IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 February 2024, accepted 22 February 2024, date of publication 26 February 2024, date of current version 7 March 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3370158*

RESEARCH ARTICLE

# CREAMY: Cross-Modal Recipe Retrieval By Avoiding Matching Imperfectly

**ZHUOYANG ZOU, XINGHUI ZHU, QINYING ZHU, YI LIU, AND LEI ZHU, (Member, IEEE)**
College of Information and Intelligence, Hunan Agricultural University, Changsha 410128, China

Corresponding author: Lei Zhu (leizhu@hunau.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62202163 and Grant 62072166, in part by the Natural Science Foundation of Hunan Province under Grant 2022JJ40190, in part by the Scientific Research Project of Hunan Provincial Department of Education under Grant 22A0145, and in part by the Key Research and Development Program of Hunan Province under Grant 2020NK2033.

**ABSTRACT** State-of-the-art methods for cross-modal recipe retrieval failed to consider an underlying but challenging issue, i.e., *matching imperfectly problem* hidden in positive image-recipe pairs, which is a culprit causing over-fitting. To make up this defect, two critical questions—how to effectively recognize and filter out mismatching parts during the model training and how to pick out and preserve as much matching information as possible need to be answered. To do so, this article proposes a novel method—**C**ross-modal **R**ecipe r**E**trieval by **A**voiding **M**atching imperfectl**Y**, abbreviated as **CREAMY**, which involving a new-designed learning strategy called **N**on-**M**atching and **P**artial-**M**atching (**NMPM**) to undertake two tasks: 1) no longer forcibly aligning each positive image-recipe pair but rather capturing the complementary information from negative pairs; 2) delicately picking up and aligning the matchable part in each pair. To the best of our knowledge, this attempt is a pioneer to defeat the matching imperfectly issue for cross-modal recipe retrieval task. Empirical analysis conducted on Recipe1M dataset validates the advantages of CREAMY over several state-of-the-arts. The code is available at: https://github.com/pouqual/CREAMY.

**INDEX TERMS** Cross-modal recipe retrieval, matching imperfectly, non-matching, partial-matching.

## I. INTRODUCTION

Thanks for the prosperity of social networks, e-commerce platforms and online recommendation system [1], people enjoy the delights of cooking by easily following plentiful cooking tutorials shared on Internet. Food computing [2], [3], [4], as a results, has been proposed and studied so as to make recipes/food searching, recommending, sharing online more effectively, efficiently, and robustly.

This work concentrates on a hot spot in food computing area, namely *cross-modal recipe retrieval* that aims to retrieve the corresponding food images by queries of recipes or vice versa. Unlike the simple image-text pair in traditional cross-modal retrieval [5], [6], [7], [8], [9], the samples in cross-modal recipe retrieval are much more complex. Specifically, the images are photos of cooked food according to the recipes consisting of three textual components: (1) *title*, a single sentence naming the food; (2) *ingredients*, a list of sentences to presents the needed ingredients for the food; (3) *instructions*, a list of sentences to describe the cooking steps in detail.

The main challenge of cross-modal recipe retrieval is mitigating the heterogeneity between food images and recipes, which is more difficult than conventional image-text retrieval task [10], [11], [12], [13], [14], [15], [16], to some extent, due to more intricate data. The usual treatment is employing independent neural networks to encode images and their corresponding recipes so as to align them in a common feature subspace. To do so, several well-known CV models (e.g. CNN [17], [18], [19] and ViT [20], [21], [22]) and NLP models (e.g. LSTM [23], [24], [25] and Transformer [20], [21], [22]) are coupled with triplet loss

The associate editor coordinating the review of this manuscript and approving it for publication was Rongbo Zhu.

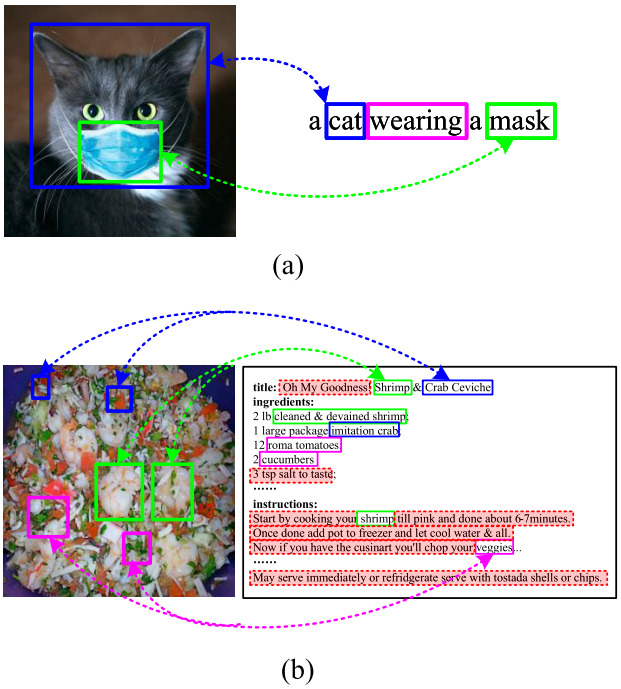



**FIGURE 1. Differences between (a) perfectly matching image-text pair and (b) image-recipe pair with matching imperfectly problem. The texts in red dotted boxes are the mismatching parts. Best view in color.**

to excavate essential features for images-recipes alignment. Object detection [26] and image reconstruction [27], [28], [29] techniques are adopted by previous works to implement a strong-sighted model so that the key visual details can be focused. To better understand the complex semantics from recipes, some works [26], [30] devote to find the key terms in texts, while others [22], [31] attempt to explore the hidden consistent information between different components in recipes, or even capture the interaction of two modalities via cross-modal attention [32], [33], [34] to enhance cross-modal alignment [35], [36], [37], [38].

### A. MOTIVATION

In spite of their extraordinary accomplishments, a critical issue hidden in cross-modal recipes data, namely *matching imperfectly problem*, seriously hinder image-recipes alignment. In short, this problem refers to a text sample (a sentence or a paragraph) in an image-text pair cannot perfectly match the content of its corresponding image. Compared to general cross-modal retrieval involving simple image-text pairs (e.g., an image and a simple sentence), cross-modal recipe retrieval with complex image-text pair (e.g., an image containing various visual contents coupled with a text with rich semantics) is more susceptible to suffering from matching imperfectly. Taking Figure 1 as an intuitive example,[1] Nouns "cat" and "mask", shown in Figure 1 (a), are matching well to the visual objects in the image; two quantifiers "a" are matching to the quantity of the corresponding objects; the static verb "wearing" is corresponds to the relationship between the cat and the mask in the image. Unfortunately, as a complex pair, the image-recipe pair shown in Figure 1 (b) is unable to avoid matching imperfectly problem. No doubt, the title "Oh My Goodness! Shrimp & Crab Ceviche" carries the emotion of the food maker, which is not indeed presented in the food image. Furthermore, some ingredients (e.g., salt) are hardly recognized from the image due to being mixed up after cooking, let alone their quantities. Worse still, the cooking steps in instructions never appear in the image, and some other information, such as the last sentence "May serve immediately or refridgerate serve with tostada shells or chips." describes the eating method of the food, which has almost no relation to the visual content.

Cross-modal alignment between a food image and its recipe, beyond all doubt, seriously suffers from the matching imperfectly problem, especially under the current learning setting. In specific, the prevailing solutions that use triplet loss are based upon the following assumption: given a positive image-recipe pair, the food image and its recipe are treated as perfectly matching. Undoubtedly, this assumption is far from correct when handling cross-modal recipe pairs, if which is naively adopted during model training, over-fitting will be inevitable. To address this serious but not widely concerned issue, we attempt to design an effective matching strategy to enhance images-recipes alignment: (1) To prevent incorrect matching, we no longer forcibly align the positive image-recipe pairs but rather capture the complementary information from negative pairs. In this way, the optimization direction could be led by the negative pairs; (2) To prevent the loss of partially matching information among positive image-recipe pairs, we try to delicately pick up and align the matching part in each pair. As mentioned above, a food image is the cooking result of the ingredients. In other words, ingredients are the components directly associated with the food image even though they may be hard to recognize after cooking. Following such fact, we therefore solely align the features of ingredients and food images in positive image-recipe pairs to precisely retain the matching information.

### B. OUR METHOD

To this end, we propose a novel cross-modal recipe retrieval method, termed as **C**ross-modal **R**ecipe r**E**trieval by **A**voiding **M**atching imperfectl**Y**, abbreviated as **CREAMY**, shown in Figure 2. Apart from a set of transformer-based encoders to capture the semantic features from food images and recipes seperately, at the heart of CREAMY is a skillful-designed learning strategy, termed as **N**on-**M**atching and **P**artial-**M**atching strategy (**NMPM** for short), served as a protector for the model from matching imperfectly problem. Specifically, NMPM treats the negative and positive image-recipe pairs in different manners: (1) for the negative

[1]The picture in Figure 1(a) is downloaded from: https://www.sohu.com/a/450519741_120051368. All the food pictures in this article are the samples from Recipe1M dataset, which can be downloaded from http://im2recipe.csail.mit.edu

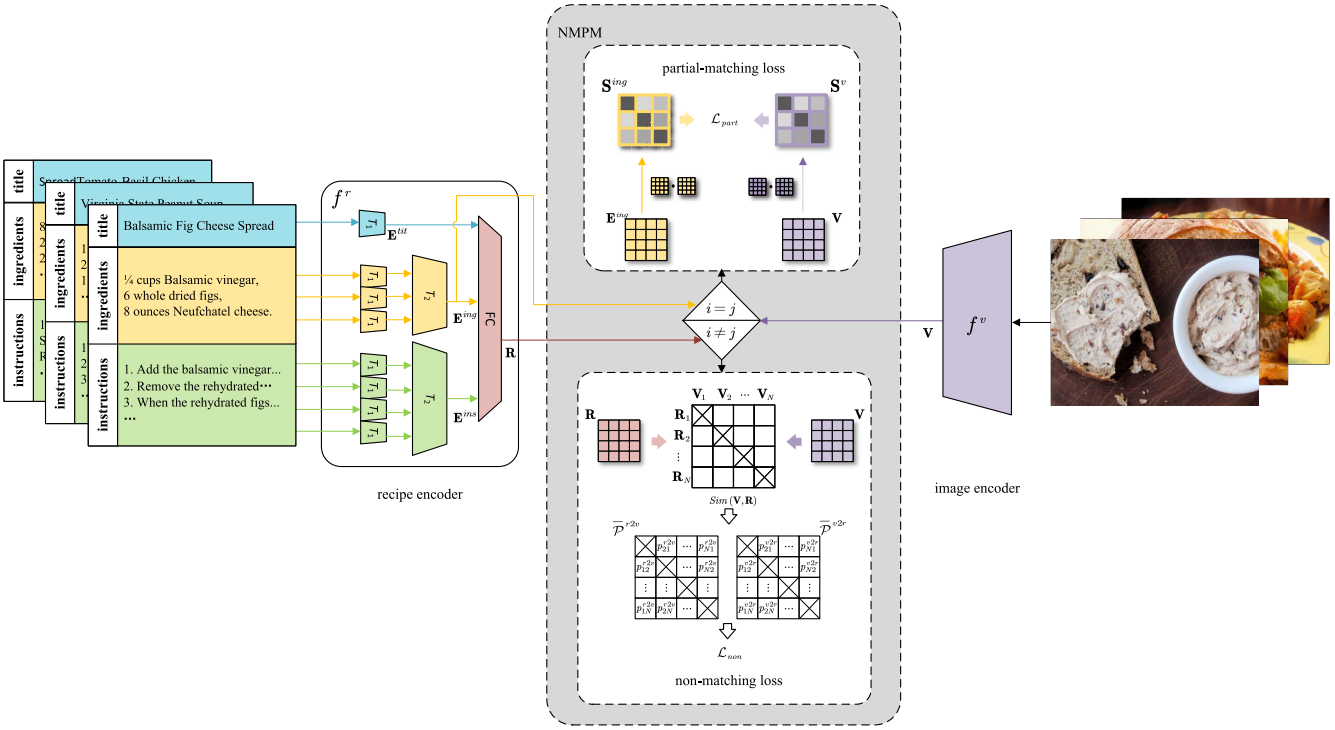



**FIGURE 2.** **The framework of CREAMY. The left is recipe inputs and the recipe encoder $f^r$, the right is image inputs and the image encoder, and the middle is our NMPM strategy, which include non-matching loss and partial-matching loss.**

pairs, the distances between them are enlarged so as to learn the complementary information by *a non-matching loss*, which is called *Non-Matching strategy*; (2) for the positive pairs, the mismatching parts are filtered out from the recipes and then reserve the matchable parts, i.e., the ingredients which then are aligned to image features by *a partial-matching loss* such that the consistent semantics can be learned, which is the *Partial-Matching strategy*. In a nutshell, this best-of-both-worlds learning strategy skillfully avoids matching imperfectly problem among cross-modal recipe data while sufficiently captures semantic correlation in each positive pair, the boost of cross-modal recipe retrieval performance brought by which has been empirically verified via extensive experiments.

### C. CONTRIBUTIONS

In summary, the main contributions of this article are three folds, listed as follows:

- We propose a novel method named CREAMY aiming to avoid the matching imperfectly problem effectively in cross-modal recipe retrieval. Different from existing methods, this work is the pioneer to improve image-recipes feature alignment by addressing the challenge of matching imperfectly.
- We introduce a novel learning strategy, NMPM, consisting of two loss functions: non-matching loss and partial-matching loss, which effectively avoids interference from matching imperfectly while maintaining matchable information in positive image-recipe pairs.
- We conduct extensive experiments on the challenging dataset Recipe1M. The results demonstrate that the proposed technique outperforms the state-of-the-arts by a significant margin.

### D. ROADMAP

The remainder of this article is organized as follows. The related works are summarized in Section II. We introduce our method in Section III. Section IV discusses the experiments and we conclude this article in Section V.

## II. RELATED WORK

This section reviews the prevailing studies concerning cross-modal recipe retrieval task and food image/recipe generation, which are related to our work.

### A. CROSS-MODAL RECIPE RETRIEVAL

As a particular case of cross-modal retrieval, cross-modal recipe retrieval aims to search the corresponding recipes by food image queries or vice versa. The main challenge of cross-modal recipe retrieval, similar to most cross-modal tasks, is to eliminate heterogeneity between different modalities (image and recipe). To this end, several well-known CV models and NLP models are employed to generate high-quality embeddings from food images and recipe texts so as to achieve cross-modal alignment. For example, deep

convolutional neural networks, such as VGG [17], [39], [40] and ResNet [18], [19], [41], [42], are used in several works for visual information embedding. To further focus on essential visual features, Faster-R-CNN is involved [43] to detect food object. More recently, ViT [20], [21], [22] is applied to enhance the food image recognition. For recipes, along with the proposed of various attention mechanism, Bert [44], [45], [46] and Transformer [47], [48], [49], [50] are utilized to implement stronger textual encoder than sequential models such as skip-thought [26], [51] and LSTM [23], [24], [25] involved in early works. Furthermore, cross-modal attention [34], [52], [53] mechanism and large vision-language pre-training models [54], [55], [56], [57], [58] are employed in cross-modal recipe understanding, which further narrow heterogeneous gap via cross-modal interaction.

The difficulty of recipe retrieval mainly stems from complex recipe sample including title, ingredients and instructions, other than a simple phrase or a sentence. These three components play different roles to present a recipe. Therefore, a more reasonable way for recipe feature embedding is to explore latent semantic information from different parts discriminately. For example, Cao et al. [43] extracted contextual information from title and ingredients, and utilized them to highlight the key elements in images and instructions. Fontanellaz et al. [31] used an attention mechanism to focus on the words or single instruction in instructions which have the strongest connection to ingredients. Fu et al. [32] employed an attention-based RNN decoder to capture the correlations between instructions and ingredients. Xie et al. [30] leveraged TF-RDF to rerank the ingredients to improve the quality of recipe embedding. Li et al. [33] used enhanced ingredients and instructions information to do the local and global alignment separately. Salvador et al. [22] proposed a self-supervised loss function to leverage semantic relationships within recipes. Xu et al. [59] regarded cross-modal recipe retrieval as a ternary image-text retrieval problem, and utilized high-level associations between these three components via bi-directional triplet loss. Other studies such as [27], [29] and [28] attempted to generate synthetic images by GAN to facilitate the understanding of recipe characteristics.

Despite significant progress made so far, a critical issue, namely matching imperfectly within each positive image-recipe pair, is ignored in all the previous works which naively treat positive image-recipe pairs in training set as perfectly matching. Unfortunately, matching imperfectly always occurs in positive pairs owing to the fact that the semantic information contained in recipe texts (including title, ingredient and instruction) generally exceeds the range of information contained in the food image. If forcibly we align the features matching imperfectly, over-fitting would be caused inevitably. Bearing such stand-out limitation, we are inspired by complementary contrastive learning [60] and attempt to design a novel loss function to reduce adverse effect caused by this issue. Meanwhile, we did not completely abandon using positive image-recipe pairs due to consistent semantic information existing in each positive pair, but attempted to enhance cross-modal alignment by exploiting partially matching information in positive pairs, which has barely been explored in prior works.

### B. FOOD IMAGE/RECIPE GENERATION

Food image/recipe generation are another two vision-language tasks related to cross-modal recipe retrieval. The main challenge of them lie in capturing the details of food images and understanding the variation of ingredients caused by cooking. For food image generation task, Han et al. [61] built an attention-based ingredients-images association model to generate images from ingredients. Wang et al. [62] introduced a cycle-consistency training method, which improved image generation by optimizing the inverted latent codes. ChefGAN [63] involved a joint image-recipe embedding model to GANs before and during the stage of generate images. CookGAN [64] mimicked visual effect of instructions and preserved the fine-grained details of images. For recipe generation task, [65] and [66] predicted ingredients firstly and then generated whole recipes using ingredients and images. Other solutions [67], [68], [69] generated instructions by exploiting the structure information of text.

As what mentioned above, ingredients after cooking are mixed together, which bring significant obstacles to determine whether a certain ingredient should appear in the image. On the other hand, without the perfect alignment between ingredients and visual information, the bridge across instruction and its corresponding image is lost. That means the matching imperfectly problem discussed above is also seriously interfere with food image/recipe generation. Unfortunately, this issue has not been addressed well right now.

## III. METHOD

This section introduces our method CREAMY. We first present notations and problem formulation in subsection III-A, followed by the technique details, including framework overview in subsection III-B, a novel learning strategy NMPM in subsection III-C. We end this section with model optimization in III-D.

### A. NOTATIONS AND PROBLEM FORMULATION

For the sake of discussion, firstly we introduce the notations involved in this paper, then giving the problem formulation of cross-modal recipe retrieval.

#### 1) NOTATIONS

Without losing generality, sets and matrices are denoted as uppercase handwritten letters (e.g. $\mathcal{D}$) and bold uppercase letters (e.g. $\mathbf{A}$), respectively. The $i$-th row of $\mathbf{A}$ is denoted as $\mathbf{A}_i$, and the element located in the $j$-th column of $i$-th row of $\mathbf{A}$ is denoted as $\mathbf{A}_{ij}$. $\|\cdot\|_2$ denotes the L2 norm of a matrix. The transpose of matrix $\mathbf{A}$ is denoted as $\mathbf{A}^\top$. Suppose that $\mathcal{P}$ is a

**TABLE 1. The summary of notations.**

| Notation | Definition |
|---|---|
| $\mathcal{D}$ | a cross-modal recipe dataset |
| $\mathbf{X}_i^v$ | the $i$-th image |
| $\mathbf{X}_i^r$ | the $i$-th recipe |
| $\mathbf{X}_i^{tit}$ | the title of $i$-th recipe |
| $\mathbf{X}_i^{ing}$ | a list of ingredients of $i$-th recipe |
| $\mathbf{X}_i^{ins}$ | a list of instructions of $i$-th recipe |
| $\mathbf{V}$ | the image embeddings |
| $\mathbf{R}$ | the recipe embeddings |
| $\theta^v$ | the parameters of image encoder |
| $\theta^r$ | the parameters of recipe encoder |
| $\mathcal{L}_{non}$ | the non-matching loss function |
| $\mathcal{L}_{part}$ | the partial-matching loss function |
| $\mathbf{S}$ | the similarity matrix |

probability of event $\mathbb{Z}$, $\overline{\mathcal{P}} = 1 - \mathcal{P}$ represents the probability of the complementary event of $\mathbb{Z}$. Table 1 summarizes the frequently used notations through this article.

#### 2) PROBLEM FORMULATION

Let $\mathcal{D} = \{\mathbf{X}_i^v, \mathbf{X}_i^r\}_{i=1}^n$ be a cross-modal recipe dataset with $n$ image-recipe pairs, where $\mathbf{X}_i^v$ and $\mathbf{X}_i^r = \{\mathbf{X}_i^{tit}, \mathbf{X}_i^{ing}, \mathbf{X}_i^{ins}\}$ represent the $i$-th sample of image and recipe. $\mathbf{X}_i^{tit}$, $\mathbf{X}_i^{ing}$ and $\mathbf{X}_i^{ins}$ represent a title, a list of ingredients and a list of instructions of the recipe, respectively (Note that, each title is a single sentence, while both ingredients and instructions consist of several sentences). Given a recipe $\mathbf{X}_i^r$ as query, cross-modal recipe retrieval is aiming to search the most similar food images $\{\mathbf{X}_i^v\}$ or vise versa. To address the matching imperfectly problem, we attempt to introduce a novel learning strategy consisting of two loss functions: a non-matching loss function $\mathcal{L}_{non}$ to avoid matching imperfectly and a partial-matching loss function $\mathcal{L}_{part}$ to preserve matchable information such that two modality-specific embedding functions, $\mathbf{V} = f^v(\mathbf{X}^v; \theta^v)$ for image modality and $\mathbf{R} = f^r(\mathbf{X}^r; \theta^r)$ for recipe modality can be learned correctly:

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

where $\mathbf{V}$ and $\mathbf{R}$ are the image and recipe embeddings, $\theta^v$ and $\theta^r$ are the learnable parameters, $\lambda$ is a parameter to balance two parts. The similarity between the $i$-th image $\mathbf{X}_i^v$ and the $j$-th recipe $\mathbf{X}_j^r$ can be denoted by $Sim(\mathbf{V}_i, \mathbf{R}_j)$.

### B. FRAMEWORK OVERVIEW

Figure 2 is an overview of our method CREAMY. Similar to the prevailing solutions, the backbone consists of two branches: an image encoder and a recipe encoder, which are the implementations of embedding functions $f^v(\cdot; \theta^v)$ and $f^r(\cdot; \theta^r)$, respectively. The technique details of them are listed as follows.

#### 1) IMAGE ENCODER

The base size model of Vision Transformer(ViT-B) [70] initialized with the weights pre-trained on ImageNet [71] is served as image encoder. Given an image set $\{\mathbf{X}_i^v\}_{i=1}^n$, each of the image embedding is denoted as $\mathbf{V}_i = f^v(\mathbf{X}_i^v)$. To better evaluate the performance of this setting, we also conduct experiment with ResNet-50 [72] pre-trained on ImageNet as the image encoder. The implementation details are introduced in IV-A.

#### 2) RECIPE ENCODER

Inspired by [22], we employ a hierarchical transformer encoder consisting of two level transformers (the one is token-level and the other is sentence-level) with the same architecture as recipe encoder. Given a recipe set $\{\mathbf{X}_i^r\}_{i=1}^n = \{\mathbf{X}_i^{tit}, \mathbf{X}_i^{ing}, \mathbf{X}_i^{ins}\}_{i=1}^n$, the first level transformer $T_1$ receives the tokens of every words of every sentences, following an average pooling layer to output the average embedding of every sentences, denoted as $((\mathbf{X}_i^{tit})', (\mathbf{X}_i^{ing})', (\mathbf{X}_i^{ins})') = T_1(\mathbf{X}_i^{tit}, \mathbf{X}_i^{ing}, \mathbf{X}_i^{ins})$. The second level transformer $T_2$ receives the output of $T_1$, following an average pooling layer to output the average embedding of these two components, denoted as $(\mathbf{E}_i^{tit}, \mathbf{E}_i^{ing}, \mathbf{E}_i^{ins}) = ((\mathbf{X}_i^{tit})', T_2(\mathbf{X}_i^{ing})', T_2(\mathbf{X}_i^{ins})')$. Note that, as a single sentence, the embedding of title is just obtained from $T_1$ directly. Lastly, we concatenate the outputs of the three components and feed them to a linear layer to get the final output $\mathbf{R}_i = FC([\mathbf{E}_i^{tit}; \mathbf{E}_i^{ing}; \mathbf{E}_i^{ins}]; \theta^l)$, where $FC(\cdot; \theta^l)$ is a learnable linear layer, $[\cdot; \cdot; \cdot]$ denotes embedding concatenation, $\theta^l$ is the parameter vector. The entire process is denoted as $\mathbf{R}_i = f^r(\mathbf{X}_i^r; \theta^r)$, where $\theta^r = [\theta^{tit}; \theta^{ing}; \theta^{ing}; \theta^l]$. The implementation details are introduced in IV-A.

### C. NMPM STRATEGY

As a common-used strategy in existing methods, triplet loss aims to learn a common subspace where the positive image-recipe pairs exhibit greater similarity (smaller distance) than negatives. Taking Fig. 3 as an example, given an anchor image vector, the triplet loss (Fig. 3 (a)) attempts to maximize the distance between the positive pair while minimizing the distance between a hard negative pair, and forcing the gap between them to be larger than a specified margin. However, due to the presence of matching imperfectly in the positive recipe (Fig. 3 (b)), blindly reducing the distance between the anchor and the positive recipe may lead to a suboptimal optimization direction. To this end, we propose the NMPM strategy (Fig. 3 (c)) comprising a non-matching loss $\mathcal{L}_{non}$ to avoid matching imperfectly problem and a partial-matching loss $\mathcal{L}_{part}$ to preserve matchable information. For the non-matching loss $\mathcal{L}_{non}$, we refrain from reducing the distance between positive pairs, and focus on maximizing the distance between all negative pairs. In this way, a better optimization direction will be led by the negative recipes, enabling us to overcome the issue of matching imperfectly. Meanwhile, we do not completely abandon the information in the positive recipe, instead, we select and align the matchable features in the

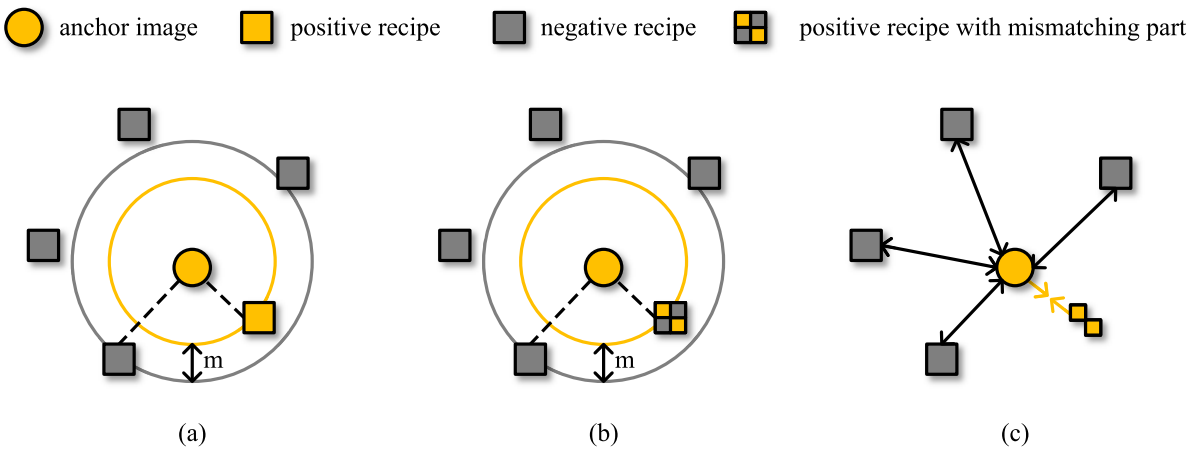



**FIGURE 3.** Comparison between triplet loss and ours. (a) shows the triplet loss ideally, m is a margin constant. (b) is the triplet loss for cross-modal recipe retrieval in actual, there are mismatching features in positive sample. (c) are the losses of NMPM, which enlarge the distance between negative pairs, and align the most matching part in positive pair.

positive pairs using a partial-matching loss $\mathcal{L}_{part}$. The details of the NMPM strategy are discussed below.

### 1) NON-MATCHING LOSS

The core idea of the non-matching loss is to find a common feature subspace where the negative image-recipe pairs have large distance as much as possible. Firstly, given a food image query $\mathbf{V}_i$, we define the cross-modal matching probability of recipe sample $\mathbf{R}_j$ w.r.t. $\mathbf{V}_i$ as $p_{ij}^{v2r}$, which can be calculated by:

$$p_{ij}^{v2r} = \frac{exp(\frac{Sim(\mathbf{V}_i,\mathbf{R}_j)}{\tau})}{\sum_{n=1}^{|\mathbf{R}|} exp(\frac{Sim(\mathbf{V}_i,\mathbf{R}_n)}{\tau})}, \tag{2}$$

where the similarity $Sim(\mathbf{V}_i, \mathbf{R}_j)$ is implemented by Cosine similarity, $\tau$ is the temperature parameter, $|\mathbf{R}|$ is the size of recipe sets. Since it is too expensive to compute the denominator of Eq.(2), we employ Monte Carlo [73] to approximate the value of $p_{ij}^{v2r}$ as:

$$\begin{aligned} p_{ij}^{v2r} &= \frac{exp(\frac{Sim(\mathbf{V}_i,\mathbf{R}_j)}{\tau})}{\sum_{n=1}^{|\mathbf{R}|} exp(\frac{Sim(\mathbf{V}_i,\mathbf{R}_n)}{\tau})} \\ &\simeq \frac{exp(\frac{Sim(\mathbf{V}_i,\mathbf{R}_j)}{\tau})}{\frac{|\mathbf{R}|}{N}\sum_{k=1}^{N} exp(\frac{Sim(\mathbf{V}_i,\mathbf{R}_{j_k})}{\tau})}, \end{aligned} \tag{3}$$

where $\mathbf{R}_{j_k}$ is a random subset sampled from training set, $\{j_k\}_{k=1}^{N}$ is the index, and $N$ is the batch size.

Then the non-matching loss for image-to-recipe could be defined as:

$$\mathcal{L}_{non}^{v2r} = -\frac{1}{N}\sum_{k=1}^{N}\sum_{p\in\overline{\mathcal{P}}_k^{v2r}} \log(1-p), \tag{4}$$

where $\overline{\mathcal{P}}_k^{v2r} = \{p_{kj}^{v2r} | j \neq k; j = 1, 2, \ldots, N\}$ is a probability set of negative image-recipe pairs. Similarity, non-matching loss for recipe-to-image could be defined as:

$$\mathcal{L}_{non}^{r2v} = -\frac{1}{N}\sum_{k=1}^{N}\sum_{p\in\overline{\mathcal{P}}_k^{r2v}} \log(1-p), \tag{5}$$

where $\overline{\mathcal{P}}_k^{r2v} = \{p_{ik}^{r2v} | i \neq k; i = 1, 2, \ldots, N\}$. To equally consider the two above retrieval tasks, we define the overall non-matching loss as:

$$\mathcal{L}_{non} = \mathcal{L}_{non}^{v2r} + \mathcal{L}_{non}^{r2v}. \tag{6}$$

By means of minimizing the non-matching loss, we can learn the complementary information from the negative pairs and avoid the matching imperfectly problem.

### 2) PARTIAL-MATCHING LOSS

It should be noted that applying the non-matching loss alone cannot be regarded as an optimal scheme because all the information (including the matchable parts) contained in positive pairs are lost under this setting. This fact raises another critical question—how to utilize the effective information in positive pairs while do not disturb the convergence. Note that, among the three components in a recipe, only the ingredients would appear in the corresponding image. It indicates that ingredients have the strongest association with the visual content, which is not only intuitive but can be seen from the data distribution [51]. Building on this, we extract the correlation information of features of ingredients and align it to the corresponding image features.

Given image embeddings $\mathbf{V} = \{\mathbf{V}_i\}_{i=1}^{n}$, and the ingredients embeddings $\mathbf{E}^{ing} = \{\mathbf{E}_i^{ing}\}_{i=1}^{n}$ of the corresponding recipes, where $n$ is the number of the image-recipe pairs. We define the similarity matrix of them respectively as:

$$\begin{cases} \mathbf{S}^{v} = \mathbf{V}\mathbf{V}^{\top} \\ \mathbf{S}^{ing} = \mathbf{E}^{ing}(\mathbf{E}^{ing})^{\top} \end{cases}, \tag{7}$$

where $\mathbf{V}^\top$ and $(\mathbf{E}^{ing})^\top$ denote the transpose of matrix $\mathbf{V}$ and $\mathbf{E}^{ing}$. Each element $\mathbf{S}^v_{ij}$ in $\mathbf{S}^v$ indicates the relationship between the $i$-th and the $j$-th feature of $\mathbf{V}$. Similarity, $\mathbf{S}^{ing}_{ij}$ indicates the adjacent information of $\mathbf{E}^{ing}$. Thereby, we define the partial-matching loss by:

$$\mathcal{L}_{part} = \left\| \mathbf{S}^v - \mathbf{S}^{ing} \right\|_2, \tag{8}$$

where $\|\cdot\|_2$ is the L2 norm. Using this loss helps to encourage the model to attend and align the matching information among positive pairs. So that we could make up for the deficiency of non-matching loss and combine these two losses to solve the matching imperfectly problem.

#### 3) TOTAL LOSS

Considering non-matching and partial-matching together, the total loss function of NMPM strategy can be formed as:

$$\mathcal{L}_{total} = \mathcal{L}_{non} + \lambda \mathcal{L}_{part}, \tag{9}$$

where $\lambda$ is a hyper-parameter to balance the preference of two losses.

### D. OPTIMIZATION

Our method is optimized in an end-to-end fashion. The optimization procedure is presented in Algorithm 1.

**Algorithm 1** Optimization Procedure for Our Method

**Input:** image-recipe pairs $\{\mathbf{X}^v_i, \mathbf{X}^r_j\}^n_{i,j=1}$, number of epoch $T$.
**Output:** parameters $\theta^v$, $\theta^r$.

```
Initialize parameters;
for t = 1 to T do
    repeat
        Compute V and R;
        for i, j = 1 to n do
            if i = j then
                Regularize V and E^ing by Eq.(7);
            else
                Calculate the matching probability between V_i
                and R_j using Eq.(4) and Eq.(5);
            end if
        end for
        Update the parameters θ^v, θ^r by Eq.(9) via gradient
        descent algorithm.
    until convergence
end for
```

## IV. EXPERIMENTS

In this section, extensive experiments are carried out to evaluate the performance of our method. In the following, the experiment settings are introduced firstly. Then, we discuss the experimental results in detail.

### A. EXPERIMENT SETTINGS

#### 1) DATASET

We conduct experiments on the largest cross-modal recipe dataset Recipe1M [51], which is collected over 1M cooking recipes and 800K food images from more than 24 popular cooking websites. We follow the official data splits: 238,399 image-recipe pairs for training, 51,119 pairs for validation and 51,303 pairs for testing. Within the dataset, each recipe, on average, contains 9.3 ingredients and 10.5 instructions. All recipes are composed in English.

#### 2) BASELINES

We compare our method with the following state-of-the-art baselines:

- JE [51] learns a joint embedding for different modalities and incorporates a classifier to predict food categories.
- AdaMin [41] uses a double triplet loss and proposes an adaptive strategy for informative triplet mining.
- R2GAN [28] adopts a GAN-based model with one generator and dual discriminators to lean compatible embeddings for cross-modal similarity measurement.
- MCEN [32] obtains modality-consistent embeddings by capturing the correlations between two modalities with latent variables.
- SN [44] applies three attention networks to enhance sentence-level information and uses an adversarial learning strategy to enhance modality alignment.
- SCAN [23] regularizes the embeddings of two modalities through aligning output semantic probabilities.
- HF-ICMA [40] considers intra- and inter- modal fusion and jointly derives the final image-recipe similarity from both local and global perspectives.
- SEJE [30] extracts additional semantic information through a two-phase deep feature engineering framework, which preprocesses data and trains model separately.
- M-SIA [45] learns multi-subspace information using multi-head attention networks to bridge the semantic gap between the two modalities.
- X-MRS [19] utilizes multilingual translations to regularize the model and jointly align the latent representations of images and recipes.
- H-T [22] applies a self-supervised loss to three components of recipes, leveraging the semantic relationships within them.
- T-Food [34] adopts a transformer decoder to capture interactions between recipe components.

#### 3) METRICS

Following prior works [22], [34], [51], we evaluate the retrieval performance (both image-to-recipe task and recipe-to-image task) using median rank (MedR), which is the median index of the retrieved samples for each query, and recall rate at top-$k$, representing the percentage of queries for which the correct sample index belongs to the top-$k$ retrieved

**TABLE 2. Main results. The cross-modal recipe retrieval results of methods evaluated with MedR (lower is better) and R@K (higher is better). The best results are presented in bold font.**

| 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 |
| JE [51] | 5.2 | 24.0 | 51.0 | 65.0 | 5.1 | 25.0 | 52.0 | 65.0 | 41.9 | - | - | - | 39.2 | - | - | - |
| AdaMin [41] | 2.0 | 39.8 | 69.0 | 77.4 | 2.0 | 40.2 | 68.1 | 78.7 | 13.2 | 14.9 | 35.3 | 45.2 | 12.2 | 14.8 | 34.6 | 46.1 |
| R2GAN [28] | 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 [32] | 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 [27] | 1.0 | 51.8 | 80.2 | 87.5 | 1.0 | 52.8 | 80.2 | 87.6 | 6.7 | 22.9 | 46.8 | 57.9 | 6.0 | 24.4 | 47.9 | 59.0 |
| SN [44] | 1.0 | 52.7 | 81.7 | 88.9 | 1.0 | 54.1 | 81.8 | 88.9 | 7.0 | 22.1 | 45.9 | 56.9 | 7.0 | 23.4 | 47.3 | 57.9 |
| SCAN [23] | 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 |
| HF-ICMA [40] | 1.0 | 55.1 | 86.7 | 92.4 | 1.0 | 56.8 | 87.5 | 93.0 | 5.0 | 24.0 | 51.6 | 65.4 | 4.2 | 25.6 | 54.8 | 67.3 |
| SEJE [30] | 1.0 | 58.1 | 85.8 | 92.2 | 1.0 | 58.5 | 86.2 | 92.3 | 4.2 | 26.9 | 54.0 | 65.6 | 4.0 | 27.2 | 54.4 | 66.1 |
| M-SIA [45] | 1.0 | 59.3 | 86.3 | 92.6 | 1.0 | 59.8 | 86.7 | 92.8 | 4.0 | 29.2 | 55.0 | 66.2 | 4.0 | 30.3 | 55.6 | 66.5 |
| X-MRS [19] | 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 [22] | 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) [22] | 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) [34] | 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) [34] | 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(ResNet-50) | 1.0 | 65.7 | 88.6 | 93.0 | 1.0 | 65.6 | 88.8 | 93.2 | 3.0 | 35.1 | 62.6 | 73.0 | 3.0 | 35.7 | 62.8 | 72.9 |
| CREAMY(ViT) | **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** |

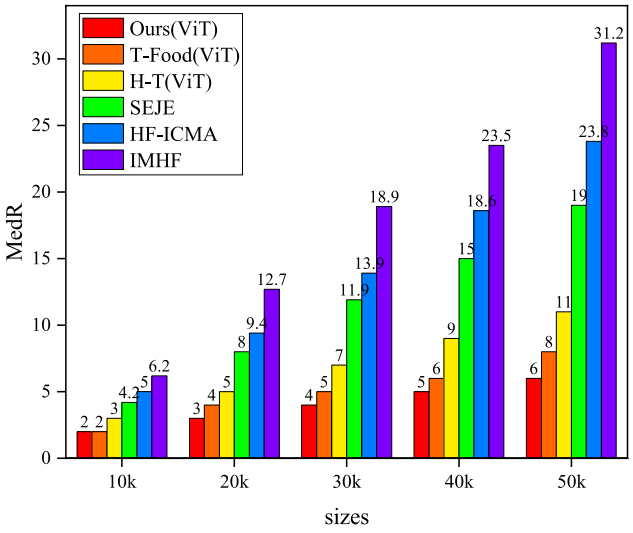



**FIGURE 4. Scalability analysis.**

samples. In our experiments, we use top-1, top-5 and top-10 in our experiments, represented as R@1, R@5 and R@10. We sample 1,000 pairs (1K set) and 10,000 (10K set) pairs on test partition, and repeat the process 10 times to report the mean results.

#### 4) IMPLEMENTATION DETAILS

Similar to previous work [22], we resize image to 256 pixels in their shorter dimension and crop them to $224 \times 224$ pixels. The image encoder is implementation using pre-trained ResNet-50 and ViT-base model, with an output size of 1024. For recipes, we truncate all components to a maximum length of 15, and the maximum number of sentences is set to 20. The hierarchical recipe encoder is implemented using two transformer encoders, each comprising two layers, with four attention heads in each layer. The dimension of the unit component is 512 and the size of final outputs is 1024. The model is trained using Adam optimizer with a batch size of 256, a learning rate of $\eta = 0.0001$, and a balance parameter of $\lambda = 0.001$.

#### 5) EXPERIMENTAL ENVIRONMENT

All our experiments are implemented using Python 3.7 on PyTorch 1.31.1 framework, running on a deep learning workstation with Intel(R) Core i9-12900K 3.9GHz, 128GB RAM, 1TB SSD and 2TB HDD storage, 2 NVIDIA GeForce RTX 3090Ti GPUs with Ubuntu-22.04.1 operating system.

### B. COMPARISON WITH STATE-OF-THE-ARTS

The performance comparison of the proposed method CREAMY with the baselines are reported in Table 2. It can be noticed that, our method outperforms others with a large margin both in 1K and 10K size over the listed metrics. Concretely, when using ResNet-50 as image encoder, our method achieves 1.7, 1.2, 0.6 R{1, 5, 10} improvement for recipe-to-image in 1K size, and 2.7, 2.4, 2.2 R{1, 5, 10} improvement for recipe-to-image in 10K size than X-MRS [19], the SOTA method using a CNN-based image encoder. When ResNet-50 is replaced by ViT, our method achieves 9.1, 3.4, 2.2 R{1, 5, 10} improvement for image-to-recipe in 1K size than SOTA method H-T(ViT) [22], 4.6, 4.6, 4.5 R{1, 5, 10} improvement for image-to-recipe in 10K size than SOTA method T-Food(ViT) [34]. These results confirm that our method is not susceptible to the variation of image encoding techniques. Whether using CNN or ViT to encode food images, the best performance can be achieved. Additionally, compared with the strongest competitor, T-Food(CLIP-ViT) [74] fine-tuned on a large-scale dataset, our method still shows obvious superiority: 1.0, 1.8, 2.2 R{1, 5, 10} improvement for image-to-recipe for 1K size, and 1.2, 0.9, 0.7 R{1, 5, 10} improvement for image-to-recipe in 10K size, which indicates that our method can obtain better performance than the state-of-the-arts with larger models. In addition, either ResNet-50 or ViT we adopt, our approach obtain a larger improvement in 10K size than 1K (2.22 point improvement on average in 1K size and 6.13 in 10K size

**TABLE 3. Ablation study. Evaluation of the impact of different parts in NMPM. The best results are presented in bold font.**

| | $\mathcal{L}_{tri}$ | $\mathcal{L}_{non}$ | $\mathcal{L}_{part}$ | Image-to-Recipe | | | | Recipe-to-Image | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | MedR | R@1 | R@5 | R@10 | MedR | R@1 | R@5 | R@10 |
| 1K | √ | | | 1.0 | 58.3 | 86.2 | 91.8 | 1.0 | 59.6 | 86.1 | 92.2 |
| | | √ | | 1.0 | 72.3 | 92.4 | 95.6 | 1.0 | **73.3** | 92.4 | 95.7 |
| | √ | | √ | 1.0 | 68.5 | 90.8 | 94.9 | 1.0 | 68.5 | 90.6 | 94.7 |
| | | √ | √ | **1.0** | **73.3** | **92.5** | **95.6** | **1.0** | 73.2 | **92.5** | **95.8** |
| 10K | √ | | | 4.1 | 26.8 | 54.7 | 66.5 | 4.0 | 37.5 | 55.1 | 66.8 |
| | | √ | | 2.0 | 43.2 | 70.5 | 79.6 | 2.0 | 43.8 | 70.4 | 79.5 |
| | √ | | √ | 2.0 | 38.1 | 66.2 | 76.3 | 2.0 | 38.6 | 66.3 | 76.0 |
| | | √ | √ | **2.0** | **44.6** | **71.6** | **80.4** | **2.0** | **45.0** | **71.4** | **80.0** |


**FIGURE 5. Examples of recipe-to-image retrieval on 10K test set. The top row are the query images, the second row are the retrieved recipes using our method, which are correctly matched with the ground truth, the third row are the retrieved recipes using H-T(ViT) [22].**

than H-T [22] using ResNet-50, 4.78 point improvement on average in 1K size and 9.33 point in 10K size than H-T(ViT) [22] using ViT). That is to say, our approach retrieves the plausible matches much better than previous methods do especially in a larger sampling size.

### C. SCALABILITY ANALYSIS

To investigate the scalability of CREAMY, we test it on different dataset sizes beyond 10K. As show in Figure 4, the gap between other solutions and ours increased when the test size increased. Actually, a larger test size means a larger number of negative pairs in candidates, it increases the difficulty to retrieve the positive samples, so the performance of other approaches decreased dramatically. However, CREAMY trains models by minimize the similarities between all the negative pairs, so that the positive samples who have the maximum similarities could be retrieved, even the test size increased. As a result, we gained a stable MedR performance even test size was enlarge.

### D. ABLATION STUDIES

We carry out an extensive ablation study to tease apart the effect by varying learning strategy. With ViT as the image encoder and hierarchical transformer as recipe encoder, three different loss functions, i.e., triplet loss $\mathcal{L}_{tri}$, non-matching loss $\mathcal{L}_{non}$ and partial-matching loss $\mathcal{L}_{part}$ are independently or collaboratively adopted to guide the cross-modal learning. In specific, we carry out the basic framework using $\mathcal{L}_{tri}$, then replace it with $\mathcal{L}_{non}$. Besides, we coupled the $\mathcal{L}_{part}$ with $\mathcal{L}_{tri}$ or $\mathcal{L}_{non}$, respectively. From Table 3 we can notice that the performance is significantly improved by replacing the triplet loss with non-matching loss. We advocate that


**FIGURE 6.** Examples of recipe-to-image retrieval on 10K test set. On the left are the query recipes, and the right are the top5 results of retrieved images. Among them, the matched images are highlight in a red box.

this improvement is obtained from eliminating the matching imperfectly problem in positive pairs by non-matching loss so as to lead a more correct optimization direction than triplet loss. Furthermore, partial-matching loss also brings additional enhancement upon both triplet loss and non-matching loss, which validates the partial-matching loss could further promote the alignment between two modalities. Note that, in the results of recipe-to-image in 1K size, the R@1 decreased 0.1 point when assembling partial-matching loss and non-matching loss together. We conjecture that the slight performance decrease is caused by the variety of ingredients (Even ingredients are most probably related to the food image directly, there may still be bits of mismatching features involved). This phenomenon, however, not appear in 10K setting, which indicates the matching information is more instrumental when the dataset is larger.

### *E. QUALITATIVE RESULTS*

#### 1) QUALITATIVE RESULTS ON IMAGE-TO-RECIPE RETRIEVAL

To further analyze the typical results on image-to-recipe retrieval of our method comparing to the strongest competitor H-T (ViT) [22], we choose four food images as queries. As presented in Figure 5, from left to right they are ''Tomato-Basil Chicken'', ''Tapenade Recipe'', ''Cornmeal Cranberry Bread'' and ''Low Cal Butter Sauce (Vegan Option)''. In the first two results, some of the ingredients are easy to recognize (e.g. softened tomatoes and leaves) directly while some are not (e.g. cubed chicken and blended anchovy fillets). Our method has found out the correct recipes, while H-T (ViT) retrieved the ingredients that are obvious in food image yet stumped by the hardly recognized ones. In the last two examples, most ingredients are invisible in the images, among which ''Cornmeal Cranberry Bread'' is more discernable.

H-T (ViT) returned the similar ingredients and instructions but failed to capture the exclusive ingredients: cornmeal and cranberry. The last query, particularly, is even hard for human to recognize the food from the image. It brought huge challenge to H-T (ViT) so that marked difference exists between the results of H-T (ViT) and the ground truth. Naturally, the ingredients in the above cases are out of shape after cooking, mixed together and even invisible, making it hard to match the features between the images and recipes. Even so, the proposed method captured the complementary information from negative pairs rather than matched the positive pairs forcibly, therefore, the accuracy of matching can be raised.

#### 2) QUALITATIVE RESULTS ON RECIPE-TO-IMAGE RETRIEVAL

Figure 6 visualizes the results of our method and H-T (ViT) on recipe-to-image retrieval. We picked up three different recipes ''Moroccan Skirt Steak Roasted Pepper Couscous'', ''Fresh Okara Cookies That Won't Crumble'' and ''Edamame Dip'' as the queries and displayed the top-5 results. In the first case, H-T (ViT) failed to match the ground truth while our method obtained the correct image at top-3. In the second test, CREAMY retrieved the correct image at top-1 and H-T (ViT) retrieved at top-2. However, all the top-5 images of our results are cookies while the results of H-T (ViT) are not, the reason behind which we conjectured is our CREAMY method successfully avoided the mismatching information in the query and correctly understood the goal. In the last comparison, both two methods returned the correct image at top-1. Note that, the ingredients in the second image of our results are stacked up similar to the ground truth. This phenomenon indicates that our method tried to align the relationships of visual details to the semantics in the query recipe (e.g. ''Spread onto bottom of…'', and ''Top with layers of…'') to the image, which is admittedly a correct manner to understand the data.

## V. CONCLUSION AND FUTURE WORK

In this paper, we propose a method for cross-modal recipe retrieval, named CREAMY, based on a new learning strategy NMPM. The proposed method focus on negative image-recipe pairs so that the matching imperfectly problem could be avoid. Concretely, we introduce the non-matching loss to maximize the distance between negative image-recipe pairs in a batch, and introduce partial-matching loss to do a regularize. We conducted experiments on Recipe1M dataset, the experimental results for MedR and Recall rate demonstrate the effectiveness of CREAMY. Additionally, we performed ablation studies and qualitative analysis on it, the evaluation results further confirm the utility of our NMPM strategy.

In the follow-up work, we will investigate the matching imperfectly problem in cross-modal image/recipe generation task, which remains to be significant but more challenging goals to achieve. In addition, another interesting and unsolved problem in cross-modal recipe retrieval is that the same ingredients may play different roles in different recipes. To address this issue, transfer learning will be attended by us. Specifically, external semantic knowledge about the ingredients will be token to improve the generalization of image-recipe matching. We expect that our work could be expanded to food engineering and nutritional science.

## REFERENCES


[1] C. Zhang, Y. Wang, L. Zhu, J. Song, and H. Yin, ''Multi-graph heterogeneous interaction fusion for social recommendation,'' *ACM Trans. Inf. Syst.*, vol. 40, no. 2, pp. 1–26, Apr. 2022.

[2] Y. Yamakata, A. Ishino, A. Sunto, S. Amano, and K. Aizawa, ''Recipe-oriented food logging for nutritional management,'' in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 6898–6904.

[3] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, and S. Jiang, ''Large scale visual food recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9932–9949, Aug. 2023.

[4] J. N. Bondevik, K. E. Bennin, Ö. Babur, and C. Ersch, ''A systematic review on food recommender systems,'' *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 122166.

[5] X. Ma, T. Zhang, and C. Xu, ''Multi-level correlation adversarial hashing for cross-modal retrieval,'' *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3101–3114, Dec. 2020.

[6] L. Zhu, J. Song, X. Wei, H. Yu, and J. Long, ''CAESAR: Concept augmentation based semantic representation for cross-modal retrieval,'' *Multimedia Tools Appl.*, vol. 81, no. 24, pp. 34213–34243, Oct. 2022.

[7] S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen, ''Cross-modal scene graph matching for relationship-aware image-text retrieval,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1497–1506.

[8] J. Wei, Y. Yang, X. Xu, J. Song, G. Wang, and H. T. Shen, ''Less is better: Exponential loss for cross-modal matching,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 5271–5280, Sep. 2023.

[9] C. Zhang, Z. Zhong, L. Zhu, S. Zhang, D. Cao, and J. Zhang, ''M2GUDA: Multi-metrics graph-based unsupervised domain adaptation for cross-modal hashing,'' in *Proc. Int. Conf. Multimedia Retr.*, Aug. 2021, pp. 674–681.

[10] M. Li, J. Fan, and Z. Lin, ''Non-Co-Occurrence enhanced multi-label cross-modal hashing retrieval based on graph convolutional network,'' *IEEE Access*, vol. 11, pp. 16310–16322, 2023.

[11] F. Ott, D. Rügamer, L. Heublein, B. Bischl, and C. Mutschler, ''Auxiliary cross-modal representation learning with triplet loss functions for online handwriting recognition,'' *IEEE Access*, vol. 11, pp. 94148–94172, 2023.

[12] K. Ouenniche, R. Tapu, and T. Zaharia, ''Vision-text cross-modal fusion for accurate video captioning,'' *IEEE Access*, vol. 11, pp. 115477–115492, 2023.

[13] L. Zhu, C. Zhang, J. Song, L. Liu, S. Zhang, and Y. Li, ''Multi-graph based hierarchical semantic fusion for cross-modal representation,'' in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2021, pp. 1–6.

[14] R.-C. Tu, X.-L. Mao, Q. Lin, W. Ji, W. Qin, W. Wei, and H. Huang, ''Unsupervised cross-modal hashing via semantic text mining,'' *IEEE Trans. Multimedia*, vol. 25, pp. 8946–8957, Feb. 2023.

[15] Y. Liu, Q. Wu, Z. Zhang, J. Zhang, and G. Lu, ''Multi-granularity interactive transformer hashing for cross-modal retrieval,'' in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 893–902.

[16] L. Zhu, L. Cai, J. Song, X. Zhu, C. Zhang, and S. Zhang, ''MSSPQ: Multiple semantic structure-preserving quantization for cross-modal retrieval,'' in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2022, pp. 631–638.

[17] H. Wang, G. Lin, S. Hoi, and C. Miao, ''Paired cross-modal data augmentation for fine-grained image-to-text retrieval,'' in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 5517–5526.

[18] Y.-C. Lien, H. Zamani, and W. B. Croft, ''Recipe retrieval with visual query of ingredients,'' in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1565–1568.

[19] R. Guerrero, H. X. Pham, and V. Pavlovic, ''Cross-modal retrieval and synthesis (X-MRS): Closing the modality gap in shared subspace learning,'' in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3192–3201.

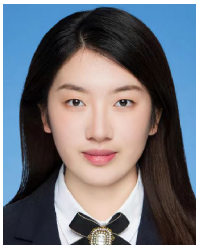
[20] D. P. Papadopoulos, E. Mora, N. Chepurko, K. W. Huang, F. Ofli, and A. Torralba, ''Learning program representations for food images and cooking recipes,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16538–16548.

[21] J. Chen, Y. Yin, and Y. Xu, "RecipeSnap—A lightweight image-to-recipe model," 2022, *arXiv:2205.02141*.
[22] A. Salvador, E. Gundogdu, L. Bazzani, and M. Donoser, "Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15475–15484.
[23] H. Wang, D. Sahoo, C. Liu, K. Shu, P. Achananuparp, E.-P. Lim, and S. C. H. Hoi, "Cross-modal food retrieval: Learning a joint embedding of food images and recipes with semantic consistency and attention mechanism," *IEEE Trans. Multimedia*, vol. 24, pp. 2515–2525, 2022.
[24] H. X. Pham, R. Guerrero, V. Pavlovic, and J. Li, "CHEF: Cross-modal hierarchical embeddings for food domain retrieval," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, 2021, pp. 2423–2430.
[25] Y. Tian, C. Zhang, Z. Guo, Y. Ma, R. Metoyer, and N. V. Chawla, "Recipe2Vec: Multi-modal recipe representation learning with graph neural networks," 2022, *arXiv:2205.12396*.
[26] J.-J. Chen, C.-W. Ngo, F.-L. Feng, and T.-S. Chua, "Deep understanding of cooking procedure for cross-modal recipe retrieval," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1020–1028.
[27] H. Wang, D. Sahoo, C. Liu, E.-P. Lim, and S. C. H. Hoi, "Learning cross-modal embeddings with adversarial networks for cooking recipes and food images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11572–11581.
[28] B. Zhu, C.-W. Ngo, J. Chen, and Y. Hao, "R$^2$GAN: Cross-modal recipe retrieval with generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11477–11486.
[29] Y. Sugiyama and K. Yanai, "Cross-modal recipe embeddings by disentangling recipe contents and dish styles," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2501–2509.
[30] Z. Xie, L. Liu, Y. Wu, L. Zhong, and L. Li, "Learning text-image joint embedding for efficient cross-modal retrieval with deep feature engineering," *ACM Trans. Inf. Syst.*, vol. 40, no. 4, pp. 1–27, Oct. 2022.
[31] M. Fontanellaz, S. Christodoulidis, and S. Mougiakakou, "Self-attention and ingredient-attention based model for recipe retrieval from image queries," in *Proc. 5th Int. Workshop Multimedia Assist. Dietary Manage.*, Oct. 2019, pp. 25–31.
[32] H. Fu, R. Wu, C. Liu, and J. Sun, "MCEN: Bridging cross-modal gap between cooking recipes and dish images with latent variable model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14570–14580.
[33] J. Li, X. Xu, W. Yu, F. Shen, Z. Cao, K. Zuo, and H. T. Shen, "Hybrid fusion with intra-and cross-modality attention for image-recipe retrieval," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2021, pp. 244–254.
[34] M. Shukor, G. Couairon, A. Grechka, and M. Cord, "Transformer decoders with multimodal regularization for cross-modal food retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 4566–4577.
[35] L. Castrejón, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba, "Learning aligned cross-modal representations from weakly aligned data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2940–2949.
[36] L. Zhu, J. Song, X. Zhu, C. Zhang, S. Zhang, and X. Yuan, "Adversarial learning-based semantic correlation representation for cross-modal retrieval," *IEEE MultimediaMag.*, vol. 27, no. 4, pp. 79–90, Oct. 2020.
[37] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, and S. Marchand-Maillet, "Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 4, pp. 1–23, Nov. 2021.
[38] P. Zeng, L. Gao, X. Lyu, S. Jing, and J. Song, "Conceptual and syntactical cross-modal alignment with cross-level consistency for image-text matching," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2205–2213.
[39] Z. Xie, L. Liu, L. Li, and L. Zhong, "Efficient deep feature calibration for cross-modal joint embedding learning," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 43–51.
[40] J. Li, J. Sun, X. Xu, W. Yu, and F. Shen, "Cross-modal image-recipe retrieval via intra- and inter-modality hybrid fusion," in *Proc. Int. Conf. Multimedia Retr.*, Aug. 2021, pp. 173–182.
[41] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, and M. Cord, "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 35–44.
[42] M. Fain, N. Twomey, A. Ponikar, R. Fox, and D. Bollegala, "Dividing and conquering cross-modal recipe retrieval: From nearest neighbours baselines to SoTA," 2019, *arXiv:1911.12763*.
[43] D. Cao, J. Chu, N. Zhu, and L. Nie, "Cross-modal recipe retrieval via parallel- and cross-attention networks learning," *Knowl.-Based Syst.*, vol. 193, Apr. 2020, Art. no. 105428.
[44] Z. Zan, L. Li, J. Liu, and D. Zhou, "Sentence-based and noise-robust cross-modal retrieval on cooking recipes and food images," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 117–125.
[45] L. Li, M. Li, Z. Zan, Q. Xie, and J. Liu, "Multi-subspace implicit alignment for cross-modal retrieval on cooking recipes and food images," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2021, pp. 3211–3215.
[46] B. Zhu, C.-W. Ngo, J. Chen, and W.-K. Chan, "Cross-lingual adaptation for recipe retrieval with mixup," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2022, pp. 258–267.
[47] Y. Tian, C. Zhang, R. Metoyer, and N. V. Chawla, "Recipe representation learning with networks," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2021, pp. 1824–1833.
[48] Z. Xie, L. Li, L. Zhong, J. Liu, and L. Liu, "Cross-modal retrieval between event-dense text and image," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2022, pp. 229–238.
[49] J. Yang, J. Chen, and K. Yanai, "Transformer-based cross-modal recipe embeddings with large batch training," in *Proc. Int. Conf. Multimedia Model.* Bergen, Norway: Springer, 2023, pp. 471–482.
[50] M. Wahed, X. Zhou, T. Yu, and I. Lourentzou, "Fine-grained alignment for cross-modal recipe retrieval," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2024, pp. 5584–5593.
[51] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, "Learning cross-modal embeddings for cooking recipes and food images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3020–3028.
[52] Y. Chen, D. Zhou, L. Li, and J.-M. Han, "Multimodal encoders for food-oriented cross-modal retrieval," in *Proc. 5th Int. Joint Conf. Web Big Data*, Guangzhou, China. Cham, Switzerland: Springer, 2021, pp. 253–266.
[53] L. Li, C. Hu, H. Zhang, and A. M. V. V. Sai, "Cross-modal image-recipe retrieval via multimodal fusion," in *Proc. ACM Multimedia Asia*, Dec. 2023, pp. 1–7.
[54] J. Sun and J. Li, "PBLF: Prompt based learning framework for cross-modal recipe retrieval," in *Proc. Int. Symp. Artif. Intell. Robot.* Shanghai, China: Springer, 2022, pp. 388–402.
[55] M. Shukor, N. Thome, and M. Cord, "Vision and structured-language pretraining for cross-modal food retrieval," 2022, *arXiv:2212.04267*.
[56] M. Shukor, N. Thome, and M. Cord, "Vision and structured-language pretraining for cross-modal food retrieval," *J. SSRN*, 2023. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4511116
[57] B. Prakash Voutharoja, P. Wang, L. Wang, and V. Guan, "MALM: Mask augmentation based local matching for food-recipe retrieval," 2023, *arXiv:2305.11327*.
[58] X. Huang, J. Liu, Z. Zhang, and Y. Xie, "Improving cross-modal recipe retrieval with component-aware prompted CLIP embedding," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 529–537.
[59] X. Xu, J. Sun, Z. Cao, Y. Zhang, X. Zhu, and H. T. Shen, "TFUN: Trilinear fusion network for ternary image-text retrieval," *Inf. Fusion*, vol. 91, pp. 327–337, Mar. 2023.
[60] P. Hu, Z. Huang, D. Peng, X. Wang, and X. Peng, "Cross-modal retrieval with partially mismatched pairs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9595–9610, Aug. 2023.
[61] F. Han, R. Guerrero, and V. Pavlovic, "CookGAN: Meal image synthesis from ingredients," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1450–1458.
[62] H. Wang, G. Lin, S. C. H. Hoi, and C. Miao, "Cycle-consistent inverse GAN for text-to-image synthesis," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 630–638.
[63] S. Pan, L. Dai, X. Hou, H. Li, and B. Sheng, "ChefGAN: Food image generation from recipes," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4244–4252.
[64] B. Zhu and C.-W. Ngo, "CookGAN: Causality based text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5518–5526.
[65] A. Salvador, M. Drozdzal, X. Giro-I-Nieto, and A. Romero, "Inverse cooking: Recipe generation from food images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10445–10454.

[66] M. Zhang, G. Tian, H. Gao, S. Liu, and Y. Zhang, "Multimodal feature fusion and exploitation with dual learning and reinforcement learning for recipe generation," *Appl. Soft Comput.*, vol. 126, Sep. 2022, Art. no. 109281.
[67] H. Wang, G. Lin, S. C. H. Hoi, and C. Miao, "Decomposing generation networks with structure prediction for recipe generation," 2020, *arXiv:2007.13374*.
[68] H. Wang, G. Lin, S. C. Hoi, and C. Miao, "Structure-aware generation network for recipe generation from images," in *Proc. 16th Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 359–374.
[69] H. Wang, G. Lin, S. C. H. Hoi, and C. Miao, "Learning structural representations for recipe generation and food retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3363–3377, Mar. 2023.
[70] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[71] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[73] F. Liang, C. Liu, and R. J. Carroll, "Stochastic approximation in Monte Carlo computation," *J. Amer. Stat. Assoc.*, vol. 102, no. 477, pp. 305–320, Mar. 2007.
[74] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

**ZHUOYANG ZOU** received the B.Eng. degree in communication engineering from the School of Electronic Information and Automation, Guilin University of Aerospace Technology, in 2018. She is currently pursuing the Ph.D. degree in agricultural information engineering with Hunan Agricultural University, China. Her main research interests include agricultural information, deep learning, natural language processing, and computer vision.

**XINGHUI ZHU** received the M.Eng. degree in computer science and technology from the School of Computer Science, National University of Defense Technology, in 2004, and the Ph.D. degree in land resources and information technology from the School of Resources and Environment, Hunan Agricultural University, China, in 2018. He is currently a Professor with the College of Information and Intelligence, Hunan Agricultural University. He has published more than 60 research articles. His main research interests include agricultural information, the Internet of Things, and distributed computing.



**QINYING ZHU** received the bachelor's degree in film and television production from the Film and Television Academy, Yunnan Arts University, in 2023. She is currently pursuing the master's degree with the College of Information and Intelligence, Hunan Agricultural University, China. Her research interests include computer vision and image recognition.

**YI LIU** received the master's degree from the School of Philosophy, Changsha University of Science and Technology, in 2021. He is currently a Lecturer with the School of Information and Intelligent Science and Technology, Hunan Agricultural University. He holds the position of a Special Researcher with the Key Research Base of Philosophy and Social Sciences in Hunan Province-the Innovative Research Center for Philosophy of Science and Technology and Ethical Governance. He is also a Researcher with the Research Center for Science and Society Development, Hunan Normal University. With more than ten published articles, he has also led eight research projects, receiving recognition through more than 30 awards. Additionally, he has authored a monograph. His primary research interests include philosophy of science and technology, and ethics of science and technology.

**LEI ZHU** (Member, IEEE) received the M.Sc. degree in control engineering from the School of Information Science and Engineering, Central South University, in 2014, and the Ph.D. degree in computer science and technology from the School of Computer Science and Engineering, Central South University, in 2020. He is currently a Lecturer with the College of Information and Intelligence, Hunan Agricultural University, China. He has published more than 30 research articles, some of them have appeared at the competitive venues, including *ACM TOIS*, *ACM TOMM*, ACM SIGIR, ACM ICMR, IEEE ICME, and IEEE MULTIMEDIA. He served as a Program Committee Member for the IJCAI 2020 and IJCAI 2021, as one of the Special Session Co-Chairs for the ICCSI 2023, and as an Invited Journal Reviewer for over some leading journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS, *ACM TOIS*, *ACM TOMM*, *ACM TKDD*, *ESWA*, *Signal, Image and Video Processing*, *Chinese Journal of Electronics*, and *Signal Processing: Image Communication*. His main research interests include machine learning, deep learning, pattern recognition, and computer vision.