A Large-Scale Benchmark for Food Image Segmentation (2021)

Paper: https://arxiv.org/pdf/2105.05409

1. Dataset Creation: FoodSeg103

The authors introduced **FoodSeg103** (and its extension, FoodSeg154), a comprehensive benchmark for food semantic segmentation. The images are sourced from the **Recipe1M** dataset, providing a bridge between visual data and detailed cooking information.

2. Methodology: ReLeM (Recipe Learning Model)

The core technical contribution is **ReLeM**, a multi-modality knowledge transfer approach.

- **How it works:** It integrates language embeddings (derived from ingredients and cooking instructions) with visual representations.
- **The Goal:** By linking visual features to common language embeddings, the model "connects" the appearance of the same ingredient across different dishes, making the feature space more cohesive.

3. Experimental Results

The authors tested ReLeM by integrating it into existing state-of-the-art (SOTA) models like **CCNet**, **Sem-FPN**, and **SeTR**.

- **Versatility:** The method is "generic," meaning it works across both CNN-based and Transformer-based backbones.
- **Efficiency:** The study found that ReLeM provides the most significant performance boosts when applied to already strong models (like CCNet), demonstrating high efficiency for advanced multimedia tasks.

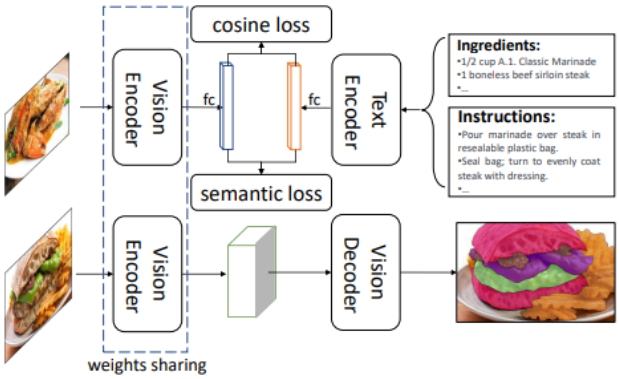

Figure 6: Our food image segmentation framework consists of two modules: Recipe Learning Module (ReLeM) and Image Segmentation Module (Segmenter). For ReLeM, we encode the recipe information into the visual representation of the food image. We deploy the cosine similarity to compute the distance between two distinct-modality models, together with a semantic loss [41]. After training, we use the trained encoder to initialize the encoder of the Segmenter. The decoder of the Segmenter is trained with the segmentation masks from a random initialization.

Text Preprocessing

To handle raw recipe data, the authors convert unstructured text into mathematical vectors:

- **Ingredients:** Cleaned of redundant words and converted into embeddings using **word2vec** and **Bi-directional LSTM** (which analyzes text in both directions).
- **Instructions:** Because cooking steps are often long, they use **skip-instructions** to prevent the model from "forgetting" information (solving the **gradient vanishing** problem) and to ensure a consistent output length.

In this paper, we conduct experiments using three representative frameworks of these three types, respectively, i.e., CCNet (Dilation) [17], FPN [22] and SeTR (Transformer) [54]. Note that the encoder of Segmenter is pre-trained by our ReLeM. With LSTM and transformer-based text encoding, we arrive at 6 different ReLeM models, i.e., ReLeM-{ CCNet, FPN, SeTR}× ({ LSTM, Transformer}). We use the standard pixel-wise cross-entropy loss to optimize segmentation models.

Food Image Segmentation based on Deep and Shallow Dual-branch Network (2024)

Paper:
https://www.researchgate.net/publication/384765579_Food_Image_Segmentation_based_on_Deep_and_Shallow_Dual-branch_Network

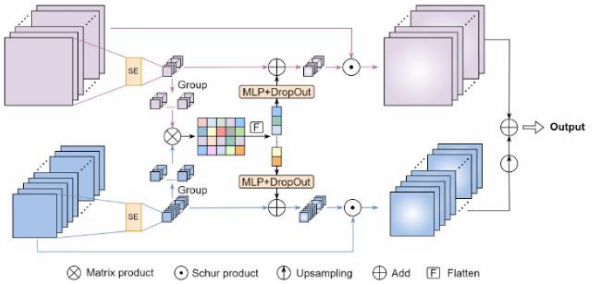

Fig. 5: Schematic diagram of the multi-scale relation-aware feature fusion module. The purple branch and the blue branch represent the high-resolution feature map and the low-resolution feature map, respectively.

Dual-branch architecture:

- *Deep branch*: extracts high-level semantic features for distinguishing food categories.
- *Shallow branch*: preserves fine-grained spatial details (edges, textures) crucial for accurate segmentation.
- Fusion of both branches enhances recognition of foods with similar appearances.

Table 2: The comparison results of FDSNet with other methods on the FoodSeg103 dataset, where a short horizontal line represents that the model is not open-sourced or the metric has not been evaluated in the referenced articles.

Methods	Epochs	Resolution	mIoU	mAcc	FLOPs(G)	Params(M)
FPN	120	512×1024	27.80	38.20	277.84	33.07
CCNet [37]	120	512×1024	35.50	45.30	615.28	71.36
PVT/S-FPN	120	512×1024	31.30	43.00	117.61	26.66
Swin/B-UperNet	120	512×1024	41.20	53.90	260.33	88.52
ViT/B-Naive	120	768×768	41.30	52.70	198.56	87.76
ViT/B-PUP	120	768×768	38.50	49.10	459.31	91.37
ViT/B-MLA	120	768×768	45.10	57.40	257.12	102.59
ViT/L-MLA	120	768×768	44.50	56.60	759.18	321.41
SegFormer/b5	120	640×640	42.50	54.64	101.04	84.67
FoodSAM	120	768×768	46.42	58.27	460.13	632.75
UperNet [38]	120	768×768	39.80	52.37	154.27	31.53
CANet [39]	120	512×1024	37.21	47.33	-	-
STPPN [40]	120	512×1024	40.30	53.98	-	-
DeeplabV3+(BYOL+FeaSC) [41]	120	512×512	36.22	48.87	-	-
FDSNet(ViT)	**120**	**768×768**	**46.38**	**58.17**	**187.46**	**98.13**
FDSNet(Swin)	**120**	**768×768**	**47.34**	**60.04**	**182.74**	**101.93**

https://www.researchgate.net/publication/391101002_MoEMASeg_An_Enhanced_DeepLab_V3_Combining_MobileNet_V2_and_EMA

The paper *"MoEMASeg: An Enhanced DeepLab V3 Combining MobileNet V2 and EMA"* proposes a lightweight yet accurate semantic segmentation model tailored for food and real-time applications. It integrates **MobileNetV2** (for efficiency) with **Exponential Moving Average (EMA)** optimization to stabilize training and improve feature representation.

 MobileNetV2 backbone**:**

- Provides lightweight feature extraction.
- Reduces computational cost compared to heavier backbones like ResNet or Xception.

 EMA optimization**:**

- Smooths parameter updates during training.
- Improves convergence stability and generalization.

 Enhanced DeepLabV3 framework**:**

- Retains atrous spatial pyramid pooling (ASPP) for multi-scale context.
- Combines MobileNetV2 + EMA to balance speed and accuracy.