Estimating Food Calories from RGB and Depth Images using Deep Neural Networks

Ziqi Wen and Haoyu Wang
The University of Melbourne
Email: {ziqi.wen, haowang3}@student.unimelb.edu.au

I. Introduction

Estimating the caloric content of food from images poses a significant challenge for machine learning systems. This task requires a model to not only recognize objects and their semantic categories related to caloric content but also to infer the three-dimensional structure of the meal to estimate its volume. Since the caloric value of food is strongly correlated with both its type and physical volume, effective estimation demands a joint understanding of appearance and geometry. Accurate calorie prediction has broad applications in healthcare, particularly in facilitating dietary monitoring and personalized nutrition management.

In this project, we employ the *Nutrition5K* dataset [1], which provides paired RGB and depth images of meals along with ground-truth calorie labels. Using this dataset, our goal is to build models that can accurately estimate meal calories by learning both visual and geometric cues directly from image data. To address this problem, we design and train four neural

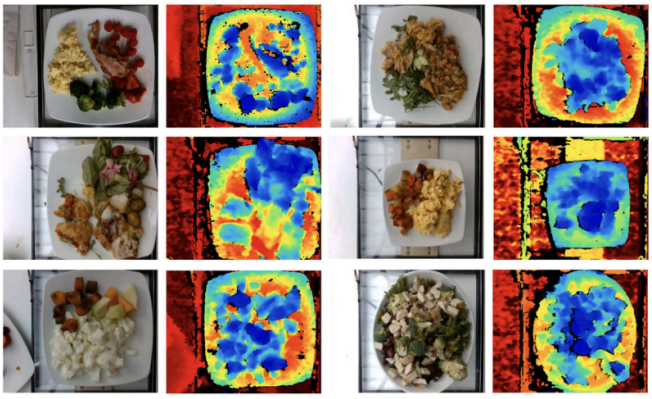


Fig. 1: Sample images from the *Nutrition5K* dataset. Each pair shows the RGB image (left) and its corresponding depth map (right). The dataset captures diverse dishes with varying shapes, volumes, and color compositions, providing both visual and geometric cues crucial for calorie prediction.

network architectures of increasing complexity, each trained from scratch on the *Nutrition5K* training set. We begin with a **baseline single-stream CNN** that operates solely on RGB inputs and serves as a fundamental benchmark for visual feature learning. Building upon this foundation, we introduce three **fusion-based architectures** that process RGB and depth modalities in parallel through independent convolutional pathways before merging their learned representations. Such multimodal fusion enables the model to leverage both color and geometric cues, yielding richer and more discriminative feature representations for calorie prediction.

The first fusion model adopts a **late-fusion** strategy, where RGB and depth features are concatenated only at the final convolutional stage before regression. This approach provides a simple yet effective means of integrating modalities while maintaining architectural simplicity.

Next, we implement two enhanced variants: a **fusion-based RGB-D network** (hereafter referred to as **RGB-D FusionNet**, following the design of Shao *et al.* [2]) and a cross-modality attention-augmented **RGB-D FusionCAB** [3]. These models address a key limitation of the late-fusion approach, whose deeper layers tend to emphasize high-level semantic information (*"what"*) while neglecting fine-grained spatial structure (*"where"*). By jointly learning from RGB and depth features throughout the network, our fusion-based models retain both semantic identity and volumetric structure, leading to more accurate and robust calorie estimation.

Finally, we explore **model ensembling** techniques based on our best-performing architecture, **RGB-D FusionCAB**, to further enhance predictive performance. Through a comprehensive series of experiments, we demonstrate that the proposed fusion-based architectures consistently outperform the single-stream baseline in calorie prediction accuracy.

II. Methodology

A. ResNet-18 with RGB as Single Input

We begin with a single-modality baseline that takes only RGB images as input. We reimplement a standard **ResNet-18** as our backbone network, which consists of four stages with output channels $[64, 128, 256, 512]$. Each stage contains two residual blocks, and each block includes two 3×3 convolutional layers followed by batch normalization and ReLU activation. The identity skip connections across blocks help mitigate vanishing gradients and enable stable training.

In our later multimodal designs, this ResNet module (`MyResNet`) will serve as a reusable building block. However, as briefly discussed in the introduction, this baseline model processes only RGB information, and thus it can capture appearance-based correlations between visual cues and calories but fails to model 3D volumetric relationships or cross-

modal interactions that are often crucial for estimating food quantity and density.

B. Late-Fusion Based Model: First Attempt

To incorporate geometric information, we next design a **late-fusion RGB-D model** that processes both RGB and depth modalities using separate convolutional encoders. Each encoder produces a high-level feature representation from the final convolutional layer, capturing semantic abstraction within its respective modality. These representations are then concatenated and passed through additional fully connected layers for regression.

This approach enables the model to jointly leverage color and geometric cues, allowing for more informative calorie estimation than the single-stream baseline. However, since the fusion occurs only at the final stage, the model cannot effectively capture fine-grained spatial correspondences between modalities. In other words, it mainly answers the *"what"* question (semantic identity) while neglecting the *"where"* aspect (spatial and volumetric localization), which is critical for accurate calorie estimation. To address this limitation, we re-implement the more sophisticated **RGB-D FusionNet** and **RGB-D FusionCAB** described below.

C. RGB-D FusionNet

The original RGB-D FusionNet proposed by Shao *et al.* [2] employs two pretrained ResNet-101 backbones to extract hierarchical features from RGB and depth modalities, respectively. In our implementation, we adopt a lightweight custom ResNet-18 backbone (`MyResNet`) for each stream to reduce memory and computational cost while maintaining sufficient representational capacity. Each residual stage outputs a feature map, forming a feature pyramid $\{C_2, C_3, C_4, C_5\}$ that encodes progressively deeper semantic representations.

a) Feature Pyramid Network (FPN).: We implement an **FPN** module to merge hierarchical representations across scales. A 1×1 lateral convolution first aligns the channel dimensions of each feature map, and a top-down pathway then upsamples high-level semantic features and merges them with lower-level spatial features via element-wise addition. A subsequent 3×3 convolution refines each fused feature map. The FPN facilitates the integration of both high-level contextual cues and low-level spatial details, enabling the network to reason jointly about food shape, structure, and texture, which are attributes directly related to portion size and thus calorie estimation.

b) Balanced Feature Pyramid (BFP).: Next, the **BFP** module aggregates information across multiple pyramid levels by resizing all feature maps to a uniform spatial resolution and averaging them into a global semantic representation. This representation is then refined through a global self-attention mechanism, allowing the network to model long-range dependencies and contextual relations among different regions. The refined global features are subsequently upsampled and redistributed back to each pyramid level to enhance local representations with holistic context. By combining global context with local multi-scale features, BFP improves spatial coherence and object-level understanding, enabling the network to better capture continuous food boundaries and volumetric consistency.

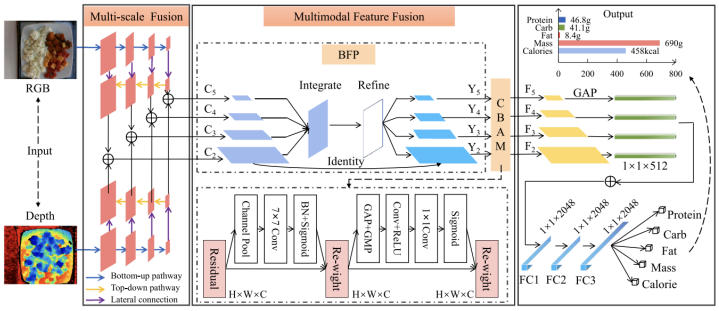


Fig. 2: Overview of the proposed **RGB-D FusionNet** architecture. The network consists of three major components: (1) *Multi-scale Fusion* for hierarchical feature extraction from RGB and depth inputs, (2) *Multimodal Feature Fusion* using BFP and CBAM to integrate and refine cross-modal representations, and (3) a regression head that outputs nutritional predictions. FPN merges features across scales, BFP integrates global contextual information, and CBAM adaptively emphasizes salient channels and spatial regions, enabling robust calorie estimation.

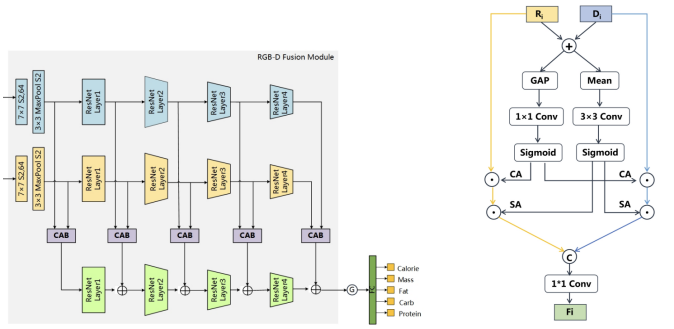
c) Convolutional Block Attention Module (CBAM).: We further refine the fused features using a **CBAM**, which sequentially applies channel and spatial attention mechanisms. The channel attention module employs global average and max pooling followed by a shared multilayer perceptron (MLP) to adaptively reweight feature channels according to their importance. The spatial attention module then applies a 7×7 convolution to concatenated average- and max-pooled feature maps to highlight salient spatial regions. By jointly emphasizing the most informative channels and spatial locations, CBAM enables the network to focus on calorie-relevant visual cues such as dense versus sparse food regions and texture variations.

d) Regression Head.: After multimodal fusion and refinement, the resulting feature maps are globally average-pooled and passed through a fully connected regression head to predict calorie values. In contrast to the original RGB-D FusionNet, which predicted five nutritional attributes (calories, mass, fat, carbohydrate, and protein) simultaneously, our implementation predicts solely on calorie estimation.

D. RGB-D FusionCAB

The **RGB-D FusionCAB** architecture replaces the multi-stage fusion modules (FPN, BFP, and CBAM) with a more compact **Cross-Modality Attention Block (CAB)** that directly extracts and integrates complementary features from RGB and depth streams.

The **CAB** operates by first aggregating RGB and depth feature maps through element-wise addition to form a joint repre-



(a) Overall architecture of the proposed **RGB-D FusionCAB** network.

(b) Internal structure of the **CAB**.

Fig. 3: Overview of the **RGB-D FusionCAB** architecture and its Cross-Modality Attention Block (CAB).

sentation that captures shared visual patterns. Two lightweight attention mechanisms are then applied:

- **Channel attention** computes global statistics to identify the most informative feature channels shared across both modalities.
- **Spatial attention** highlights salient spatial regions that are consistent between RGB textures and depth structures.

The attention maps are used to reweight the RGB and depth features, adaptively emphasizing regions that contribute most to calorie estimation (e.g., dense food areas or elevated surfaces). Finally, the attended features from both modalities are concatenated and fused through a 1×1 convolution to produce a unified cross-modal representation.

This design enables the model to jointly reason about both the semantic content and geometric structure of food items in a direct and efficient manner, enhancing the fusion of complementary information while reducing architectural complexity compared to the multi-stage FusionNet.

III. Training Configuration and Implementation

A. Experimental Setup

All experiments were conducted on the *Nutrition5K* dataset using RGB and raw depth images (`depth_color` was used for visualization only). The dataset consists of dish IDs ranging from 0–3489, of which 0–3300 were used for training and 3301–3489 for testing. We evaluated three architectures of increasing complexity: (1) an RGB-only ResNet-18, (2) a late-fusion CNN, and (3) the RGB-D FusionNet and (4) FusionCAB.

B. Data Preprocessing

To minimize runtime overhead and avoid expensive I/O operations associated with repeatedly traversing the dataset directory, we preprocessed all samples into a memory-mapped (`mmap`) format for efficient loading. The provided *Nutrition5K* training set was further divided into a **training** (95%) and **validation** (5%) split. Corrupted or unreadable samples were identified and excluded prior to training (e.g., `dish_2368` was removed due to invalid image data).

a) Image Normalization.: RGB images were normalized to the $[0, 1]$ range by dividing pixel intensities by 255, ensuring consistent dynamic range across samples. Depth maps were linearly rescaled according to:

$$I'_{\text{depth}} = \frac{I_{\text{depth}} - I_{\min}}{I_{\max} - I_{\min}},$$

where $I_{\min} = 0.0$ and $I_{\max} = 65535.0$ represent the global minimum and maximum depth values computed over the training set. This normalization ensures that both modalities lie within comparable numerical ranges, facilitating stable optimization during multimodal fusion.

b) Label Normalization.: The calorie values in the Nutrition5K dataset exhibit a highly skewed, long-tailed distribution, where the frequency decays approximately exponentially as the calorie value increases. Such imbalance introduces instability during regression training, as the model becomes biased toward samples with lower calorie values. To mitigate this issue, we experimented with a logarithmic transformation of the labels:

$$y' = \log(1 + y),$$

which compresses the dynamic range and yields a distribution closer to Gaussian, leading to more stable optimization and faster convergence during training. However, performing regression in the log-transformed space implicitly downweights large errors in the original value space. As a result, the model produced smoother predictions but failed to align accurately with high-calorie samples, ultimately degrading test performance. Therefore, we reverted to training directly on the original calorie values using the mean squared error (MSE) loss without label normalization.

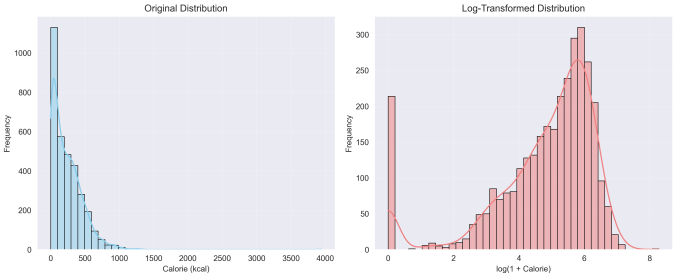


Fig. 4: Effect of label normalization on the calorie value distribution. (**Left**) The original calorie distribution is highly skewed, with frequency decreasing roughly exponentially as calorie values increase. (**Right**) After applying the logarithmic transformation $y' = \log(1+y)$, the distribution becomes more symmetric and approximately Gaussian, improving training stability but reducing sensitivity to large calorie values.

C. Training & Hyperparameters tuning

We performed a grid search over batch sizes $\{32, 64\}$ and learning rates $\{5 \times 10^{-5}, 1 \times 10^{-4}\}$, while fixing the number of epochs to 50. During training, we recorded the training loss every 10 steps and the validation loss every 50 steps for convergence monitoring.

All models were optimized using the **AdamW** optimizer with a weight decay of 10^{-4} and the mean squared error (MSE) loss function. Formally, for a batch of N samples, the loss is defined as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2,$$

where $\hat{y}_i$ denotes the model-predicted calorie value and y_i the corresponding ground-truth label. During training, we continuously tracked the model checkpoint that achieved the lowest validation MSE. After training completion, this best-performing checkpoint was reloaded and used for evaluation on the test set.

To improve generalization and mitigate overfitting, we employed dropout regularization at multiple stages of the network. A dropout rate of 0.3 was applied to the prediction head, placed before the ReLU activation layer. For the **RGB-D FusionNet**, we additionally applied spatial dropout with a rate of 0.1 between major fusion stages—specifically, after the FPN module and before entering the BFP, as well as after the BFP and before the CBAM module. This hierarchical dropout strategy was designed to promote robustness by preventing co-adaptation of spatially correlated features.

All experiments were implemented in `PyTorch` and executed on the *Spartan* HPC system, utilizing 8 CPU cores and a single NVIDIA A100 GPU.

D. Quantitative Analysis

Table I summarizes the mean squared error (MSE) achieved by each model on the training and validation splits. In theory, we expected the **RGB-D FusionNet** to significantly outperform the single-stream baselines due to its explicit integration of both appearance and geometric cues. However, the empirical results show that all three models converge to comparable MSE values on both the training and validation sets, although the RGB-D model exhibits a noticeably faster convergence rate during training.

We hypothesize that this performance similarity arises from several factors. First, the Nutrition5K dataset exhibits substantial variation in lighting conditions, plating styles, and viewing angles, which may reduce the effective contribution of depth information. Second, all models in this study were trained from scratch rather than initialized with pretrained weights. In our architectures, the majority of computation resides in the encoder, which is responsible for learning high-level representations useful for downstream tasks, while the regression head remains relatively shallow. Without pretraining on large-scale image datasets such as ImageNet, the encoder must learn both low-level visual primitives and high-level semantic representations directly from the limited dataset of roughly 3,300 samples. Consequently, the learned features may not generalize well enough to fully exploit the potential of the RGB-D design for caloric estimation.

Despite the small quantitative difference, the **RGB-D FusionCAB** demonstrates qualitatively faster and more stable convergence and achieves the best validation performance, suggesting that cross-modality attention enhances training dynamics and representation learning efficiency. Future work may benefit from initializing the backbone with pretrained weights to better leverage the complementary cues from RGB and depth modalities.

TABLE I: Training and validation MSE comparison across architectures.

Model	Train MSE	Validation MSE
ResNet-18 (RGB only)	2230.44	6795.60
Late-Fusion CNN	2391.77	6480.42
RGB-D FusionNet	1759.23	7412.24
RGB-D FusionCAB	**808.33**	**6351.32**

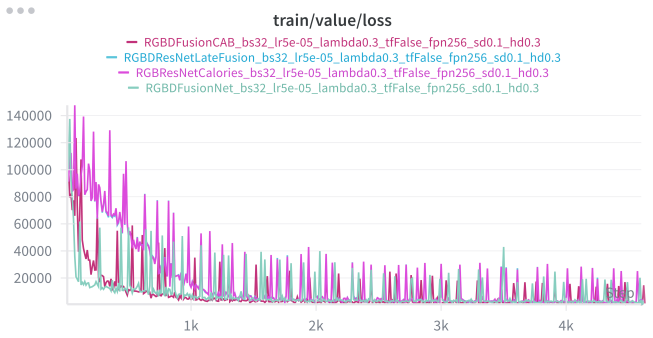


Fig. 5: Training loss curves of the three models across 50 epochs.

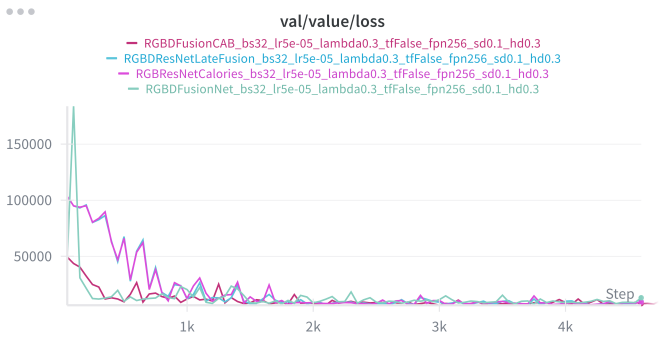


Fig. 6: Validation loss curves of the three models across 50 epochs. Although all models eventually converge to similar validation MSE values, the RGB-D FusionNet achieves faster convergence and smoother validation dynamics.

Ensembling Methods. To further improve robustness and generalization, we employed an **ensemble strategy** based on our best-performing architecture, the RGB-D FusionCAB. Specifically, three models trained with different random seeds ($\{0, 42, 1234\}$) were aggregated by averaging their predictions with equal weights. This ensemble produced our final submission and achieved the best performance on the hidden Kaggle test set, reaching a mean squared error of **5395.33**.

These results highlight that ensembling enhances stability and predictive accuracy.

REFERENCES

[1] Q. Thames, A. Karpur, W. Norris, F. Xia, L. Panait, T. Weyand, and J. Sim, "Nutrition5k: Towards automatic nutritional understanding of generic food," 2021. [Online]. Available: https://arxiv.org/abs/2103.03375

[2] W. Shao, W. Min, S. Hou, M. Luo, T. Li, Y. Zheng, and S. Jiang, "Vision-based food nutrition estimation via rgb-d fusion network," *Food Chemistry*, vol. 424, p. 136309, 2023.

[3] Y. Han, Q. Cheng, W. Wu, and Z. Huang, "Dpf-nutrition: Food nutrition estimation via depth prediction and fusion," 2023. [Online]. Available: https://arxiv.org/abs/2310.11702