NutritionVerse: Empirical Study of Various Dietary Intake Estimation Approaches

Chi-en Amy Tai[1], Matthew Keller[1], Saeejith Nair[1], Yuhao Chen*[1], Yifan Wu[1], Olivia Markham[1], Krish Parmar[1], Pengcheng Xi[2], Heather Keller[1], Sharon Kirkpatrick[1], Alexander Wong[1]

[1] University of Waterloo

[2] National Research Council Canada

MADiMa 2023

Introduction

- Accurate dietary intake estimation is critical to support healthy eating
- Automated nutrition tracking requires a large, comprehensive dataset with diverse viewpoints, modalities, and food annotations
- Existing food image datasets don't satisfy these requirements, but *what if we could generate such a perfect dataset?*

Existing Datasets Are Limited

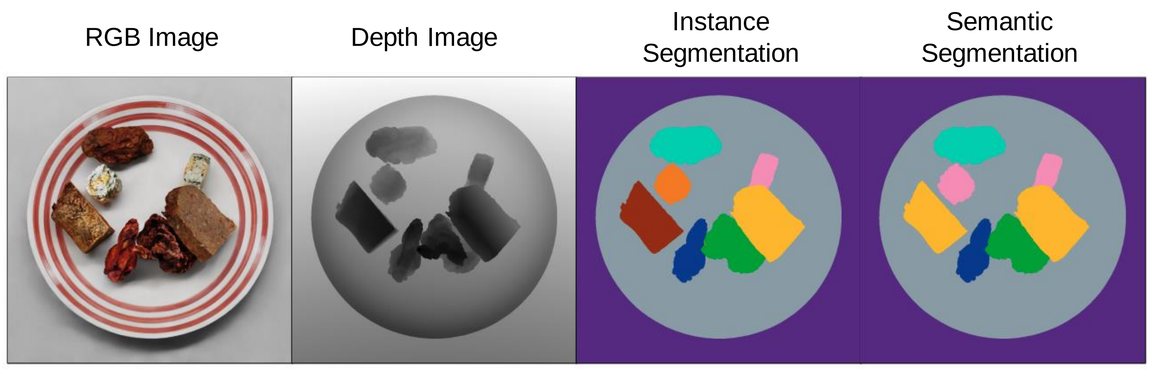
- Comparison of existing dietary intake estimation datasets to ours.
- Mixed refers to whether multiple food item types are present in an image, and CL refers to calories, M to mass, P to protein, F to fat, and CB to carbohydrate.

Work	Public	Data							Dietary Info				
		# Img	# Items	Real	Mixed	# Angles	Depth	Annotation Masks	CL	M	P	F	CB
DepthCalorieCam	✓	18	3	Y	N	1			✓				
Menu-Match	✓	646	41	Y	Y	1			✓				
Im2Calories	✓	50,374	201	Y	Y	1			✓				
Computer vision-based food calorie estimation	✓	2,978	160	Y	N	2			✓	✓			
Nutrition5k	✓	5,006	555	Y	Y	4	✓		✓	✓	✓	✓	✓
Measuring Calorie and Nutrition from Food Image		3000	8	Y	Y	2		✓	✓	✓			
NV-Real (Ours)	✓	889	45	Y	Y	4		✓	✓	✓	✓	✓	✓
NV-Synth (Ours)	✓	84,984	45	N	Y	12	✓	✓	✓	✓	✓	✓	✓

NutritionVerse Datasets

NV-Synth

- Dataset with 84k+ synthetically generated food images and associated dietary information and multimodal annotations

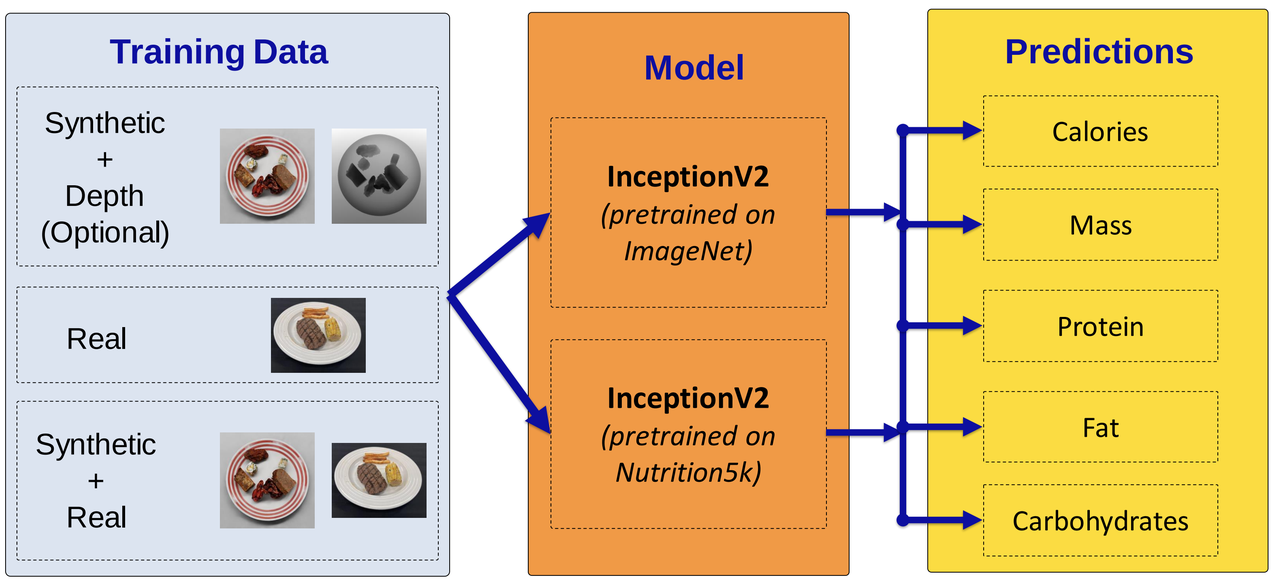


NV-Real

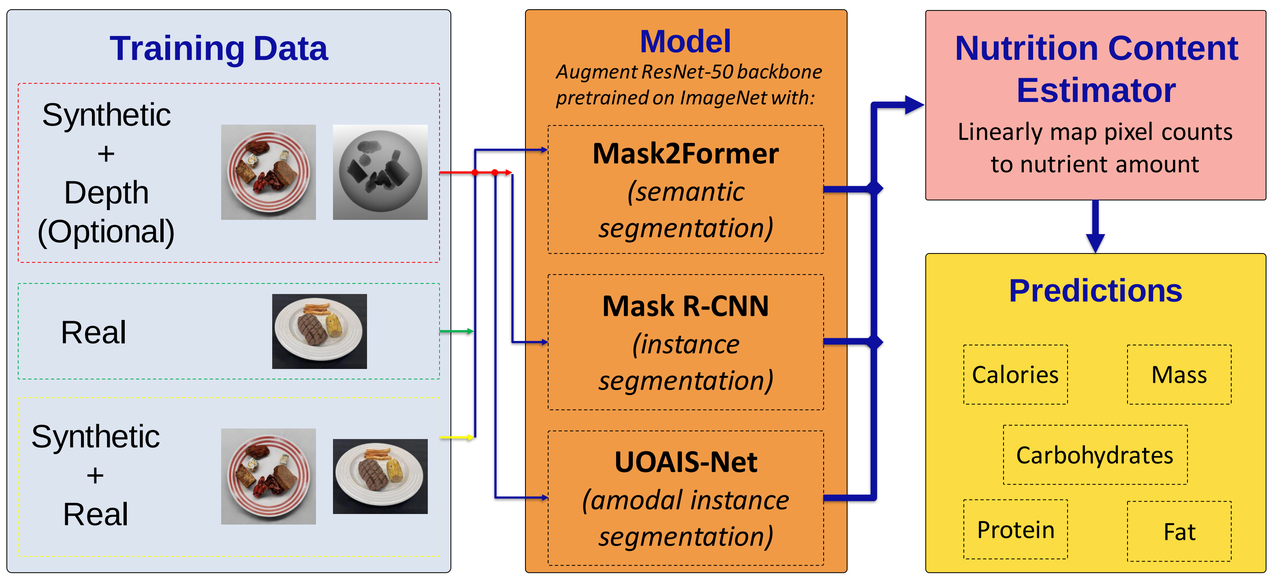
- Dataset with 889 real food images across 251 distinct dishes along with human annotated instance segmentation masks.

Methodology

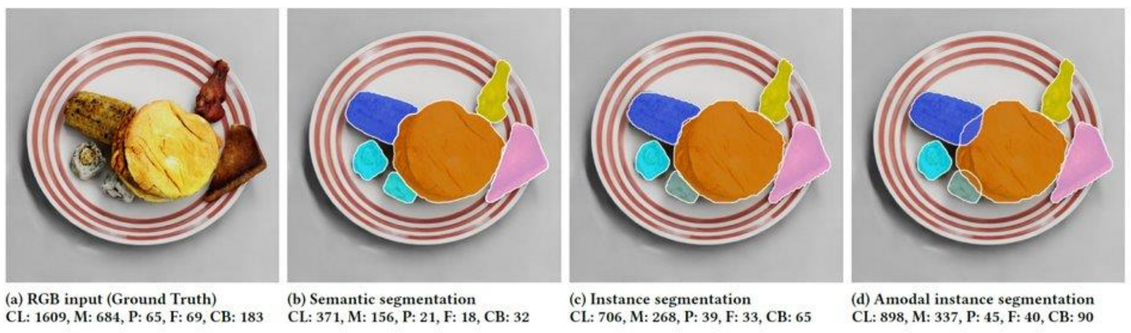
Direct Prediction



Indirect Prediction



Experimental Results



(a) RGB input (Ground Truth) CL: 1609, M: 684, P: 65, F: 69, CB: 183 — (b) Semantic segmentation CL: 371, M: 156, P: 21, F: 18, CB: 32 — (c) Instance segmentation CL: 706, M: 268, P: 39, F: 33, CB: 65 — (d) Amodal instance segmentation CL: 898, M: 337, P: 45, F: 40, CB: 90

Figure 8: Segmentation and prediction results of models trained with RGB input where CL refers to calories, M to mass, P to protein, F to fat, and CB to carbohydrate.

What is the best approach for dietary assessment?

Model (RGB)	Eval Dataset	Calories MAE	Mass MAE	Protein MAE	Fat MAE	Carb MAE
Semantic	NV-Synth	418.1	185.4	39.0	23.5	32.3
Instance	NV-Synth	430.9	191.4	39.3	24.1	34.4
Amodal Instance	NV-Synth	451.3	202.8	39.6	24.8	38.5
Direct Prediction (ImageNet)	NV-Synth	229.2	102.6	56.0	12.0	**19.4***
Direct Prediction (Nutrition5k)	NV-Synth	**128.7***	**77.2***	**18.5***	**9.1***	21.5

Table 3: Evaluation of model architectures using NV-Synth (RGB images) with the lowest MAE value for each column bolded with an * next to it.

- When given perfect labels from simulation, direct prediction gives the best nutrition estimation

Does depth information improve model performance?

Model (RGBD)	Eval Dataset	Calories MAE	Mass MAE	Protein MAE	Fat MAE	Carb MAE
Semantic	NV-Synth	418.3	185.3	39.0	23.5	32.2
Instance	NV-Synth	432.9	194.1	39.1	24.1	35.2
Amodal Instance	NV-Synth	462.0	208.1	39.7	25.2	40.4
Direct Prediction (ImageNet)	NV-Synth	371.7	317.6	34.8	**19.2***	**25.2***
Direct Prediction (Nutrition5k)	NV-Synth	**202.0***	**78.8***	**23.5***	30.1	33.3

Table 4: Investigation of depth information using NV-Synth (RGBD images) with the lowest MAE value for each column bolded with an * next to it.

- The addition of depth information results in a worse performance for direct prediction models
- For indirect methods, depth does not make any significant differences

What is the impact of using synthetic data?

Model Description	Trained	Fine-Tuned	Calories MAE	Mass MAE	Protein MAE	Fat MAE	Carb MAE
(A) Direct Prediction (Nutrition5k)	NV-Synth	N/A	525.9	188.4	39.1	27.4	54.6
(B) Direct Prediction (ImageNet)	NV-Synth	NV-Real	**229.8***	**63.3***	**24.6***	**13.5***	70.8
(C) Semantic	NV-Real	N/A	442.7	221.0	40.1	23.0	**42.4***

Table 8: Comparison of the best model from the three scenarios evaluated on the NV-Real dataset, with the lowest MAE value for each column bolded with an * next to it.

- The best model is Direct Prediction (ImageNet) model trained on the NV-Synth train set and fine-tuned on the NV-Real train set

Acknowledgements

This work was supported by the National Research Council Canada (NRC) through the Aging in Place (AiP) Challenge Program, project number AiP-006. The authors also thank the graduate student partner in the Kinesiology and Health Sciences department Meagan Jackson and undergraduate research assistants Tanisha Nigam, Komal Vachhani, and Cosmo Zhao.