Dining on Details: LLM-Guided Expert Networks for Fine-Grained Food Recognition

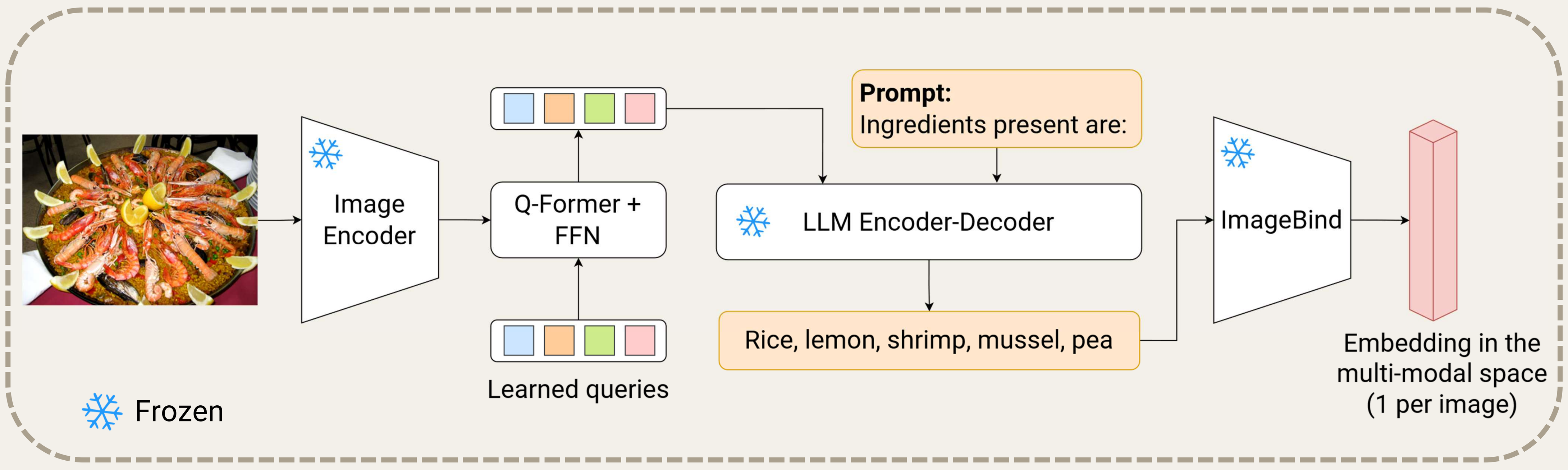
Jesús M. Rodríguez-de-Vera[1], Pablo Villacorta[1], Imanol G. Estepa[1], Marc Bolaños[2], Ignacio Sarasúa[3], Bhalaji Nagarajan[1] and Petia Radeva[1,4]
Universitat de Barcelona[1], AIGecko Technologies SL[2], NVIDIA Corporation[3], Computer Vision Center[4]

MADiMa 2023

Summary

- Food recognition is a **fine-grained problem**: high inter-class similarity and intra-class variance.
- We propose **Dining on Details** (DoD), a **subset expert-based** approach in fine-grained food recognition.
  - With power of recent **LLMs** and the robustness of **ImageBind** to find similar classes in the **multi-modal space**.
  - **End-to-end multi-task** learning process, enhancing **performance** especially **with highly similar classes**.
  - It is a **universal add-on** to any existing classifier.
- Obtain **competitive results** in various food benchmarks with different backbones, and **state-of-the-art in Food-101**.

Our Proposal: Dining on Details



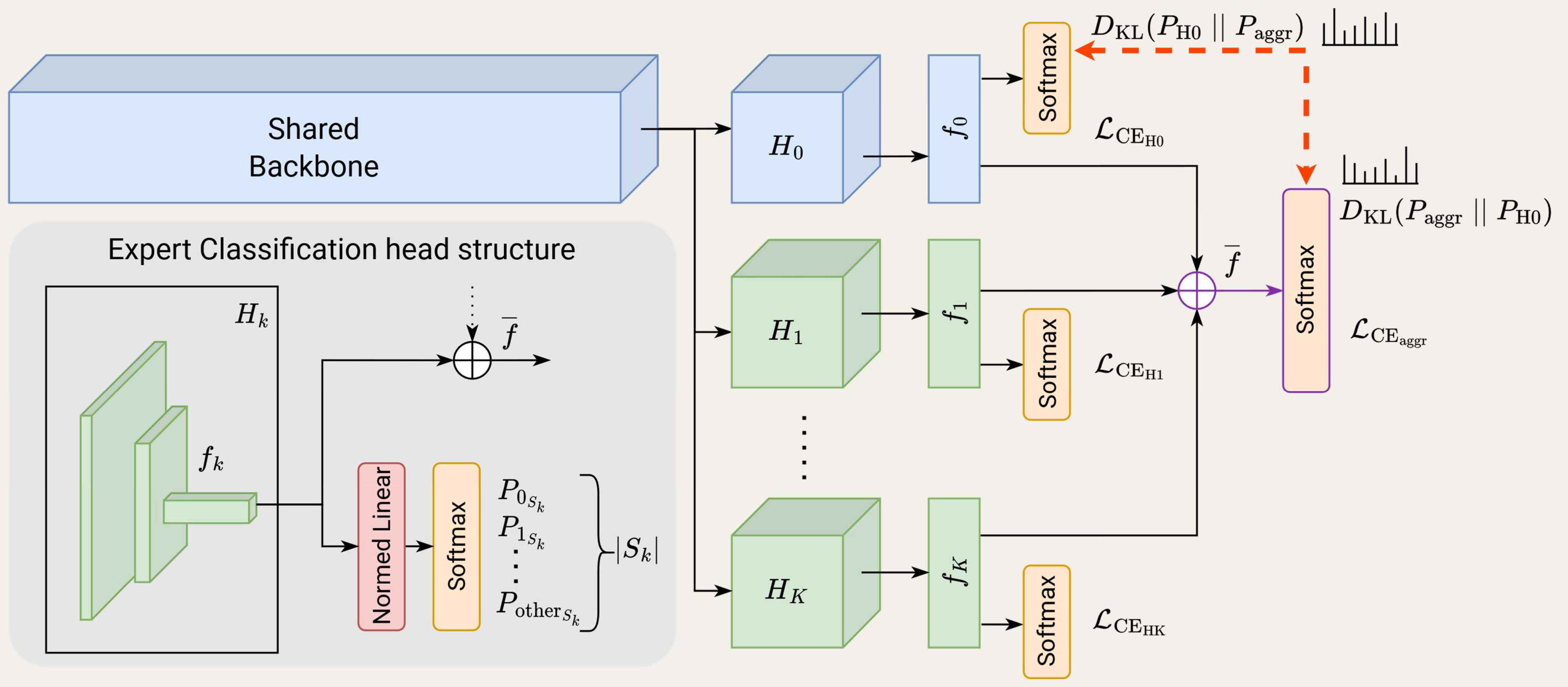
- Using BLIP-2, we use a pretrained image encoder and LLM to obtain the **ingredient list of every image**.
- ImageBind **projects** every list of ingredients to the **multi-modal latent space**.
- Get the **average** vector for each **class**.

Hierarchical agglomerative **clustering** to find **similar classes** in the multi-modal space.

$$\cos(\theta_{ij}) = \frac{L_i \cdot L_j}{\|L_i\|_2 \|L_j\|_2}$$

- For **each cluster** of classes, we append an **expert classifier sub-network** after the baseline backbone.
  - Trained to distinguish specific in that cluster or "other".
- We average the last pre-classifier vector of every head (including the original) and train a combined or **aggregated classifier** from that **regularized** vector.
- To speed up the learning, we use **mutual** knowledge **distillation** between the original classifier H0 and the aggregated classifier.
- Everything is trained jointly in an **end-to-end multi-task** fashion:

$$\mathcal{L} = \lambda_1(\mathcal{L}_{\mathrm{CE_{H0}}} + \mathcal{L}_{\mathrm{ML_{H0}}}) + \lambda_2 \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_{\mathrm{CE_{Hk}}} + \lambda_3(\mathcal{L}_{\mathrm{CE_{aggr}}} + \mathcal{L}_{\mathrm{ML_{aggr}}})$$



Results

- Improves the baseline in a wide variety of datasets and backbones (CNN and transformers).
- Improves Food-101 SOTA by more than 1 point.

Table 1: Test accuracy (%) of baseline and proposed DoD.

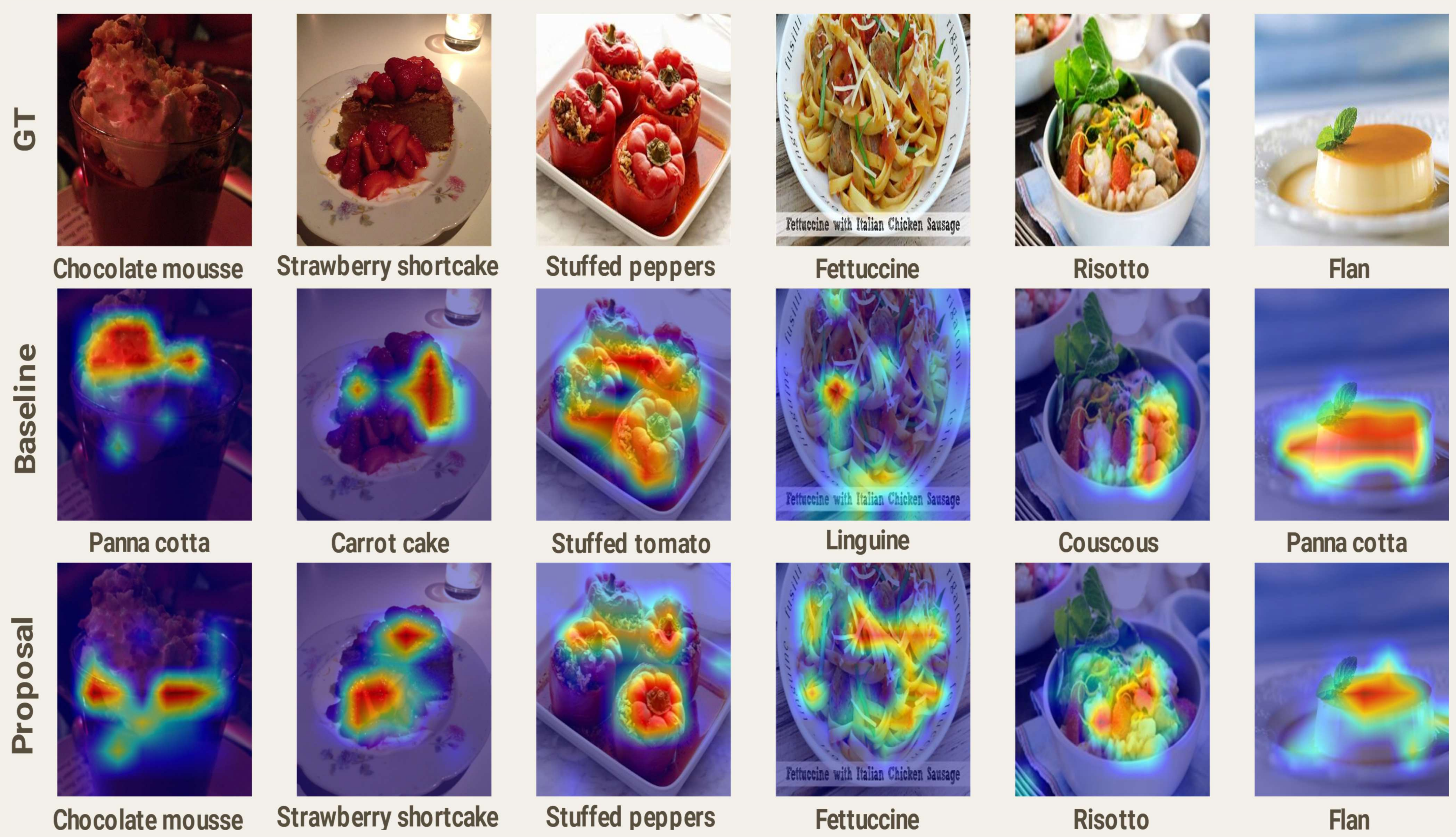
Dataset	Backbone	Baseline	DoD (Ours)	Gain
UECFood-100	EfficientNet-B0	78.43	**79.58**	+1.15
UECFood-100	ResNet-50	77.24	**78.85**	+1.61
UECFood-100	SwinV2-T	77.94	**78.44**	+0.50
Food-101	SwinV2-T	89.96	**90.70**	+0.74
FoodX-251	SwinV2-T	72.89	**74.25**	+1.36

Table 2: Comparison of DoD with SoTA methods in Food-101. † =bigger image size. § =subset-based method.

Method	Test Accuracy (%)
Grafit (ICCV'21)	93.7
EffNet-B7 (ICML'19)†	93.0
PMG (CVPR'21)†§	87.5
FGFR (Madima'22)§	93.8
DoD + SwinV2-B§	**94.9**

Analysis

- The method mainly **improves** in previously highly confused images (**very similar**).
- GradCAM shows that DoD **focuses** characteristics and differentiating parts of the images.



Acknowledgement

This work was partially funded by the Horizon EU project MUSAE (No. 01070-421), 2021-SGR-01094 (AGAUR), Icrea Academia'2022 (Generalitat de Catalunya), Robo STEAM (2022-1-BG01-KA220-VET-000089434, Erasmus+ EU), DeepSense (ACE053/22/000029, ACCIÓ), DeepFoodVol (AEI-MICINN, PDC-2022-133642-I00), PID2022-141566NB-I00 (AEI-MICINN), and CERCA Programme / Generalitat de Catalunya.B. Nagarajan acknowledges the support of FPI Becas, MICINN., Spain. The authors thankfully acknowledge the computer resources at FinisTerrae III and the technical support provided by the Galician Supercomputing Center (CESGA) (RES-IM-2023-2-0025).