G2: 3D-FoodCalorie: Geometry-Aware Calorie Estimation from signal image

Haodong Zhang
Brown University
Providence, RI 02912
haodong_zhang@brown.edu

Xinye Yang
Brown University
Providence, RI 02912
xinye_yang@brown.edu

Rui Zhou
Brown University
Providence, RI 02912
rui_zhou1@brown.edu

Abstract

We present a novel 3D-aware framework that addresses the limitations of 2D-based calorie estimation from food images. Our three-stage approach first performs fine-grained semantic segmentation to identify food items, then leverages self-supervised learning for food-specific depth estimation, and finally employs a fusion model that implicitly learns 3D food information while integrating 2D features to predict nutrition content. By capturing geometric and volumetric properties from a single image, our method overcomes challenges posed by occluded or irregularly shaped foods. This solution bridges computer vision and diet science to support related nutrition science applications. Our work has been proposed at https://github.com/Yangxinyee/3D-FoodCalorie.

1. Introduction

Poor dietary habits constitute a significant global health challenge, accounting for approximately one-fifth of all deaths worldwide [7]. Accurate assessment of the nutritional content of food plays a pivotal role in promoting healthier eating behaviors and informing public health policies. Although nutrition professionals offer critical expertise, their services are not always accessible or practical for routine dietary monitoring. In response, a range of mobile applications and computational tools for dietary assessment have been developed [36]. However, these tools often depend heavily on manual input and lack the precision required for reliable, automated nutritional analysis [45].

A fundamental challenge in automated nutrition estimation lies in accurately predicting caloric content from food images. Most existing approaches rely heavily on 2D visual features, which are inherently inadequate for capturing the three-dimensional geometric and volumetric characteristics of food items. This deficiency significantly affects the estimation of food mass and volume—key factors for calorie prediction. The limitations of 2D perception are particularly evident in scenarios involving occlusions, self-overlapping structures, or irregularly shaped food items, where scale ambiguity and perspective distortion further exacerbate prediction errors. To overcome the constraints of pure 2D methods, some recent approaches [33, 44, 49] have introduced depth information into the prediction pipeline to enhance 3D perception. These methods often assume access to accurate depth maps at inference time. While this integration improves estimation performance under controlled conditions, it imposes unrealistic assumptions for real-world deployment. In everyday settings, users typically capture food images with standard monocular cameras, making it impractical to obtain reliable ground-truth depth maps. As a result, such methods lack scalability and are unsuitable for widespread adoption in consumer-facing applications.

To address these challenges, we propose a novel geometry-aware framework that incorporates 3D structural reasoning into the nutrition estimation pipeline. Given a single RGB image as input, our method performs the following three stages:

1. **Semantic segmentation**: We perform fine-grained segmentation to isolate individual food items and separate them from background elements such as the plate.
2. **Depth estimation via self-supervised learning**: A pre-trained self-supervised depth estimation module is fine-tuned on food imagery to infer depth maps that capture volumetric cues critical for nutrition estimation.
3. **RGB-D fusion**: We integrate RGB features, depth and segmentation information through a fusion network to estimate nutritional attributes such as calorie content.

By introducing depth-aware reasoning into the estimation process, our RGB-D Fusion Network substantially improves the accuracy and robustness of calorie predictions from monocular images. This approach bridges the gap between computer vision and nutritional science, providing a scalable and practical solution in real-world setting.

2. Background & Related Work

Nutrition Estimator With the advancement of computer vision, vision-based methods for food recognition [37, 52]

and nutrition estimation [13, 49] have gained traction as scalable tools for monitoring eating habits and estimating nutrient intake, contributing to the field of precision nutrition [9]. Ege et al.[11] demonstrated that multi-task CNNs could jointly predict food categories, ingredients, and calories, outperforming single-task models. Liu et al.[30] further explored CNN-based approaches combined with non-destructive detection to analyze complex food matrices.

Despite these advances, many methods incorporate depth in a simplistic fashion. Thames et al. [33, 49], for instance, appended depth as a fourth channel to RGB inputs and treated all modalities identically. This approach neglects the distinct roles of RGB and depth, limiting the model's ability to fully exploit multimodal features. Uniform fusion of RGB and depth features often underutilizes geometric cues, reducing both prediction accuracy and robustness.

Depth Estimator To address limited annotated food-depth data and generalization concerns, we adopt self-supervised monocular depth estimation. These approaches frame depth learning as view synthesis, where networks jointly predict depth and pose by minimizing photometric reconstruction losses. Zhou et al.[62] introduced this paradigm using separate depth and pose networks, leveraging image reconstruction as supervision. Godard et al.[18] improved robustness through minimum reprojection loss and auto-masking. Johnston and Carneiro [24] further incorporated self-attention [50] and discrete disparity volumes [25] to enhance spatial reasoning and boundary sharpness.

A persistent issue in these models is scale inconsistency between predicted depth and pose, resulting in scale drift. While constraints [2, 15] can partially mitigate this, they cannot fully resolve the problem. Zhao et al. [60] addressed scale ambiguity by eliminating the pose network, instead estimating the fundamental matrix from optical flow and using triangulation for 3D reconstruction. Though effective, this method increases computational overhead.

Semantic Segmentation Mask R-CNN [17], an extension of Faster R-CNN [41], remains one of the most widely adopted instance segmentation methods. It employs a ResNet-50 backbone and Feature Pyramid Network to extract multi-scale features, followed by a Region Proposal Network and ROIAlign for spatial precision. Three branches output object class, bounding boxes, and binary masks, with the segmentation path structured as a lightweight Fully Convolutional Network.

Although powerful, Mask R-CNN is commonly trained on datasets like COCO, which lack fine-grained food annotations. This domain gap undermines performance on food segmentation tasks that require recognizing subtle category distinctions in visually similar or densely composed meals.

3. Method

3.1. Overall

The proposed framework comprises three stages. An input image is first passed through a segmentation network and a depth estimator to generate a mask and depth map. These outputs are combined to isolate the 3D structure of food while suppressing irrelevant background. The resulting food-specific depth and RGB image are then fused to estimate nutritional content.

3.2. Semantic Segmentation

The standard Mask R-CNN model is typically pretrained on the COCO dataset [28], which contains only a limited number of food-related categories. This introduces a significant domain gap between the COCO label space and the rich, fine-grained food representation space encountered in real-world applications. As a result, the COCO-pretrained Mask R-CNN [17] struggles to generalize effectively to complex food scenes.

To bridge this gap and adapt the model to our domain, we fine-tune it on the FoodSeg103 dataset [55], a large benchmark that has more than 11,000 images labeled with 103 fine-grained food classes and pixel-level segmentation masks. The benchmark contains a large number of real-life food compositions, allowing subtle visual differences in different food categories to be learned.

Mask R-CNN casts segmentation as a multitask learning task for which it defines a composite loss consisting of classification loss, bounding box regression loss, and mask loss. Mask loss is computed as an average binary cross-entropy over predicted mask pixels that operates only on the ground-truth class of each Region of Interest (RoI) [17]. The instance-specific object masks with high spatial resolution are learned through it.

3.3. Self-supervised Depth Estimator

Common PoseNet-based methods often encounter three core issues: 1) erroneous supervision signals caused by scale inconsistency between pose and depth [2, 15, 60], 2) photometric error for supervision in an implicit manner is susceptible to textureless regions and data noise, and 3) CNN-based pose estimation is hard to generalize [42, 61]. FlowNet-based methods [20, 48, 60] solve these issues, but they also introduce excessive computational costs. Therefore, our depth estimator aims to design a lightweight FlowNet-based model to address the above problems simultaneously while maintaining performance.

3.3.1. Estimate relative pose from flow prediction

Unlike PoseNet, which directly regresses camera pose via a neural network, our FlowNet-based approach offers better geometric interpretability and enforces scale consistency explicitly. We adopt PWCNet [47] to estimate dense

optical flow between adjacent frames, yielding pixel correspondences. To filter out unreliable matches caused by occlusion or textureless regions, we follow [60] and select non-occluded pixels (based on occlusion mask M_o) with top 20% forward-backward consistency scores M_s. From these reliable correspondences, we compute the fundamental matrix $\mathbf{F}$ using the normalized 8-point algorithm [21] with RANSAC [12]. $\mathbf{F}$ encodes the epipolar geometry between views and is decomposed into four candidate poses $([R, t])^4$, from which the correct one is selected using the cheirality check [60].

Although the recovered pose is still scale-ambiguous, we align the predicted depth map accordingly. Specifically, we compute a geometric consistency score map M_r by measuring each pixel's distance to its corresponding epipolar line. Combining M_r, M_o, and M_s, we retain the most reliable correspondences for two-view triangulation, yielding a sparse, up-to-scale 3D structure D^{tri}. This structure supervises the predicted depth D^{pred} through a scale-aligned loss:

$$L_d = \left(\frac{D^{tri} - s \cdot D^{pred}}{D^{tri}} \right)^2 \tag{1}$$

where s is a learnable scale factor. This design ensures consistent scale between depth and pose. Fig.6 shows the specific structure of the FlowNet-based framework.

3.3.2. Lightweight DepthNet

Typical PoseNet-based frameworks often adopt a simple ResNet architecture as the depth network, or enhance the depth prediction capability by incorporating additional components and modalities [3, 24, 32, 40, 58, 63]. However, since our AI system already includes additional computational overhead due to the pose estimation method described in Section 3.3.1, our model has a lightweight depth network to balance the overall computational cost.

Our encoder consists of three stages. Each stage contains multi-dilated convolution blocks and attention mechanisms. This block employs several convolutions [5] with different dilation rates to expand the receptive field across multiple scales without increasing the parameter count. For the attention mechanism, we use cross-covariance attention [1], which models global context along the channel dimension and significantly reduces computational and memory overhead compared to traditional self-attention [50]. To preserve spatial information, each downsampling layer is concatenated with pooled features from the input image. In addition, skip connections are applied between encoder and decoder features to facilitate feature flow and fusion. This architecture strengthens feature representation, enabling accurate depth estimation and efficient inference with a compact model. The decoder is the same as DispNet [18]. Fig.7 shows the overall pipeline.

3.3.3. Loss Function

The loss function of training the depth estimator is:

$$L = w_1 L_f + w_2 L_d + w_3 L_p + w_4 L_s \tag{2}$$

L_f is the unsupervised loss for optical flow in [47], L_d is Eq.1 in Section 3.3.1, L_p is reprojection error for image pairs, and L_s is depth smoothness loss in [2]. Each loss function can be found in details in Appendix .2.

3.4. RGB-D Feature Fusion Module

To improve food nutrition estimation from RGB and depth images, we adopt a dual-branch feature fusion framework inspired by Shao et al. [44], combining multi-scale and multimodal information.

3.4.1. Model Structure

Multi-scale Fusion The model consists of two ResNet-101 backbones [22] for RGB and depth modalities. Each branch extracts multi-level features, which are then fused at corresponding scales using a Feature Pyramid Network (FPN) [29]. This process yields a set of multi-scale feature maps $\{C_i\}_{i=2}^5$ that integrate both visual appearance and geometric cues through element-wise addition:

$$C_i = R_i \oplus D_i, \quad i = 2, 3, 4, 5 \tag{3}$$

Multimodal Feature Fusion (MMFF). The fused features are fed into a Multimodal Feature Fusion module consisting of two components. First, a Balanced Feature Pyramid (BFP) [38] resizes and aggregates the multi-scale features to reinforce semantic consistency. Then, non-local attention [53] is applied to capture global context. Finally, a CBAM [54] block refines the features using sequential channel and spatial attention, enhancing important regions.

Nutrition Predictor The output of MMFF forms a unified feature representation that encapsulates both fine-grained and high-level semantic information. This refined feature is then passed to the prediction head, which regresses five nutritional attributes: calories, mass, fat, carbohydrate, and protein. Overall pipeline has been shown in Fig.8.

3.4.2. Loss Function

We employ a normalized multi-task loss to jointly optimize all five targets:

$$\mathcal{L} = \sum_{t \in \{\text{cal,mass,fat,carb,protein}\}} \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{y}_i^t - y_i^t|}{\frac{1}{N}\sum_{i=1}^{N} y_i^t} \tag{4}$$

This formulation balances training across targets with different value scales.

4. Result

4.1. Mask R-CNN

Two standard metrics are used to evaluate segmentation performance: **Mean Intersection over Union (mIoU)** and **Mean Average Precision (mAP)**. mIoU measures the average overlap between predicted and ground truth regions across all classes, and is widely used in semantic segmentation. mAP is computed at IoU thresholds of 0.50 and 0.75 (mAP@0.50 and mAP@0.75), reflecting detection and instance segmentation accuracy. While mAP@0.50 is more lenient, mAP@0.75 requires stricter overlap, providing a more challenging evaluation.

We infer our model on the FoodSeg103 test set from both our fine-tuned model and COCO-pretrained Mask R-CNN. As shown in Table 1, the fine-tuned model significantly outperforms the baselines for all three metrics, recording a mIoU of **0.537**, mAP@0.50 of **0.395**, and mAP@0.75 of **0.264**. On the same domain-specific dataset, COCO-pretrained model fares poorly with mIoU of 0.0455 and both mAP values approaching 0. Compared against FoodSeg103 baselines [10, 23, 26, 31] shown in Fig.10, our model is outperforming while applying a lightweight architecture. Fig. 5 shows the visualization results of segmentation for food.

4.2. Self-supervised Depth Estimator

We evaluate monocular depth estimation using several standard metrics. Assume D^{pred} is the predicted depth and D^{gt} is the ground truth. **Absolute Relative Error (AbsRel)** measures the average of the absolute difference between D^{pred} and D^{gt}, normalized by the D^{gt}. **Squared Relative Error (SqRel)** emphasizes larger errors by computing the squared difference relative to the D^{gt}. **Root Mean Square Error (RMSE)** calculates the overall difference between D^{pred} and D^{gt}, while **Log RMSE** performs this calculation in log space to reduce sensitivity to absolute scale. Finally, **Threshold Accuracy** δ reports the percentage of predicted depths that fall within a certain threshold of the ground truth. Lower error values and higher threshold accuracy indicate better performance.

The model is validated on KITTI [16]. Then, it is pretrained on NYUv2 [46], followed by fine-tuning on Nutrition5K [49] to reduce the domain gap between indoor and food-centric images and enhance generalization to real-world dietary scenarios. Inference on the KITTI test set follows the standard protocol from [6], with results shown in Table 6. Our method achieves the best or second-best performance across most metrics, outperforming both the baseline and prior methods. Visualization results are presented in Fig. 1. Table 4 shows NYUv2 pretraining results, where our model performs competitively despite limited training due to dataset size. Table 3 compares fine-tuning strategies, including tuning the decoder only or both encoder and decoder, with either L_2 or Scale-Invariant Logarithmic Loss. Fine-tuning both encoder and decoder with L_2 loss performs best. Few-shot evaluation results in Table 2 further demonstrate strong generalization. Depth visualizations for food images are shown in Fig. 4.

4.3. RGB-D Feature Fusion Module

Percentage of Mean Absolute Error (PMAE) [49] is the main evaluation metric being adopted. It normalizes MAE [8] by the mean ground truth, allowing fair comparison across nutritional components.

The backbone network is initialized with weights pretrained on the Food2k dataset which is known for its large-scale diversity in food categories and image quantity [37], serving as an effective source for model pretraining. Prior to finetuning, we apply several preprocessing techniques to the dataset, including image resizing, random horizontal flipping, and center cropping. During the finetuning process, we adopted a multi-scale strategy: every 10 iterations, the input resolution for each batch was randomly selected from the following set — (256×352), (288×384), (320×448), (352×480), and (384×512).

Our reproduced model achieved promising results, with Percent Mean Absolute Error (PMAE) values of 15.1% for calories, 11.2% for mass, 23.4% for fat, 21.4% for carbohydrate, and 23.0% for protein. Table 5 shows the comparison of results between our reproduced version and the original paper. Compared to traditional methods that rely solely on 2D images or simply fuse depth maps as an additional input channel (e.g., using them as a fourth channel), our approach achieves superior performance. Table 7 shows the results.

Different models' training setups can be found in Table 8. User Interface is shown from Fig.11 to Fig.14.

5. Conclusion

Our work presents a food nutrition estimation system that integrates segmentation, monocular depth estimation, and multi-scale 2D-3D feature fusion under a multi-task learning framework. The system enables calorie and nutrient estimation from a single RGB image, offering practical value for personal dietary monitoring and clinical nutrition support. Despite its effectiveness, the system faces challenges in cases of food occlusion and visually similar items, which can lead to missing detections or misclassifications. These limitations largely stem from error propagation across the modular pipeline. To address these issues, future work will explore the integration of additional modalities such as recipe text, and incorporate large language models to enhance contextual understanding and robustness. We also plan to make the model learning become end-to-end. Overall, our system provides a promising step toward practical and scalable nutrition estimation.

References

[1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34:20014–20027, 2021. 3

[2] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in neural information processing systems*, 32, 2019. 2, 3, 9

[3] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *International Journal of Computer Vision*, 129(9):2548–2564, 2021. 3

[4] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8001–8008, 2019. 9

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[6] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7063–7072, 2019. 4, 9

[7] GBD 2015 Obesity Collaborators. Health effects of overweight and obesity in 195 countries over 25 years. *New England journal of medicine*, 377(1):13–27, 2017. 1

[8] Arnaud De Myttenaere, Boris Golden, Bénédicte Le Grand, and Fabrice Rossi. Mean absolute percentage error for regression models. *Neurocomputing*, 192:38–48, 2016. 4

[9] Juan de Toro-Martín, Benoit J Arsenault, Jean-Pierre Després, and Marie-Claude Vohl. Precision nutrition: a review of personalized nutritional approaches for the prevention and management of metabolic syndrome. *Nutrients*, 9(8):913, 2017. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[11] Takumi Ege and Keiji Yanai. Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 367–375, 2017. 2

[12] MA FISCHLER AND. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 3

[13] Emma Foster and Jennifer Bradley. Methodological considerations and future insights for 24-hour dietary recall assessment in children. *Nutrition Research*, 51:1–11, 2018. 2

[14] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 8

[15] Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T Barron. Learning single camera depth estimation using dual-pixels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7628–7637, 2019. 2

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013. 4, 9

[17] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. `https://github.com/facebookresearch/detectron`, 2018. 2

[18] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 2, 3, 9

[19] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8977–8986, 2019. 9

[20] Vitor Guizilini, Kuan-Hui Lee, Rareş Ambruş, and Adrien Gaidon. Learning optical flow, depth, and scene flow without real-world labels. *IEEE Robotics and Automation Letters*, 7 (2):3491–3498, 2022. 2

[21] Richard I Hartley. In defense of the eight-point algorithm. *IEEE Transactions on pattern analysis and machine intelligence*, 19(6):580–593, 1997. 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[23] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 4

[24] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 4756–4765, 2020. 2, 3

[25] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 2

[26] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 4

[27] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1119–1127, 2015. 8

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 2

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[30] Yao Liu, Hongbin Pu, and Da-Wen Sun. Efficient extraction of deep image features using convolutional neural network (cnn) for applications in detecting and analysing complex food matrices. *Trends in Food Science & Technology*, 113: 193–204, 2021. 2

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[32] Zhong Liu, Ran Li, Shuwei Shao, Xingming Wu, and Weihai Chen. Self-supervised monocular depth estimation with self-reference distillation and disparity offset refinement. *IEEE transactions on circuits and systems for video technology*, 33(12):7565–7577, 2023. 3

[33] Ya Lu, Thomai Stathopoulou, Maria F Vasiloglou, Stergios Christodoulidis, Zeno Stanga, and Stavroula Mougiakakou. An artificial intelligence-based system to assess nutrient intake for hospitalised patients. *IEEE transactions on multimedia*, 23:1136–1147, 2020. 1, 2

[34] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2624–2641, 2019. 9

[35] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5667–5675, 2018. 9

[36] Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P Murphy. Im2calories: towards an automated mobile vision food diary. In *Proceedings of the IEEE international conference on computer vision*, pages 1233–1241, 2015. 1

[37] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9932–9949, 2023. 1, 4

[38] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 821–830, 2019. 3

[39] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12240–12249, 2019. 9

[40] Chaopeng Ren. Eite-mono: an extreme lightweight architecture for self-supervised monocular depth estimation. *IEEE Access*, 2024. 3

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015. 2

[42] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3302–3312, 2019. 2

[43] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008. 8

[44] Wenjing Shao, Weiqing Min, Sujuan Hou, Mengjiang Luo, Tianhao Li, Yuanjie Zheng, and Shuqiang Jiang. Vision-based food nutrition estimation via rgb-d fusion network. *Food Chemistry*, 424:136309, 2023. 1, 3

[45] Jee-Seon Shim, Kyungwon Oh, and Hyeon Chang Kim. Dietary assessment methods in epidemiologic studies. *Epidemiology and health*, 36:e2014009, 2014. 1

[46] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 4, 8

[47] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2, 3

[48] Yiyang Sun, Zhiyuan Xu, Xiaonian Wang, and Jing Yao. Flowdepth: Decoupling optical flow for self-supervised monocular depth estimation. *arXiv preprint arXiv:2403.19294*, 2024. 2

[49] Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, and Jack Sim. Nutrition5k: Towards automatic nutritional understanding of generic food. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8903–8911, 2021. 1, 2, 4, 8

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia

Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[51] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2022–2030, 2018. 9

[52] Wei Wang, Weiqing Min, Tianhao Li, Xiaoxiao Dong, Haisheng Li, and Shuqiang Jiang. A review on vision-based analysis for automatic dietary assessment. *Trends in Food Science & Technology*, 122:223–237, 2022. 1

[53] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[54] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3

[55] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven CH Hoi, and Qianru Sun. A large-scale benchmark for food image segmentation. In *Proceedings of ACM international conference on Multimedia*, 2021. 2

[56] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5354–5362, 2017. 8

[57] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 9

[58] Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Wordepth: Variational language prior for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9708–9719, 2024. 3

[59] Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18537–18546, 2023. 9

[60] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9161, 2020. 2, 3, 9

[61] Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. To learn or not to learn: Visual localization from essential matrices. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3319–3326. IEEE, 2020. 2

[62] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 2, 9

[63] Zhongkai Zhou, Xinnan Fan, Pengfei Shi, and Yuanxue Xin. R-msfm: Recurrent multi-scale feature modulation for monocular depth estimating. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12777–12786, 2021. 3, 9

[64] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 36–53, 2018. 9

Supplementary Material

.1. Mask R-CNN

.1.1. Tables and Figures

Model	mIoU	mAP@0.50	mAP@0.75
Original MRCNN	0.0455	0.0000	0.0000
Our MRCNN	**0.5372**	**0.3952**	**0.2641**

Table 1. Comparison of our fine-tuned Mask R-CNN model and the original COCO-pretrained model on FoodSeg103 test set.

.2. Self Supervised Depth Estimator

Unsupervised loss for optical flow L_f is:

$$L_f = \sum_{l=l_0}^{L} \alpha_l \sum_{\mathbf{x}} \left\| \mathbf{w}_\Theta^l(\mathbf{x}) - \mathbf{w}_{\mathrm{GT}}^l(\mathbf{x}) \right\|_2 + \gamma \left\| \Theta \right\|_2, \quad (5)$$

$\mathbf{w}_\Theta^l$ is the flow field at the lth pyramid level, $\mathbf{w}_{GT}^l$ is the supervised signal.

Dense reprojection error L_p computed by transformed depth can be formulated in the equation:

$$L_p = w_{31} L_{pf} + w_{32} L_{pd} \quad (6)$$

L_{pf} is the 2D error between optical flow and rigid flow generated by depth reprojection:

$$\begin{aligned} p_{bd} &= \phi\left(\mathbf{K}\left[\mathbf{T}_{ab} D_a(p_a) \mathbf{K}^{-1}(h(p_a))\right]\right) \\ p_{bf} &= p_a + F_{ab}(p_a) \\ L_{pf} &= \frac{1}{|M_r|} \sum_{p_a} M_r(p_a) \left| p_{bd} - p_{bf} \right| + |D_{epi}| \end{aligned} \quad (7)$$

where p_a is the pixel coordinate (x, y) in the image I_a, and $h(p_a)$ indicates the homogeneous coordinates of p_a.

The error of depth reprojection L_{pd} is defined as:

$$L_{pd} = \frac{1}{|M_o M_r|} \sum_{p_a} M_o(p_a) M_r(p_a) \left| 1 - \frac{D_b^a(p_{bd})}{D_b^s(p_{bd})} \right| \quad (8)$$

where M_o is the occlusion mask from optical flow, M_r is the inlier score map described in Sec 3.3.1. D_b^a is the reprojected depth map by D_a and T_{ab}. D_b^s is the interpolated depth map of D_b.

The depth smoothness loss L_s is defined as

$$L_s = \sum_{p} \left(e^{-\nabla I_a(p)} \cdot \nabla D_a(p) \right)^2 \quad (9)$$

where ∇ is the first derivative along spatial directions

Samples	Error (↓) rel	sq rel	rms	Accuracy δ (↑) < 1.25	$< 1.25^2$	$< 1.25^3$
0	0.238	0.296	0.109	0.543	0.863	0.972
5	0.226	0.308	0.107	0.592	0.838	0.952
10	0.226	0.307	0.106	0.591	0.838	0.953
20	0.224	0.304	0.105	0.592	0.840	0.955
50	0.222	0.301	0.104	0.597	0.844	0.958
100	0.218	0.297	0.103	0.607	0.848	0.961
300	**0.184**	**0.249**	**0.087**	**0.669**	**0.897**	**0.986**

Table 2. Few-shot depth estimation performance under varying sample sizes on Nutrition5K [49] dataset.

Method	Error (↓) rel	sq rel	rms	Accuracy δ (↑) < 1.25	$< 1.25^2$	$< 1.25^3$
De+Sil	0.089	0.122	0.045	0.926	0.994	0.999
EnDe+Sil	0.061	0.083	0.032	0.984	0.999	1.000
De+L_2	0.068	0.093	0.036	0.978	0.998	1.000
EnDe+L_2	**0.053**	**0.071**	**0.028**	**0.992**	**0.999**	**1.000**

Table 3. Comparison of finetuning strategies using different loss functions and components on Nutrition5K [49] dataset. De is trained for Decoder only, and EnDe is training for both Encoder and Decoder. Sil is Scale-Invariant Logarithmic Loss, and L_2 is MSE loss.

Method	Error (↓) rel	sq rel	rms	Accuracy δ (↑) < 1.25	$< 1.25^2$	$< 1.25^3$
Make3D [43]	0.349	-	1.214	0.447	0.745	0.897
Li *et al.* [27]	0.232	0.094	0.821	0.621	0.886	0.968
MS-CRF [56]	0.121	0.052	0.586	0.811	0.954	0.987
DORN [14]	0.115	0.051	0.509	0.828	0.965	0.992
Zhou *et al.*	0.208	0.086	0.712	0.674	0.900	0.968
PoseNet	0.283	0.122	0.867	0.567	0.818	0.912
PoseFlow	0.221	0.091	0.764	0.659	0.833	0.939
Zhao *et al.*	0.201	0.085	0.708	0.687	0.903	0.968
Ours	0.217	0.093	0.776	0.648	0.880	0.960

Table 4. Comparison of monocular depth estimation methods on the NYUv2 [46] dataset.

.3. RGB-D Feature Fusion Module

Target	Ori. PMAE (%)	Rep. PMAE (%)	Rep. MAE
Calories	15.0	15.116	38.841
Mass	10.8	11.198	22.414
Fat	23.5	**23.416**	3.020
Carbohydrate	22.4	**21.362**	4.230
Protein	21.0	23.035	4.049
Average	**18.5**	18.8	–

Table 5. Comparison of PMAE and MAE between original (Ori.) and reproduced (Rep.) results.

.4. More tables and figures

Method	Error (↓) AbsRel	SqRel	RMS	RMSlog	Accuracy, δ (↑) < 1.25	$< 1.25^2$	$< 1.25^3$
Zhou *et al.* [35]	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Mahjourian *et al.* [62]	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Geonet [57]	0.155	1.296	5.857	0.233	0.793	0.931	0.973
DDVO [51]	0.151	1.257	5.583	0.228	0.810	0.936	0.974
DF-Net [64]	0.150	1.124	5.507	0.223	0.806	0.933	0.973
CC [39]	0.140	1.070	5.326	0.217	0.826	0.941	0.975
EPC++ [34]	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2Depth [4]	0.141	1.026	5.291	0.215	0.816	0.945	0.979
GLNet [6]	0.135	1.070	5.230	0.210	0.841	0.948	0.980
SC-SfMLearner [2]	0.137	1.089	5.439	0.217	0.830	0.942	0.975
Gordon *et al.* [19]	0.128	0.959	5.230	0.212	0.845	0.947	0.976
Monodepth2-Resnet18 [18]	0.132	1.044	5.142	0.210	0.845	0.948	0.977
Monodepth2-Resnet50 [18]	0.131	1.023	5.064	0.206	0.849	0.951	0.979
R-MSFM3 [63]	0.128	0.965	5.019	0.207	0.853	0.951	0.977
R-MSFM6 [63]	0.126	0.944	4.981	0.204	0.857	0.952	0.978
Zhao *et al.* [60]	0.130	0.893	5.062	0.205	0.832	0.949	0.981
Lite-Mono [59]	**0.121**	0.876	4.918	**0.199**	**0.859**	0.953	0.980
Ours (2nd Stage)	0.128	**0.843**	**4.875**	0.201	0.840	**0.953**	**0.982**

Table 6. Comparison of monocular depth estimation methods on KITTI [16] dataset. All models in this table are not pretrained.

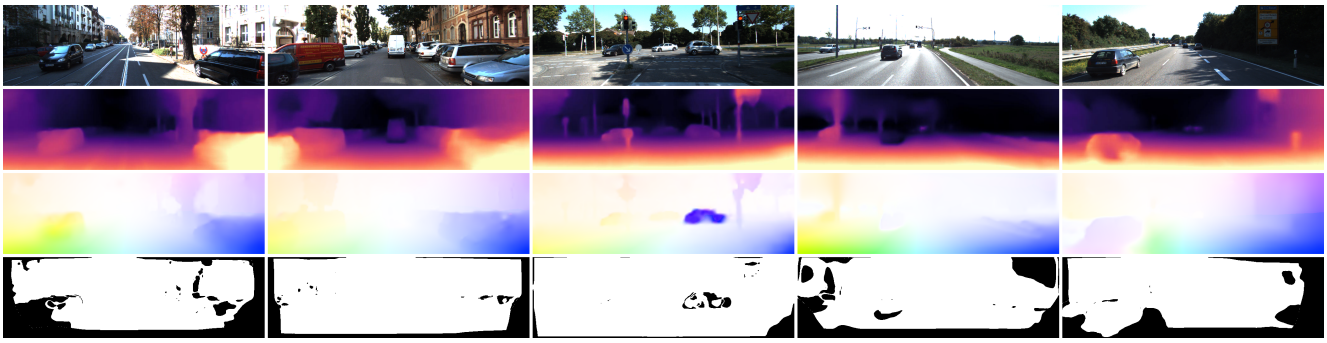

Figure 1. Qualitative visualization of the depth model pretained on KITTI. The first row is the original image, the second row is the depth map, the third row is the flow map, and the last row is the occlusion mask.

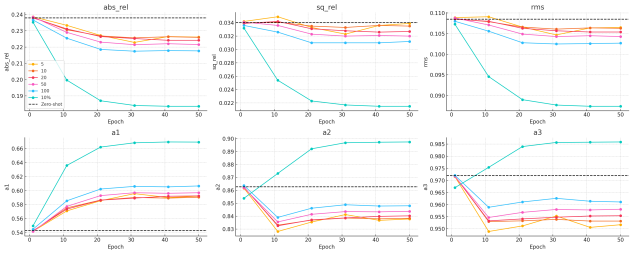

Figure 2. Few-shot finetune evaluation results of our depth model with samples = 0, 5, 10, 20, 50, 100, 300 during the training. 300 is the 10% of the nutrition dataset we used to finetune

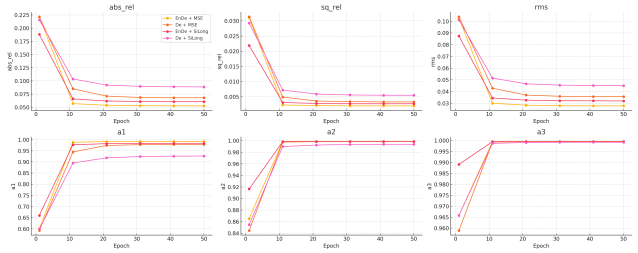

Figure 3. Evaluation results of Four different strategies for fine-tuning the depth model during the training.

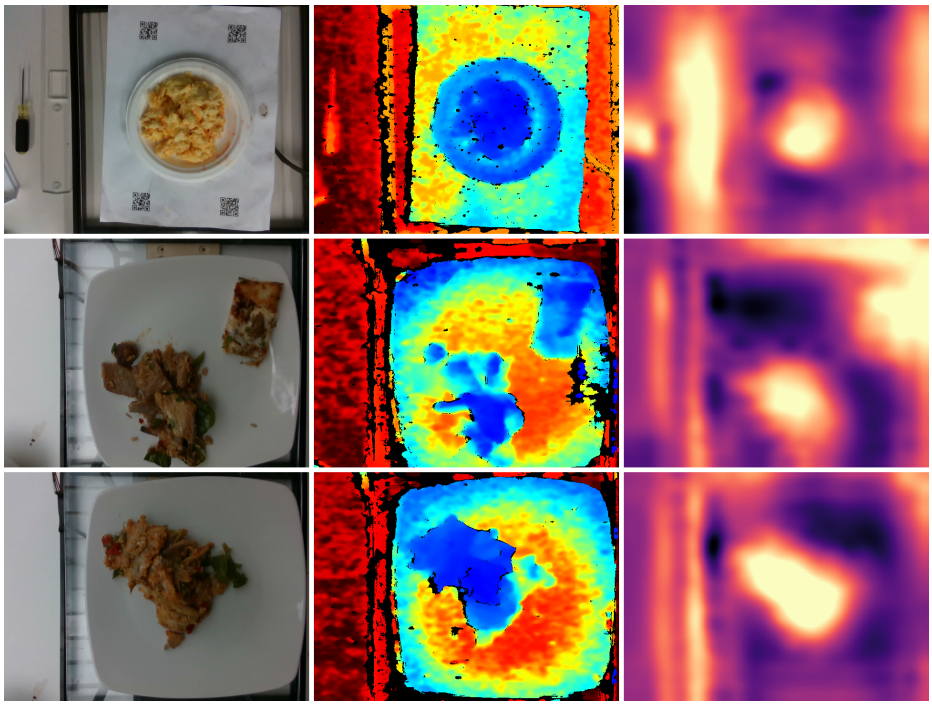

Figure 4. Qualitative visualization of the depth model on Nutrition5K dataset

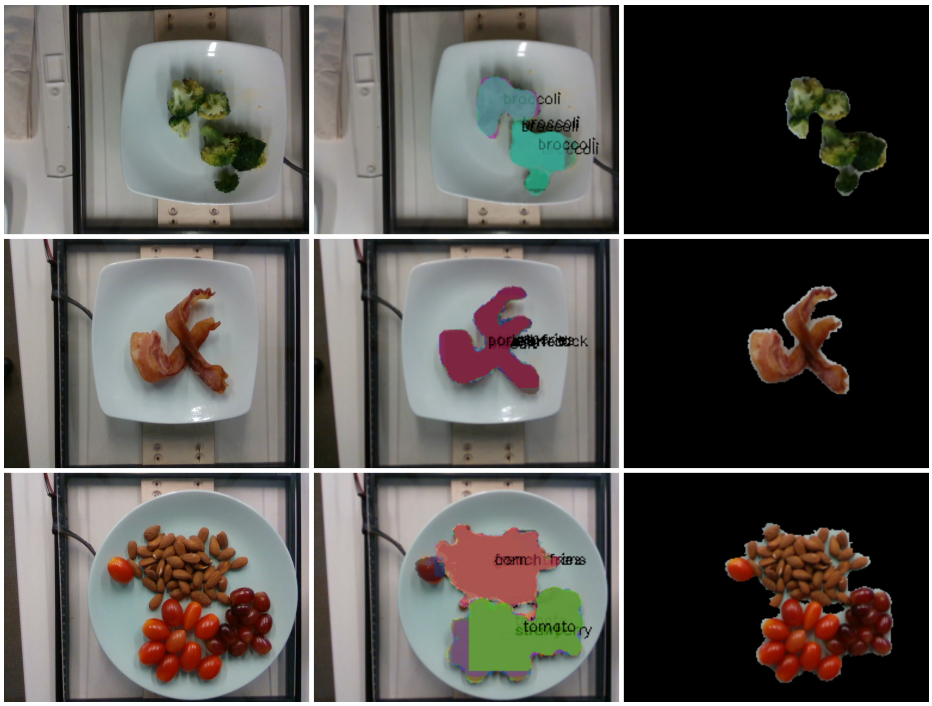

Figure 5. Qualitative visualization of the segmentation model on Nutrition5K dataset

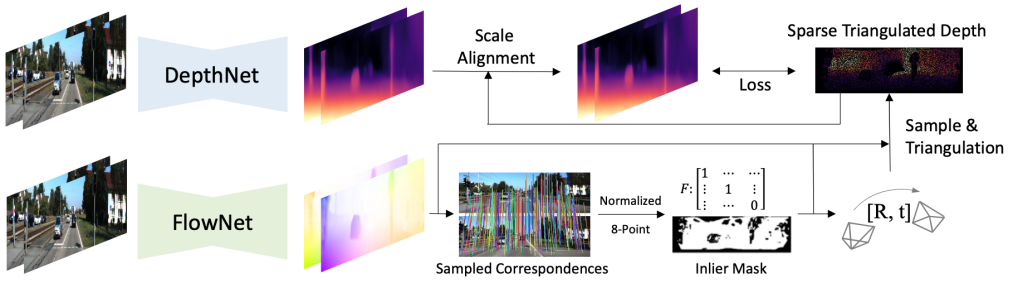

Figure 6. Overall framework of flownet-based model

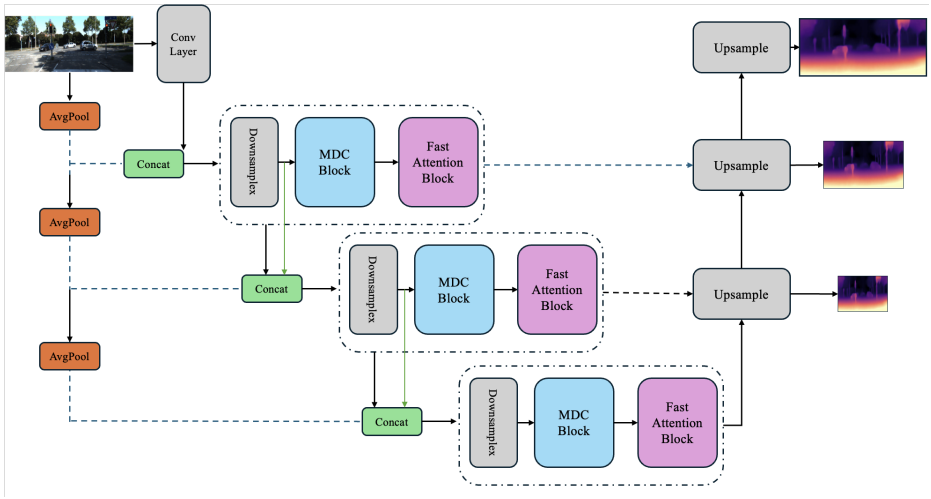

Figure 7. The pipeline of the light-weight depth net

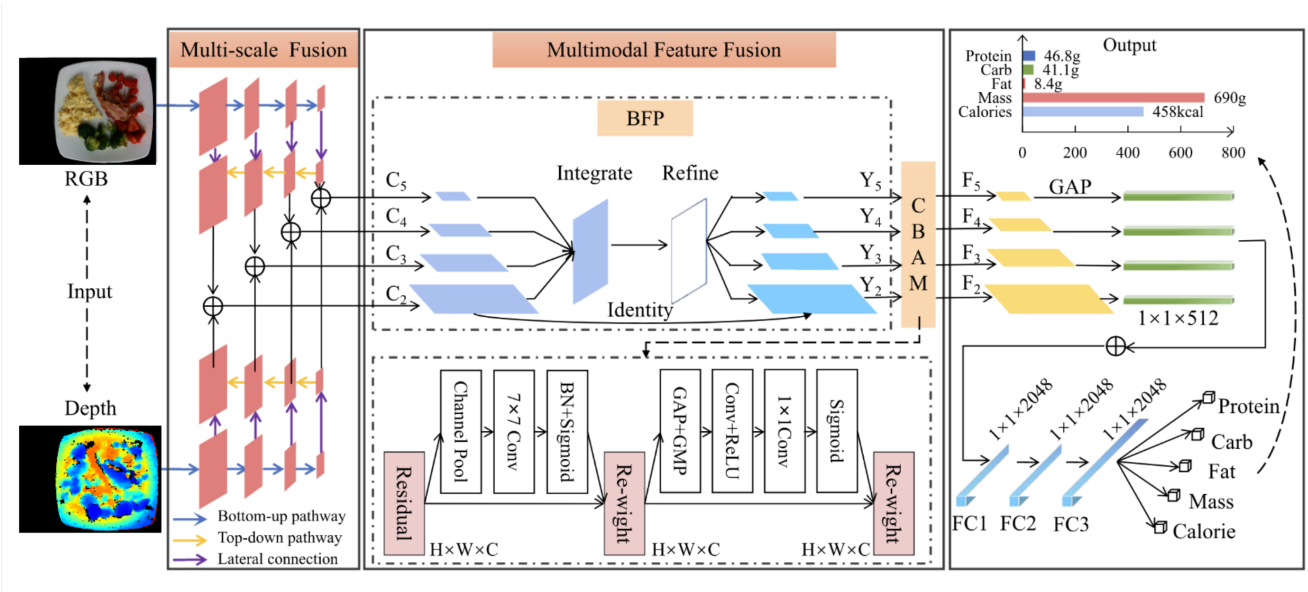

Figure 8. The pipeline of the nutrition fusion model

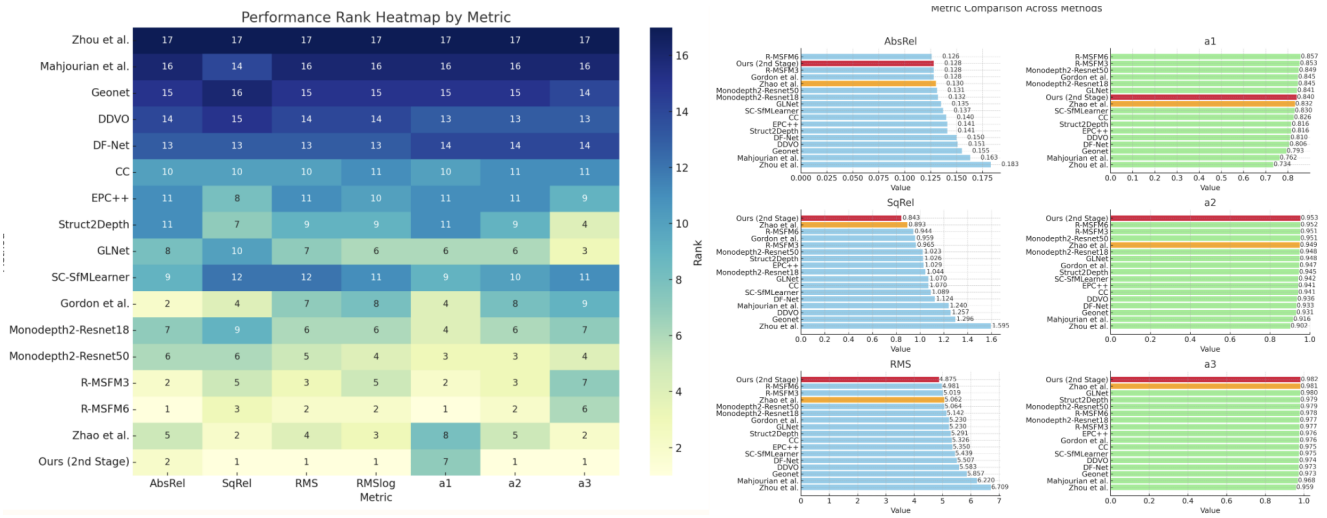

Figure 9. Comparison between baselines and depth model

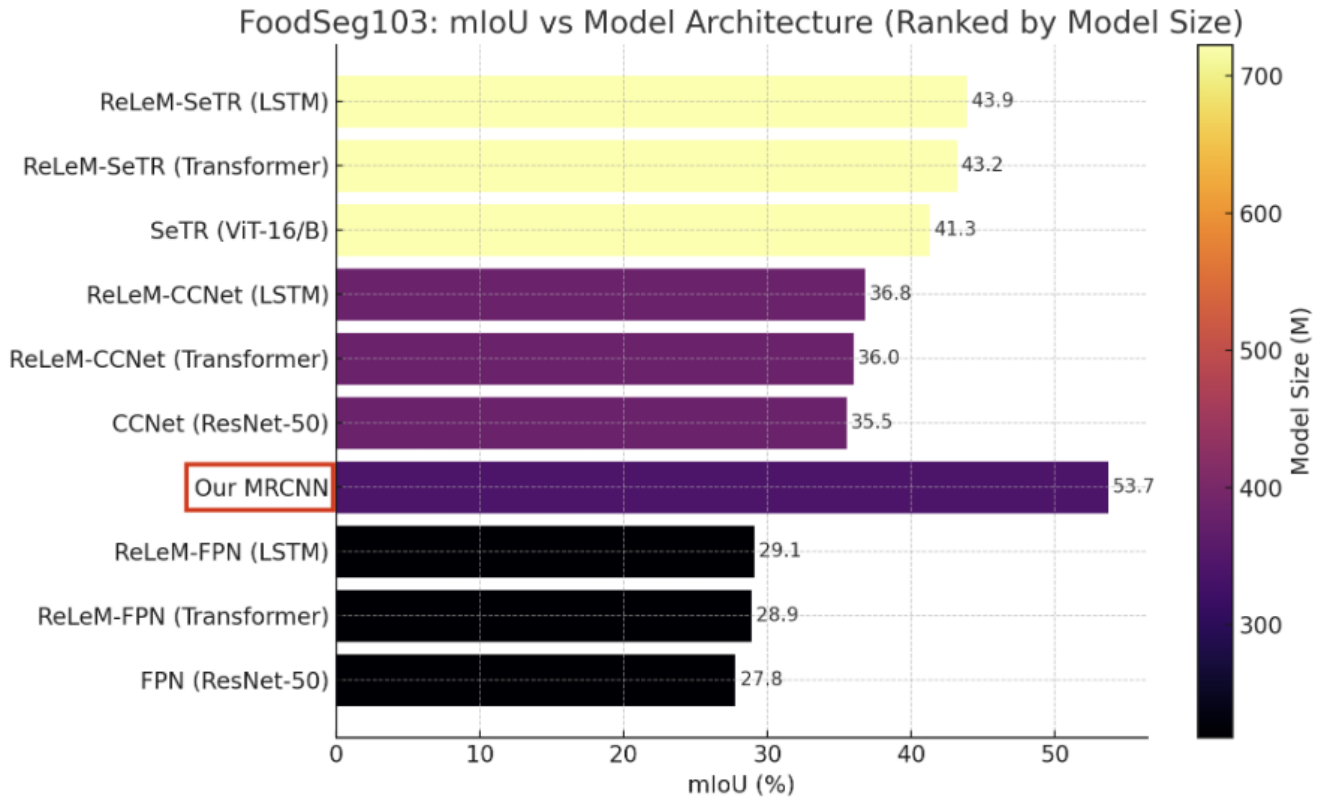

Figure 10. Comparison between baselines and segmentation model

Metric	2D	Simple Fusion	Ours
PMAE (%)			
Calories	26.1	18.8	**15.116**
Mass	18.8	18.9	**11.198**
Fat	34.2	**18.1**	23.416
Carbohydrate	31.9	23.8	**21.362**
Protein	29.5	**20.9**	23.035
MAE			
Calories	70.6	47.6	**38.841**
Mass	40.4	40.7	**22.414**
Fat	5.0	**2.27**	3.020
Carbohydrate	6.1	4.6	**4.230**
Protein	5.5	**3.7**	4.049

Table 7. Comparison of five nutrition metrics between Nutrition5k baselines (2D Direct Prediction and Depth as 4th Channel) and our method. Metrics include PMAE (%) and MAE.

Model	**Optimizer**	**Learning rate**	**GPUs**	**Epochs**	**Batchsize**	**Dataset**	**Training Time**	**Inference Time**
Masked R-CNN (Fin.)	SGD	5.00E-03	2x 3090	30	2*8	FoodSeg103	1h30mins	0.1236s
Depth Estimator (Pre. - 1st Stage)	AdamW	1.00E-04	4x 3090	20	4*32	NYUv2	24h	/
Depth Estimator (Pre. - 2nd Stage)	AdamW	1.00E-04	4x 3090	40	4*8	NYUv2	22h	/
Depth Estimator (Fin.)	AdamW	1.00E-04	2x 3090	50	2*128	Nutrition5K	10mins	0.3493s
Nutrition Fusion network (Fin.)	Adam	5.00E-05	2x A6000	300	2*8	Nutrition5K	24h	0.3829s

Table 8. Model's training settings and inference time. Pre. stands for pretrained, and Fin. stands for Finetune

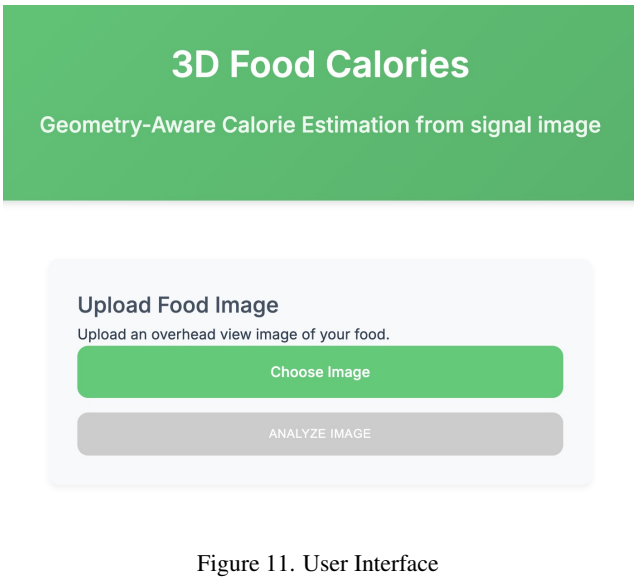

Figure 11. User Interface

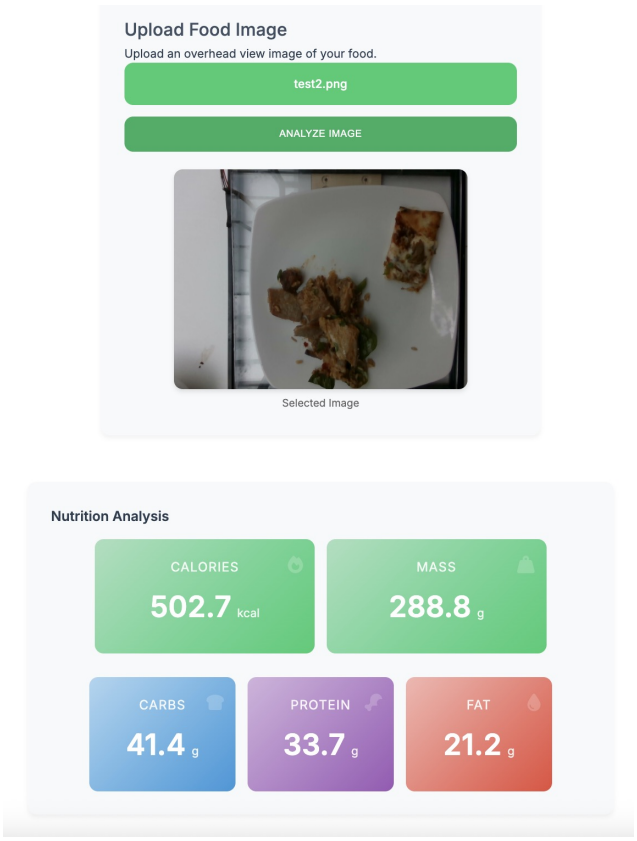

Figure 12. User Interface

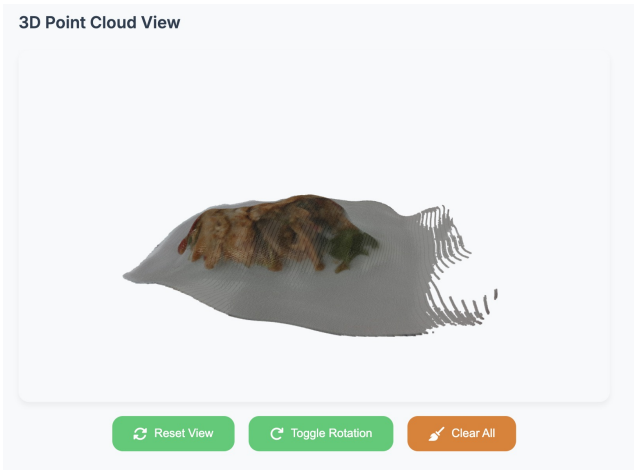

Figure 13. 3D point cloud

Figure 14. 3D point cloud

.5. Contributions Statement

1. We fine-tuned a Mask R-CNN model on the FoodSeg103 dataset to improve food segmentation performance.
2. We propose a novel depth estimation approach by integrating optical flow pathways with depth prediction, trained in a self-supervised manner.
3. We systematically integrated segmentation and depth estimation outputs into the nutrition prediction pipeline, enabling fast and accurate inference from a single RGB image.
4. We constructed a new training set using our predicted depth maps and refined segmentation masks on the Nutrition5k dataset, which significantly improved nutrition estimation in real-world scenarios.
5. We developed a lightweight system demo to support practical use and showcase the full pipeline.