# An Improved Encoder-Decoder Framework for Food Energy Estimation

*Jack Ma, Jiangpeng He, and Fengqing Zhu*
**Elmore Family School of Electrical and Computer Engineering,**
**Purdue University, USA**

MADiMa 2023

VIPER Video and Image Processing Laboratory

## Introduction

- Dietary assessment involves evaluating someone's nutritional intake (e.g. food energy consumption), towards identifying nutrient deficiencies and reducing metabolic disorder risks

**A. Underestimate**
Ground truth: 514
Estimated: 427



**C. Accurate Estimate**
Ground truth: 494
Estimated: 506



**E. Overestimate**
Ground truth: 547
Estimated: 780

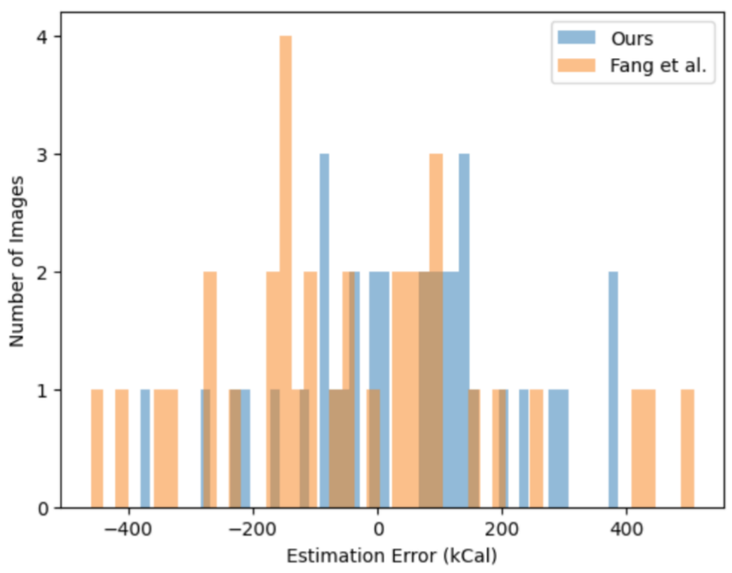
- Our work lies in the realm of automatic food energy estimation from a single monocular image due to its simplicity for the user

- Recent methods (e.g. [1]) employ an encoder-decoder model, where the encoder transforms the hard-to-extract image into a grayscale representing per-pixel caloric density that a regression decoder can extract calories from easier, but these methods have the following drawbacks:

  - ❑Inherent limitations with grayscale due to insufficient storage capacity and granularity since values are confined to integers from 0 to 255

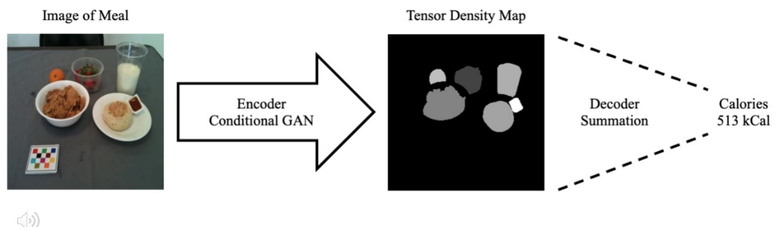
  - ❑Regression decoders add complexity to the pipeline and don't improve performance (as seen in results)

- In this work, we propose an improved encoder-decoder model that addresses the limitations of prior work and improves upon existing methods by over 10% MAPE and 30 kCal MAE




Distribution of the energy estimation errors for our method in comparison to the next best method

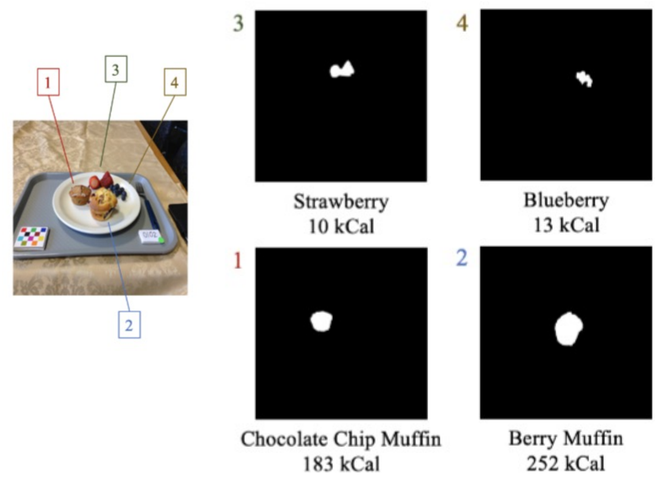
## Our Method




We employ an encoder-decoder model for caloric estimation. Encoder aims to embed caloric information into a per-pixel caloric density map for corresponding pixels in image. Decoder extracts caloric information from the encoded representation.

### Dataset

- Compile and prune different datasets to obtain 175 images with associated seg mask/calorie value for each food item in image




### Density Map Generation

- For each image, generate tensor $d_i$ from seg mask $s_i$ and calorie value $c_i$ for each food item $i$ in image having dimensions $h$ by $w$ ($w_i$ is number of foreground pixels in seg mask $s_i$):

$$d_i[x,y] = \begin{cases} 0 & s_i[x,y] = 0 \\ c_i/w_i & \text{otherwise} \end{cases}$$

- For all 1
- Obtain tensor density map $d$ by concatenating all $d_i$
- Note that the sum of all pixel values in $d$ is precisely $c = \sum c_i$

### Encoder

- Train Conditional GAN (cGAN) to learn mapping between images and their tensor density maps. The trained cGAN then generates density maps that decoder can extract calories from

### Decoder

- Employ simple summation decoder that obtains calorie value $c$ by adding up all values in tensor density map $d'$ generated by the encoder. We can directly use this decoder, unlike the regression decoders in previous work that require training

$$c = \sum_{x=1}^{h} \sum_{y=1}^{w} d'[x,y]$$

## Experimental Results

➢ **Comparison of methods in terms of mean absolute error (MAE) and mean absolute percent error (MAPE).**

| | MAE (kCal) | MAPE (%) |
|---|---|---|
| Grayscale [1] | 183.5 | 48.5 |
| Image Only [2] | 287.7 | 61.2 |
| Density Map + Image, LN + GN [2] | 219.1 | 54.9 |
| Density Map + Image, LN [2] | 208.4 | 58.3 |
| **Ours** | **150.5** | **35.7** |

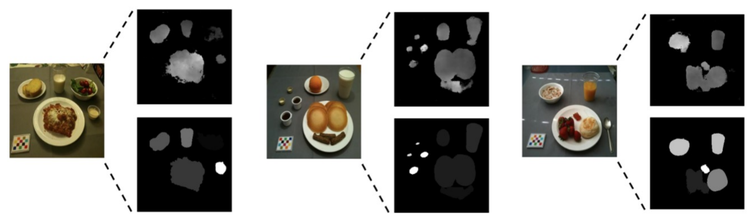
➢ **Our summation decoder compared against different regression decoders employing VGG16, Resnet18, and Resnet50**

| | Pretrained | MAE (kCal) | MAPE (%) |
|---|---|---|---|
| VGG16 | ✓ | 166.3 | 38.5 |
| VGG16 | ✗ | 155.5 | 37.9 |
| Resnet18 | ✓ | 231.8 | 54.7 |
| Resnet18 | ✗ | **149.3** | 35.4 |
| Resnet50 | ✓ | 173.3 | 37.52 |
| Resnet50 | ✗ | 154.0 | **34.5** |
| **Ours** | N/A | 150.5 | 35.7 |

➢ **Comparison between the tensor density map and grayscale as the encoded representation. Experiments are run using the pipeline from [1] because it achieves the best performance out of all methods using a grayscale.**

| | MAE (kCal) | MAPE (%) |
|---|---|---|
| **Tensor Density Map (Ours)** | **166.3** | **38.5** |
| Grayscale | 183.5 | 48.5 |

➢ **Exemplary tensor density maps generated by the cGAN encoder (top right), along with the ground truth tensor density map (bottom right) and original image (left) for comparison**



## Conclusion and Future Work

- We proposed an improved encoder-decoder framework for food energy estimation from a single monocular image, which achieves big performance improvements over existing work
- Our future work will focus on improving the decoding portion of the pipeline because the decoder employed in our method is simple and still has a large room for more complexity

[1] Fang et al., An End-to-End Image-Based Automatic Food Energy Estimation Technique Based on Learned Energy Distribution Images: Protocol and Methodology, *Nutrients* 2019
[2] Shao et al., Towards Learning Food Portion From Monocular Images With Cross-Domain Feature Adaptation, *MMSP* Workshop 2021