foods

MDPI

Article

Disambiguity and Alignment: An Effective Multi-Modal Alignment Method for Cross-Modal Recipe Retrieval

Zhuoyang Zou, Xinghui Zhu, Qinying Zhu, Hongyan Zhang and Lei Zhu *

College of Information and Intelligence, Hunan Agricultural University, Changsha 410128, China; zzy@stu.hunau.edu.cn (Z.Z.); zhuxh@hunau.edu.cn (X.Z.); zhuqy@stu.hunau.edu.cn (Q.Z.); hongyan_zhang@hunau.edu.cn (H.Z.)
* Correspondence: leizhu@hunau.edu.cn

Abstract: As a prominent topic in food computing, cross-modal recipe retrieval has garnered substantial attention. However, the semantic alignment across food images and recipes cannot be further enhanced due to the lack of intra-modal alignment in existing solutions. Additionally, a critical issue named food image ambiguity is overlooked, which disrupts the convergence of models. To these ends, we propose a novel Multi-Modal Alignment Method for Cross-Modal Recipe Retrieval (MMACMR). To consider inter-modal and intra-modal alignment together, this method measures the ambiguous food image similarity under the guidance of their corresponding recipes. Additionally, we enhance recipe semantic representation learning by involving a cross-attention module between ingredients and instructions, which is effective in supporting food image similarity measurement. We conduct experiments on the challenging public dataset Recipe1M; as a result, our method outperforms several state-of-the-art methods in commonly used evaluation criteria.

Keywords: cross-modal recipe retrieval; multi-modal alignment; food image ambiguity; deep learning

Citation: Zou, Z.; Zhu, X.; Zhu, Q.; Zhang, H.; Zhu, L. Disambiguity and Alignment: An Effective Multi-Modal Alignment Method for Cross-Modal Recipe Retrieval. *Foods* **2024**, *13*, 1628. https://doi.org/10.3390/foods13111628

Academic Editors: Zhiming Guo and Weiqing Min

Received: 30 April 2024
Revised: 20 May 2024
Accepted: 20 May 2024
Published: 23 May 2024

1. Introduction

With rising awareness of health and sustainability, issues such as food safety [1,2] and nutrition [3] have gained unprecedented attention. Food computing [4–8] plays a crucial role in promoting healthier lifestyles, mitigating food waste, and enhancing both the quality and safety of food products. Cross-modal recipe retrieval [9,10] is one of the hot topics in food computing, leveraging artificial intelligence (AI) [11,12] which aims to retrieve the corresponding recipes by queries of food images or vice versa. In this task, food images depict finished dishes, while recipes comprise text encompassing three key components: a title, a list of ingredients, and detailed instructions outlining the cooking process.

The principal challenge in cross-modal recipe retrieval lies in mitigating the inherent heterogeneity between two distinct modalities: the recipes and the food images. To solve this challenging task, numerous studies have delved into additional interactions between the two modalities. For instance, Refs. [13–15] tried to learn the consistent feature distribution of food images and recipe texts. Refs. [16–19] boosted the interaction between two modalities through cross-modal attention. Ref. [20] employed a joint transformer encoder to promote alignment. Due to the complexity of image–recipe pairs, many existing studies focused on exploiting the latent semantic information within a modality. As typical studies, refs. [21–25] aimed to focus on the crucial term within recipes, while others [26–28] attempted to capture the salient objects or regions from food images to improve the cross-modal similarity measurement. Due to the complexity of the textual structure in recipes, other researchers [29–33] investigated the interaction among the title, ingredients, and instructions to excavate important semantics. Furthermore, some studies introduced diverse augmentation mechanisms to enhance cross-modal feature representations. For example, Refs. [34–38] employed various Generative Adversarial Networks (GANs) to reconstruct information from food images and recipes to bridge the heterogeneity gap across modalities, while refs. [39,40]

leveraged multilingual translation to enrich the recipe information. Crowdsourcing strategy is also used to construct program representations of recipes [41]. Thanks to the flourishing development of visual language pre-training recently, some pioneers [42–45] have further embedded complex semantic relationship information into common feature subspace by leveraging the pre-trained Contrastive Language–Image Pre-Training model (CLIP).

Despite the significant progress made so far, there is still room for further improvement in semantic distribution alignment across food images and recipes. To be specific, the prevailing efforts [18,29,30] concentrate on exploring inter-modal semantic alignment using conventional metric learning strategies, such as triplet loss. As shown in Figure 1a, the conventional metric learning strategy is devoted to reducing the distance between positive image–recipe pairs (the circles and squares with the same color) and enlarging the distance between negative samples (the circles and the gray squares) and is proficient in learning similarity relations within each image–recipe pair. However, semantic relations exist not only within each image–recipe pair, but extensively between different pairs. For example, the two image–recipe pairs in Figure 1a belong to the same food (chilli sauce), indicating strong semantic relations (highlighted by red lines) between the two food images as well as the two recipes. The conventional metric learning method (e.g., triplet loss), however, fails to capture this relation information. To be sure, there are lots of image pairs belonging to the same food in practical time. This situation indicates that effectively enhancing intra-modal semantic alignment is significant for improving recipe retrieval performance.

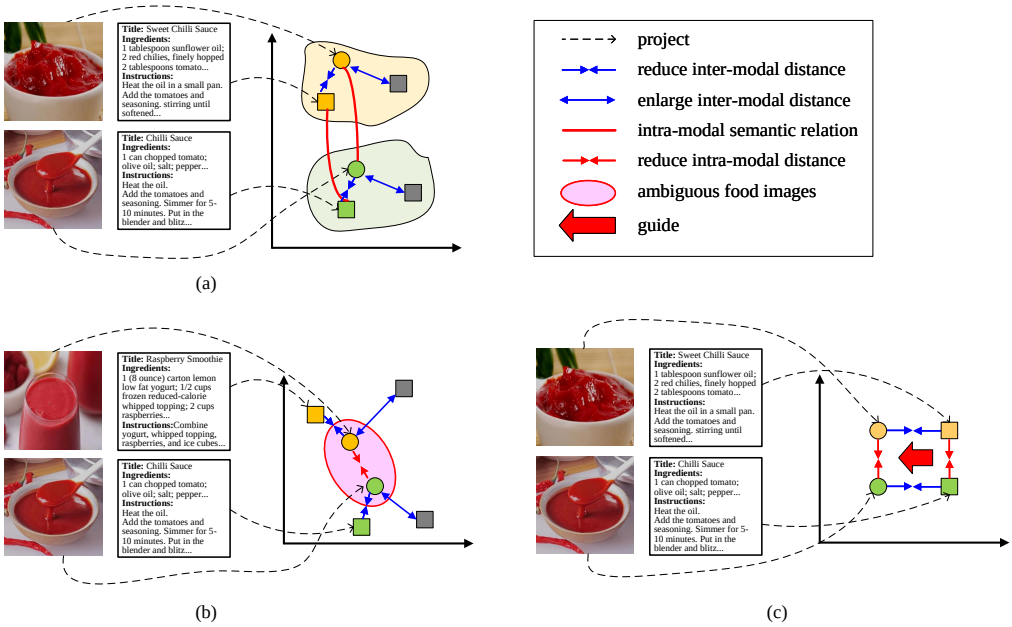

Figure 1. The demonstration of multi-modal alignment schemes for cross-modal recipe retrieval. (**a**) The prevailing learning strategy that ignores intra-modal alignment. (**b**) The food image ambiguity issue. (**c**) Our solution (negative samples are omitted). Circles represent images, and squares represent recipes. Shapes of the same color indicate positive pairs, while gray shapes indicate negative samples.

For this purpose, a straightforward method applied in lots of cross-modal retrieval tasks [46–48] is to utilize metric learning or contrastive learning strategy within each modality. However, there is a non-trivial issue, i.e., *food image ambiguity*, in cross-modal recipe retrieval that has not been considered. Specifically, foods that look similar may be made from quite different materials and via different preparation methods. Thus, these similar food images correspond to significantly distinct recipes. For example, in Figure 1b, the top food image is a cup of raspberry smoothie, and the bottom one is a bowl of chilli sauce. Theses two foods are visually similar to each other, yet they are crafted from distinct ingredients and have undergone quite different instructions. Unfortunately, existing methods embed their semantics from the two modalities to the common subspace

independently. Due to the resemblance in appearance, these two images will be close in the common space, while their corresponding recipes will not. This leads to a dilemma; embeddings that have a large similarity (small distance) in the visual modality may have a small similarity (large distance) in the text modality. As a result, the two modalities are hard to align with each other, and the models are difficult to converge, which reduces the accuracy of retrieval. Stumped by this stand-out drawback, we observe that recipes are the more reliable modality. In other words, foods prepared using similar recipes will have similar visual appearances. Therefore, this study aims to answer the following two questions:

- **Q1:** How can we measure the similarity between ambiguous food images guided by their corresponding recipes?
- **Q2:** How can we further improve the fine-grained semantic alignment between ingredients and instructions within each recipe to support food image similarity measurement?

To this end, we propose a novel cross-modal recipe retrieval method called the **M**ulti-**M**odal **A**lignment Method for **C**ross-**M**odal **R**ecipe Retrieval (**MMACMR**). To answer **Q1**, we design a novel strategy, the **M**ulti-Modal **D**isambiguity and **A**lignment strategy (**MDA** for short), which calculates the intra-modal similarity of recipes and guides the distances between corresponding images. As shown in Figure 1c, the green square is a recipe (chilli sauce) similar to the one shown by the orange square (sweet chilli sauce). Our MDA strategy attempts to pull them close and guide the distance between the green and orange circles (their corresponding images). For **Q2**, considering ingredients play a significant role within instructions, we introduce sentence-level cross-attention to focus on important ingredients in the instructions and further enhance the representations of recipes. In a nutshell, this work is a pioneering effort to further narrow cross-modal heterogeneity between food images and recipes by considering both multi-modal (inter-modal and intra-modal) alignment while mitigating the impact of food image ambiguity.

To sum up, the main contributions of this article are four fold:

- We propose a novel framework called MMACMR which addresses the problem of ambiguous food images in cross-modal recipe retrieval;
- We introduce a novel deep learning strategy named MDA which promotes the alignment of two modalities without adding new parameters;
- We enhance the representation of recipes by focusing on important ingredients within instructions at the sentence level;
- We conduct extensive experiments on the challenging dataset Recipe1M. The results demonstrate that the proposed technique outperforms several state-of-the-art methods.

The remainder of this article is organized as follows. The technical details and specific learning process are outlined in Section 2, the experimental particulars are discussed in Section 3, and we conclude the paper in Section 4.

2. Method

In this section, we first present the notations involved in this paper and provide the problem formulation for cross-modal recipe retrieval in Section 2.1. Then, we elaborate on the technique details of our method MMACMR, including the models in Section 2.2, the strategy in Section 2.3, and the algorithm in Section 2.4.

2.1. Notations and Problem Formulations

2.1.1. Notations

Without loss of generality, we denote sets as uppercase, handwritten, bold letters (e.g., $\mathcal{D}$) and matrices as uppercase letters (e.g., $\mathbf{W}$). The i-th row of $\mathbf{W}$ is denoted by $\mathbf{W}_i$, and the element found in the j-th column of i-th row in $\mathbf{W}$ is denoted as $\mathbf{W}_{ij}$. We represent the transpose of a matrix $\mathbf{W}$ as $\mathbf{W}^\top$. Notation $\|\cdot\|_2$ denotes the L2 norm of a matrix. We

use $softmax(\cdot)$ to represent the softmax function. To ease reading, we summarize the frequently used notations in Table 1.

Table 1. Summary of notations.

Notation	Definition
$\mathcal{D}$	A cross-modal recipe dataset
$\mathbf{X}_i^v$	The food image of the i-th pair
$\mathbf{X}_i^r$	The recipe of the i-th pair
$\mathbf{X}_i^{tit}$	The title of the recipe $\mathbf{X}_i^r$
$\mathbf{X}_i^{ing}$	The ingredients of the recipe $\mathbf{X}_i^r$
$\mathbf{X}_i^{ins}$	The instructions of the recipe $\mathbf{X}_i^r$
$\mathbf{E}_i^{tit}$	The embedding of the title in a recipe $\mathbf{X}_i^r$
$\mathbf{E}_i^{ing}$	The embedding of the ingredients in a recipe $\mathbf{X}_i^r$
$\mathbf{E}_i^{ins}$	The embedding of the instructions in a recipe $\mathbf{X}_i^r$
$\mathbf{R}$	The recipe embedding
$\mathbf{V}$	The food image embedding
f^r	The recipe encoder
f^v	The image encoder
$\boldsymbol{\theta}^r$	The parameters of recipe encoder
$\boldsymbol{\theta}^v$	The parameters of image encoder
$\mathcal{L}_{tri}$	The N-pairs triplet loss function
$\mathcal{L}_{RGI}$	The RGI loss function

2.1.2. Problem Formulations

Let $\mathcal{D} = \{\mathbf{X}_i^v, \mathbf{X}_i^r\}_{i=1}^n$ denote a cross-modal recipe dataset comprising n image–recipe pairs, where $\mathbf{X}_i^v$ and $\mathbf{X}_i^r = \langle \mathbf{X}_i^{tit}, \mathbf{X}_i^{ing}, \mathbf{X}_i^{ins} \rangle$ represent the food image and recipe of the i-th pair, respectively. $\mathbf{X}_i^{tit}$, $\mathbf{X}_i^{ing}$, and $\mathbf{X}_i^{ins}$ denote the title, list of ingredients, and list of instructions of the recipe, respectively. Note that each title comprises a single sentence, while both ingredients and instructions consist of several sentences. Given a recipe $\mathbf{X}_i^r$ as a query, cross-modal recipe retrieval aims to search for the most similar food image $\mathbf{X}_i^v$ from this dataset $\mathcal{D}$, or vise versa. To enhance consistent feature distribution alignment across food images and recipes, we attempt to optimize an improved recipe encoder $\mathbf{R} = f^r(\mathbf{X}^r; \boldsymbol{\theta}^r)$ and an image encoder $\mathbf{V} = f^r(\mathbf{X}^v; \boldsymbol{\theta}^v)$ under the guidance of a novel learning strategy dubbed MDA. This strategy integrates two losses: an N-pairs triplet loss $\mathcal{L}_{tri}$ to focus on inter-modal semantic alignment and an RGI loss $\mathcal{L}_{RGI}$ to focus on semantic consistency within the same modality. By considering both inter-modal and intra-modal alignment, this approach effectively avoids the harmful effects of food image ambiguity. Therefore, the objective function is formulated as follows:

$$(\hat{\boldsymbol{\theta}}^v, \hat{\boldsymbol{\theta}}^r) = \arg\min_{\boldsymbol{\theta}^v, \boldsymbol{\theta}^r} (\mathcal{L}_{tri} + \lambda \mathcal{L}_{RGI}), \tag{1}$$

where $\boldsymbol{\theta}^v$ and $\boldsymbol{\theta}^r$ are two learnable parameter vectors for image and recipe encoders, and λ is a pre-defined balance parameter.

2.2. Framework Overview

An overview of our method MMACMR is depicted in Figure 2. Following prevailing solutions [29,49], the backbone of MMACMR comprises an image encoder $f^v(\cdot; \boldsymbol{\theta}^v)$ and a recipe encoder $f^r(\cdot; \boldsymbol{\theta}^r)$ which project food images and recipes into a common feature subspace. In this subspace, the cross-modal features can be aligned effectively so that the similarity between images and recipes can be measured with accuracy. Below, we provide details of them.

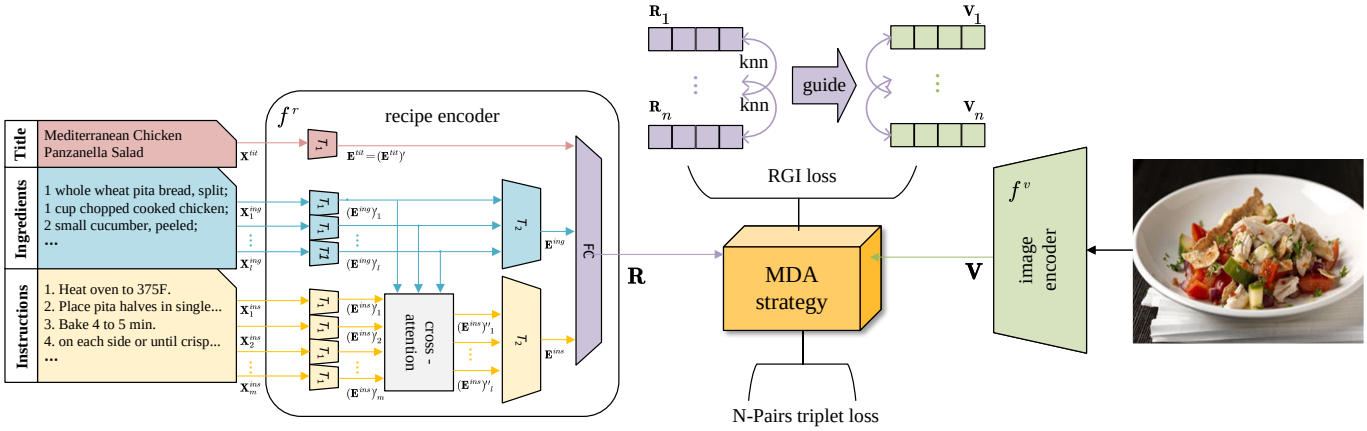

Figure 2. The framework of MMACMR, which comprises two branches of modality encoder, f^r for recipe texts and f^v for food images, along with the MDA strategy.

2.2.1. Image Encoder

To fully capture the global semantic relations between fine-grained features in the content of each food image, we adopt the base-size model of Vision Transformer (ViT-B) [50] as the image encoder $f^v(\cdot;\boldsymbol{\theta}^v)$. It is initialized with the weights pre-trained on ImageNet [51] and fine-tuned on the cross-modal recipe dataset. Given a food image $\mathbf{X}_i^v$, the embedding of $\mathbf{X}_i^v$ is denoted as $\mathbf{V}_i = f^v(\mathbf{X}_i^v;\boldsymbol{\theta}^v)$.

2.2.2. Improved Recipe Encoder

To focus on the consistent fine-grained semantics between ingredients and instructions, we improve the hierarchical transformer-based recipe encoder [29]. This encoder consists of two levels of transformers, denoted as T_1 and T_2, with identical architectures. The first level encodes the title $\mathbf{X}_i^{tit}$, ingredients $\mathbf{X}_i^{ing}$, and instructions $\mathbf{X}_i^{ins}$ at word level and then outputs their sentence-level embeddings, while the second level encoder receives the sentence-level embeddings of ingredients and instructions and produces component-level embeddings. Such a widely adopted recipe embedding scheme unfortunately overlooks a fundamental yet crucial rule in a recipe; the instructions are steps tailored to the ingredients, with the ingredients playing a determining role in shaping the instructions to some extent. To obey this rule, we plug a cross-attention module for instructions between the two transformers for two purposes: (1) to focus on the salient ingredient and (2) to highlight semantic relationships between ingredients and instructions.

Specifically, given a recipe set $\{\mathbf{X}_i^r\}_{i=1}^n$, as shown in Figure 2, the first level module T_1 receives the word-level tokens of the three components separately and outputs the average embeddings of every sentence of every components, denoted as $((\mathbf{E}_i^{tit})', (\mathbf{E}_i^{ing})', (\mathbf{E}_i^{ins})') = T_1(\mathbf{X}_i^{tit}, \mathbf{X}_i^{ing}, \mathbf{X}_i^{ins})$, where $(\mathbf{E}_i^{ing})' = \{(\mathbf{E}_i^{ing})_k'\}_{k=1}^l$, $(\mathbf{E}_i^{ins})' = \{(\mathbf{E}_i^{ins})_k'\}_{k=1}^m$. To highlight the effect of ingredients to instructions at sentence level and enhance the semantic relationship learning, cross-attention is carried out between $(\mathbf{E}_i^{ing})'$ and $(\mathbf{E}_i^{ins})$. Firstly, within a recipe, we construct an affinity matrix $\mathbf{W}$ as an attention map:

$$\mathbf{W} = softmax\left(\frac{((\mathbf{E}_i^{ing})'\mathbf{W}^{ing})((\mathbf{E}_i^{ins})'\mathbf{W}^{ins})^\top}{d}\right), \tag{2}$$

where $(\mathbf{E}_i^{ing})' \in R^{l\times d}$, $(\mathbf{E}_i^{ins})' \in R^{m\times d}$, and d is the dimension of each ingredient and each instruction. $\mathbf{W}^{ing} \in R^{d\times d}$ and $\mathbf{W}^{ins} \in R^{d\times d}$ are learnable weight matrices. Each element $\mathbf{W}_{jk}$ means the normalized correlation between the j-th ingredient and the k-th instruction. Thereby, the embedding of instructions can be enhanced by focusing on the consistent semantics between instructions and ingredients as follows:

$$(\mathbf{E}_i^{ins})'' = \mathbf{W}(\mathbf{E}_i^{ins})'. \tag{3}$$

After the second-level processing, we obtain the component-level features of title, ingredients, and instructions: $\mathbf{E}_i^{tit}$, $\mathbf{E}_i^{ing} = T_2(\mathbf{E}_i^{ing})'$ and $\mathbf{E}_i^{ins} = T_2(\mathbf{E}_i^{ins})''$. Finally, these three component embeddings are concatenated and fed into a linear layer; thus, we obtain the final recipe feature, $\mathbf{R}_i = FC([\mathbf{E}_i^{tit}; \mathbf{E}_i^{ing}; \mathbf{E}_i^{ins}]; \boldsymbol{\theta}^l)$, where $FC(\cdot; \boldsymbol{\theta}^l)$ is a linear layer, $\boldsymbol{\theta}^l$ are the parameters of it, and symbol $[\cdot;\cdot;\cdot]$ denotes the concatenation operation.

2.3. Multi-Modal Disambiguity and Alignment

To enhance the consistent feature distribution alignment across food images and recipes, we extend the prevailing learning scheme (only inter-modal metric learning, e.g., triplet loss) by considering both inter- and intra-modal alignment. To do so, we employ N-pairs triplet loss to realize inter-modal alignment within each batch, while we propose a novel RGI loss to steer the model towards capturing intra-modal consistent semantics effectively by preventing the misrecognition of ambiguous food images.

2.3.1. Inter-Modal Alignment: N-Pairs Triplet Loss

Given an anchor food image $\mathbf{V}_i$, a positive recipe $\mathbf{R}_i^+$, and a negative recipe $\mathbf{R}_j^-$, where $i \neq j$, the N-pairs triplet loss for visual modality can be defined as follows:

$$\mathcal{L}_{tri}^v = \sum_{|V|}[(\mathbf{V}_i, \mathbf{R}_i^+, \mathbf{R}_j^-) = (S(\mathbf{V}_i, \mathbf{R}_j^-) - S(\mathbf{V}_i, \mathbf{R}_i^+) + m)]_+, \tag{4}$$

where $[\cdot]_+ = max(0, \cdot)$, $S(\cdot)$ is the similarity function (we use cosine similarity here), $|V|$ is the number of the image sample in the batch, and m is a pre-defined margin (we set $m = 0.3$ in this work). Similarly, the N-pairs triplet loss for text modality can be written in the same way. Consequently, we formulate the whole N-pairs triplet loss as follows:

$$\mathcal{L}_{tri} = \mathcal{L}_{tri}^v + \mathcal{L}_{tri}^r. \tag{5}$$

2.3.2. Intra-Modal Alignment with Disambiguity: RGI Loss

As discussed above, N-pairs triplet loss is a satisfactory scheme for reducing heterogeneity between images and recipes. Using it within each modality, however, is far from a suitable intra-modal alignment solution due to the disturbance of food image ambiguity. Nor is this all; the prevailing recipe retrieval approaches [18,29] only consider cross-modal similarity measurement, which narrows the distance between anchor and positive samples while enlarging the distances between the anchor and negative samples. Such a limitation, on the one hand, leads to a discrepancy between the two modalities, making it difficult for model convergence. On the other hand, it is easy to match one of the ambiguous images, resulting in low retrieval performance.

Fortunately, recipes, or, more rigorously, text, are the more reliable modality owing to their ability to abstract semantic expression word by word. Thus, inspired by [52], we design a novel learning strategy termed RGI loss which chooses the similarity relations between recipes as guidance to determine the relations between corresponding food images. Specifically, if we assume that $< \mathbf{R}_i, \mathbf{R}_j >$ is a recipe pair in a batch, we aim to preserve the similarity relation for it and project this relation to the corresponding image pair $< \mathbf{V}_i, \mathbf{V}_j >$. Given a recipe $\mathbf{R}_i$, we first rank other recipes in this batch by the similarity to $\mathbf{R}_i$ using the K-nearest neighbors (KNN) algorithm [53]. From the ranked recipes, we select the nearest neighbor as the positive sample $\mathbf{R}_j^+$ and a randomly selected recipe that is not among the top 10 neighbors as the negative recipe $\mathbf{R}_k^-$, $i \neq j \neq k$. Inspired by the angular loss [54], our RGI loss for text modality is defined as follows:

$$\mathcal{L}_{RGI}^r = \left[\left\| \mathbf{R}_i - \mathbf{R}_j^+ \right\|_2^2 - 4\tan^2\alpha \left\| \mathbf{R}_k^- - C_i \right\|_2^2 \right]_+, \tag{6}$$

$$C_i = \frac{\mathbf{R}_i + \mathbf{R}_j^+}{2}, \tag{7}$$

where $\tan^2 \alpha = 1$ is a pre-defined upper bound. For the visual modality, we no longer compute the KNN for images, while we adopt the rank of the neighbors of corresponding recipes directly. The RGI loss for the visual modality is defined in the same way:

$$\mathcal{L}_{RGI}^{v} = \left[\left\| \mathbf{V}_i - \mathbf{V}_j^{+} \right\|_2^2 - 4\tan^2\alpha \left\| \mathbf{V}_k^{-} - C_i \right\|_2^2 \right]_+, \tag{8}$$

$$C_i = \frac{\mathbf{V}_i + \mathbf{V}_j^{+}}{2}, \tag{9}$$

where $\tan^2 \alpha = 1$ is a pre-defined upper bound. Note that the indices of the visual modality are the same as the text modality. Thus, the entire RGI loss is formulated as follows:

$$\mathcal{L}_{RGI} = \lambda_1 \mathcal{L}_{GRI}^{r} + \lambda_2 \mathcal{L}_{RGI}^{v}, \tag{10}$$

where λ_1 and λ_2 are hyper-parameters for adjusting the relation projection.

2.3.3. Total Loss

Finally, the total loss can be written as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{tri} + \lambda \mathcal{L}_{RGI}, \tag{11}$$

where λ is a balance hyper-parameter for adjusting the performance of the two loss functions.

2.4. Optimization

Our method undergoes end-to-end optimization. The optimization procedure is outlined in Algorithm 1.

Algorithm 1 Optimization procedure of MMACMR

Input: cross-modal recipe dataset $\mathcal{D} = \{\mathbf{X}_i^v, \mathbf{X}_j^r\}_{i,j=1}^n$, number of epoch T.
Output: parameters $\boldsymbol{\theta}^v$, $\boldsymbol{\theta}^r$ of modality encoders.

```
Initialize θ^v, θ^r;
for t = 1 to T do
  repeat
    Compute embeddings V and R;
    for i = 1 to n do
      Calculate Equation (5)
      Rank the recipes neighbors via KNN algorithm;
      Rank the images neighbors follow recipes;
      Calculate Equation (10);
    end for
    Update the parameters θ^v, θ^r by Equation (11) via gradient descent algorithm.
  until Convergence
end for
```

3. Experiments and Discussion

This section presents extensive experiments conducted to assess our method's performance. We begin by introducing the experiment settings, followed by a detailed discussion of the experimental results.

3.1. Experiment Settings

3.1.1. Dataset

We implement experiments on the Recipe1M [9] dataset, which is by far the largest public multi-modal recipe dataset available. Recipe1M comprises over 1 million cooking recipe texts and 800 K food images which are collected from more than 24 popular cooking

websites. We adhere to the official splits for data, with 238,399 image–recipe pairs allocated for training, 51,119 pairs for validation, and 51,303 pairs for testing.

3.1.2. Baselines

We benchmark our approach against the state-of-the-art baselines below:

- CCA [9] stands for Canonical Correlation Analysis, a classical statistical method used to learn a joint embedding space;
- JE [9] was the first to conduct the cross-modal recipe retrieval task on the Recipe1M dataset. It uses a joint encoder and a classifier to learn the information from food images and recipes;
- AdaMin [10] combines the retrieval loss and classifies the loss to improve the robustness of models and proposes a novel strategy to mine the significant triplets;
- R2GAN [35] promotes the modality alignment by employing a GAN mechanism equipped with two discriminators and one generator;
- MCEN [14] bridges the semantic gap between the two modalities using stochastic latent variable models;
- SN [16] employs three attention mechanisms on three components of recipes to capture the relationship between sentences;
- SCAN [13] introduces semantic consistency loss to regularize the representations of images and recipes;
- HF-ICMA [20] exploits the global and local similarity between the two modalities by considering inter- and intra-modal fusion;
- SEJE [22] constructs a two-phase feature framework and divides the processes of data pre-processing and model training to extract additional semantic information;
- M-SIA [17] argues that multiple aspects in recipes are related to multiple regions in food images and leverages multi-head attention to bridge them;
- X-MRS [39] augments recipe representations by utilizing multilingual translation;
- LCWF-GI [31] employs latent weight factors to fuse the three components of recipes by considering their complex interaction;
- H-T [29] captures the latent semantic information in recipes by applying self-supervised loss to push components sourced from the same close recipe;
- LMF-CSF [30] introduces a low-rank fusion strategy to combine the components in recipes and generate superior representations.

3.1.3. Evaluation Criteria

Similar to the majority of previous studies [9,29,44], we sample 1 K and 10 K image–recipe pairs from the test partition and assess the retrieval performance for image-to-recipe and and recipe-to-image tasks using median rank (MedR) and recall rate at top k (R@k). Among these metrics, MedR represents the median index of the retrieved samples for each query, measuring the ability of models to understand the semantic correlation between two modalities and the accuracy of retrieval. A lower MedR value indicates better performance. R@k indicates that the percentage of the ground truth index is among the first k retrieved samples, which is also known as sensitivity or the true positive rate, measuring the ability of models to correctly identify all relevant instances. A higher R@k value indicates better performance. Here, we evaluate the top 1 (R@1), top 5 (R@5), and top 10 (R@10). By using these two metrics, we can evaluate the comprehensive performance of the models. Every evaluation is repeated 10 times, and the mean results are returned.

3.1.4. Implementation Details

In line with prior research [49], we use food images with a depth of three channels in the RGB color space. All the images in our experiments are resized to 256 pixels in their shorter dimension and then cropped to 224 $\times$ 224 pixels. The image encoder utilizes a pre-trained ViT-based model, yielding an output size of 1024. Regarding recipes, sentences in three components are truncated to a maximum length of 15, and every ingredients

or instructions list has a maximum of 20 sentences. Each transformer in the hierarchical transformer recipe encoder comprises two layers, and each layer has four attention heads. Every component in the recipes is embedded as 512 dimensions, and the final embedding of a recipe is output as 1024 dimensions. The model is trained utilizing the Adam optimizer, the batch size is set as 128, and the learning rate is $\eta = 10^{-4}$. The balance parameters $\lambda_1 = 0.09$, $\lambda_2 = 0.1$, and $\lambda = 0.01$.

3.1.5. Experimental Environment

Our experiments are conducted using Python 3.7 with the PyTorch 1.31.1 framework. We utilize a deep learning workstation equipped with an Intel(R) Core i9-12900K 3.9 GHz processor, 128 GB of RAM, 1 TB SSD, and 2 TB HDD storage. The workstation runs on the Ubuntu-22.04.1 operating system and is powered by two NVIDIA GeForce RTX 3090Ti GPUs (NVIDIA, Palo Alto, CA, USA).

3.2. Comparison with State-of-the-Art Methods

We compare the performance of our method with the baselines mentioned above. The results are reported in Table 2. It is easy to see that MMACMR is superior to the best results of existing works using all the metrics. Concretely, our method achieves a 3.3, 1.1, 0.6 R{1, 5, 10} improvement for image to recipe and a 3.7, 1.2, 0.7 R{1, 5, 10} improvement for recipe to image in the 1 K size compared to the SOTA method LMF-CSF [30] and achieves a 3.5, 3.1, 2.7 R{1, 5, 10} improvement for image to recipe and a 4.0, 3.1, 2.8 R{1, 5, 10} improvement for recipe to image in the 10 K size compared to the SOTA method LMF-CSF [30]. In addition, the MedR of our method in the 10 K size dataset decreases to 2.1 for image to recipe and 2.2 for recipe to image compared to 3.0 in LMF-CSF [30]. These results demonstrate the effectiveness of our MMACMR. In other words, our approach to addressing the questions mentioned above is effective for cross-modal recipe retrieval.

Table 2. Comparison with SOTA methods. MedR(↓) and R@k(↑) in 1 K and 10 K size. The best results are marked in bold font.

	Methods	**Image to Recipe**				**Recipe to Image**			
		MedR	**R@1**	**R@5**	**R@10**	**MedR**	**R@1**	**R@5**	**R@10**
1 K	CCA [9]	15.7	14.0	32.0	43.0	24.8	9.0	24.0	35.0
	JE [9]	5.2	24.0	51.0	65.0	5.1	25.0	52.0	65.0
	AdaMin [10]	2.0	39.8	69.0	77.4	2.0	40.2	68.1	78.7
	R2GAN [35]	2.0	39.1	71.0	81.7	2.0	40.6	72.6	83.3
	MCEN [14]	2.0	48.2	75.8	83.6	1.9	48.4	76.1	83.7
	ACME [34]	1.0	51.8	80.2	87.5	1.0	52.8	80.2	87.6
	SN [16]	1.0	52.7	81.7	88.9	1.0	54.1	81.8	88.9
	SCAN [13]	1.0	54.0	81.7	88.8	1.0	54.9	81.9	89.0
	HF-ICMA [20]	1.0	55.1	86.7	92.4	1.0	56.8	87.5	93.0
	SEJE [22]	1.0	58.1	85.8	92.2	1.0	58.5	86.2	92.3
	M-SIA [17]	1.0	59.3	86.3	92.6	1.0	59.8	86.7	92.8
	X-MRS [39]	1.0	64.0	88.3	92.6	1.0	63.9	87.6	92.6
	H-T [29]	1.0	60.0	87.6	92.9	1.0	60.3	87.6	93.2
	LCWF-GI [31]	1.0	59.4	86.8	92.5	1.0	60.1	86.7	92.7
	H-T(ViT) [29]	1.0	64.2	89.1	93.4	1.0	64.5	89.3	93.8
	LMF-CSF [30]	1.0	65.8	89.7	94.3	1.0	65.5	89.4	94.3
	Ours	**1.0**	**69.1**	**90.8**	**94.9**	**1.0**	**69.2**	**90.6**	**95.0**

Table 2. *Cont.*

	Methods	Image to Recipe				Recipe to Image			
		MedR	R@1	R@5	R@10	MedR	R@1	R@5	R@10
10 K	JE [9]	41.9	-	-	-	39.2	-	-	-
	AdaMin [10]	13.2	14.9	35.3	45.2	12.2	14.8	34.6	46.1
	R2GAN [35]	13.9	13.5	33.5	44.9	12.6	14.2	35.0	46.8
	MCEN [14]	7.2	20.3	43.3	54.4	6.6	21.4	44.3	55.2
	ACME [34]	6.7	22.9	46.8	57.9	6.0	24.4	47.9	59.0
	SN [16]	7.0	22.1	45.9	56.9	7.0	23.4	47.3	57.9
	SCAN [13]	5.9	23.7	49.3	60.6	5.1	25.3	50.6	61.6
	HF-ICMA [20]	5.0	24.0	51.6	65.4	4.2	25.6	54.8	67.3
	SEJE [22]	4.2	26.9	54.0	65.6	4.0	27.2	54.4	66.1
	M-SIA [17]	4.0	29.2	55.0	66.2	4.0	30.3	55.6	66.5
	X-MRS [39]	3.0	32.9	60.6	71.2	3.0	33.0	60.4	70.7
	H-T [29]	4.0	27.9	56.4	68.1	4.0	28.3	56.5	68.1
	LCWF-GI [31]	4.0	27.9	56.0	67.8	4.0	28.6	55.8	67.5
	H-T(ViT) [29]	3.0	33.5	62.1	72.8	3.0	33.7	62.2	72.7
	LMF-CSF [30]	3.0	34.6	62.7	73.2	3.0	34.3	62.5	72.8
	Ours	**2.1**	**38.1**	**65.8**	**75.9**	**2.2**	**38.3**	**65.6**	**75.6**

3.3. Scalability Analysis

In order to investigate the scalability of our method, we conduct experiments on datasets larger than 10 K in size. As shown in Figure 3, the MedR results of MMACMR are consistently lower than those of all other methods across all dataset sizes. In addition, it can be seen that, with the increase in test size, the performance gap between our method and others also widens. We argue that, on the one hand, the enhancement of recipe embedding promotes the alignment between the two modalities. On the other hand, as the dataset size increases, so does the number of ambiguous food images, leading to a higher probability of matching incorrect recipes. By effectively addressing this issue, our method demonstrates improved robustness and scalability as the dataset size enlarges.

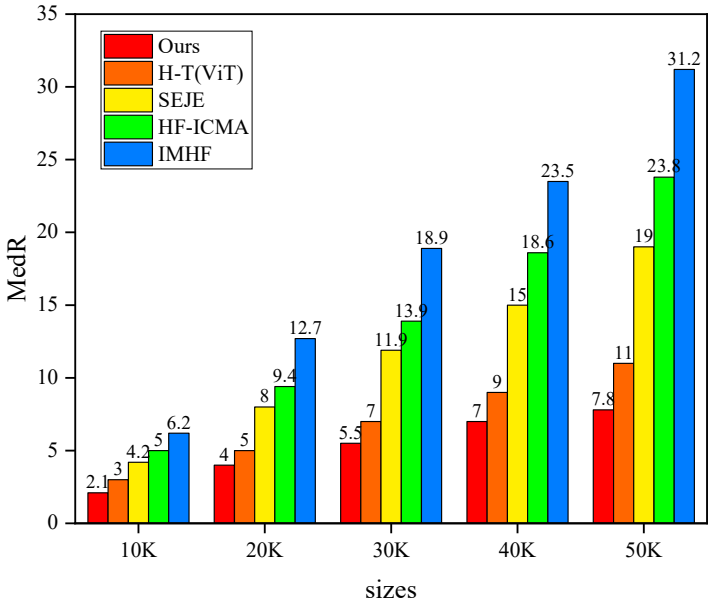

Figure 3. Scalability analysis. The abscissa represents the dataset size ranging from 10 K to 50 K, while the ordinate represents the MedR value.

3.4. Ablation Studies

In this subsection, we conduct ablation experiments to assess the contribution of each part of our model to the overall performance. Table 3 reports the image-to-recipe

retrieval results of different parts of MMACMR in 1 K and 10 K test size. In Table 3, Base is the baseline framework consisting of the food image encoder (ViT-B) and the original hierarchical transformer recipe encoder coupled with the N-pairs triplet loss. IR means introduction of the improved recipe encoder, and $\mathcal{L}_{RGI}$ is our RGI loss. A $\surd$ symbol under the columns Base, IR, and $\mathcal{L}_{RGI}$ indicates the use of that part. On the right, we list the MedR, R@1, R@5, and R@10 results for the image-to-recipe and recipe-to image tasks. We first evaluate the Base framework, then introduce the improved recipe encoder and RGI loss separately. Finally, we combine all three parts. It can be observed that the addition of both IR and $\mathcal{L}_{RGI}$ boosts the baseline model. This indicates that the solutions we propose to address the questions mentioned above are effective. When employing all subassemblies, we achieve the best performance, further validating the effectiveness of each element in our approach. Note that the method without IR obtains the same scores as the full method in R@5 and R@10 for image to recipe, and R@5 for recipe to image, for the 10 K size dataset. Additionally, it achieves better performance in MedR for recipe to image in 10 K size. Therefore, we attribute the main contribution to the MDA strategy.

Table 3. Ablation study. MedR (↓) and R@k (↑) in 1 K and 10 K size. The best results are marked in bold font. A $\surd$ symbol indicates that the corresponding part in this column is being used.

	Base	**I R**	$\mathcal{L}_{RGI}$	**Image to Recipe**				**Recipe to Image**			
				MedR	**R@1**	**R@5**	**R@10**	**MedR**	**R@1**	**R@5**	**R@10**
1 K	$\surd$			1.0	58.3	86.2	91.8	1.0	59.6	86.1	92.2
	$\surd$	$\surd$		1.0	67.2	90.0	94.5	1.0	67.6	90.0	94.5
	$\surd$		$\surd$	1.0	68.6	90.5	94.7	1.0	68.2	90.3	94.7
	$\surd$	$\surd$	$\surd$	**1.0**	**69.1**	**90.8**	**94.9**	**1.0**	**69.2**	**90.6**	**95.0**
10 K	$\surd$			4.1	26.8	54.7	66.5	4.0	37.5	55.1	66.8
	$\surd$	$\surd$		3.0	35.9	64.5	74.7	3.0	36.6	64.7	74.9
	$\surd$		$\surd$	2.2	37.7	65.8	75.9	**2.0**	38.0	65.6	75.4
	$\surd$	$\surd$	$\surd$	**2.1**	**38.1**	**65.8**	**75.9**	2.2	**38.3**	**65.6**	**75.6**

3.5. Qualitative Results

3.5.1. Qualitative Results on Image-to-Recipe Retrieval

To more intuitively analyze the representative results of MMACMR in image-to-recipe retrieval, we select four food images as queries to retrieve the recipes from the test set using our method and the SOTA method H-T (ViT) [29]. As shown in Figure 4, from left to right, the queries are "Chickpeas and Spinach with Smoky Paprika", "Blue Ribbon Apple Crumb Pie", "Apricot Nectar Cake", and "Sweet and Spicy Grilled Pork Tenderloin". In the first two samples, the categories of food are relatively easy to distinguish; both of these methods retrieve approximate recipes. However, in the first example, H-T (ViT) [29] does not retrieve the main ingredient, apricot nectar, while our method successfully retrieves it. The same situation occurs in the second example, where H-T (ViT) [29] retrieves a recipe whose corresponding image is similar to the query image but it misrecognizes the pork tenderloin as chicken thighs. In contrast, MMACMR retrieves the ground truth recipe. We attribute this to our MDA strategy, which can better address the problem of ambiguous food images and recognize the ingredients correctly. In the third example, H-T (ViT) [29] identifies some beans and vegetable leaf in the image but misclassifies their types, and the retrieved entire recipe deviates significantly from the ground truth. In the last example, the food image is difficult to recognize by human eye. H-T (ViT) [29] retrieves a recipe whose corresponding image has a similar color to the query (actually, it is a shortcut for models to classify objects which have not been seen before). However, our method retrieves the correct recipe even though the query image is ambiguous. We believe this is because MMACMR can reduce the distances between images with similar recipes, allowing the correct sample to be retrieved even when the query is hard to distinguish.

Figure 4. Examples of image-to-recipe retrieval results for the 10 K test set. The first row contains the query images, the second row shows the recipes retrieved using our method (all of which are the ground truth recipes; therefore, the first row is their corresponding food images), the third row displays the recipes retrieved using H-T (ViT) [29], and the last row presents the corresponding food images of the recipes from the third row. The key ingredients not retrieved by H-T [29] but retrieved by our method are highlighted in red.

3.5.2. Qualitative Results on Recipe-to-Image Retrieval

We also conduct experiments to visualize the results of recipe-to-image retrieval for the 1 K test set, which are presented in Figure 5. From top to bottom, the query recipes are titled "Fruit Salad", "Italian Beef Roast", and "Pesto Salmon", followed by the top five retrieved images using our method and the SOTA method H-T (ViT) [29]. In the first example, both methods retrieve five food images of fruit salad, but our method retrieves the ground truth as the top one, while H-T (ViT) [29] retrieves it in the top three. In the second example, the two methods retrieve the correct image in the top two. However, all the food images MMACMR retrieves are roast beef, while the third and fifth retrieved images of H-T (ViT) [29] do not match the recipe query. In the last example, our method retrieves the ground truth image as the top one, while H-T (ViT) [29] fails to retrieve the correct food image. At the same time, the second image retrieved by MMACMR is similar to the correct one, while the first and third images retrieved by H-T (ViT) [29] deviate significantly from the ground truth. We attribute these achievements to the capability of our method to

address the problem of ambiguous food images, allowing MMACMR to retrieve images that have similar recipes.

Figure 5. Examples of recipe-to-image retrieval results for the 10 K test set. The left side shows the query recipes, while the right side displays the top 5 retrieved images using our method or H-T (ViT) [29]. The ground truth images are marked by a red box.

4. Conclusions

In this paper, we propose a novel cross-modal recipe retrieval method named MMACMR which addresses the problem of ambiguous food images in retrieval using a novel training strategy, MDA, that guides the similarity within food images by recipe. Additionally, we improve the recipe encoder to ensure the precision of recipe embeddings. We conduct extensive experiments on the challenging public dataset Recipe1M, and the experimental results demonstrate the effectiveness of our method. Given the necessity of analyzing vast numbers of food data, our method could offer significant practical value in the food industry by enhancing user convenience and efficiency.

However, due to the complexity of recipe texts, some information representing the dish preparation program is still not captured by our method. In the future, we aim to focus on the fine-grained information in recipes using visual language pre-training models.

Author Contributions: Methodology, Z.Z.; Software, Z.Z.; Validation, Q.Z.; Investigation, Z.Z.; Resources, X.Z. and H.Z.; Data curation, Q.Z.; Writing—original draft, Z.Z.; Writing—review & editing, L.Z.; Supervision, X.Z., H.Z. and L.Z.; Project administration, X.Z.; Funding acquisition, X.Z. and L.Z. All authors have read and agreed to the published version of the manuscript

Funding: This work was supported in part by the National Natural Science Foundation of China (62202163), the Natural Science Foundation of Hunan Province (2022JJ40190), the Scientific Research Project of Hunan Provincial Department of Education (22A0145), the Key Research and Development Program of Hunan Province (2020NK2033), the Hunan Provincial Department of Education Scientific Research Outstanding Youth Project (21B0200), and the Hunan Provincial Natural Science Foundation Youth Fund Project (2023JJ40333).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank the *Foods* journal for providing this opportunity to submit the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data.

Abbreviations

The following abbreviations are used in this manuscript:

MMACMR	Multi-Modal Alignment Method for Cross-Modal Recipe Retrieval
MDA	Multi-Modal Disambiguity and Alignment
RGI	Recipe Guide Image
AI	Artificial intelligence
LSTM	Long Short-Term Memory
GANs	Generative Adversarial Networks
ViT	Vision Transformer
CLIP	Contrastive Language–Image Pre-training
KNN	K-Nearest Neighbors
SOTA	State Of The Art
MedR	Median Rank
SSD	Solid-State Disk
RAM	Random-Access Memory
HDD	Hard Disk Drive
CCA	Canonical Correlation Analysis

References

1. Guo, Z.; Jayan, H. Fast Nondestructive Detection Technology and Equipment for Food Quality and Safety. *Foods* **2023**, *12*, 3744. [CrossRef] [PubMed]
2. Guo, Z.; Wu, X.; Jayan, H.; Yin, L.; Xue, S.; El-Seedi, H.R.; Zou, X. Recent developments and applications of surface enhanced Raman scattering spectroscopy in safety detection of fruits and vegetables. *Food Chem.* **2023**, *434* , 137469. [CrossRef] [PubMed]
3. Thames, Q.; Karpur, A.; Norris, W.; Xia, F.; Panait, L.; Weyand, T.; Sim, J. Nutrition5k: Towards automatic nutritional understanding of generic food. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8903–8911.
4. Min, W.; Wang, Z.; Liu, Y.; Luo, M.; Kang, L.; Wei, X.; Wei, X.; Jiang, S. Large scale visual food recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9932–9949. [CrossRef] [PubMed]
5. Min, W.; Wang, Z.; Yang, J.; Liu, C.; Jiang, S. Vision-based fruit recognition via multi-scale attention CNN. *Comput. Electron. Agric.* **2023**, *210*, 107911. [CrossRef]
6. Min, W.; Liu, C.; Xu, L.; Jiang, S. Applications of knowledge graphs for food science and industry. *Patterns* **2022**, *3*, 100484. [CrossRef] [PubMed]
7. Wang, W.; Min, W.; Li, T.; Dong, X.; Li, H.; Jiang, S. A review on vision-based analysis for automatic dietary assessment. *Trends Food Sci. Technol.* **2022**, *122*, 223–237. [CrossRef]
8. Liu, Y.; Min, W.; Jiang, S.; Rui, Y. Convolution-Enhanced Bi-Branch Adaptive Transformer with Cross-Task Interaction for Food Category and Ingredient Recognition. *IEEE Trans. Image Process.* **2024**, *33*, 2572–2586. [CrossRef] [PubMed]

9. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 3020–3028.
10. 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 Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 35–44.
11. Min, W.; Zhou, P.; Xu, L.; Liu, T.; Li, T.; Huang, M.; Jin, Y.; Yi, Y.; Wen, M.; Jiang, S.; Jain, R. From Plate to Production: Artificial Intelligence in Modern Consumer-Driven Food Systems. *arXiv* **2023**, arXiv:2311.02400.
12. Guo, Z.; Zhang, Y.; Wang, J.; Liu, Y.; Jayan, H.; El-Seedi, H.R.; Alzamora, S.M.; Gómez, P.L.; Zou, X. Detection model transfer of apple soluble solids content based on NIR spectroscopy and deep learning. *Comput. Electron. Agric.* **2023**, *212*, 108127. [CrossRef]
13. Wang, H.; Sahoo, D.; Liu, C.; Shu, K.; Achananuparp, P.; Lim, E.P.; Hoi, S.C. Cross-modal food retrieval: Learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. *IEEE Trans. Multimed.* **2021**, *24*, 2515–2525. [CrossRef]
14. 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, Seattle, WA, USA, 13–19 June 2020; pp. 14570–14580.
15. Chen, Y.; Zhou, D.; Li, L.; Han, J.M. Multimodal encoders for food-oriented cross-modal retrieval. In Proceedings of the Web and Big Data: 5th International Joint Conference, APWeb-WAIM 2021, Guangzhou, China, 23–25 August 2021; Proceedings, Part II 5; Springer International Publishing: Cham, Switzerland, 2021; pp. 253–266.
16. 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, Dublin, Ireland, 8–11 June 2020; pp. 117–125.
17. 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, Virtual Event, 1–5 November 2021; pp. 3211–3215.
18. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 4567–4578.
19. Li, L.; Hu, C.; Zhang, H.; Maradapu Vera Venkata sai, A. Cross-modal Image-Recipe Retrieval via Multimodal Fusion. In Proceedings of the 5th ACM International Conference on Multimedia in Asia, Taiwan, China, 6–8 December 2023; pp. 1–7.
20. 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, Taipei, Taiwan, 21–24 August 2021; pp. 173–182.
21. Chen, J.J.; Ngo, C.W.; Feng, F.L.; Chua, T.S. Deep understanding of cooking procedure for cross-modal recipe retrieval. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1020–1028.
22. Xie, Z.; Liu, L.; Wu, Y.; Zhong, L.; Li, L. Learning text-image joint embedding for efficient cross-modal retrieval with deep feature engineering. *ACM Trans. Inf. Syst. (TOIS)* **2021**, *40*, 1–27. [CrossRef]
23. Xie, Z.; Liu, L.; Li, L.; Zhong, L. Efficient Deep Feature Calibration for Cross-Modal Joint Embedding Learning. In Proceedings of the 2021 International Conference on Multimodal Interaction, Montréal, QC, Canada, 18–22 October 2021; pp. 43–51.
24. Xie, Z.; Liu, L.; Li, L.; Zhong, L. Learning joint embedding with modality alignments for cross-modal retrieval of recipes and food images. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Virtual Event, 1–5 November 2021; pp. 2221–2230.
25. Xie, Z.; Liu, L.; Wu, Y.; Li, L.; Zhong, L. Learning tfidf enhanced joint embedding for recipe-image cross-modal retrieval service. *IEEE Trans. Serv. Comput.* **2021**, *15*, 3304–3316. [CrossRef]
26. Cao, D.; Chu, J.; Zhu, N.; Nie, L. Cross-modal recipe retrieval via parallel-and cross-attention networks learning. *Knowl.-Based Syst.* **2020**, *193*, 105428. [CrossRef]
27. Li, J.; Xu, X.; Yu, W.; Shen, F.; Cao, Z.; Zuo, K.; Shen, H.T. Hybrid fusion with intra-and cross-modality attention for image-recipe retrieval. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 11–15 July 2021; pp. 244–254.
28. Xie, Z.; Li, L.; Zhong, L.; Liu, J.; Liu, L. Cross-Modal Retrieval between Event-Dense Text and Image. In Proceedings of the 2022 International Conference on Multimedia Retrieval, Newark, NJ, USA, 27–30 June 2022; pp. 229–238.
29. 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, Nashville, TN, USA, 20–25 June 2021; pp. 15475–15484.
30. Zhao, W.; Zhou, D.; Cao, B.; Zhang, K.; Chen, J. Efficient low-rank multi-component fusion with component-specific factors in image-recipe retrieval. *Multimed. Tools Appl.* **2024**, *83*, 3601–3619. [CrossRef]
31. Zhao, W.; Zhou, D.; Cao, B.; Liang, W.; Sukhija, N. Exploring latent weight factors and global information for food-oriented cross-modal retrieval. *Connect. Sci.* **2023**, *35*, 2233714. [CrossRef]
32. 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, Waikoloa, HI, USA, 3–8 January 2024; pp. 5584–5593.

33. Wang, H.; Lin, G.; Hoi, S.C.; Miao, C. Learning structural representations for recipe generation and food retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 3363–3377. [CrossRef]
34. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 11572–11581.
35. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 11477–11486.
36. Sugiyama, Y.; Yanai, K. Cross-modal recipe embeddings by disentangling recipe contents and dish styles. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 2501–2509.
37. 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, Lisboa, Portugal, 10–14 October 2022; pp. 5517–5526.
38. Yang, J.; Chen, J.; Yanai, K. Transformer-Based Cross-Modal Recipe Embeddings with Large Batch Training. In Proceedings of the International Conference on Multimedia Modeling, Bergen, Norway, 9–12 January 2023; Springer: Cham, Switzerland, 2023; pp. 471–482.
39. 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, Virtual Event, 20–24 October 2021; pp. 3192–3201.
40. Zhu, B.; Ngo, C.W.; Chen, J.; Chan, W.K. Cross-lingual adaptation for recipe retrieval with mixup. In Proceedings of the 2022 International Conference on Multimedia Retrieval, Newark, NJ, USA, 27–30 June 2022; pp. 258–267.
41. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 16559–16569.
42. 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, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 529–537.
43. Sun, J.; Li, J. PBLF: Prompt Based Learning Framework for Cross-Modal Recipe Retrieval. In Proceedings of the International Symposium on Artificial Intelligence and Robotics, Shanghai, China, 21–23 October 2022; Springer: Singapore, 2022; pp. 388–402.
44. Shukor, M.; Thome, N.; Cord, M. Vision and Structured-Language Pretraining for Cross-Modal Food Retrieval. *arXiv* **2022**, arXiv:2212.04267.
45. Voutharoja, B.P.; Wang, P.; Wang, L.; Guan, V. MALM: Mask Augmentation based Local Matching for Food-Recipe Retrieval. *arXiv* **2023**, arXiv:2305.11327.
46. Zhang, C.; Song, J.; Zhu, X.; Zhu, L.; Zhang, S. Hcmsl: Hybrid cross-modal similarity learning for cross-modal retrieval. *ACM Trans. Multimed. Comput. Commun. Appl. (Tomm)* **2021**, *17*, 1–22. [CrossRef]
47. Zhu, L.; Zhang, C.; Song, J.; Liu, L.; Zhang, S.; Li, Y. Multi-graph based hierarchical semantic fusion for cross-modal representation. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
48. Yi, Z.; Zhu, X.; Wu, R.; Zou, Z.; Liu, Y.; Zhu, L. Multi-Label Weighted Contrastive Cross-Modal Hashing. *Appl. Sci.* **2023**, *14*, 93. [CrossRef]
49. Zou, Z.; Zhu, X.; Zhu, Q.; Liu, Y.; Zhu, L. CREAMY: Cross-Modal Recipe Retrieval by Avoiding Matching Imperfectly. *IEEE Access* **2024**, *12*, 33283–33295. [CrossRef]
50. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
51. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
52. Thomas, C.; Kovashka, A. Preserving semantic neighborhoods for robust cross-modal retrieval. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XVIII 16; Springer International Publishing: Cham, Switzerland, 2020; pp. 317–335.
53. Guo, G.; Wang, H.; Bell, D.; Bi, Y.; Greer, K. KNN model-based approach in classification. In Proceedings of the On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, 3–7 November 2003; Springer: Berlin/Heidelberg, Germany, 2003; pp. 986–996.
54. Wang, J.; Zhou, F.; Wen, S.; Liu, X.; Lin, Y. Deep metric learning with angular loss. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2593–2601.