LLM-Generated Semantic Co-occurrences for Multi-label Food Recognition

Daniel Ponte[1(✉)], Eduardo Aguilar[1,2], Mireia Ribera[1], and Petia Radeva[1,3]

[1] Deptartment de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, Barcelona 08007, Spain
dponteva163@alumnes.ub.edu, {eduardo.aguilar,ribera,petia.ivanova}@ub.edu
[2] Deptartment de Ingeniería de Sistemas y Computación, Universidad Católica del Norte, Angamos 0610, Antofagasta 1270398, Chile
[3] Computer Vision Center Campus UAB, Edifici O, Cerdanyola, 08193 Barcelona, Spain

Abstract. Multi-label learning in food image recognition presents a promising avenue for understanding the visual composition of meals through joint ingredient prediction. In this article, we improve an existing GCN-based framework by replacing its standard co-occurrence matrix with a novel semantic variant, constructed using large language models (LLMs). Unlike traditional approaches that derive co-occurrence statistics solely from the training data which often introducing biases and limiting generalization, our method leverages prior knowledge extracted from LLMs to build an adjacency matrix that captures broader and more contextually grounded ingredient relationships. We evaluated our approach on the MAFood-121 and VireoFood-172 datasets, significantly outperforming the benchmark method that relies on dataset-conditioned co-occurrence graphs. On MAFood-121, our model improved the mean average precision (mAP) from 82.77% to 87.46%, while on VireoFood-172, it increased from 60.88% to 65.28%. The results demonstrate the effectiveness of integrating LLM-derived semantic structure into graph-based multi-label models for structured food recognition.

Keywords: Multi-label · LLMs · Co-occurrence matrix · GCN

1 Introduction

Multi-label learning for food image recognition has gained increasing attention, driven by the need to automatically understand the complex composition of meals in real-world scenarios. Unlike traditional single-label classification, food images often require the simultaneous prediction of multiple ingredients, many of which co-occur naturally or share semantic relationships. Early efforts focused on dish classification, but limitations in capturing internal ingredient structures led to a shift toward fine-grained multi-label frameworks, accelerated by the success of convolutional neural networks (CNNs) [6].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
M. Castrillón-Santana et al. (Eds.): CAIP 2025, LNCS 15622, pp. 220–230, 2026.
https://doi.org/10.1007/978-3-032-05060-1_19

The emergence of vision-language models (VLMs) (e.g. CLIP [12]) has expanded image recognition by aligning visual and textual information in a shared space, allowing flexible prompt-based recognition. Nevertheless, in food analysis, VLMs often treat labels independently, ignoring ingredients relationships (e.g. the presence of "rice" may imply "seafood" or "vegetable", a dependency not directly modeled in traditional approaches).

To capture label dependencies, GCNs [15] have been introduced into multi-label pipelines. DualCoOp [16] and SCPNet [9] combine VLMs with graph structures to refine predictions. Nevertheless, a major limitation persists: the co-occurrence matrices used are derived from training data, reflecting dataset-specific biases and limiting generalization across diverse cuisines and contexts.

To overcome this, we propose LLM-MLR (Large Language Model Multi-Label Recognition), a novel framework that incorporates external semantic knowledge extracted from LLMs. Instead of relying solely on training data, we generate a semantic co-occurrence matrix mined from over three million global recipes, guided by empirical co-occurrence frequency, culinary compatibility, shared cooking techniques, and multicultural diversity.

This matrix allows the GCN module to refine initial predictions from a CLIP-based backbone more effectively, resulting in improved performance and robustness. We validate LLM-MLR on two public datasets, MAFood-121 [2] and VireoFood-172 [5], which present challenges due to inter-class similarity, intra-class variability, and limited data. Our experiments show consistent improvements compared to baseline methods that rely solely on co-occurrence statistics extracted from the training data, highlighting the benefits of integrating external structured knowledge into food ingredient recognition.

Our main contributions in this paper are as follows:

1. We propose LLM-MLR, a novel framework that leverages large language models to build a semantic ingredient co-occurrence matrix, mitigating biases inherent in dataset-conditioned graphs.
2. We demonstrate the effectiveness of LLM-MLR through comprehensive experiments on two public food datasets, achieving superior generalization compared to the Multi-label Recognition (MLR) [13] baseline method.
3. We introduce a carefully designed prompt-based approach, explicitly defining culinary categories and their relationships, enabling LLMs to generate precise and semantically coherent ingredient co-occurrence probabilities.

2 Methodology

We present the LLM-MLR framework for multi-label ingredient recognition in food images. Our method integrates a frozen CLIP-based backbone for visual and textual representation extraction with a GCN for semantic refinement, an externally constructed semantic co-occurrence matrix derived using LLMs.

2.1 Overall Architecture

Our work builds upon the architecture proposed in the Multi-label Recognition (MLR) [13], which operates in two stages: feature extraction and semantic refinement. In the first stage, visual and textual features are extracted using a frozen CLIP backbone, which preserves the pretrained vision-language alignment. Instead of static prompts, MLR employs learnable positive and negative prompt learners, following recent advances in prompt tuning for vision-language models. These learnable prompts allow the model to adapt contextually to the specifics of the ingredient recognition task. The extracted features are then processed by a Graph Convolutional Network (GCN) that refines the initial logits, modeling dependencies between ingredient labels. The key contribution of our method lies in the construction of the semantic co-occurrence matrix (see Table 1) used by the GCN. While MLR relies on co-occurrence statistics derived directly from the training data, our approach replaces this with a semantically enriched matrix generated through prompt-based querying of a large language model (LLM), offering broader generalization and mitigating dataset-induced biases.

2.2 Semantic Co-occurrence Matrix Construction

Traditional approaches often derive co-occurrence matrices directly from training data, making them susceptible to dataset biases. To overcome this, we introduce a semantic co-occurrence matrix generated using a LLM. The matrix construction is based on information extracted from more than three million culinary recipes across international platforms like Allrecipes [3] and Epicurious [11], according to specific culinary and statistical criteria.

We used OpenAI's ChatGPT (GPT-4) [1] to generate the semantic co-occurrence matrix via prompt-based querying. The LLM was instructed to assign co-occurrence probabilities between ingredients based on culinary and statistical principles:

- Frequency of co-occurrence in recipe databases (e.g., meat + pasta with 0.82; meat + rice with 0.70).
- Culinary compatibility (e.g., soup + vegetable at 0.86; egg + bread at 0.60).
- Shared cooking techniques (e.g., friedfood + seafood at 0.76).
- Restrictions or dietary preferences, penalizing rare or redundant combinations (e.g., bread + dumpling at 0.24; rice + bread at 0.12).
- Multicultural and the global diversity (e.g. noodle + soup at 0.66).

The resulting matrices for MAFood-121 (10×10) and VireoFood-172 (18×18) (see Fig. 1) are normalized between 0 and 1, symmetric, and enforce 1 s on the diagonal. A further normalization is applied to stabilize graph propagation: row sums (excluding diagonals) are scaled to 0.2, while the diagonal is adjusted to 0.8. This ensures that the adjacency structure is numerically stable for GCN processing without overwhelming node self-representations.

Table 1. Prompt developed to instruct the LLM for matrix generation

Calculate the probability that two food groups appear together in the same dish or culinary preparation.

Food groups and what they include:

- *bread: wheat flour, bread, pizza dough, etc.*
- *dumpling: stuffed pasta, wheat semolina, etc.*
- *egg: whole eggs, egg whites, etc.*
- *friedfood: any fried food (tempura, croquettes, French fries, etc.).*
- *meat: red meats, chicken, turkey, bacon, sausage, etc.*
- *noodle/pasta: spaghetti, noodles, macaroni, ramen, etc.*
- *rice: rice in all its variations.*
- *seafood: fish, seafood, octopus, shrimp, etc.*
- *soup: light or dark poultry, beef or vegetable broths.*
- *vegetable: vegetables, legumes, mushrooms, avocado and fruits.*

Rules when generating the matrix:

a. The matrix is 10×10 and symmetric.
b. The main diagonal is always 1.00.
c. All values must be between 0.00 and 1.00 with two decimal places.
d. Base the numbers on plausible culinary data and compatibility.
e. Return only the matrix without explanations.
f. Use tabs or spaces to separate columns.

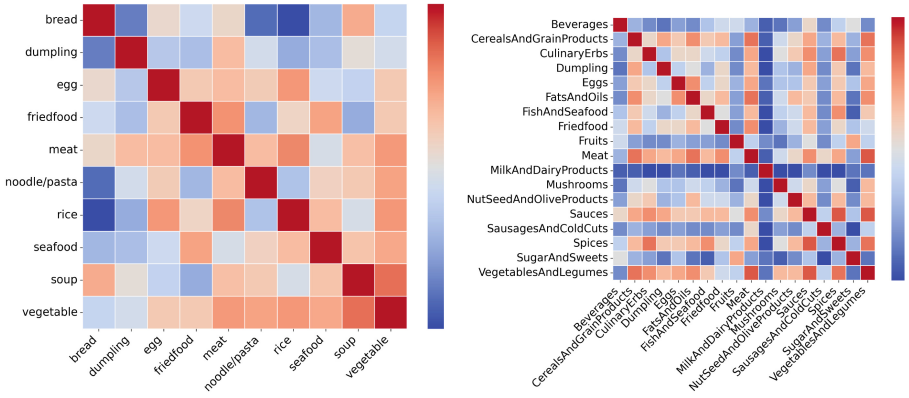

Fig. 1. Semantic Co-occurrence Matrices generated for MAFood-121 (left) and VireoFood-172 (right), based on structured knowledge extracted using a LLM. These matrices serve as the relational backbone of the graph-based ingredient prediction model.

2.3 GCN-Based Refinement Module

The semantic refinement stage employs a three-layer GCN architecture inspired by recent advances in graph-based multi-label learning [4]. The GCN processes

the initial CLIP-based logits, integrating information from related ingredients through the adjacency matrix. Specifically, the architecture consists of:

- A first graph convolution layer (GCL) maps the input dimension (1) to an intermediate representation (4).
- A second GCL maintains the same intermediate dimension (4).
- A final GCL reduces the dimension back to 1.

Each of the first two layers applies a LeakyReLU activation with a negative slope of 0.2, introducing non-linearity and improving information propagation through the graph. No dropout is applied to preserve information across the small number of nodes. The propagation rule for each layer follows:

$$H^{(l+1)} = \sigma(\hat{A}H^{(l)}W^{(l)})$$

where $H^{(l)}$ is the feature matrix at layer l, $W^{(l)}$ is the learnable weight matrix, $\hat{A}$ is the normalized adjacency matrix, and $\sigma(\cdot)$ denotes the LeakyReLU activation function. The final refined logits are added residually to the original logits from CLIP to preserve strong visual signals.

2.4 Loss Function and Training Strategy

Given the class imbalance inherent in ingredient recognition datasets, we adopt the Asymmetric Loss function [14] to optimize the model. This loss introduces distinct focusing parameters for positive (γ^+) and negative (γ^-) classes to prioritize rare ingredient detection while controlling over-prediction of absent classes.

The loss is defined as:

$$\mathcal{L}(x, y) = -y \log p^+ - (1 - y) \log p^-$$

where:

- y is the ground truth label (1 for presence, 0 for absence of an ingredient),
- p^+ is the predicted probability for the positive class after asymmetric clipping and focusing,
- p^- is the predicted probability for the negative class,
- $\log p^+$ and $\log p^-$ are the log-likelihoods emphasizing correct predictions for positives and negatives, respectively.

The asymmetric clipping reduces the impact of correctly predicted negative examples, while the asymmetric focusing dynamically scales the contribution of hard-to-classify positive examples.

The model is trained using Stochastic Gradient Descent (SGD) with momentum. A cosine annealing learning rate schedule with a warmup phase is employed. Only the prompt learners, GCN parameters, and optionally the attention heads are updated during training. Data augmentation strategies such as resizing, Cutout [8], and RandAugment [7] are used to enhance robustness.

3 Experimental Setting

We first describe the datasets used, followed by the evaluation metrics, the experimental setup, and finally, the model settings employed to validate our approach.

3.1 Datasets

The two public datasets selected to validate our approach are MAFood-121 and VireoFood-172, each supporting multi-label food ingredient recognition with distinctive characteristics.

MAFood-121. consists of 21,175 images distributed across 121 traditional dishes from 11 global cuisines. The dishes are organized into 10 major food groups, such as "bread", "meat", and "seafood", enabling analysis across both common and culturally significant dishes.

VireoFood-172. contains 110,241 images covering 172 Chinese dishes. Unlike MAFood-121, this dataset provides ingredient-level annotations, allowing a detailed analysis of the components present in each dish. Ingredients (353 types) were grouped into 18 food categories based on the HELIS ontology [10] (Healthy Eating and Lifestyle ontology), which provides a structured categorization for food ingredients, facilitating consistent grouping of related items into broader food groups. For instance, in HELIS, ingredients such as Bread, Rice, and Spaghetti are grouped under *CerealsAndGrainProducts*; Pork chunks, Chicken legs, and Beef slices are grouped under *Meat*; and Asparagus, Zucchini slices, and Green soybeans are grouped under *VegetablesAndLegumes*. This structured organization enables the ingredient recognition task in VireoFood-172 to be aligned with the grouping structure used in MAFood-121.

For both datasets, the images were partitioned into 80% for training and 20% for testing. In all experiments, input images were resized to 224×224 pixels to match the requirements of the model backbone.

3.2 Metrics

We evaluate our models using standard multi-label classification metrics, computed at both macro and micro levels based on the per-class predictions. Specifically, we report **Mean Average Precision (mAP)**, **Precision**, **Recall**, **F1-Score**, and **Jaccard Index (J)**.

3.3 Experimental Setup

We trained end-to-end a CLIP-based architecture enhanced with a GCN (see Sect. 2). The visual encoder uses a frozen ResNet-101 backbone, with only the prompt learners and GCN parameters being optimized during training.

Two model variants were compared:

- MLR using a **train-conditioned co-occurrence matrix** (biased to training data statistics).

Table 2. Comparison between train-conditioned co-occurrence matrix and LLM-derived semantic co-occurrence matrix (LLM-MLR) on MAFood-121 and VireoFood-172 datasets.

		MACRO		MICRO		
Dataset	**Method**	**F1**	**J**	**F1**	**J**	**mAP**
MAFood-121	MLR	72.48	65.85	57.38	60.46	82.77
	LLM-MLR	**78.49**	**74.35**	**64.89**	**67.81**	**87.46**
VireoFood-172	MLR [13]	57.42	68.36	42.89	61.24	60.88
	LLM-MLR	**60.94**	**73.61**	**46.72**	**66.30**	**65.28**

– LLM-MLR using an **LLM-derived semantic co-occurrence matrix** (based on external culinary knowledge).

All models were initialized from pretrained CLIP weights. The training process used SGD with a momentum of 0.9 and a weight decay of 5×10^{-4}. The initial learning rate was set to 0.001, following a cosine annealing schedule with a warmup phase during the first epoch. Models were trained for 150 epochs, using a batch size of 32 for training and 100 for testing. No validation set was used; only training and testing splits were considered.

All experiments were conducted on NVIDIA GPUs with CUDA acceleration using PyTorch.

4 Results and Discussion

Table 2 presents the quantitative evaluation of our proposed LLM-MLR model compared to the MLR baseline, which relies solely on a co-occurrence matrix derived from the training set. The evaluation was conducted on the official test splits of MAFood-121 and VireoFood-172, two benchmark datasets for multi-label food ingredient recognition. Results show consistent improvements across all macro and micro metrics when replacing the biased co-occurrence graph with our LLM-derived semantic prior. In MAFood-121, macro-F1 improves from 72.48% to 78.49%, and mAP from 82.77% to 87.46%. For VireoFood-172, gains are also notable, with macro-F1 increasing from 57.42% to 60.94% and micro-mAP from 60.88% to 65.28%. The greater improvements in macro metrics demonstrate better handling of minority classes, which are typically harder to model under heavy class imbalance.

To understand the behavior of each model in more depth, Fig. 2 shows the per-class multi-label confusion matrices for MAFood-121 and VireoFood-172, respectively. Each matrix visualizes true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) for all food groups. The matrices on the top correspond to the baseline (MLR), and those on the bottom to our model (LLM-MLR). In MAFood-121, LLM-MLR improves both precision and recall across most food groups. For instance, *bread* increases from 975

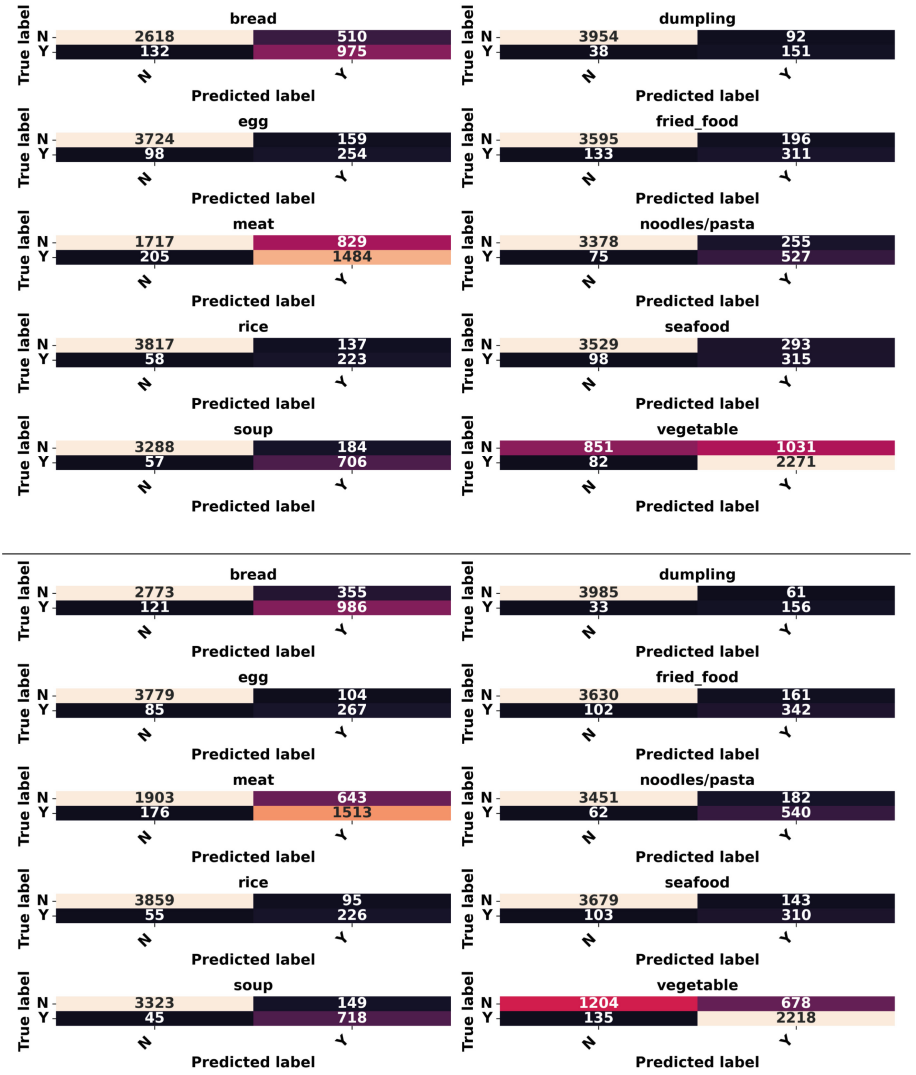

Fig. 2. Per-class multi-label confusion matrices for MAFood-121. Top: MLR baseline using a train-conditioned co-occurrence matrix. Bottom: LLM-MLR with an LLM-derived semantic matrix. The proposed model yields fewer false positives and improved true positive rates across most food groups. Notable gains are observed in *bread*, *meat* and *vegetable*, demonstrating stronger discriminative capacity and better context modeling.

to 986 TP while reducing FP from 510 to 355. Similar trends appear in *meat* and *vegetable*, the latter seeing a substantial FP reduction from 1,031 to 678 while maintaining high TP counts. These shifts suggest that the semantic prior

improves discriminative capacity, especially for food groups that are visually or contextually confusable. On the other hand, Table 3 summarizes the qualitative differences between the baseline and the proposed method on the MAFood-121 and VireoFood-172 datasets.

Beyond per-class metrics, we performed a per-sample agreement analysis to determine how often each model succeeded or failed independently. Table 4 categorizes the predictions into 4 cases: both correct (YY), both incorrect (NN), only LLM-MLR correct (Y/N), and only MLR correct (N/Y). As expected, most test samples fall into the YY group, especially in high-frequency classes. However, in both datasets, the Y/N group is substantially larger than N/Y, demonstrating that LLM-MLR recovers harder examples more effectively than the baseline (e.g. in MAFood-121, there are 811 samples correctly predicted only by LLM-MLR, versus just 462 where the baseline was superior). This indicates that semantic priors help resolve visually ambiguous cases more reliably.

Furthermore, we observe that certain food groups exhibit stronger inter-class disentanglement under the semantic graph. In MAFood-121, ingredient pairs like *bread* and *fried_food*, which often co-occur in recipes but are visually distinct, are better separated by LLM-MLR. Likewise, in VireoFood-172, confusions between *meat* and *CerealsAndGrainProducts*, commonly caused by overlapping textures

Table 3. Comparison of model predictions across representative food groups. Classes highlighted where LLM-MLR improves over the baseline (MLR), as well as where performance remains stable or slightly degraded.

Dataset	MLR (Baseline)	LLM-MLR (Ours)
MAFood-121 [2]	*Bread*: 975 TP, 510 FP, 132 FN	*Bread*: 986 TP, 355 FP, 121 FN
	Vegetable: 2,271 TP, 1,031 FP	*Vegetable*: 2,218 TP, 678 FP
	Meat: 1,484 TP, 829 FP	*Meat*: 1,513 TP, 643 FP
VireoFood-172 [5]	*Meat*: 10,065 TP, 4,886 FP, 887 FN	*Meat*: 10,821 TP, 3,978 FP, 671 FN
	FishAndSeafood: 2,647 TP, 1,490 FP	*FishAndSeafood*: 2,688 TP, 1,186 FP
	Spices: 7,430 TP, 4,925 FP	*Spices*: 7,560 TP, 4,165 FP
Case Agreement	Correct: both TP in frequent classes (e.g., *Eggs*, *Dairy*) Incorrect: both FP in ambiguous pairs (e.g., *Sauces* vs *Spices*)	Improves rare/overlapping classes (e.g., *Vegetables*, *Fish*) Few degradation cases, often on borderline predictions

Table 4. Per-image agreement analysis between MLR (baseline) and LLM-MLR. YY: both correct, NN: both incorrect, Y/N: only LLM-MLR correct, N/Y: only baseline correct.

Dataset	YY both correct	NN both incorrect	Y/N only ours	N/Y only baseline
MAFood-121	3,942	728	811	462
VireoFood-172	15,184	4,367	2,143	1,002

in composite dishes, are significantly reduced. These outcomes reinforce our core hypothesis: integrating a linguistically structured prior into a GCN enables better generalization and contextual reasoning, especially in scenarios affected by class imbalance and semantic overlap. These findings are consistent with broader trends in vision-language research, where pretrained language models provide strong inductive biases for downstream recognition tasks.

5 Conclusions

In this paper, we proposed LLM-MLR, a novel approach that integrates semantic ingredient relationships derived from LLMs into GCN for multi-label food ingredient recognition. Unlike traditional models that rely solely on training-data-specific co-occurrence matrices, our method leverages external structured knowledge, effectively mitigating dataset biases and significantly enhancing generalization capabilities.

Extensive experiments on the MAFood-121 and VireoFood-172 datasets demonstrate consistent and substantial improvements in multiple metrics such as macro-F1, Jaccard Index, and mean Average Precision. Specifically, our method shows notable advantages for minority classes, effectively reducing common misclassifications between visually similar food groups. These results clearly highlight the importance and efficacy of embedding linguistically informed semantic priors into recognition models, reinforcing their potential as robust inductive biases in complex multi-label recognition scenarios.

In future work, we aim to extend our framework beyond the prediction of food groups by simultaneously incorporating ingredient-level and dish-level tasks within a multi-task learning approach. This multi-task strategy could further exploit the structured semantic knowledge, benefiting from shared representations across closely related tasks and leveraging the inherent semantic relationships between dishes and ingredients. Additionally, exploring hierarchical food ontologies integrated directly into the GCN structure could yield even finer-grained and more accurate predictions, enhancing not only model accuracy, but also interpretability in food image recognition tasks.

Acknowledgments. This work has been partially supported by the Spanish project PID2022-136436NB-I00 (AEI-MICINN), Horizon EU project MUSAE (No. 01070421), 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, PDC2022-133642-I00), PID2022-141566NB-I00 (AEI-MICINN), CERCA Programme / Generalitat de Catalunya and Beatriu de Pinós Programme (2022 BP 00257). D. Ponte acknowledges the support of Secretaría Nacional de Ciencia, Tecnología e Innovación Senacyt Panamá (Scholarship No. 270-2022-125).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achiam, J., et al.: Gpt-4 Technical Report (2023). arXiv preprint arXiv:2303.08774
2. Aguilar, E., Bolaños, M., Radeva, P.: Regularized uncertainty-based multi-task learning model for food analysis. JVCI **60**, 360–370 (2019)
3. Allrecipes, I.: Allrecipes (2023). https://www.allrecipes.com/
4. Bei, Y., et al.: Correlation-aware graph convolutional networks for multi-label node classification (2024). arXiv preprint arXiv:2411.17350
5. Chen, J.J., Ngo, C.W.: Deep-based ingredient recognition for cooking recipe retrieval. ACM Multimedia (2016)
6. Chen, J., et.al.: Zero-shot ingredient recognition by multi-relational graph convolutional network. In: AAAI CAI. vol. 34, pp. 10542–10550 (2020)
7. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: practical automated data augmentation with a reduced search space. In: CVPRW, pp. 702–703 (2020)
8. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout (2017). arXiv preprint arXiv:1708.04552
9. Ding, Z., et.al.: Exploring structured semantic prior for multi label recognition with incomplete labels. In: CVPR, pp. 3398–3407 (2023)
10. Donadello, I., Dragoni, M.: Ontology-driven food category classification in images. In: ICIAP, Part II 20, pp. 607–617. Springer (2019)
11. Jimenez-Mavillard, A., Suarez, J.L.: Diffusion of elbulli's innovation: rate of adoption in allrecipes and epicurious. IJGFS **22**, 100243 (2020)
12. Radford, A., et.al.: Learning transferable visual models from natural language supervision. In: ICML, pp. 8748–8763. PmLR (2021)
13. Rawlekar, S., Bhatnagar, S., Srinivasulu, V.P., Ahuja, N.: Improving multi-label recognition using class co-occurrence probabilities. In: ICPR, pp. 424–439 (2025)
14. Ridnik, T., et.al.: Asymmetric loss for multi-label classification. In: ICCV, pp. 82–91 (2021)
15. Singh, I.P., Ghorbel, E., Oyedotun, O., Aouada, D.: Multi-label image classification using adaptive graph convolutional networks: from a single domain to multiple domains. Comput. Vis. Image Underst. (2024)
16. Sun, X., Hu, P., Saenko, K.: Dualcoop: Fast adaptation to multi-label recognition with limited annotations. NIPS **35**, 30569–30582 (2022)