. In *2024 27th International Conference on Pattern Recognition (ICPR), 9th International Workshop on Multimedia Assisted Dietary Management.* arXiv preprint arXiv:2411.10492.

[10] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. 2023. RealFusion: 360 Reconstruction of Any Object from a Single Image. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 8517–8527.

[11] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 4460–4470.

[12] Shu Naritomi and Keiji Yanai. 2021. Hungry Networks: 3D mesh reconstruction of a dish and a plate from a single dish image for estimating food volume. In *Proc. of the 2nd ACM International Conference on Multimedia in Asia.*

[13] Koichi Okamoto and Keiji Yanai. 2016. An Automatic Calorie Estimation System of Food Images on a Smartphone. In *Proc. of the 2nd International Workshop on Multimedia Assisted Dietary Management.* ACM, 63–70.

[14] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 165–174.

[15] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *International Conference on Learning Representations (ICLR).*

[16] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. In *arXiv preprint arXiv:2408.00714.* arXiv:2408.00714 [cs.CV]

[17] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhang. 2024. Grounding DINO 1.5: Advance the "Edge" of Open-Set Object Detection. *arXiv preprint arXiv:2405.10300* (2024). arXiv:2405.10300 [cs.CV]

[18] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. *arXiv preprint arXiv:2401.14159* (2024). arXiv:2401.14159 [cs.CV]

[19] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. 2024. MVDream: Multi-view Diffusion for 3D Generation. In *International Conference on Learning Representations (ICLR).*

[20] Hikaru Tanabe and Keiji Yanai. 2025. CalorieVoL: Integrating Volumetric Context Into Multimodal Large Language Models for Image-Based Calorie Estimation. In *Proc. of International Conference on MultiMedia Modeling.*

[21] Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, and Jack Sim. 2021. Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food. In *Proc. of IEEE Computer Vision and Pattern Recognition.* 8903–8911.

[22] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. 2018. Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images. In *European Conference on Computer Vision (ECCV).* 52–67.