Enhancing Recipe Retrieval with Foundation Models: A Data Augmentation Perspective

Fangzhou Song[1*], Bin Zhu[2*], Yanbin Hao[1†], and Shuo Wang[1]

[1] University of Science and Technology of China
[2] Singapore Management University

Abstract. Learning recipe and food image representation in common embedding space is non-trivial but crucial for cross-modal recipe retrieval. In this paper, we propose a new perspective for this problem by utilizing foundation models for data augmentation. Leveraging on the remarkable capabilities of foundation models (i.e., Llama2 and SAM), we propose to augment recipe and food image by extracting alignable information related to the counterpart. Specifically, Llama2 is employed to generate a textual description from the recipe, aiming to capture the visual cues of a food image, and SAM is used to produce image segments that correspond to key ingredients in the recipe. To make full use of the augmented data, we introduce **D**ata **A**ugmented **R**etrieval framework (**DAR**) to enhance recipe and image representation learning for cross-modal retrieval. We first inject adapter layers to pre-trained CLIP model to reduce computation cost rather than fully fine-tuning all the parameters. In addition, multi-level circle loss is proposed to align the original and augmented data pairs, which assigns different penalties for positive and negative pairs. On the Recipe1M dataset, our DAR outperforms all existing methods by a large margin. Extensive ablation studies validate the effectiveness of each component of DAR. Code is available at https://github.com/Noah888/DAR.

Keywords: Recipe retrieval · Data augmentation · Foundation models

1 Introduction

With the rapidly increasing amount of multimodal food data (e.g., recipes, food images and cooking videos) from various sources, the demand for food computing [19] to analyze the food data has grown, ranging from food recognition [5,16,20], cross-modal recipe retrieval [25,26,42], food recommendation [39] and food logging [21,24]. In this paper, we focus on cross-modal recipe retrieval, which aims to search for recipe given a food image as query and vice versa.

Great progress has been achieved in cross-modal recipe retrieval by advancing the architectures from convolutional neural network and LSTM [26,38,46,47] to transformer [25] and pre-trained vision and language models [22,27,34]. While

* Co-first author. Email: fangzhousong@mail.ustc.edu.cn, binzhu@smu.edu.sg
† Corresponding author. Email: haoyanbin@hotmail.com.

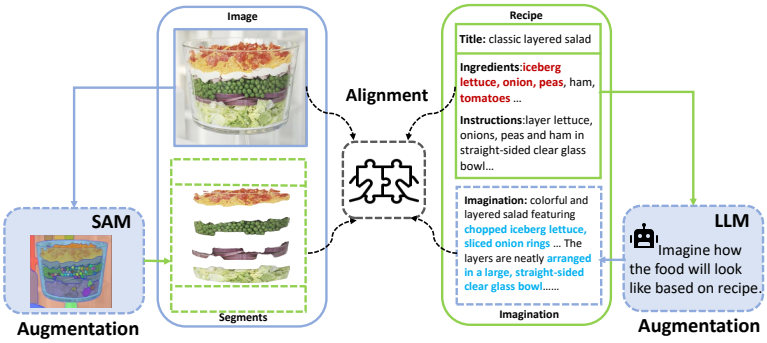

Fig. 1: Illustration of our proposed data augmentation paradigm using foundation models. LLM generates textual descriptions from the recipe to capture the dish's visual cues, while SAM produces image segments aligned with recipe ingredients.

encouraging, these works pay little attention to a fundamental perspective, i.e., the data misalignment. To be specific, a recipe and its corresponding image could contain misaligned information to each other, which is potentially harmful to retrieval performance. On the one hand, a recipe is a long document describing the process of cooking a dish, while a food image is the consequence of following the recipe to cook. As a result, the recipe usually contains redundant information that is irrelevant to the visual appearance of the corresponding image, e.g., "*preheat the oven* 350^0F" and "*Marinate the chicken in the refrigerator for 1 hour*". On the other hand, the food image usually contains irrelevant information to the recipe as well, such as the background and the food container, etc.

This paper proposes a novel data augmentation paradigm by leveraging foundation models to address the limitation. Inspired by the remarkable capabilities of foundation models [1,17,31,32], we introduce language and visual foundation models to augment the recipe and food image data, respectively. Specifically, as shown in Fig. 1, to augment the recipe, large language model Llama2 [32] is instructed to act as a helpful assistant to generate a description of the visual appearance of the dish based on recipe as input, which we call "visual imagination description". Compared to the original recipe, the visual imagination description concentrates on the content relevant to the visual appearance of a dish image, eliminating the redundant and cumbersome information in the recipe. To augment food image, segment anything (SAM) [17] is utilized to segment the image into multiple segments, which usually correspond to the key ingredients in the recipe. In this way, a fine-grained relationship between food image segments and ingredients in the recipe could be established for cross-modal alignment, while alleviating the impact of irrelevant visual information, such as the background in the image. It is worth noting that the augmented image segments and visual imagination description can be used not only for training but also for evaluation.

We propose a Data Augmented Retrieval (DAR) framework that utilizes data augmentation to enhance cross-modal recipe retrieval performance. We aim to take advantage of the powerful ability of pre-trained CLIP to encode original

image-recipe pairs as well as augmented data with efficient computational cost. As a result, instead of fully fine-tuning the CLIP model as in [27, 34], we keep the parameters of CLIP frozen and inject lightweight adapter layers [8] into the CLIP. In addition, we propose a multi-level circle loss to learn cross-modal alignment across original and augmented data during training. The multi-level circle loss is based on the circle loss [29], which assigns different penalties for positive and negative pairs, thus making the training more flexible compared to the widely used triplet loss. The circle loss is adopted to both original and augmented data pairs. Consequently, our DAR manages to make full use of the data augmented by the foundation model and achieve state-of-the-art retrieval performance on Recipe1M [26] dataset.

In summary, our contributions are as follows:

- We propose a new data augmentation paradigm for cross-modal recipe retrieval with foundation models. The recipe is augmented as visual imagination description to focus on visual cues by LLM, and the food image is augmented as segments capturing the key ingredients by SAM. The augmentation can be adopted in both training and testing phrases.
- We introduce Data Augmented Retrieval framework to fully utilize the augmented data using CLIP encoder with lightweight adapter layers and more flexible multi-level circle loss for cross-modal alignment.
- Our proposed model outperforms existing methods by a large margin on the Recipe1M [26] dataset. Extensive ablation studies verify the effectiveness of the proposed techniques.

2 Related Work

2.1 Cross-Modal Recipe Retrieval

Cross-modal recipe retrieval aims at mutual retrieval between recipes and corresponding food images. Earlier works [4, 14, 23, 26, 41, 45, 46] use a pre-trained textual representation (e.g., word2vec [18] for word embedding, skip-thought [13] for sentence embedding), followed by LSTM to obtain the final recipe embeddings. [36] introduces StyleGAN2 [11] to generate images and aligns them in latent space with text for data enhancement. The recipe retrieval performance is further advanced by transformer-based works [7, 25, 28, 37]. [27, 28, 34] have introduced CLIP-based models and achieved promising performance, but they all fully fine-tuned the image encoder of the pre-trained CLIP to encode food images with extensive computation cost. Different from existing methods, we propose a new data augmentation paradigm to enhance the cross-modal recipe retrieval performance.

2.2 Foundation Model

Large language models (LLMs) [1–3, 31, 32] have demonstrated strong language abilities and achieved great success by training with large-scale data. In computer

vision, the recently proposed Segment Anything Model (SAM) [12], a visual foundation model can segment the specified content of an image for a given prompt (e.g., box, point, or mask), thus providing semantic visual information for downstream tasks. These foundation models have been attempted for data augmentation. For example, SAM is used in [44] to obtain priority maps for medical image segmentation enhancement. [40] explores data augmentation on multilingual commonsense datasets with powerful instruction-tuned LLMs, and diffusion model is used to enhance image diversity in [33]. In this paper, we employ Llama2 [32] and SAM to generate visual imagination description from recipe to capture visual cues of food images and image segments corresponding to ingredients in recipe respectively.

3 Method

3.1 Overview

Given an image as a query, image-to-recipe retrieval aims to find the most relevant recipe from a recipe corpus, and vice versa. Assume a set of N image-recipe pairs $\{(i_t, r_t)\}_t^N$ are given, where i_t is a food image and r_t is the corresponding recipe. $i_t \in V$ and $r_t \in R$, where V, R represent the visual and recipe modal spaces, respectively. As the image and recipe belong to different modalities, they cannot be directly compared. As a result, the key to cross-modal recipe retrieval is to learn a mapping function $\Psi(V, R) \rightarrow (E_V, E_R)$ that maps recipes and images into a common embedding space for similarity measurement, where E_V and E_R are image and recipe embeddings, respectively.

This paper aims to explore a new data augmentation paradigm using foundation models to enhance cross-modal recipe retrieval performance. The overview of our proposed model is depicted in Fig. 2. Firstly, we propose to employ foundation models to produce augmented data from recipe and image respectively. Specifically, on the one hand, image segments are extracted from the original food image by the visual foundation model SAM, which corresponds to the key ingredients in the recipe at the semantic level. On the other hand, for the recipe, Llama2 is instructed to imagine the visual description of the food based on the recipe, the result of which we refer to "visual imagination description", denoted as d. Furthermore, inspired by [30], we propose to adopt adapters based on pretrained CLIP model to encode both original and augmented data, then we further introduce a multi-level circle loss function to align original and augmented recipe and image data in the common embedding space.

3.2 Data Augmentation by Foundation Models

Augment recipe with LLM. A recipe contains a list of ingredients and a set of instructions to prepare a particular dish. As a food image is the consequence of the corresponding recipe rather than a caption to describe the image, some information in the recipe is redundant with respect to the visual appearance of

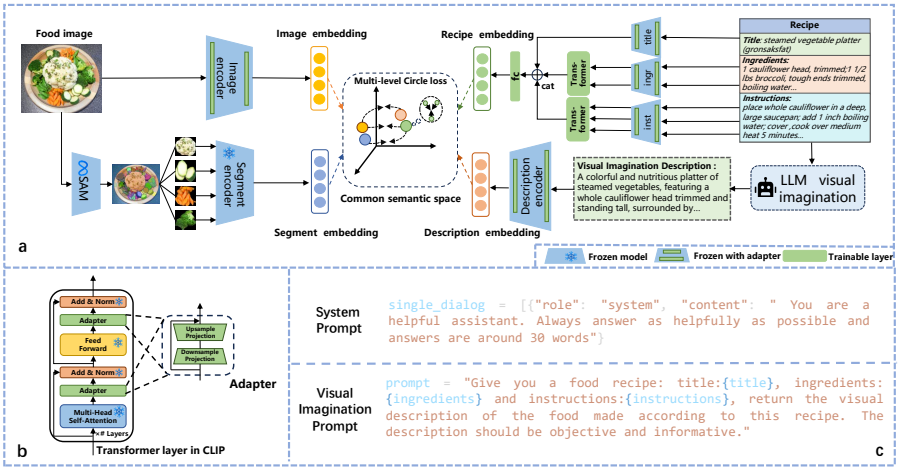

Fig. 2: **(a)** Overview of the DAR framework architecture. **(b)** Architecture of the adapter in the Transformer layer of CLIP. **(c)** Prompt of the LLM to generate visual imagination description.

the image, for instance, "*preheat the oven to 400 degrees F*" and "*Place yogurt in the strainer and allow it to drain for 15 to 20 minutes*". The misalignment between recipe and image could hinder cross-modal recipe retrieval. In order to address this issue, unlike previous approaches in designing better encoders for recipe [15, 37], we propose a new paradigm by employing *Llama2-13b-chat* [32], one of the latest Large Language Models (LLM) to extract the visually aligned information from the recipe for textual data augmentation.

By leveraging the powerful language generation capabilities of the LLM, we aim to produce a description that is visually aligned with a food image, using a recipe as input. As shown in Fig. 2 (c), we carefully design a "visual imagination" prompt to instruct LLM to play the role of a helpful assistant who can imagine what the food will look like based on a recipe. The generated visual imagination description is limited to approximately 30 words to fit the maximum number of input tokens of CLIP. Furthermore, to mitigate the hallucination, we instruct the LLM to generate "objective" and "informative" responses "according to this recipe" in the prompt.

Augment food image with SAM. In a food image, ingredients are the most informative component, which can be matched to the ingredient section in a recipe. However, irrelevant information with the corresponding recipe is inevitable to be introduced to food images, such as the food plate and background. To mitigate this issue, the visual foundation model SAM [12] is adopted to segment the key ingredients from the food image for visual data augmentation.

We use the "everything" † mode to obtain multiple segments from each food image from SAM. i.e., Zero-Shot object proposal generation. SAM samples a

† https://segment-anything.com/demo

large number of points on the image as the prompt, then filters and de-duplicates the generated prediction masks, and finally generates all the prediction masks for the whole image. Note that not all the segments are useful. For instance, there are a large number of background and decorative noises that are not related to food, and there are also many segments that are too small with less semantic information. As a result, we filter out valuable segments by setting an area threshold and semantically consistent matching of segments with text embeddings of *{a picture of food}* using the CLIP model. As shown in Fig. 2, the image segments extracted from the food image are the key ingredients to cook the dish. We use the top-n segments as the image augmented data based on the filtering scores.

3.3 Data Augmented Retrieval(DAR)

To fully make use of augmented data, i.e., visual imagination description and image segments, we propose Data Augmented Retrieval (DAR) framework to enhance cross-modal recipe retrieval performance. We aim to make the model efficient without significantly increasing computational cost and introduce a flexible cross-modal alignment learning strategy for the rich set of data.

Fine-tune CLIP with lightweight adapter. CLIP has demonstrated strong capabilities in various vision-language tasks via contrastive learning on large-scale image-text pairs. There has been some work [27, 34] all fine-tuning CLIP with lots of parameters. To reduce the cost, we propose injecting lightweight adapter layers into the CLIP while keeping all the pre-trained CLIP parameters frozen. As a result, the trainable parameters can be reduced significantly and the adapter layers are trained to enhance the ability of CLIP for cross-modal recipe retrieval. As shown in Fig. 2 (a), we introduce CLIP image and text encoders with adapter layers to obtain image and recipe embeddings, respectively, i.e., $E_V = \phi_{\text{img\&A}}(i), E_R = \phi_{\text{rec\&A}}(r)$, where $\&A$ stands for adding adapter to the encoders. Fig. 2 (b) presents the details of one transformer layer in CLIP with adding the adapter layers. Specifically, the adapter consists of a downsampling projection and an upsampling projection, as well as a residual connection.

The CLIP image encoder can be directly applied to encode the food image. In contrast, a recipe is a long document with three parts: title, ingredients and instructions, denoted as $r = (r_{\text{tit}}, r_{\text{ing}}, r_{\text{ins}})$, CLIP text encoder is incapable to model a recipe in one shot. To address this issue, we propose to use three independent CLIP text encoders to encode the three parts separately. As title is usually one sentence, we can directly use title encoder $\phi_{\text{tit\&A}}$ to obtain title features $E_{R_{\text{tit}}}$. Different from title, the ingredients and instructions sections generally consist of multiple sentences, thus we encode each sentence independently with the CLIP text encoder with adapter layers to get the sentence-level features. Moreover, to learn the compact features of ingredients and instructions, we propose to use a two-layer transformer encoder to learn the interactions among the sentences in ingredient and instruction as follows:

$$E_{R_{\text{ing}}} = \mathit{Trans}\left(\phi_{\text{ing\&A}}\left([r^1_{\text{ing}}, \ldots, r^{M_{\text{ing}}}_{\text{ing}}]\right)\right), \tag{1}$$

$$E_{R_{\text{ins}}} = Trans\left(\phi_{\text{ins\&A}}\left([r^1_{\text{ins}}, \ldots, r^{M_{\text{ins}}}_{\text{ins}}]\right)\right), \tag{2}$$

where M_{ing} and M_{ins} are the maximum number of sentences that can be acceptable for ingredients and instructions, $[r^1_*, \cdots, r^{M_*}_*]$ form the list of sentences for ingredients or instructions, and *Trans* represents the transformer structure.

The features of the three parts of the recipe are then concatenated together, which is subsequently fed into a fully-connected (FC) layer to get the final recipe embedding E_R, which can be formalized as follows:

$$E_R = Tanh(FC(Concat(E_{R_{\text{tit}}}, E_{R_{\text{ing}}}, E_{R_{\text{ins}}}))). \tag{3}$$

Similar to the operation for the recipe, we employ CLIP text encoder with the adapter to compute the output of the visual imagination description d as follows:

$$E_D = \phi_{\text{dec\&A}}(d). \tag{4}$$

Considering that there are several image segments and to prevent the effect of partially noisy samples from increasing, we use the fully frozen CLIP image encoder ϕ_{seg} to encode each image segment first. The final segment embedding E_S is obtained by averaging the embeddings of the n segments. The formula can be written as:

$$E_S = \frac{1}{n}\sum_{i=1}^{n} \phi^*_{\text{seg}}(s_i), \tag{5}$$

where s_i is the image segment and n is the number of segments. ϕ^*_{seg} refers to frozen segment encoder.

Multi-level circle loss with multiple embeddings. The cross-modal recipe retrieval performance depends on the effectiveness of common embedding space between recipe and image. We propose multi-level alignment using circle loss [29] to regulate cross-modal embedding space based on raw and augmented data, including recipe (E_R), food image (E_V), image segments (E_S) and visual imagination description (E_D).

Existing works typically employ triplet loss and its variants for cross-modal embedding space learning [6, 25, 27, 38, 47], Given a query, triplet loss aims to push distance between the query and positive samples larger than that of negative pairs with a pre-defined margin, but this means that increasing the cosine similarity score c_p between a query and its positive sample is equivalent to decreasing the cosine similarity score c_n between the query and negative sample, e.g. when c_p is small and c_n approaches 0, it keeps on penalizing c_n with a large gradient, thus the optimization lacks flexibility.

To address this issue, we first introduce circle loss [29] for cross-modal recipe retrieval. Specifically, the key idea is to employ different penalties for c_p and c_n as follows:

$$L_{\text{circle}}(A, B) = \log[1 + \textstyle\sum_{j=1}^{L} e^{\gamma[c^j_n + m]_+ \cdot (c^j_n - m)} \sum_{i=1}^{K} e^{\gamma[1+m-c^i_p]_+ \cdot (1-m-c^i_p)}], \tag{6}$$

where $m \in [0, 1]$ represents the relaxation factor, and γ is the scale factor used to rescale the cosine similarity score. For the set of data pairs (A, B), K and L

denote the number of positive and negative pairs for all queries. In addition, to improve the robustness of the model, we use the symmetric bidirectional circle loss, with the following equation:

$$L(A, B) = L_{\text{circle}}(A, B) + L_{\text{circle}}(B, A). \tag{7}$$

We then build the cross-modal embedding space with multiple alignment objectives, including recipe and image $L(E_V, E_R)$, image segment and recipe $L(E_S, E_R)$, and image and visual imagination $L(E_V, E_D)$. Furthermore, as there are three sections in a recipe, i.e., title, ingredients and instructions, we are inspired by [25] to align any two sections of a recipe in a self-supervised manner to encourage semantic consistency within the recipe. We define the recipe loss L_{rec} using circle loss as follows:

$$L_{\text{rec}} = \frac{1}{6} \sum_{a} \sum_{b} L(E_{R_a}, LN(E_{R_b}))\delta(a, b), \tag{8}$$

where $a, b \in [tit, ing, ins]$, $\delta(a, b) = 1$ if $a \neq b$ otherwise 0 and $LN(\cdot)$ is a linear projection. E_{R_a} and E_{R_b} are two different content embeddings in a recipe.

Finally, we can obtain the overall multi-level circle loss by combining the above losses as follows:

$$L_{\text{multi-circle}} = L(E_V, E_R) + \alpha L(E_S, E_R) + \beta L(E_V, E_D) + \sigma L_{\text{rec}}, \tag{9}$$

where α, β and σ are hyperparameters to balance the losses.

4 Experiments

4.1 Setups

Dataset. In line with previous works, Recipe1M [26] dataset is used to train and evaluate our method. The recipes with corresponding images are split into 238,408, 51,119 and 51,303 for training, validation and testing respectively. In addition to utilizing these pairs of data, following [25], we also utilize unpaired 482,231 training recipes from the rest of the dataset for the recipe loss L_{rec}.

Evaluation. Following previous works [25, 26, 41], median rank (medR) and recall rate at top K (R@K) are employed to evaluate the performance of our model. MedR represents the median position of the true positives in the distance ranking in the database, and R@K measures the ranking of the percentage of the top K (with $K \in \{1, 5, 10\}$) containing true positive results. During testing, we randomly sample 1,000 image-recipe pairs and 10,000 image-recipe pairs in the test set as the test subset of two scales ($1k$ set up and $10k$ set up). In the test subset, the embedding of one modality is treated as a query to compute the cosine distance with the candidate embeddings of the other modality in the database, and finally the retrieval results are obtained based on the distance ranking. The final reported results are averaged over the 10 sampled subsets.

For evaluation protocols, the augmented data can be not only used for training but also pre-computed for evaluation. Existing works [28,34] generally compute the cosine similarity between recipe and image as a measure to evaluate the retrieval performance. In contrast, we introduce two extra evaluation protocols based on the augmented visual imagination description and image segments as follows:

- **DAR.** Similar to previous works, DAR only utilizes the distance between image and recipe for evaluation, i.e., $dist = dist_{i-r}$.
- **DAR**+ adds visual imagination description into the evaluation, which calculates the distance between image and visual imagination description $dist_{i-d}$, and multiplies by $dist_{i-r}$ to get the distance, i.e., $dist = dist_{i-d} \cdot dist_{i-r}$.
- **DAR**++ adds both visual imagination and image segments for evaluation, which computes the distance $dist_{s-r}$ between segments embeddings and recipe embeddings, then multiplies by the distance metric of DAR+, i.e., $dist = dist_{s-r} \cdot dist_{i-d} \cdot dist_{i-r}$.

Implementation Details. The downsampled dimension of the bottleneck adapter architecture is set to 64. In addition, following [25,27], the CLIP encoder used in our model is the pre-trained CLIP ViT-B/16 model and the output embedding dimension $d = 512$. For the recipe encoder, ingredients and instructions text can accept a maximum number of sentences $M_{\text{ing}}, M_{\text{ins}} = 15$, and the maximum length of each sentence is 20 tokens. The trainable transformer is 2 layers with 4 attention heads. The dimensionality of the Recipe embedding obtained after the FC layer is also 512.

In the training phase, we set the relaxation factor $m = 0.25$ and scale factor $\gamma = 32$ for circle loss. The weight factors for multi-level circle loss are set to be $\alpha = 1$, $\beta = 1$, $\sigma = 1$. Following the baseline settings, we train the model with batch size =128 and optimize the parameters using Adam optimizer. The initial learning rate lr is set to 10^{-4} and every 30 epochs step decays to 0.1 of the previous lr. The model is trained for a total of 100 epochs, and the model parameter with the highest R@1 in validation is selected for testing.

4.2 Performance Comparison

As shown in Table 1, we compare the cross-modal recipe retrieval performance of our DAR with state-of-the-art methods. Our DAR outperforms all existing methods for the majority of the metrics in both 1k and 10k setups, including CLIP-based methods [9,27,28,34,35]. All of these methods simply fine-tune the entire image encoder of the CLIP to improve the performance. In contrast, by inserting lightweight adapter layers, the number of trainable parameters in our image encoder is only 8% of theirs. Compared to MALM [34], the image-to-recipe retrieval performance in 1k testing is boosted by 0.9%, 3.4% and 3.2% in terms of R@1, R@5 and R@10 respectively. The results show that our DAR learns more discriminative recipe and image embeddings, enhancing cross-modal recipe retrieval through data augmentation.

Table 1: Cross-modal recipe retrieval performance comparison with existing methods. The results are reported in terms of medR (↓) and R@K (↑).

Methods	1k								10k							
	image-to-recipe				recipe-to-image				image-to-recipe				recipe-to-image			
	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10
Adamine [4]	2.0	40.2	68.1	78.7	2.0	39.8	69.0	77.4	13.2	14.8	34.6	46.1	14.2	14.9	35.3	45.2
R2GAN [47]	2.0	39.1	71.0	81.7	2.0	40.6	72.6	83.3	13.9	13.5	33.5	44.9	12.6	14.2	35.0	46.8
MCEN [6]	2.0	48.2	75.8	83.6	1.9	48.4	76.1	83.7	7.2	20.3	43.3	54.4	6.6	21.4	44.3	55.2
ACME [38]	1.0	51.8	80.2	87.5	1.0	52.8	80.2	97.6	6.7	22.9	46.8	57.9	6.0	24.4	47.9	59.0
SCAN [37]	1.0	54.0	81.7	88.8	1.0	54.9	81.9	89.0	5.9	23.7	49.3	60.6	5.1	25.3	50.6	61.6
X-MRS [7]	1.0	64.0	88.3	92.6	1.0	63.9	87.6	92.6	3.0	32.9	60.6	71.2	3.0	33.0	60.4	70.7
H-T [25]	1.0	60.0	87.6	92.9	1.0	60.3	87.6	93.2	4.0	27.9	56.4	68.1	4.0	28.3	56.5	68.1
H-T (ViT) [25]	1.0	64.2	89.1	93.4	1.0	64.5	89.3	93.8	3.0	33.5	62.1	72.8	3.0	33.7	62.2	72.7
T-Food (ViT) [27]	1.0	68.2	87.9	91.3	1.0	68.3	87.8	91.5	2.0	40.0	67.0	75.9	2.0	41.0	67.3	75.9
T-Food (CLIP-ViT) [27]	1.0	72.3	90.7	93.4	1.0	72.6	90.6	93.4	2.0	43.4	70.7	79.7	2.0	44.6	71.2	79.7
CREAMY(ViT) [48]	1.0	73.3	92.5	95.6	1.0	73.2	92.5	95.8	2.0	44.6	71.6	80.4	2.0	45.0	71.4	80.0
VLPCook [28]	1.0	73.6	90.5	93.3	1.0	74.7	90.7	93.2	2.0	45.3	72.4	80.8	2.0	46.4	73.1	80.9
FARM [35]	1.0	73.7	90.7	93.4	1.0	73.6	90.8	93.5	2.0	44.9	71.8	80.0	2.0	44.3	71.5	80.0
MALM [34]	1.0	74.0	91.3	94.3	1.0	73.0	91.0	93.9	2.0	45.9	72.3	80.5	2.0	44.2	71.7	80.1
CIP(no-Rerank) [9]	1.0	73.1	94.1	97.1	1.0	72.5	93.7	97.0	-	-	-	-	-	-	-	-
CIP [9]	1.0	77.1	94.2	97.2	1.0	77.3	94.4	97.0	2.0	44.9	72.8	82.0	2.0	45.2	73.0	81.8
DAR	**1.0**	74.9	94.7	97.5	**1.0**	75.7	**95.4**	**97.9**	**2.0**	44.2	73.2	82.4	**2.0**	44.8	73.9	83.1
DAR+	**1.0**	76.9	94.9	97.4	**1.0**	**77.7**	**95.4**	**97.9**	**2.0**	47.4	75.3	83.8	**2.0**	**48.3**	**75.9**	**84.4**
DAR++	**1.0**	**77.3**	**95.3**	**97.7**	**1.0**	77.1	**95.4**	**97.9**	**2.0**	**47.8**	**75.9**	**84.3**	**2.0**	47.4	75.5	84.1

In addition, we also report the results with data augmentation in test time. By adding visual imagination description, the retrieval performance of DAR+ can be further improved with noticeable margins in terms of all the metrics than DAR. Furthermore, by adding both augmented image segments and visual imagination description during testing, i.e., "DAR++", image-to-recipe performance is slightly boosted than DAR+, but the recipe-to-image retrieval performance is inferior to DAR+. We believe the reason is the limitation of image segments, which is analyzed in Sec.4.4. Regarding CLP [9], the re-ranking step during inference is essential for achieving higher R@1 than our DAR and DAR+, yet it remains inferior to our DAR++. Moreover, our DAR, DAR+ and DAR++ consistently outperform CLP.

4.3 Ablation Study

We conduct the ablation study to validate the key components of our proposed method DAR, including the adapters in CLIP encoders, augmented data from foundation models and training losses. Unless otherwise specified, all the results are reported in 10K testing for evaluation with original image-recipe pairs.

Effect of adapter layers. Our DAR is built upon CLIP encoders with adapter layers for cross-modal recipe retrieval. As shown in Table 2, we firstly show the zero-shot cross-modal recipe retrieval performance using pre-trained CLIP model, where the recipe embedding is averaged over the title, ingredients and instructions embeddings. By adding adapter layers to either CLIP image encoder ϕ_{img} or recipe encoder ϕ_{rec}, the retrieval performance is boosted significantly than frozen CLIP. The results show that the introduced adapter layers to CLIP

Table 2: Ablation study for adapters in CLIP encoders for cross-modal recipe retrieval. &A refers to adding adapter layers to CLIP encoder. The operations are all added to the Zero-shot Retrieval model based on CLIP. All the results are reported in 10K testing with original recipe-image pairs for evaluation.

Model	Operation	image-to-recipe				recipe-to-image			
		medR	**R@1**	**R@5**	**R@10**	**medR**	**R@1**	**R@5**	**R@10**
Zero-shot Retrieval		17.0	14.9	32.7	42.4	20.2	12.6	29.7	39.5
	+ &A to ϕ_{img}	5.9	24.8	49.7	61.2	4.9	27.5	53.1	64.6
	+ &A to ϕ_{rec}	4.0	28.2	55.4	66.7	4.0	28.2	55.4	66.7
	+ &A in ϕ_{img}, ϕ_{rec}	2.0	39.2	68.7	79.0	2.0	40.7	69.7	79.8
Baseline	+ &A in ϕ_{img}, ϕ_{rec}, + L_{rec}	**2.0**	**42.3**	**71.8**	**81.5**	**2.0**	**43.4**	**72.8**	**82.3**

Table 3: Ablation study for augmented data. The operations are all added to the Baseline model. w/ (with) ϕ^* and $\phi_{\&\text{A}}$ represent the introduction of augmentation data with a frozen CLIP encoder or a CLIP encoder added to the adapter layers respectively. All the results are reported in 10K testing with original recipe-image pairs for evaluation.

Model	Operation	image-to-recipe				recipe-to-image			
		medR	**R@1**	**R@5**	**R@10**	**medR**	**R@1**	**R@5**	**R@10**
Baseline		2.0	42.3	71.8	81.5	2.0	43.4	72.8	82.3
	+ description w/ ϕ^*_{dec}	2.0	42.9	72.2	81.7	2.0	43.4	72.7	82.3
	+ description w/ $\phi_{\text{dec\&A}}$	2.0	43.7	72.6	82.0	2.0	44.7	73.9	83.1
	+ segments w/ ϕ^*_{seg}	2.0	43.0	72.3	81.9	2.0	43.6	73.0	82.6
	+ segments w/ $\phi_{\text{seg\&A}}$	2.0	42.1	71.2	81.0	2.0	43.0	72.3	81.9
	+ segment w/ ϕ^*_{seg}, description w/ $\phi_{\text{dec\&A}}$ (**DAR**)	**2.0**	**44.2**	**73.2**	**82.4**	**2.0**	**44.8**	**73.9**	**83.1**

Table 4: Performance comparison with different number of segments from SAM. All the results are reported in 10K testing with the DAR++ evaluation.

Number of segments	image-to-recipe				recipe-to-image			
	medR	**R@1**	**R@5**	**R@10**	**medR**	**R@1**	**R@5**	**R@10**
n = 1	2.0	47.7	75.5	84.1	2.0	46.2	74.6	83.7
n = 2	2.0	47.7	75.7	**84.3**	2.0	46.7	75.0	84.0
n = 4	**2.0**	**47.8**	**75.9**	**84.3**	**2.0**	**47.4**	**75.5**	**84.1**
n = 6	2.0	47.7	75.5	84.1	2.0	46.9	75.2	**84.1**

Table 5: Performance comparison between circle loss and triplet loss based on Baseline model setup.

Loss function	image-to-recipe				recipe-to-image			
	medR	**R@1**	**R@5**	**R@10**	**medR**	**R@1**	**R@5**	**R@10**
Triplet loss	2.4	36.9	66.9	77.7	2.0	37.3	67	77.8
Circle loss	**2.0**	**42.3**	**71.8**	**81.5**	**2.0**	**43.4**	**72.8**	**82.3**

model is effective for cross-modal recipe retrieval. Further gains can be obtained by employing adapter layers to both image and recipe encoders with recipe loss with both paired and unpaired recipes, which is regarded as our baseline model.
Effect of data augmentation by foundation models. We next examine the effectiveness of augmented data from recipe and image in Table 3. First, retrieval performance improvement can be observed by adding visual imagination description with frozen encoder (i.e., + description w/ ϕ^*_{dec}) compared to baseline. The performance is further boosted with adapter layers in the encoder (i.e., + description w/ $\phi_{\text{dec\&A}}$). The results show the effectiveness of visual imagination description for cross-modal recipe retrieval. Second, we further add image segments with frozen encoder (i.e., + segments w/ ϕ^*_{seg}) with performance gains. Nevertheless, injecting the adapter layers for segment encoder (i.e., + segments w/ $\phi_{\text{seg\&A}}$) is inferior to frozen one. We believe the reason is the image segments obtained from SAM still contain noise though we have filtered the outputs, which could do harm to the segment embeddings with adapter layers. The best performance is achieved by combining both visual imagination description with adapter layers in encoder and image segments with frozen encoder, i.e., our DAR.

In addition, we conduct experiments by instructing LLM to generate a summary of recipe for data augmentation as well. The results show that visual imagination description manages to outperform summary by 0.7% and 0.9% in image-to-recipe and recipe-to-image retrieval respectively. Finally, as listed in Table 4, we examine the performance with different numbers of segments from SAM. When segments are too few (n=1 or 2), key ingredients may be excluded. Conversely, a large number of segments (n=6) not only raises computational costs but also increases the likelihood of introducing noise. Based on the results, our DAR sets segment numbers to 4.
Effect of circle loss. As shown in Table 5, we compare the retrieval performance between triplet loss and circle loss under the same setting as our baseline model. Circle loss significantly outperforms the triplet loss across all the metrics.

4.4 Qualitative Analysis

Fig. 3 and Fig. 4 show the qualitative examples of recipe-to-image retrieval and image-to-recipe retrieval respectively. The results are reported based on DAR++, which includes extra augmented image segments and visual imagination description for testing.

First, we show the result of retrieving the corresponding image using recipe as query in Fig. 3. During inference, we can generate visual imagination description for each recipe query, which is combined as query to retrieve food images. In the first example, the query recipe is "feta chicken bake". We can see that all the Top-5 retrieval images are chicken, which are semantically and visually similar. Though our model DAR ranks the corresponding image in third position, our model DAR++ manages to improve the rank to first place. The reason is that our visual imagination description is capable of providing auxiliary visual-related information "topped with fresh parsley" and "red pepper flakes ..." to further distinguish the similar image returned by DAR but where the chicken is topped

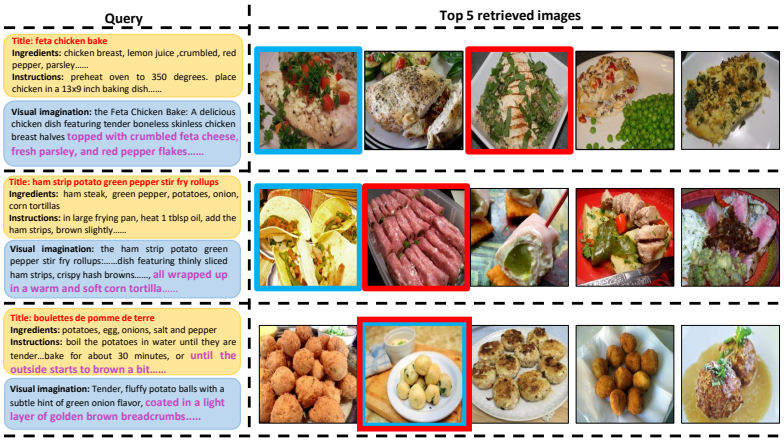

Fig. 3: Qualitative examples of recipe-to-image retrieval. The query in the left column shows the recipe and the corresponding visual imagination description produced by LLM. The right column shows the retrieved Top-5 food images from DAR++, where blue boxes represent the ground truth. We also highlight the Top-1 retrieved results of DAR with red bounding boxes.

with scallions. The same phenomenon can be observed in the second example as well where "rollups" is interpreted as "...wrapped up in corn tortilla" to correct the result that retrieves a picture of the rollups made of meat by DAR. As the visual imagination description contains visual cues that do not exist in recipes, the recipe-to-image retrieval could be more interpretable by incorporating visual imagination description in the retrieval system. In the third example "boulettes de pomme de terre", which means "potato dumplings", compared to DAR, our DAR++ is not able to rank higher than DAR, it is because the visual imagination provides misleading information "coated in a light layer of golden brown breadcrumbs" whereas in the recipe the color is described as "brown a bit" and there are no ingredients for "breadcrumbs".

As shown in Fig. 4, we present examples of Top-5 retrieved recipes using food image as query of DAR++. Following [25], the recipes are shown with word clouds for better visualization. Given a query image, our DAR++ employs SAM to obtain image segments, which are subsequently combined as query as well. Multiple segments are combined as one image for visualization purpose. It can be observed that there is a correspondence between the segments and key ingredients in the corresponding recipe. For instance, as the most notable ingredients, "ribs" in the first example and "beef", "carrots" in the second example can be found in the segment picture and the first ranked retrieved recipes. As a result, our method potentially provides better interpretability for cross-modal recipe retrieval with the image segments.

Limitation of image segments. Though image segments produced by SAM manage to boost the retrieval performance in both training and testing time, we

Fig. 4: Qualitative examples of image-to-recipe retrieval. The first two columns are image query and segments from SAM. The DAR++'s Top-5 retrieved recipes are represented by word clouds. The blue boxes represent the ground truth and red boxes represent the Top-1 retrieved recipe of DAR.

are aware of the limitations as well. Similar to [10, 43], we also note that SAM may miss small and irregular objects. Moreover, the "everything" mode samples a large number of point prompts for automatic segmentation, which results in SAM sometimes over-segmenting the fine-grained features of large and complex objects while neglecting their wholeness. In the third example in Fig. 4, the "cream" with a large area is missing from the segments, which causes DAR++ to focus on the "lemon", thus DAR++ fails to rank the ground-truth recipe in the first place. Please see supplementary material for more details.

5 Conclusion

We have presented a new data augmentation paradigm for cross-modal recipe retrieval via foundation models. Leveraging the augmented data from SAM and Llama2, we propose DAR framework, which achieves state-of-the-art performance on Recipe1M dataset. We introduce lightweight adapter layers in CLIP to encode the original and augmented recipe and image data. Furthermore, we propose multi-level circle loss to perform multi-objective optimization to regularize the common embedding space. Importantly, we demonstrate that augmented data can not only be beneficial during training, but also can be used for test-time augmentation. While encouraging, it is inevitable to generate imperfect image segments which could limit the boost of cross-modal retrieval performance, which will be our future work.

Acknowledgement. This work was partially supported by the National Natural Science Foundation of China (No. 62101524 and No. 62202439), and by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (No. MSS23C018). It was also supported by the advanced computing resources provided by the Supercomputing Center of the USTC.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Carvalho, M., Cadène, R., Picard, D., Soulier, L., Thome, N., Cord, M.: Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 35–44 (2018)
5. Chen, J., Zhu, B., Ngo, C.W., Chua, T.S., Jiang, Y.G.: A study of multi-task and region-wise deep learning for food ingredient recognition. IEEE Transactions on Image Processing **30**, 1514–1526 (2020)
6. Fu, H., Wu, R., Liu, C., Sun, J.: Mcen: Bridging cross-modal gap between cooking recipes and dish images with latent variable model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14570–14580 (2020)
7. Guerrero, R., Pham, H.X., Pavlovic, V.: Cross-modal retrieval and synthesis (x-mrs): Closing the modality gap in shared subspace learning. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3192–3201 (2021)
8. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
9. Huang, X., Liu, J., Zhang, Z., Xie, Y.: Improving cross-modal recipe retrieval with component-aware prompted clip embedding. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 529–537 (2023)
10. Ji, W., Li, J., Bi, Q., Li, W., Cheng, L.: Segment anything is not always perfect: An investigation of sam on different real-world applications. arXiv preprint arXiv:2304.05750 (2023)
11. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
13. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. Advances in neural information processing systems **28** (2015)
14. Li, J., Sun, J., Xu, X., Yu, W., Shen, F.: Cross-modal image-recipe retrieval via intra-and inter-modality hybrid fusion. In: Proceedings of the 2021 International Conference on Multimedia Retrieval. pp. 173–182 (2021)
15. Li, L., Li, M., Zan, Z., Xie, Q., Liu, J.: Multi-subspace implicit alignment for cross-modal retrieval on cooking recipes and food images. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 3211–3215 (2021)

16. Liu, G., Jiao, Y., Chen, J., Zhu, B., Jiang, Y.G.: From canteen food to daily meals: Generalizing food recognition to more practical scenarios. IEEE Transactions on Multimedia (2024)
17. Ma, J., Wang, B.: Segment anything in medical images. arXiv preprint arXiv:2304.12306 (2023)
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
19. Min, W., Jiang, S., Liu, L., Rui, Y., Jain, R.: A survey on food computing. ACM Computing Surveys (CSUR) **52**(5), 1–36 (2019)
20. Min, W., Wang, Z., Liu, Y., Luo, M., Kang, L., Wei, X., Wei, X., Jiang, S.: Large scale visual food recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
21. Ming, Z.Y., Chen, J., Cao, Y., Forde, C., Ngo, C.W., Chua, T.S.: Food photo recognition for dietary tracking: System and experiment. In: MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part II 24. pp. 129–141. Springer (2018)
22. Papadopoulos, D.P., Mora, E., Chepurko, N., Huang, K.W., Ofli, F., Torralba, A.: Learning program representations for food images and cooking recipes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16559–16569 (2022)
23. Pham, H.X., Guerrero, R., Pavlovic, V., Li, J.: Chef: cross-modal hierarchical embeddings for food domain retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2423–2430 (2021)
24. Sahoo, D., Hao, W., Ke, S., Xiongwei, W., Le, H., Achananuparp, P., Lim, E.P., Hoi, S.C.: Foodai: Food image recognition via deep learning for smart food logging. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2260–2268 (2019)
25. Salvador, A., Gundogdu, E., Bazzani, L., Donoser, M.: Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15475–15484 (2021)
26. Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., Torralba, A.: Learning cross-modal embeddings for cooking recipes and food images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3020–3028 (2017)
27. Shukor, M., Couairon, G., Grechka, A., Cord, M.: Transformer decoders with multimodal regularization for cross-modal food retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4567–4578 (2022)
28. Shukor, M., Thome, N., Cord, M.: Vision and structured-language pretraining for cross-modal food retrieval. Available at SSRN 4511116 (2023)
29. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6398–6407 (2020)
30. Sung, Y.L., Cho, J., Bansal, M.: Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5227–5237 (2022)
31. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

32. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
33. Trabucco, B., Doherty, K., Gurinas, M., Salakhutdinov, R.: Effective data augmentation with diffusion models. arXiv preprint arXiv:2302.07944 (2023)
34. Voutharoja, B.P., Wang, P., Wang, L., Guan, V.: Malm: Mask augmentation based local matching for food-recipe retrieval. arXiv preprint arXiv:2305.11327 (2023)
35. Wahed, M., Zhou, X., Yu, T., Lourentzou, I.: Fine-grained alignment for cross-modal recipe retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5584–5593 (2024)
36. Wang, H., Lin, G., Hoi, S., Miao, C.: Paired cross-modal data augmentation for fine-grained image-to-text retrieval. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5517–5526 (2022)
37. Wang, H., Lin, G., Hoi, S.C., Miao, C.: Learning structural representations for recipe generation and food retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3363–3377 (2022)
38. Wang, H., Sahoo, D., Liu, C., Lim, E.p., Hoi, S.C.: Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11572–11581 (2019)
39. Wang, W., Duan, L.Y., Jiang, H., Jing, P., Song, X., Nie, L.: Market2dish: Health-aware food recommendation. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **17**(1), 1–19 (2021)
40. Whitehouse, C., Choudhury, M., Aji, A.F.: Llm-powered data augmentation for enhanced crosslingual performance. arXiv preprint arXiv:2305.14288 (2023)
41. Xie, Z., Liu, L., Wu, Y., Li, L., Zhong, L.: Learning tfidf enhanced joint embedding for recipe-image cross-modal retrieval service. IEEE Transactions on Services Computing (2021)
42. Zan, Z., Li, L., Liu, J., Zhou, D.: Sentence-based and noise-robust cross-modal retrieval on cooking recipes and food images. In: Proceedings of the 2020 International Conference on Multimedia Retrieval. pp. 117–125 (2020)
43. Zhang, C., Liu, L., Cui, Y., Huang, G., Lin, W., Yang, Y., Hu, Y.: A comprehensive survey on segment anything model for vision and beyond. arXiv preprint arXiv:2305.08196 (2023)
44. Zhang, Y., Zhou, T., Wang, S., Liang, P., Zhang, Y., Chen, D.Z.: Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 129–139. Springer (2023)
45. Zhu, B., Ngo, C.W., Chan, W.K.: Learning from web recipe-image pairs for food recognition: Problem, baselines and performance. IEEE Transactions on Multimedia **24**, 1175–1185 (2021)
46. Zhu, B., Ngo, C.W., Chen, J.j.: Cross-domain cross-modal food transfer. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 3762–3770 (2020)
47. Zhu, B., Ngo, C.W., Chen, J., Hao, Y.: R2gan: Cross-modal recipe retrieval with generative adversarial network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11477–11486 (2019)
48. Zou, Z., Zhu, X., Zhu, Q., Liu, Y., Zhu, L.: Creamy: Cross-modal recipe retrieval by avoiding matching imperfectly. IEEE Access (2024)