Cross-Modal Food Retrieval: Learning a Joint Embedding of Food Images and Recipes with Semantic Consistency and Attention Mechanism

Hao Wang, Doyen Sahoo, Chenghao Liu, Ke Shu, Palakorn Achananuparp, Ee-peng Lim, and Steven C. H. Hoi, *Fellow, IEEE*

***Abstract*—Food retrieval is an important task to perform analysis of food-related information, where we are interested in retrieving relevant information about the queried food item such as ingredients, cooking instructions, etc. In this paper, we investigate cross-modal retrieval between food images and cooking recipes. The goal is to learn an embedding of images and recipes in a common feature space, such that the corresponding image-recipe embeddings lie close to one another. Two major challenges in addressing this problem are 1) large intra-variance and small inter-variance across cross-modal food data; and 2) difficulties in obtaining discriminative recipe representations. To address these two problems, we propose Semantic-Consistent and Attention-based Networks (SCAN), which regularize the embeddings of the two modalities through aligning output semantic probabilities. Besides, we exploit a self-attention mechanism to improve the embedding of recipes. We evaluate the performance of the proposed method on the large-scale Recipe1M dataset, and show that we can outperform several state-of-the-art cross-modal retrieval strategies for food images and cooking recipes by a significant margin.**

***Index Terms*—Deep Learning, Cross-Modal Retrieval, Vision-and-Language.**

I. Introduction

Food plays an essential role in human daily life. To discover the relationship between cross-modal food data, i.e. food images and recipes, we aim to address the problem of cross-modal food retrieval based on a large amount of heterogeneous food dataset [1]. Specifically, we take the cooking recipes (ingredients & cooking instructions) as the query to retrieve the food images, and vice versa. Recently, there have been many works [1], [2], [3] on cross-modal food retrieval, where they mainly learn the joint embeddings for recipes and images with vanilla pair-wise loss to achieve the cross-modal alignment. Despite those efforts, cross-modal food retrieval remains challenging mainly due to the following two reasons: 1) the large intra-class variance across food data pairs, and 2) the difficulties of obtaining discriminative recipe representation.

In cross-modal food data, given a recipe, we may have many food images that are cooked by different chefs. Besides, the images from different recipes can look very similar because they have similar ingredients. Hence, the data representation from the same food can be different, but different food may have similar data representations. This leads to large intra-class variance but small inter-class variance in food data. Existing studies [2], [4], [5] only address the small inter-class variance problem by utilizing triplet loss to measure the similarities between cross-modal data. Specifically, the objective of triplet loss is to make inter-class feature distance larger than intra-class feature distance by a predefined margin [6]. Therefore, cross-modal instances from the same class may form a loose cluster with a large average intra-class distance. As a consequence, it eventually results in less-than-optimal ranking, i.e., irrelevant images are closer to the queried recipe than relevant images. See Figure 1 for an example.

Besides, many recipes share common ingredients in different food. For instance *fruit salad* has ingredients of *apple*, *orange* and *sugar* etc., where *apple* and *orange* are the main ingredients, while *sugar* is one of the ingredients in many other foods. If the embeddings of ingredients are treated equally during training, the features learned by the model may not be discriminative enough. In addition, the cooking instructions crawled from cooking websites tend to be noisy, some instructions turn out irrelevant to cooking e.g. *'Enjoy!'*, which convey no information for cooking instruction features but degrade the performance of cross-modal retrieval task. In order to find the attended ingredients, Chen et al. [5] apply a two-layer deep attention mechanism, which learns joint features by locating the visual food regions that correspond to ingredients. However, this method relies on high-quality food images and essentially increases the computational complexity.

To resolve those issues, we propose a novel unified framework of **S**emantic-**C**onsistent and **A**ttention-Based **N**etwork (SCAN) to improve the cross-modal food retrieval performance. The pipeline of the framework is shown in Figure 2. To reduce the intra-class variance, we introduce a semantic consistency loss, which imposes Kullback-Leibler (KL) Divergence to minimize the difference between the output semantic probabilities of paired image and recipe, such that the image and recipe representations would follow similar distributions. In order to obtain discriminative recipe representations, we combine the self-attention mechanism [7] with LSTM to find the key ingredients and cooking instructions for each recipe. Without requiring food images or adding extra layers, we can learn better discriminative recipe embeddings, compared to

Work done in Singapore Management University.

Hao Wang is with Nanyang Technological University; e-mail: hao005@ntu.edu.sg.

Doyen Sahoo and Chenghao Liu are with Salesforce Research Asia; e-mail: {doyensahoo, twinsken}@gmail.com

Ke Shu, Palakorn Achananuparp, and Ee-peng Lim are with Singapore Management University; e-mail: {keshu, palakorna, eplim}@smu.edu.sg.

Steven C. H. Hoi is with Singapore Management University and Salesforce Research Asia; e-mail: chhoi@smu.edu.sg.

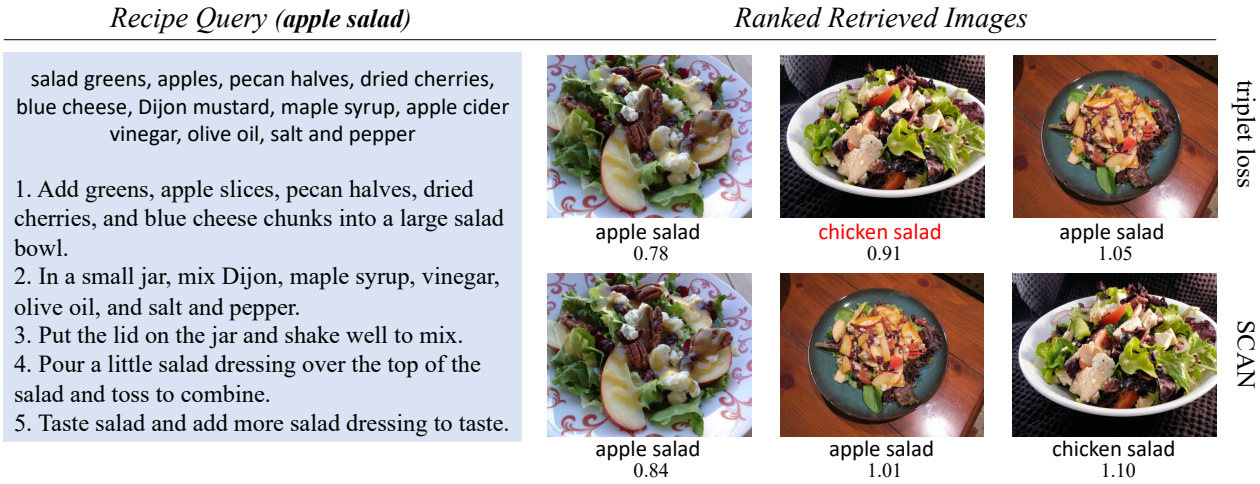


Fig. 1. **Recipe-to-image retrieval ranked results:** Take an *apple salad* recipe as the query, which contains ingredients and cooking instructions, we show the retrieval results based on **Euclidean distance** (as the numbers indicated in the figure) for 3 different food images with large intra-class variance and small inter-class variance, i.e. images of *apple salad* have different looks, while *chicken salad* image is more similar to *apple salad*. We rank the retrieved results of using (i) vanilla triplet loss; and (ii) our proposed SCAN model. It shows vanilla triplet loss outputs a wrong ranking order, while SCAN can provide more precise ranking results.

that trained with plain LSTM.

Our work makes two major contributions as follows:

- We introduce a semantic consistency loss to cross-modal food retrieval task. The result shows that it can align cross-modal matching pairs and reduce the intra-class variance of food data representations.
- We integrate the self-attention mechanism with LSTM, and learn discriminative recipe features without requiring the food images. It is useful to discriminate samples of similar recipes.

We perform extensive experimental analysis on Recipe1M, which is the largest cross-modal food dataset and available in the public. We find that our proposed cross-modal food retrieval approach SCAN outperforms state-of-the-art methods. Finally, we show some visualizations of the retrieved results.

II. Related Work

A. Cross-modal Retrieval

Our work is closely related to the general cross-modal retrieval task, which aims to retrieve the corresponding instance of different modalities based on the given query. The general idea of cross-modal retrieval is to correlate heterogeneous data, mapping the data from different modalities to the common space. As an early work for multi-media, Canonical Correlation Analysis (CCA) [8] utilizes global alignment to allow the data mapping of different modalities with similar semantics to be close in the common space, by maximizing the correlation between cross-modal similar pairs. However, CCA-based approaches model the cross-modal data only by linear projections, when it comes to large-scale complex real-world data, it is difficult for CCA to fully model the correlations.

Many recent works [9], [10], [11] utilize deep architectures for cross-modal retrieval, which have the advantage of capturing complex non-linear cross-modal correlations. Specifically, to improve the efficiency in retrieval process, hashing has been introduced to multimedia retrieval [10], [11]. With the advent of generative adversarial networks (GANs) [12], which are helpful to model the data distributions, some adversarial training methods [13], [14], [15] are frequently used for modality fusion. Peng et al. [14] utilize two kinds of discriminative models to simultaneously conduct intra-modality and inter-modality discrimination, and model the joint distribution over the data of different modalities. [13], [16] incorporate generative processes into the cross-modal feature embedding. Specifically, Gu et al. [13] try to generate the images from text features and generate corresponding captions from the image features. In this way, they learn not only the global abstract features but also the local grounded features. To address the challenge that unpaired data may exist in the cross-modal dataset, Jing et al. [16] propose to learn modality-invariant representations with autoencoders, which are further dual-aligned at the distribution level and the semantic level. To learn fine-grained phrase correspondence, Liu et al. [17] construct textual and visual graph, they learn the cross-modal correspondence by node-level and structure-level matching.

B. Food Computing

Food computing [18] utilizes computational methods to analyze the food data including the food images and recipes. Many food-related computational tasks have been widely researched, like food recognition [19], [20], [21], [22], retrieval [23], [1], recommendation [24], [25] and recipe generation [26], etc. In this paper, we mainly investigate the cross-modal food retrieval problem based on Recipe1M dataset [1].

Recipe1M [1] is currently the largest cross-modal food dataset, which was scraped from over two dozen popular cooking websites, and contains rich cooking instructions, ingredient information, and the corresponding cooked food

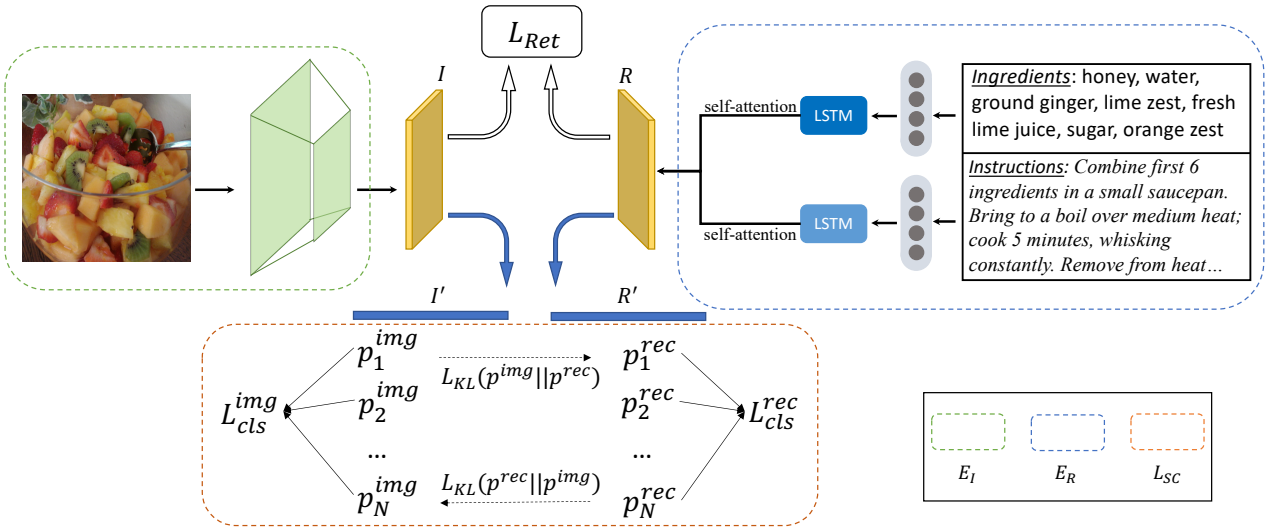


Fig. 2. Our proposed framework for cross-modal retrieval task. We have two branches to encode food images and recipes respectively. One embedding function E_I is designed to extract food image representations I, where a CNN is used. The other embedding function E_R is composed of two LSTMs with self-attention mechanism, designed for obtaining discriminative recipe representations R. I and R are fed into retrieval loss (triplet loss) L_{Ret} to do cross-modal retrieval learning. We add another FC transformation on I and R with the output dimensionality as the number of food categories, to obtain the semantic probabilities p^{img} and p^{rec}, where we utilize semantic consistency loss L_{SC} to correlate food image and recipe data.

images. Besides, half amount of the data in Recipe1M has semantic food category labels, which are extracted from the food titles in the cooking websites. Recipe1M is proposed mainly for the cross-modal food retrieval task.

[27], [23] are early works in cross-modal food retrieval. In [27], a multi-task deep learning architecture is proposed for simultaneous ingredient and food recognition. The learned visual features and semantic attributes of ingredients are then used for recipe retrieval, but they only test their model in a small-scale dataset, and cannot demonstrate the efficacy in real-world large-scale data. Min et al. [23] utilize a multi-modal Deep Boltzmann Machine for recipe-image retrieval. [5], [4] integrate attention mechanism into cross-modal retrieval, Chen et al. [5] introduce a stacked attention network (SAN) to learn joint space from images and recipes for cross-modal retrieval. However, SAN only considers ingredient lists and ignores the rich information provided by cooking instructions, so they have poor performance in Recipe1M dataset. Consequently, Chen et al. improve the previous work SAN in [4], where they make full use of the ingredient, cooking instruction, and title (food category) information of Recipe1M, and concatenate the three types of features above to construct the recipe embeddings. Compared with the self-attention mechanism we adopt in our model, both [5] and [4] add extra learnable parameters to compute the attended parts, which increase the computational complexity. In order to have better regularization on the shared representation space learning, [1], [2] both corporate the semantic labels with the joint training. Salvador et al. [1] develop a hybrid neural network architecture with a cosine embedding loss for retrieval learning and a cross-entropy loss for classification, such that a joint common space for image and recipe embeddings can be learned for cross-modal retrieval. [2] is an extended version of [1], providing a double-triplet strategy to express both the retrieval loss and the classification loss.

Different from existing cross-modal food retrieval work, we propose a novel semantic consistency loss with a self-attention mechanism, where we impose regularization on the output semantic probabilities of paired food image and recipe embeddings, to correlate the learned food image and recipe representations. Self-attention helps learn discriminative recipe embeddings without depending on the food images or adding some extra learnable parameters.

III. Proposed Methods

In this section, we introduce our proposed model, where we utilize food image-recipe paired data to learn cross-modal embeddings as shown in Figure 2.

A. Overview

We formulate the proposed cross-modal food retrieval with three networks, i.e. one convolutional neural network (CNN) for food image embeddings, and two LSTMs to encode ingredients and cooking instructions respectively. The food image representations I can be obtained from the output of CNN directly, while the recipe representations R come from the concatenation of the ingredient features $f_{ingredient}$ and instruction features $f_{instruction}$. Specifically, for obtaining discriminative ingredient and instruction embeddings, we integrate the self-attention mechanism [7] into the LSTM embedding. Triplet loss is used as the main loss function L_{Ret} to map cross-modal data to the common space, and semantic consistency loss L_{SC} is utilized to align cross-modal matching pairs for retrieval task, reducing the intra-class variance of food data.

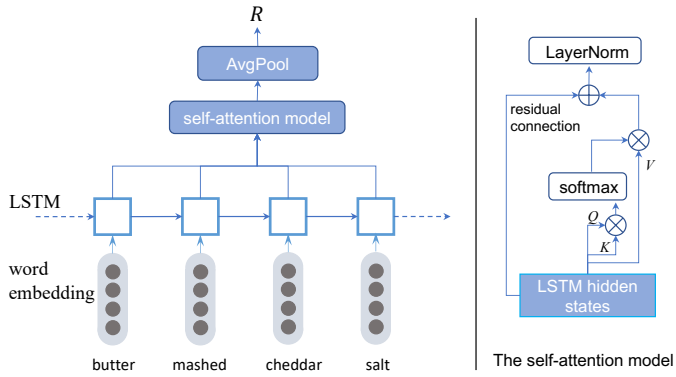


Fig. 3. The structure of ingredient (instruction) embedding model with self-attention mechanism. Q, K and V denote queries, keys and values respectively.

The overall objective function of the proposed SCAN is given as:

$$L = L_{Ret} + \lambda L_{SC}, \tag{1}$$

B. Recipe Embedding

We use two LSTMs to get ingredient and instruction representations $f_{ingredient}$, $f_{instruction}$, concatenate them and pass through a fully-connected layer to give a 1024-dimensional feature vector, as the recipe representation R.

1) Ingredient Representation Learning: Instead of word-level word2vec representations, ingredient-level word2vec representations are used in ingredient embedding. To be specific, *ground ginger* is regarded as a single word vector, instead of two separate word vectors of *ground* and *ginger*.

We integrate the self-attention mechanism with LSTM output to construct recipe embeddings. The purpose of applying the self-attention model lies in assigning higher weights to main ingredients for different food items, making the attended ingredients contribute more to the ingredient embedding, while reducing the effect of common ingredients. The self-attention structure is shown in Figure 3.

Given an ingredient input $\{z_1, z_2, ..., z_n\}$, we first encode it with pretrained embeddings from word2vec algorithm to obtain the ingredient representation Z_t. Then $\{Z_1, Z_2, ..., Z_n\}$ will be fed into the one-layer bidirectional LSTM as a sequence step by step. For each step t, the recurrent network takes in the ingredient vector Z_t and the output of previous step h_{t-1} as the input, and produces the current step output h_t by a non-linear transformation, as follow:

$$h_t = \tanh(\mathbf{W}Z_t + \mathbf{U}h_{t-1} + b), \tag{2}$$

The bidirectional LSTM consists of a forward hidden state $\overrightarrow{h_t}$ which processes ingredients from Z_1 to Z_n and a backward hidden state $\overleftarrow{h_t}$ which processes ingredients from Z_n to Z_1. We obtain the representation h_t of each ingredient z_t by concatenating $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$, i.e. $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$, so that the representation of the ingredient list of each food item is $H = \{h_1, h_2, ..., h_n\}$.

We further measure the importance of ingredients in the recipe with the self-attention mechanism which has been studied in Transformer [7], where the input comes from queries Q and keys K of dimension d_k, and values V of dimension d_v (the definition of Q, K and V can be referred in [7]), we compute the attention output as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{3}$$

Different from the earlier attention-based methods [4], we use self-attention mechanism where all of the keys, values and queries come from the same ingredient representation H. Therefore, the computational complexity is reduced since it is not necessary to add extra layers to train attention weights. The ingredient attention output H_{attn} can be formulated as:

$$\begin{aligned} H_{attn} &= \text{Attention}(H, H, H) \\ &= \text{softmax}(\frac{HH^T}{\sqrt{d_h}})H, \end{aligned} \tag{4}$$

where d_h is the dimension of H. In order to enable unimpeded information flow for recipe embedding, skip connections are used in the attention model. Layer normalization [28] is also used since it is effective in stabilizing the hidden state dynamics in recurrent network. The final ingredient representation $f_{ingredient}$ is generated from summation of H and H_{attn}, which can be defined as:

$$f_{ingredient} = \text{LayerNorm}(H_{attn} + H), \tag{5}$$

2) Instruction Representation Learning: Considering that cooking instructions are composed of a sequence of variable-form and lengthy sentences, we compute the instruction embedding with a two-stage LSTM model. For the first stage, we apply the same approach as [1] to obtain the representations of each instruction sentence, in which it uses skip-instructions [1] with the technique of skip-thoughts [29].

The next stage is similar to the ingredient representation learning. We feed the pre-computed fixed-length instruction sentence representation into the LSTM model to generate the hidden representation of each cooking instruction sentence. Based on that, we can obtain the self-attention representation. The final instruction feature $f_{instruction}$ is generated from the layer normalization function on the previous two representations, as we formulate in the last section. By doing so, we are able to find the key sentences in cooking instruction. Some visualizations on attended ingredients and cooking instructions can be found in Section IV-G.

C. Image Embedding

We use ResNet-50 [30] pretrained on ImageNet to encode food images. The dimension of the final food image features I is 1024, which is identical to that of recipe features R.

D. Cross-modal Food Retrieval Learning

Triplet loss is utilized to do retrieval learning, the objective function is:

$$L_{Ret} = \sum_{I} [d(I_a, R_p) - d(I_a, R_n) + \alpha]_+ + \sum_{R} [d(R_a, I_p) - d(R_a, I_n) + \alpha]_+ , \quad (6)$$

where $d(\bullet)$ is the Euclidean distance, subscripts a, p and n refer to anchor, positive and negative samples respectively and α is the margin. The summation symbol means that we construct triplets and do the training for all samples in the mini-batch. To improve the effectiveness of training, we adopt the *BatchHard* idea proposed in [31]. Specifically in a mini-batch, each sample can be used as an anchor, then for each anchor, we select the closest negative sample and the farthest positive sample to construct the triplet.

E. Semantic Consistency

Given the pairs of food image and recipe representations I, R, we first transform I and R into I' and R' for classification with an extra FC layer. The dimension of the output is same as the number of categories N. The probabilities of food category i can be computed by a softmax activation as:

$$p_i^{img} = \frac{exp(I'_i)}{\sum_{i=1}^{N} exp(I'_i)}, \quad (7)$$

$$p_i^{rec} = \frac{exp(R'_i)}{\sum_{i=1}^{N} exp(R'_i)}, \quad (8)$$

where N represents the total number of food categories. Given p_i^{img} and p_i^{rec}, where $i \in \{1, 2, ..., N\}$, the predicted label l^{img} and l^{rec} for each food item can be obtained. We formulate the classification (cross-entropy) loss as $L_{cls}(p^{img}, p^{rec}, c^{img}, c^{rec})$, where c^{img}, c^{rec} are the ground-truth class label for food image and recipe respectively.

In the prior work [1], $L_{cls}(p^{img}, p^{rec}, c^{img}, c^{rec})$ consists of L_{cls}^{img} and L_{cls}^{rec}, which are treated as two independent classifiers, focusing on the regularization on the embeddings from food images and recipes separately. However, food image and recipe embeddings come from heterogeneous modalities, the output probabilities of each category can be significantly different, i.e. for each food item, the distributions of p_i^{img} and p_i^{rec} remain big variance. As a result, the distance of intra-class features remains large. To improve image-recipe matching and make the probabilities predicted by different classifiers consistent, we minimize Kullback-Leibler (KL) Divergence between p_i^{img} and p_i^{rec} of paired cross-modal data for each food item, which can be formulated as:

$$L_{KL}(p^{img} \| p^{rec}) = \sum_{i=1}^{N} p_i^{img} \log \frac{p_i^{img}}{p_i^{rec}}, \quad (9)$$

$$L_{KL}(p^{rec} \| p^{img}) = \sum_{i=1}^{N} p_i^{rec} \log \frac{p_i^{rec}}{p_i^{img}}, \quad (10)$$

By aligning the output probabilities of cross-modal data representations, we minimize the intra-class variance with back-propagation. The overall semantic consistency loss L_{SC} is defined as:

Algorithm 1 Pseudocode of SCAN in a PyTorch-like style.

```
# load a minibatch containing ingredients, instructions
    and images with N samples
for (ingr, instr, img) in loader:

   # perform word embedding on ingredients
   Z = Embedding.forward(ingr)
   # compute LSTM features, Eq. (2)
   H = LSTM.forward(Z)
   # set the temperate
   t = sqrt(dimension_H)
   # compute attention scores
   attn = bmm(H, H.T) / t
   # compute attention outputs, Eq. (4)
   output = bmm(attn, H)
   # use residual connection to get the final self-
       attention outputs, Eq. (5)
   f_ingredient = LayerNorm(output + H)

   # use self-attention to get the instruction features
   f_instruction = SelfAttention.forward(instr)

   # compute the recipe features
   R = cat([f_ingredient, f_instruction], dim=1)
   # compute the image features
   I = CNN.forward(img)
   # compute triplet loss, Eq. (6)
   L_Ret = TripletLoss(R, I)

   # compute the class probabilities for R and I, Eq. (7,
       8)
   p_rec = softmax(R)
   p_img = softmax(I)
   # compute cross-entropy loss
   L_cls_rec = CrossEntropyLoss(p_rec, labels)
   L_cls_img = CrossEntropyLoss(p_img, labels)
   L_cls = (L_cls_rec + L_cls_img) / 2
   # compute KL divergence between recipes and images, Eq.
       (9, 10)
   L_KL = (KL(p_rec||p_img) + KL(p_img||p_rec)) / 2
   # compute the semantic consistency loss, Eq. (11)
   L_SC = L_cls + L_KL

   # Eq. (1)
   loss = L_Ret + L_SC

   # Adam update
   loss.backward()
```

`sqrt`: square root; `bmm`: batch matrix multiplication; `cat`: concatenation.

$$L_{SC} = \{(L_{cls}^{img} + L_{KL}(p^{rec} \| p^{img})) + (L_{cls}^{rec} + L_{KL}(p^{img} \| p^{rec}))\}/2. \quad (11)$$

IV. Experiments

A. Dataset

We conduct extensive experiments to evaluate the performance of our proposed methods in Recipe1M dataset [1], the largest cooking dataset with recipe and food image pairs available to the public. Recipe1M was scraped from over 24 popular cooking websites and it not only contains the image-recipe paired labels but also more than half the amount of the food data with semantic category labels extracted from food titles on the websites. The category labels provide semantic information for the cross-modal retrieval task, making it fit in our proposed method well. The paired labels and category labels construct the hierarchical relationships among the food. One food category (e.g. *fruit salads*) may contain hundreds of different food pairs, since there are many recipes of different *fruit salads*.

We perform the cross-modal food retrieval task based on food data pairs, i.e. when we take the recipes as the query to do the retrieval, the ground truth will be the food images in food data pairs, and vice versa. We use the original Recipe1M data

TABLE I

MAIN RESULTS. EVALUATION OF THE PERFORMANCE OF OUR PROPOSED METHOD COMPARED AGAINST THE BASELINES. THE MODELS ARE EVALUATED ON THE BASIS OF MEDR, WHERE LOWER IS BETTER, AND R@K (%), WHERE HIGHER IS BETTER.

Size of Test Set		Image-to-Recipe Retrieval				Recipe-to-Image Retrieval			
	Methods	**medR ↓**	**R@1 ↑**	**R@5 ↑**	**R@10 ↑**	**medR ↓**	**R@1 ↑**	**R@5 ↑**	**R@10 ↑**
1k	CCA [8]	15.7	14.0	32.0	43.0	24.8	9.0	24.0	35.0
	SAN [5]	16.1	12.5	31.1	42.3	-	-	-	-
	JE [1]	5.2	24.0	51.0	65.0	5.1	25.0	52.0	65.0
	AM [4]	4.6	25.6	53.7	66.9	4.6	25.7	53.9	67.1
	AdaMine [2]	2.0	39.8	69.0	77.4	1.0	40.2	68.1	78.7
	R^2GAN [32]	2.0	39.1	71.0	81.7	2.0	40.6	72.6	83.3
	MCEN [33]	2.0	48.2	75.8	83.6	2.0	48.4	76.1	83.7
	ACME [3]	1.0	51.8	80.2	87.5	1.0	52.8	80.2	87.6
	tri-pro [34]	1.0	52.7	81.0	88.1	1.0	53.8	81.1	88.3
	SCAN (Ours)	1.0	**54.0**	**81.9**	**89.2**	1.0	**54.9**	**81.9**	**89.0**
10k	JE [1]	41.9	-	-	-	39.2	-	-	-
	AM [4]	39.8	7.2	19.2	27.6	38.1	7.0	19.4	27.8
	AdaMine [2]	13.2	14.9	35.3	45.2	12.2	14.8	34.6	46.1
	R^2GAN [32]	13.9	13.5	33.5	44.9	11.6	14.2	35.0	46.8
	MCEN [33]	7.2	20.3	43.3	54.4	6.6	21.4	44.3	55.2
	tri-pro [34]	7.0	22.1	45.9	56.9	7.0	23.4	47.3	57.9
	ACME [3]	6.7	22.9	46.8	57.9	6.0	24.4	47.9	59.0
	SCAN (Ours)	5.9	**23.7**	**49.3**	**60.6**	5.1	**25.3**	**50.6**	**61.6**

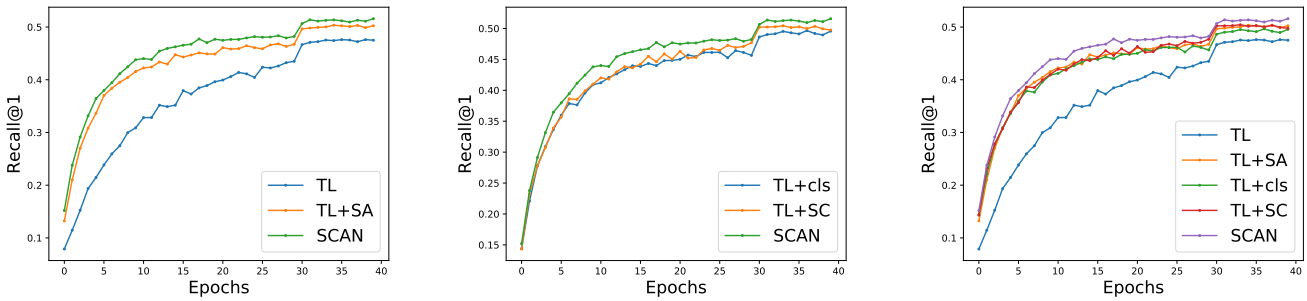


Fig. 4. The training records of our proposed model SCAN and each component of SCAN.

TABLE II

THE PERFORMANCE OF PROPOSED MODEL SCAN TRAINED WITH DIFFERENT TRADE-OFF PARAMETER λ.

λ	**MedR**	**R@1 (%)**	**R@5 (%)**	**R@10 (%)**
0.01	1.0	53.2	81.3	88.2
0.05	1.0	**54.0**	**81.9**	**89.2**
0.1	1.0	51.9	80.5	87.6
0.5	1.7	42.7	71.8	81.1

split [1], containing 238,999 image-recipe pairs for training, 51,119 and 51,303 pairs for validation and test, respectively. In total, the dataset has 1,047 categories.

B. Evaluation Protocol

We evaluate our proposed model with the same metrics used in prior works [1], [4], [2], [32], [3]. To be specific, median retrieval rank (MedR) and recall at top K (R@K) are used. MedR measures the median rank position among where true positives are returned. Therefore, higher performance comes with a lower MedR score. Given a food image, R@K calculates the fraction of times that the correct recipe is found within the top-K retrieved candidates, and vice versa. Different from MedR, the performance is directly proportional to the score of R@K. In the test phase, we first sample 10 different subsets of 1,000 pairs (1k setup), and 10 different subsets of 10,000 (10k setup) pairs. It is the same setting as in [1]. We then consider each item from food image modality in subset as a query, and rank samples from recipe modality according to L2 distance between the embedding of image and that of the recipe, which is served as image-to-recipe retrieval, and vice versa for recipe-to-image retrieval.

C. Implementation Details

We set the trade-off parameter λ in Eq. (1) based on empirical observations, where we tried a range of values and evaluated the performance on the validation set, as shown in Table II. We set the λ as 0.05. The model was trained using Adam optimizer [35] with the batch size of 64 in all our experiments. The initial learning rate is set as 0.0001, and the learning rate decreases 0.1 in the 30th epoch. We take a pretrained ResNet-50 and the bidirectional LSTMs as E_I and E_R respectively, whose output dimension is 1024. Note that we update the two sub-networks, i.e. image encoder E_I and

TABLE III
ABLATION STUDIES. EVALUATION OF BENEFITS OF DIFFERENT COMPONENTS OF THE SCAN FRAMEWORK. THE MODELS ARE EVALUATED BASED ON MEDR, WHERE LOWER IS BETTER, AND R@K (%), WHERE HIGHER IS BETTER.

L_{Ret}	Component	medR ↓	R@1 ↑	R@5 ↑	R@10 ↑
Cosine Loss	**CL**	2.0	46.9	76.5	84.9
	CL+SA	1.0	51.0	79.9	86.3
	CL+cls	1.9	47.0	75.5	83.4
	CL+SC	1.0	50.7	79.7	86.4
	CL+SC+SA	1.0	52.1	80.4	87.2
Triplet Loss	**TL**	2.0	47.5	76.2	85.1
	TL+SA	1.0	52.5	81.1	88.4
	TL+cls	1.7	48.5	78.0	85.5
	TL+SC	1.0	51.9	80.3	88.0
	SCAN	**1.0**	**54.0**	**81.7**	**88.8**

recipe encoder E_R, alternatively. It only takes 40 epochs to get the best performance with our proposed methods, while [1] requires 220 epochs to converge. Our training records can be viewed in Figure 4. We do our experiments on a single Tesla V100 GPU, which costs about 16 hours to finish the training.

D. Baselines

We compare the performance of our proposed methods with several state-of-the-art baselines, and the results are shown in Table I.

CCA [8]: Canonical Correlation Analysis (CCA) is one of the most widely-used classic models for learning a common embedding from different feature spaces. CCA learns two linear projections for mapping text and image features to a common space that maximizes their feature correlation.

SAN [5]: Stacked Attention Network (SAN) considers ingredients only (and ignores recipe instructions), and learns the feature space between ingredient and image features via a two-layer deep attention mechanism.

JE [1]: They use pairwise cosine embedding loss to find a joint embedding (JE) between the different modalities. To impose regularization, they add classifiers to the cross-modal embeddings which predict the category of a given food item.

AM [4]: Attention mechanism (AM) over the recipe is adopted in [4], applied at different parts of a recipe (title, ingredients and instructions). They use an extra transformation matrix and context vector in the attention model.

AdaMine [2]: A double triplet loss is used, where triplet loss is applied to both the joint embedding learning and the auxiliary classification task of categorizing the embedding into an appropriate category. They also integrate the adaptive learning schema (AdaMine) into the training phase, which performs adaptive mining for significant triplets.

$\mathbf{R^2GAN}$ [32]: After embedding the image and recipe information, $\mathrm{R}^2\mathrm{GAN}$ adopt GAN learning and semantic classification for cross-modal retrieval. They also introduce two-level ranking loss at embedding and image spaces.

MCEN [33]: Fu et al. adopt the generative idea, where they convert the embedding computation into a generative process. They first sample the latent variables from Gaussian distributions, based on which they use several layers to generate new feature embeddings. This method inevitably increase the computational cost.

ACME [3]: Adversarial training methods are utilized in ACME for modality alignment, to make the feature distributions from different modalities to be similar. To further preserve the semantic information in the cross-modal food data representation, Wang et al. introduce a translation consistency component.

tri-pro [34]: Zan et al. propose to use the improved triplet loss, where they enforce the embeddings of all images for a given recipe to be close to this recipe, while to be distant from other recipes. They also attempt to discover some hard negatives during training.

It has been validated that using attention can improve feature representations. Both SAN [5] and AM [4] adopt the attention mechanism to improve the recipe embeddings, while these methods add extra learnable layers to compute the attention weights and need to rely on the high-quality food images, which may affect the model performance. In contrast, without adding extra layers or requiring food images, our adopted self-attention method can find the attended ingredients and cooking instructions effectively. We show some attention results by our model in Figure 6. To improve the modality alignment, $\mathbf{R^2GAN}$ [32] and ACME [3] use adversarial learning. Specifically, Wang et al. [3] preserve the semantic consistency by transforming the feature representations to another modality. While our proposed method can achieve semantic alignment with the KL divergence, which is light-weight and effective. In summary, it can be observed our proposed model SCAN is useful on cross-modal food retrieval and outperforms all of earlier methods, as is shown in Table I.

E. Ablation Studies

Extensive ablation studies are conducted to evaluate the effectiveness of each component of our proposed model. Table III illustrates the contributions of self-attention model (**SA**), semantic consistency loss (**SC**) and their combination on improving the image to recipe retrieval performance. We test these different components based on different retrieval learning loss functions L_{Ret}, i.e. triplet loss (**TL**) and cosine embedding loss (**CL**).

TL serves as a baseline for **SA**, which adopts the *BatchHard* [31] training strategy. We then add **SA** and **SC** incrementally,

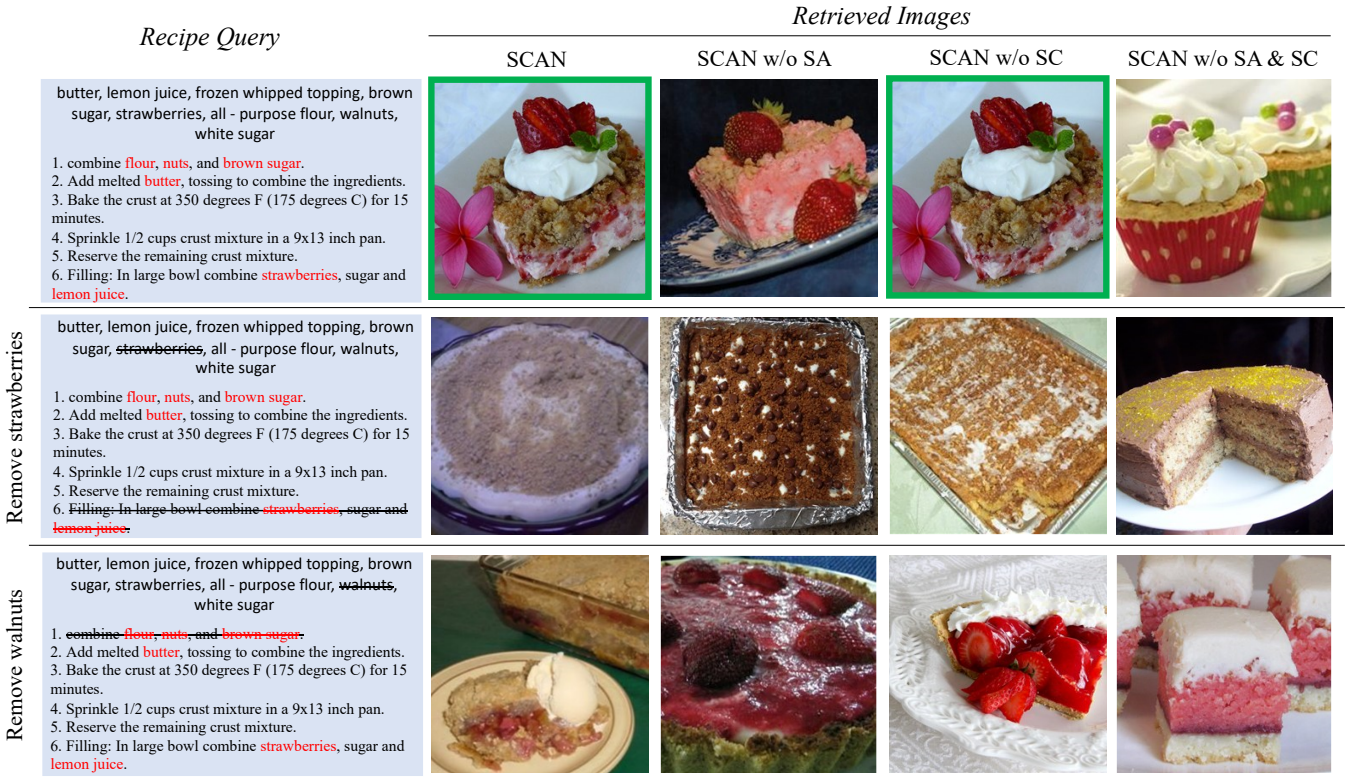


Fig. 5. Recipe-to-image retrieval results in Recipe1M dataset. We give an original recipe query from *dessert*, and then remove different ingredients of *strawberries* and *walnuts* separately to construct new recipe queries. We show the retrieved results by SCAN and different components of our proposed model.

and significant improvements can be found in both of the two components. To be specific, integrating **SA** into **TL** helps improve the performance of the image-to-recipe retrieval more than 4% in R@1, illustrating the effectiveness of the self-attention mechanism to learn discriminative recipe representations. The model trained with triplet loss and classification loss (**cls**) used in [1] is another baseline for **SC**. It shows that our proposed semantic consistency loss improves the performance in R@1 and R@10 by more than 2%, which suggests that reducing intra-class variance can be helpful in the cross-modal retrieval task. When we add **SA** and **SC** to **CL**, similar improvements can also been observed.

We show the training records in Figure 4, in the left figure, we can see that for the first 20 epochs, the performance gap between **TL** and **TL+SA** gets larger, while the performance of **TL+cls** and **TL+SC** keeps being similar, which is shown in the middle figure. But for the last 20 training epochs, the performance of **TL+SC** improves significantly, which indicates that for those hard samples whose intra-variance can hardly be reduced by **TL+cls**, **TL+SC** contributes further to the alignment of paired cross-modal data. The effect of trade-off parameter λ is shown in Table II. We illustrate the performance of models trained with four different λ, and we can see that setting λ as 0.05 can obtain the best performance.

In conclusion, we observe that each of the proposed components improves the cross-modal retrieval model, and the combination of those components yields better performance overall.

F. Recipe-to-Image Retrieval Results

We show three recipe-to-image retrieval results in Figure 5. In the top row, we select a recipe query *dessert* from Recipe1M dataset, which has the ground truth for retrieved food images. Images with the green box are the correctly retrieved ones, which come from the retrieved results by SCAN and TL+SA. But we can see that the model trained only with semantic consistency loss (TL+SC) has a reasonable retrieved result as well, which is relevant to the recipe query.

In the middle and bottom row, we remove some ingredients and the corresponding cooking instruction sentences in the recipe, and then construct the new recipe embeddings for the recipe-to-image retrieval. In the bottom row where we remove the *walnuts*, we can see that all of the retrieved images have no *walnuts*. However, only the image retrieved by our proposed SCAN reflects the richest recipe information. For instance, the image from SCAN remains visible ingredients of *frozen whipped topping*, while images from TL+SC and TL+SA have no *frozen whipped toppings*.

The recipe-to-image retrieval results indicate an interesting way to satisfy users' needs to find the corresponding food images for their customized recipes.

G. Image-to-Recipe Retrieval Results & Effect of Self-Attention model

In this section, we show some of the image-to-recipe retrieval results in Figure 6 and then focus on analyzing the effect of our self-attention model. Given images from *cheese*

Ingredients :
french - fried onions
cheddar cheese
water
whole kernel corn
mashed potatoes
milk
butter

Instructions: Preheat oven to 350 degrees Fahrenheit. Spray pan with non stick cooking spray. Heat milk, water and butter to boiling; stir in contents of both pouches of potatoes; let stand one minute. Stir in corn. Spoon half the potato mixture in pan. Sprinkle half each of cheese and onions; top with remaining potatoes. Sprinkle with remaining cheese and onions. Bake 10 to 15 minutes until cheese is melted. Enjoy!

Ingredients:
frozen chopped spinach
onion, eggs, salt
romano cheese
Greek seasoning
ground lamb
tzatziki

Instructions: If making your own sauce, prepare it the day before you are planning to make the meatloaf. In a large bowl, combine everything but the lamb/beef and tzatziki sauce. Shape into a loaf and place in a greased 11x7" baking dish. Bake uncovered at 350 for 55-60 minutes or until no pink remains and a thermostat reads 160. Let stand for 15 minutes before slicing. Drizzle with tzatziki sauce.

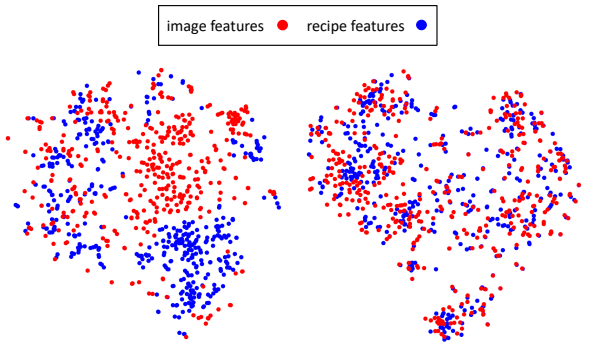
Ingredients:
salad greens
sticks
cherry tomatoes
pineapple
vinaigrette dressing

Instructions: Preheat greased grill to medium-high heat. Grill fruit 3 min, on each side or until lightly browned on both sides. Cut fruit into 2-inch sticks; place in large salad bowl. Add greens, jicama and tomatoes; toss lightly. Drizzle with dressing just before serving.

Fig. 6. Visualizations of image-to-recipe retrieval. We show the retrieved recipes of the given food images, along with the attended ingredients and cooking instruction sentences.

cake, *meat loaf* and *salad*, we show the retrieved recipe results by SCAN, which are all correct. We visualize the attended ingredients and instructions for the retrieved recipes with the yellow background, where we choose the ingredients and cooking instruction sentences of the top 2 attention weights as the attended ones. We can see that some frequently used ingredients like *water*, *milk*, *salt*, etc. are not attended with high weights, since they are not visible and shared by many kinds of food, which cannot provide enough discriminative information for cross-modal food retrieval. This is an intuitive explanation for the effectiveness of our self-attention model.

Another advantage of using the self-attention mechanism is that the image quality cannot affect the attended outputs. Obviously, the top two rows of food images *cheese cake* and *meat loaf* do not have good image quality, while our self-attention model still outputs reasonable attended results. This suggests that our proposed attention model has good capabilities to capture informative and reasonable parts for recipe embedding.



	TL	SCAN
Mean Intra-variance	1.02	**0.82**

Fig. 7. The difference on the intra-class feature distance of cross-modal paired data trained without and with semantic consistency loss. The food data is selected from the same category, *chocolate chip*. **SCAN** obtains closer image-recipe feature distance than **TL**. (Best viewed in color.)

H. Effect of Semantic Consistency

To have a concrete understanding of the ability of our proposed semantic consistency loss on reducing the mean intra-class feature distance (intra-class variance) between paired food image and recipe representations, we show the difference of the intra-class feature distance on cross-modal data trained without and with semantic consistency loss, i.e. TL and SCAN, in Figure 7. In the test set, we select the recipe and food image data from *chocolate chip*, which in total has 425 pairs. We obtain the food data representations from models trained with two different methods, then we compute the Euclidean distance between paired cross-modal data to obtain the mean intra-class feature distance. We adopt t-SNE [36] to do dimensionality reduction to visualize the food data.

It can be observed that cross-modal food data which is trained with semantic consistency loss (SCAN) has smaller intra-class variance than that trained without semantic consistency loss (TL). This means that semantic consistency loss is able to correlate paired cross-modal data representations effectively by reducing the intra-class feature distance, and also our experiment results suggest its efficacy.

V. Conclusion

In conclusion, we propose SCAN, an effective training framework, for cross-modal food retrieval. It introduces a novel semantic consistency loss and employs the self-attention mechanism to learn the joint embedding between food images and recipes for the first time. To be specific, we apply semantic consistency loss to cross-modal food data pairs to reduce the intra-class variance, and utilize the self-attention mechanism to find the important parts in the recipes to construct discriminative recipe representations. SCAN is easy to implement and can extend to other general cross-modal datasets. We

have conducted extensive experiments and ablation studies. We achieved state-of-the-art results in Recipe1M dataset.

ACKNOWLEDGEMENT

This research is supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

REFERENCES

[1] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3020–3028.

[2] 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 *ACM SIGIR*, 2018.

[3] H. Wang, D. Sahoo, C. Liu, E.-p. Lim, and S. C. Hoi, "Learning cross-modal embeddings with adversarial networks for cooking recipes and food images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 572–11 581.

[4] J.-J. Chen, C.-W. Ngo, F.-L. Feng, and T.-S. Chua, "Deep understanding of cooking procedure for cross-modal recipe retrieval," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 1020–1028.

[5] J. Chen, L. Pang, and C.-W. Ngo, "Cross-modal recipe retrieval: How to cook this dish?" in *International Conference on Multimedia Modeling*. Springer, 2017, pp. 588–600.

[6] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1335–1344.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[8] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[9] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2070–2083, 2018.

[10] Z. Li, J. Tang, L. Zhang, and J. Yang, "Weakly-supervised semantic guided hashing for social image retrieval," *International Journal of Computer Vision*, vol. 128, pp. 2265–2278, 2020.

[11] L. Jin, Z. Li, and J. Tang, "Deep semantic multimodal hashing network for scalable image-text and video-text retrievals," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[13] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," *CVPR*, 2018.

[14] Y. Peng, J. Qi, and Y. Yuan, "Cm-gans: Cross-modal generative adversarial networks for common representation learning," *arXiv preprint arXiv:1710.05106*, 2017.

[15] K. Ghasedi Dizaji, F. Zheng, N. Sadoughi, Y. Yang, C. Deng, and H. Huang, "Unsupervised deep generative adversarial hashing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3664–3673.

[16] M. Jing, J. Li, L. Zhu, K. Lu, Y. Yang, and Z. Huang, "Incomplete cross-modal retrieval with dual-aligned variational autoencoders," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3283–3291.

[17] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 921–10 930.

[18] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," *arXiv preprint arXiv:1808.07202*, 2018.

[19] W. Min, B.-K. Bao, S. Mei, Y. Zhu, Y. Rui, and S. Jiang, "You are what you eat: Exploring rich recipe information for cross-region food analysis," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 950–964, 2017.

[20] S. Horiguchi, S. Amano, M. Ogawa, and K. Aizawa, "Personalized classifier for food image recognition," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2836–2848, 2018.

[21] S. Jiang, W. Min, L. Liu, and Z. Luo, "Multi-scale multi-view deep feature aggregation for food recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 265–276, 2019.

[22] W. Min, L. Liu, Z. Wang, Z. Luo, X. Wei, X. Wei, and S. Jiang, "Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 393–401.

[23] W. Min, S. Jiang, J. Sang, H. Wang, X. Liu, and L. Herranz, "Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 1100–1113, 2016.

[24] M. Ge, M. Elahi, I. Fernaández-Tobías, F. Ricci, and D. Massimo, "Using tags and latent factors in a food recommender system," in *Proceedings of the 5th International Conference on Digital Health 2015*. ACM, 2015, pp. 105–112.

[25] W. Min, S. Jiang, and R. Jain, "Food recommendation: Framework, existing solutions, and challenges," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2659–2671, 2019.

[26] H. Wang, G. Lin, S. C. Hoi, and C. Miao, "Structure-aware generation network for recipe generation from images," in *European Conference on Computer Vision*. Springer, 2020, pp. 359–374.

[27] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 32–41.

[28] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[29] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in neural information processing systems*, 2015, pp. 3294–3302.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[32] B. Zhu, C.-W. Ngo, J. Chen, and Y. Hao, "R2gan: Cross-modal recipe retrieval with generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 477–11 486.

[33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 570–14 580.

[34] Z. Zan, L. Li, J. Liu, and D. Zhou, "Sentence-based and noise-robust cross-modal retrieval on cooking recipes and food images," in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020, pp. 117–125.

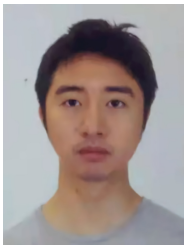
[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

Hao Wang is a PhD candidate with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include multi-modal analysis and computer vision.

Doyen Sahoo is a Senior Research Scientist at Salesforce Research Asia. Prior to joining Salesforce, Doyen was a Research Fellow at the Living Analytics Research Center at Singapore Management University (SMU). He was also serving as Adjunct Faculty in SMU. Doyen earned his PhD in Information Systems from SMU in 2018 and B.Eng in Computer Science from Nanyang Technological University in 2012. His research interests include Online Learning, Deep Learning, Computer Vision, and he also works on applied research including AIOps, Computational Finance and Cyber Security applications. He has published over 40 articles in top tier conferences and journals including ICLR, CVPR, ACL, KDD, JMLR, etc.

Chenghao Liu is currently a senior applied scientist of Salesforce Research Asia. Before, he was a research scientist in the School of Information Systems (SIS), Singapore Management University (SMU), Singapore. He received his Bachelor degree and Ph.D degrees from the Zhejiang University. His research interests include large-scale machine learning (online learning and deep learning) with application to tackle big data analytics challenges across a wide range of real-world applications.



Ke Shu is a research engineer at the Living Analytics Research Centre (LARC), Singapore Management University. His research interests include machine learning and deep learning.

Palakorn Achananuparp is a senior research scientist at the Living Analytics Research Centre (LARC), Singapore Management University. He is interested in developing and applying machine learning, natural language processing, and crowdsourcing techniques to solve problems in a variety of domains, including online social networks, politics, and public health.

Ee-peng Lim is the Lee Kong Chian Professor with the School of Computing and Information Systems at the Singapore Management University. He is also the Director of Living Analytics Research Centre in the School, a research centre focusing developing personalized and participatory analytics capabilities for smart city and smart nation relevant applications. Dr Lim received his PhD degree from University of Minnesota. His research expertise covers social media mining, social/urban data analytics, and information retrieval. He is the recipient of the Distinguished Contribution Award at the 2019 Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD), and the Test of Time award at 2020 ACM Conference on Web Search and Data Mining (WSDM).

Steven C. H. Hoi is currently the Managing Director of Salesforce Research Asia, and a Professor of Information Systems at Singapore Management University, Singapore. Prior to joining SMU, he was an Associate Professor with Nanyang Technological University, Singapore. He received his Bachelor degree from Tsinghua University, P.R. China, in 2002, and his Ph.D degree in computer science and engineering from The Chinese University of Hong Kong, in 2006. His research interests are machine learning and data mining and their applications to multimedia information retrieval, social media and web mining, and computational finance, etc. He has served as the Editor-in-Chief for Neurocomputing Journal, general co-chair for ACM SIGMM Workshops on Social Media, program co-chair for the fourth Asian Conference on Machine Learning, book editor for "Social Media Modeling and Computing", guest editor for ACM Transactions on Intelligent Systems and Technology. He is an IEEE Fellow and ACM Distinguished Member.