6-2022

Cross-lingual adaptation for recipe retrieval with mixup

Bin ZHU

Chong-Wah NGO
Singapore Management University, cwngo@smu.edu.sg

Jingjing CHEN

Wing-Kwong CHAN

Cross-lingual Adaptation for Recipe Retrieval with Mixup

Bin Zhu
University of Bristol
bin.zhu@bristol.ac.uk

Chong-Wah Ngo
Singapore Management University
cwngo@smu.edu.sg

Jingjing Chen
Fudan University
chenjingjing@fudan.edu.cn

Wing-Kwong Chan
City University of Hong Kong
wkchan@cityu.edu.hk

ABSTRACT

Cross-modal recipe retrieval has attracted research attention in recent years, thanks to the availability of large-scale paired data for training. Nevertheless, obtaining adequate recipe-image pairs covering the majority of cuisines for supervised learning is difficult if not impossible. By transferring knowledge learnt from a data-rich cuisine to a data-scarce cuisine, domain adaptation sheds light on this practical problem. Nevertheless, existing works assume recipes in source and target domains are mostly originated from the same cuisine and written in the same language. This paper studies unsupervised domain adaptation for image-to-recipe retrieval, where recipes in source and target domains are in different languages. Moreover, only recipes are available for training in the target domain. A novel recipe mixup method is proposed to learn transferable embedding features between the two domains. Specifically, recipe mixup produces mixed recipes to form an intermediate domain by discretely exchanging the section(s) between source and target recipes. To bridge the domain gap, recipe mixup loss is proposed to enforce the intermediate domain to locate in the shortest geodesic path between source and target domains in the recipe embedding space. By using Recipe 1M dataset as source domain (English) and Vireo-FoodTransfer dataset as target domain (Chinese), empirical experiments verify the effectiveness of recipe mixup for cross-lingual adaptation in the context of image-to-recipe retrieval.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**.

KEYWORDS

recipe retrieval, mixup, cross-lingual, domain adaptation

ACM Reference Format:
Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Wing-Kwong Chan. 2022. Cross-lingual Adaptation for Recipe Retrieval with Mixup. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3512527.3531375

ICMR '22, June 27–30, 2022, Newark, NJ, USA

ACM ISBN 978-1-4503-9238-9/22/06…$15.00
https://doi.org/10.1145/3512527.3531375

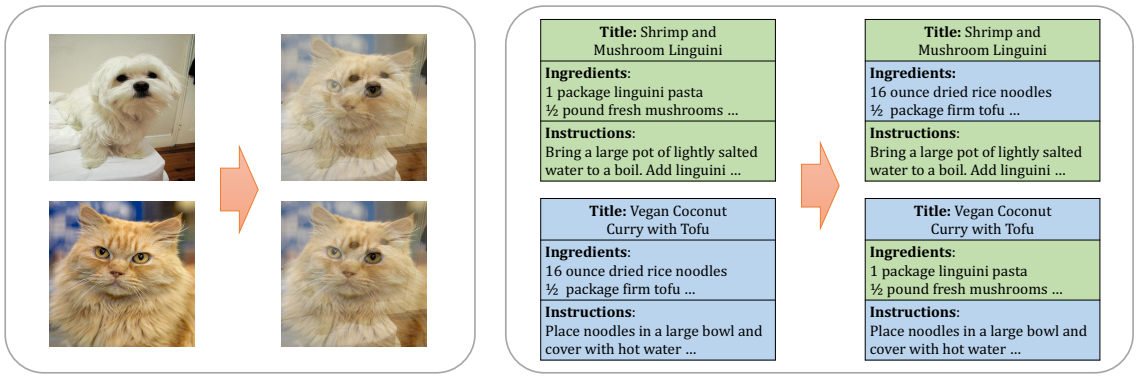
1 INTRODUCTION

The prevalence of cooking sharing websites accumulates plenty of recipe-image pairs for training cross-modal deep networks. Cross-modal recipe retrieval, i.e., searching the recipe by using the food image as query, becomes feasible with these networks [5, 8, 14, 36–38, 44, 52, 54]. With an image-to-recipe retrieval system, applications such as food recognition [3, 6, 7, 9], food recommendation [16, 29], food image generation [32, 51], nutrition estimation and food logging [1, 26, 30] will be benefited.

To generalize recipe retrieval models for different cuisines, cross-modal domain adaptation is explored in [53]. Leveraging on the abundant recipe-image pairs in a source domain, the aim is to learn an aligned embedding space for retrieving the recipes in a target domain with few or even no training pairs of data. As cross-domain model adaptation is capable of learning generalizable representation, the burden of data collection, especially for data-scare cuisine, can be alleviated. Existing works only study cross-modal retrieval in the context of monolingual recipes, e.g., English [8, 14, 37, 44, 54], Chinese [6, 53] and Japanese [23]. Nevertheless, recipes are naturally cross-lingual in the open world since the recipes of a cuisine are usually written in a mother tongue language. Transferring knowledge across languages is also becoming an issue for cross-modal recipe retrieval.

This paper studies the problem of unsupervised domain adaptation in the context of image-to-recipe retrieval. Specifically, given a source domain with abundant recipe-image pairs and a target domain with only recipes, the objective is to transfer knowledge from source to target domains for cross-modal recipe retrieval, where the recipes between domains are written in different languages. Due to missing of paired data in the target domain for performing supervised learning, this task is considered to be unsupervised domain adaptation [33]. In general, the domain gap stems from the differences in languages as well as the ingredient usage and cooking styles among cuisines. The nature of recipes makes this task more challenging than other cross-lingual problems, for instance, cross-lingual sentiment classification [11, 50] and cross-language image-sentence matching [24]. For example, in [24], the multi-lingual semantic alignment between captions written in different languages is assumed. However, the multi-lingual recipes in this paper are not linguistically or semantically aligned.

To address these problems, we propose a novel recipe mixup method. Inspired by image mixup [45, 46], which constructs new examples by linear interpolation with a ratio between two random



(a) Image Mixup (b) Recipe Mixup

Figure 1: Examples of mixup. (a) Image mixup [45, 46] constructs new examples by linearly interpolating two random image samples for data augmentation. (b) The proposed recipe mixup produces mixed recipes by discretely exchanging recipe section(s) (e.g., ingredient section) between source and target recipes for domain augmentation.

images for data augmentation (Figure 1 (a)), recipe mixup is proposed to produce mixed recipes by discretely exchanging section(s) between source and target recipes for domain augmentation (Figure 1 (b)). As a recipe usually consists of three sections including title, ingredients and instructions, recipe mixup can be performed effortlessly by different section combinations of two recipes. For example, as shown in Figure 1 (b), given two recipes from source and target domains, if the ingredient section is exchanged, two symmetric mixed recipes can be obtained. Specifically, the title and instructions are kept while ingredients are swapped between recipes. Intuitively, the mixed recipes share partial content with source and target domains, forming an intermediate domain [18] to bridge the domain gap. It has been shown that such an intermediate domain should locate in the shortest geodesic path between source and target domain on Grassmann manifold for positive knowledge transfer [12, 17, 18]. Based on these prior studies, recipe mixup loss is proposed to directly minimize the extra domain shift introduced by the intermediate domain compared with the shortest geodesic path. Together with cross-modal learning, ingredient recognition from image embedding, image generation from recipe embedding, as well as adversarial learning for feature alignment between domains, the proposed recipe mixup method outperforms baseline models with a large margin using Recipe 1M [37] as source domain (English) and Vireo-FoodTransfer [53] as target domain (Chinese).

2 RELATED WORK

2.1 Cross-modal Recipe Retrieval

The goal of cross-modal recipe retrieval is to retrieve the relevant recipes of a query dish image. Most existing works [5, 8, 14, 36–38, 44, 54] focus on learning similarity measurement between recipe and image in a common embedding space. The efforts are ranged from exploration of attention mechanism [8, 14], sample mining [5, 44, 54], generative adversarial nets [38, 44, 54] to transformers [36, 43, 47]. Nevertheless, the foundation of these works is the assumption of large-scale datasets with recipe-image pairs [37] for model training. The long-tail cuisine effect, specifically the model that is overly dominated by the training pairs from popular cuisines, is ignored. Consequently, these works are incapable to generalize well if a query image is drawn from a data-scare cuisine.

The limitation is partially addressed by cross-domain cross-modal food transfer (CCFT) [53], which is the first work to study cross-domain adaptation for recipe retrieval. Nevertheless, the focus of [53] is to address the problem of incomplete recipe transfer. This is different from our paper, where the recipes are complete with three sections but the languages are different. With this intuitively larger domain gap due to different modalities and languages, the contribution of this paper is to propose a parameter-free mixup technique for domain adaptation.

2.2 Domain Adaptation

Domain adaptation has been extensively studied in visual domain [34], which aims to transfer knowledge from a source domain to a target domain. To deal with the domain gap, one typical line of works is to explicitly measure domain discrepancy by metrics, such as maximum mean discrepancy (MMD) [27, 41]. Another line of works is to introduce a domain classifier to adversarially learn domain-invariant features [4, 15, 20, 40, 48]. In addition, the idea of mixup is explored in [28, 45] by conducting linearly interpolation of both image-level and feature-level mixup for domain adaptation. In contrast, the mixup in this paper is conducted on the highly structured but free-form written recipes in a discretely non-linear way for cross-lingual cross-modal adaptation.

Multi-modal [21, 35] and cross-lingual [10, 11, 24] transfers are also explored in the literature. Nevertheless, most works study either cross-modal adaptation without considering different languages [21, 35], or cross-lingual adaptation for single text modality [10, 11]. This paper shares similar spirit with [24] which considers both cross-lingual and cross-modal adaptation. The aim of [24] is to learn a multi-lingual embedding space for image-sentence matching. A shared language embedding is learnt by a language-specific fully-connected layer. Different from [24], this paper does not assume that the multi-lingual descriptions are aligned for model

learning. Our assumption is that there exists an intermediate domain that bridges the domain gap while the instance-level alignment between recipes of different domains is unknown.

3 METHOD

3.1 Problem Definition

Suppose we have a source domain $\mathcal{D}^s = \{(r_i^s, v_i^s)\}_{i=1}^{N^s}$ with N^s recipe-image pairs and a target domain $\mathcal{D}^t = \{(r_j^t)\}_{j=1}^{N^t}$ with only N^t recipes, where r_i^s and r_j^t refer to source and target recipes respectively, v_i^s is a source image corresponding to r_i^s. Leveraged on paired source recipe-image pairs and target recipes for model learning, the goal is to retrieve the corresponding recipes using unseen target images as queries. Note that we assume the source and target recipes are written in different languages, for example, r^s can be written in English while r^t is in Chinese.

3.2 Model Architecture

Figure 2 depicts the overview architecture of the proposed model. Following other cross-lingual works [31, 50], the target recipes are first translated to the source language by machine translation [39]. Both source and translated recipes are subsequently fed into the pre-trained multilingual Bert model [1] [13] to extract features for each recipe section respectively, i.e., title, ingredients and instructions. To be specific, the title, each ingredient with quantity and unit (e.g., a teaspoon of sugar, 50 grams pork), and each instruction written as a sentence, are fed into Bert model respectively and transformed to fixed-length vectors. The output source and target recipe Bert features are denoted as $F_R^s = \{E_{ti}^s, E_{ing}^s, E_{ins}^s\}$ and $F_R^t = \{E_{ti}^t, E_{ing}^t, E_{ins}^t\}$ respectively, where E_{ti}, E_{ing} and E_{ins} represent title, ingredient and instruction features.

The key component of our model is recipe mixup block (Section 3.3), which aims to produce mixed recipe features to form an intermediate domain by discretely exchanging section(s) between source and target recipes. Together with F_R^s and F_R^t, the mixed recipe features are passed to the recipe encoder to obtain the final recipe embeddings, including source recipe embedding E_R^s, target recipe embedding E_R^t, source mixed recipe embedding $E_{R_M}^s$ and target mixed recipe embedding $E_{R_M}^t$. Meanwhile, the source image embedding E_I^s is extracted from the image encoder and mapped into the common embedding space with recipes. As recipe-image pairs are available in the source domain, similar to other works in cross-modal recipe retrieval [8, 44, 53, 54], triplet loss is adopted for common space learning between recipe and image. The triplet loss is defined as follows:

$$\mathcal{L}_{tri} = [d(E_a^s, E_p^s) - d(E_a^s, E_n^s) + \alpha]_+, \tag{1}$$

where $d(\cdot,\cdot)$ is a distance function measured by cosine similarity, E_a^s, E_p^s and E_n^s are the anchor, positive and negative embeddings in the source domain respectively. The parameter α is the margin.

To align the recipe embeddings between source and target domains, a domain discriminator D is employed to distinguish whether the input recipe embedding comes from source or target domain [15, 53]. The adversarial domain loss $\mathcal{L}_{adv}$ is defined as follows:

$$\begin{aligned}\mathcal{L}_{adv} =& \mathbb{E}_{E_R^s \sim p_R^s}[\log D(E_R^s)] \\ &+ \mathbb{E}_{E_R^t \sim p_R^t}[\log(1 - D(E_R^t)].\end{aligned} \tag{2}$$

Similar to [53], semantic loss $\mathcal{L}_{sem}$ and image generation loss $\mathcal{L}_{gen}$ are also employed to better capture semantic information. Specifically, the source image embedding E_I^s is fed into an ingredient decoder to predict the ingredients of the food image. The ingredient decoder is a multi-label classifier and trained using cross-entropy loss, i.e., $\mathcal{L}_{sem}$. The source recipe embedding E_R^s is conditioned to reconstruct the food image. $\mathcal{L}_{gen}$ is computed by the generation error between the reconstructed and real images.

The overall objective function of the proposed method is defined as follows:

$$\mathcal{L} = \mathcal{L}_{tri} + \lambda_1 \mathcal{L}_{rm} + \lambda_2 \mathcal{L}_{adv} + \lambda_3(\mathcal{L}_{sem} + \mathcal{L}_{gen}), \tag{3}$$

where $\mathcal{L}_{rm}$ denotes recipe mixup loss (Section 3.3). The hyper-parameters λ_1, λ_2 and λ_3 balance the importance of the losses.

3.3 Recipe Mixup (RM)

The key idea of the proposed recipe mixup block is to produce mixed recipes by exchanging recipe section(s) between source and target domains. As the translation and multilingual Bert models are frozen during training, it is equivalent to exchange sections of recipe Bert features between F_R^s and F_R^t in practice. Specifically, as shown in Figure 3, given a batch of source and target recipe features, six kinds of recipe mixup strategies are derived by either exchanging one or two sections between F_R^s and F_R^t. If one section is exchanged, three groups of single recipe section mixup features can be obtained as shown in the left side of Figure 3, i.e., RM_1, RM_2 and RM_3 corresponding to the exchange of title, ingredients and instructions, respectively. On the other hand, if two sections are exchanged, three groups of dual recipe sections mixup features are derived as shown in the right side of Figure 3, i.e., RM_4, RM_5 and RM_6. As the source and target recipe mixups are symmetric, each kind of recipe mixup produces a pair of mixed recipe features, denoted as $\{(F_{R_{M_i}}^s, F_{R_{M_i}}^t)\}_{i=1}^6$, where i refers to the number of recipe mixup. To maintain the consistence of the mixed recipes and reduce the extra domain shift from intermediate domain [12], identical recipe mixup strategy (i.e., one of the six recipe mixup strategies) is employed to form the intermediate domain at one time.

The original and mixed Bert recipe features are subsequently transformed to recipe embeddings by the recipe encoder, i.e., E_R^s, E_R^t, $E_{R_M}^s$ and $E_{R_M}^t$. Inspired by [12, 17, 18], intermediate domain should lie in the shortest geodesic path between source and target domains on Grassmann manifold. In other words, the sum of the distances between intermediate domain with source and target domains should be equal to the distance between source and target domains, i.e., $d(P^s, P^{inter}) + d(P^t, P^{inter}) = d(P^s, P^t)$, where $d(\cdot,\cdot)$ is a distance measurement between two distributions. P^s, P^t and P^{inter} refer to data distribution of source, target and intermediate domains respectively. Violation of the constraint, i.e., $d(P^s, P^{inter}) + d(P^t, P^{inter}) > d(P^s, P^t)$, would introduce "extra domain shift", which is more likely to exert negative transfer. Our solution is to minimize the extra domain shift in the embedding

[1] https://github.com/google-research/bert/blob/master/multilingual.md

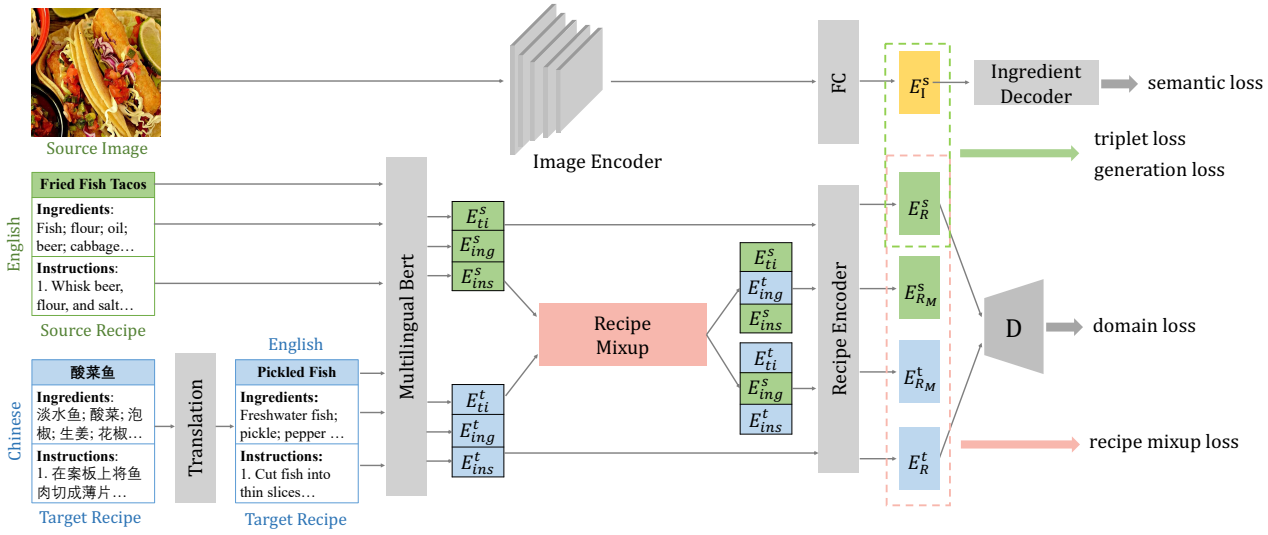


Figure 2: Overview of the proposed model architecture. The recipe mixup block aims to produce mixed recipes to construct intermediate domain by exchanging section(s) between source and target recipes. Ingredient section exchange is used as an example for illustration.

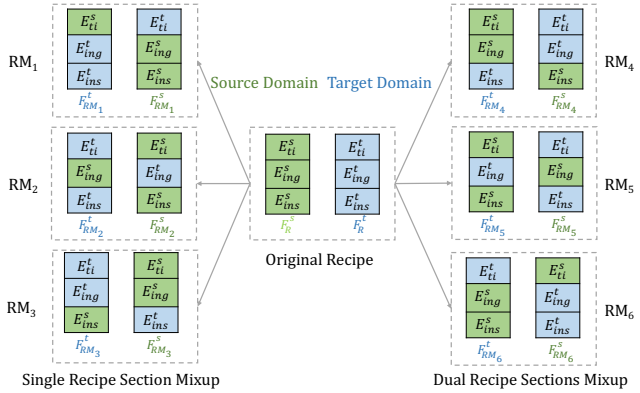


Figure 3: Six kinds of recipe mixup (RM) strategies. RM_1, RM_2 and RM_3 in the left side correspond to one recipe section exchange while RM_4, RM_5 and RM_6 in the right side perform two recipe sections exchange.

space. Following [12], L2 norm is employed to measure the distribution distance between two domains. As we have two symmetric source and target mixed recipes in each kind of recipe mixup, intermediate domain can be formed with only source mixed recipes ($F^s_{R_{M_i}}$), only target mixed recipes ($F^t_{R_{M_i}}$) or both source and target mixed recipes ($F^s_{R_{M_i}}, F^t_{R_{M_i}}$). The recipe mixup losses corresponding to these three cases are defined as follows:

$$\mathcal{L}^s_{rm} = \underbrace{\left\|E^s_R - E^s_{R_M}\right\|_2 + \left\|E^t_R - E^s_{R_M}\right\|_2 - \left\|E^s_R - E^t_R\right\|_2}_{\text{extra domain shift from source mixed recipes}}, \quad (4)$$

$$\mathcal{L}^t_{rm} = \underbrace{\left\|E^s_R - E^t_{R_M}\right\|_2 + \left\|E^t_R - E^t_{R_M}\right\|_2 - \left\|E^s_R - E^t_R\right\|_2}_{\text{extra domain shift from target mixed recipes}}, \quad (5)$$

$$\mathcal{L}^{st}_{rm} = \frac{1}{2}(\mathcal{L}^s_{rm} + \mathcal{L}^t_{rm}). \quad (6)$$

Note that if source and target mixed recipes are considered independently, the source single recipe section mixup is equivalent to the target dual recipe sections mixup in practice, and vice versa. For instance, as shown in Figure 3, $F^s_{RM_1}$ in RM_1 is the same to the $F^t_{RM_6}$ in RM_6. Hence, $\mathcal{L}^s_{rm}$ and $\mathcal{L}^t_{rm}$ are somehow interchangeable. Only the results of $\mathcal{L}^s_{rm}$ are reported to avoid redundancy.

4 EXPERIMENTS

4.1 Experiment Settings

Datasets. The experiments are conducted on Recipe 1M [37] and Vireo-FoodTransfer [53] datasets, which involve two most widely used languages: English and Chinese respectively. Recipe 1M contains 341,421 English recipe-image pairs in total, with 4,102 unique ingredients. Vireo-FoodTransfer consists of 70,985 Chinese recipe-images pairs with 1,635 ingredients. In the experiments, Recipe 1M and Vireo-FoodTransfer are regarded as the source and target domains respectively.

Note that the ingredient lists of Vireo-FoodTransfer and Recipe 1M are not fully overlapped. Nevertheless, high-level semantic information (i.e., ingredient labels) is known to be helpful for cross-modal embedding learning as also evidenced in [37, 44, 54]. To make use of ingredient semantic information and reduce the redundancy of the ingredients, K-Means is employed to obtain a set of 1,000 unified ingredients for the two datasets. Specifically, the Chinese ingredients in Vireo-FoodTransfer are first translated to English. Then

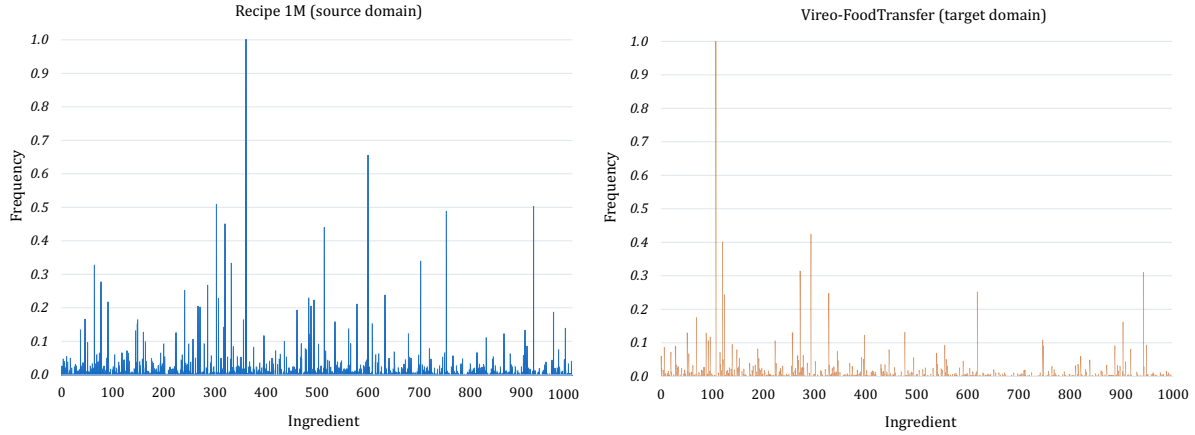


Figure 4: Ingredient usage comparison between Recipe 1M (source domain) and Vireo-FoodTransfer (target domain) datasets. Note that the frequency is normalized within dataset for better comparison.

clustering is conducted on the Bert ingredient features extracted from the ingredients of Recipe 1M and Vireo-FoodTransfer. For example, "tomato", "tomato puree" and "tomato paste" are clustered into the same class. Among the 1,000 ingredients after clustering, there are 543 common ingredients (e.g., chicken, cucumber and oil) between Recipe 1M and Vireo-FoodTransfer datasets. The numbers of unique ingredients for Recipe 1M (e.g., mayonnaise, philadelphia cheese and nutella) and Vireo-FoodTransfer (e.g., Chinese angelica and lycium barbarum) are 384 and 73 respectively.

Figure 4 further demonstrates the ingredient usage comparison between Recipe 1M and Vireo-FoodTransfer. Except for the unique ingredients, the frequency of common ingredients in the two datasets vary greatly. For example, the "soy sauce" is much more heavily used in the Vireo-FoodTransfer while "toast" is much more common in Recipe 1M. The difference in ingredient usage is one of the main origins of domain gap.

Evaluation Metrics. Similar to [53], median rank (MedR) and recall rate at top K (R@K) are adopted as the metrics to evaluate the retrieval adaptation performance. During testing, a subset of 1,000 unseen target images are formed as queries by randomly sampling from the target domain. MedR is the median rank of the ground truth recipes for all the queries, and R@K is the percentage value averaged over all the queries at the search depth of K. The reported MedR and R@K values are the average performance on 10 different sets of target images randomly drawn.

Implementation Details. The backbone of the image encoder is based on ResNet-50 [19] pre-trained on ImageNet by replacing the last fully-connected layer with 1024-dimensional output. Similar with [37, 53], the recipe encoder is composed of a Bidirectional LSTM and a hierarchical LSTM for ingredients and instructions encoding respectively. The two features are concatenated with title features and fed into a fully-connected layer to obtain the recipe embedding. The dimension of the Bert features is 768. Adam optimizer [25] is adopted for model training with a batch size of 32 in all the experiments. The initial learning rate is set to be 0.0001. The trade-off hyper-parameters λ_1, λ_2 and λ_3 in Equation 3 are set to be 0.1, 0.01 and 0.002 respectively. We set the margin α=0.3 in Equation 1.

4.2 Performance Comparison

The performances of the proposed method against the baseline source-only model, CCFT [53] and the oracle model are listed in Table 1. Note that the proposed model is essentially built upon CCFT with a recipe mixup block. The source-only and oracle models are trained by only using the paired data in source and target domains for supervised learning respectively. By directly applying the source model for the target domain with Chinese recipes represented by Bert features, the result is fairly poor, as shown in the first row of Table 1. Instead, we use the language translator [39] to translate Chinese recipes to English and then feed into Bert model. The result is boosted sharply as shown in the second row (MedR = 182.0) of the table. In the experiments, we only report results based on the translated recipes. More analysis about the visualization of language gap in recipes is presented in the section 4.3.

Among the six kinds of RM strategies and two types of RM loss functions, the best performance is achieved by RM^s_4, which exchanges title and ingredient sections of source and target recipes (Equation 4). Compared with CCFT [53], the performance of RM^s_4 surpasses CCFT with a large margin in terms of MedR and R@K. The MedR is significantly boosted by 13.4 ranks from 128.0 to 114.6, and R@50 is improved by 12.6% from 30.33 to 34.16.

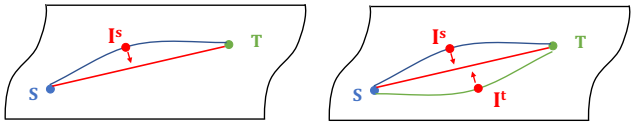
Why does RM^s work for domain adaptation? As shown in pink zone of Table 1, all the six RM strategies for recipe mixup loss $\mathcal{L}^s_{rm}$ (Equation 4) constantly outperform CCFT in terms of both MedR and R@K. When comparing among single recipe section mixup (i.e., RM^s_1, RM^s_2 and RM^s_3), the ingredient section (i.e., RM^s_2) brings the most significant improvement, and title section (i.e., RM^s_1) manages to achieve slightly better results than instruction section (i.e., RM^s_3). When comparing among dual recipe sections mixup (i.e., RM^s_4, RM^s_5 and RM^s_6), the performance of RM^s_4, which exchanges title and ingredient sections, is superior than RM^s_5 and RM^s_6. Furthermore, RM^s_4 outperforms title-only (RM^s_1) and ingredient-only

Table 1: Image-to-recipe retrieval performance comparison in terms of MedR and R@K. "E", "C" and "C(T)" represents English, Chinese and Chinese-Translated English respectively. RM^s and RMst refer to model training by using recipe mixup losses $\mathcal{L}^s_{rm}$ (in pink) and $\mathcal{L}^{st}_{rm}$ (in yellow) respectively.

Method	Domains	MedR	R@10	R@50
Source-only	E→C	395.0	2.44	9.32
	E→C(T)	182.0	10.08	26.36
CCFT [53]		128.0	11.78	30.33
RM$^s_1$		121.2	12.30	32.27
RM$^s_2$		115.9	14.01	33.61
RM$^s_3$		125.7	12.08	32.54
RM$^s_4$		**114.6**	**14.42**	**34.16**
RM$^s_5$		119.1	14.11	33.13
RM$^s_6$		120.4	12.53	31.97
RM$^{st}_1$		142.7	10.32	28.35
RM$^{st}_2$		141.9	10.62	28.49
RM$^{st}_3$		135.3	11.28	29.30
RM$^{st}_4$		139.9	10.79	29.28
RM$^{st}_5$		142.3	10.74	28.24
RM$^{st}_6$		152.6	9.76	27.33
Oracle	C(T)→C(T)	2.1	74.70	88.92

(RM$^s_2$) exchange, which shows that the title and ingredient sections are complementary with each other. Similarly, the combination of title and instructions (RM$^s_5$) exchange also demonstrates superior results compared with RM$^s_1$ and RM$^s_3$. Nevertheless, the property is not applied to the case of ingredient and instruction sections, i.e., RM$^s_6$, which is inferior than RM$^s_2$ but still better than RM$^s_3$. The reasons are two-fold. On the one hand, ingredients demonstrate the composition of a dish which is more informative and discriminative than title and instructions [36]. On the other hand, instructions tend to suffer from translation quality of machine translation, due to the naming convention and the free-form writing styles across languages. In contrast, the ingredient section contains the quantity, units and the ingredients, which shows a more universal format across languages, leading to better translation quality.

Compared with CCFT, the RM^s benefits from the constraint imposed by RM loss, which is aware of the shortest geodesic path between source and target domains in constructing the intermediate space. Specifically, the recipe embedding space is regularized to accommodate the mixed recipe features which are artificially generated. The mixed recipe embeddings are also optimized to maintain a structure obeying the constraint of the shortest geodesic path. In other words, during the learning process to construct a desired intermediate domain in the recipe embedding space, the recipe encoder is steered to bridging the domain gap and learn transferable features between source and target domains. This can be validated from the reduced gap between the two domains. Take RM$^s_1$ as an example, compared with CCFT, the distribution distance of source and target recipes (measured by L2 norm with 10K samples) is reduced by 0.19%. Furthermore, as shown in Table 1, the MedR of RM$^s_1$ is also boosted by 6.8 ranks. This result is also consistent with the distribution distance between target recipe-image pairs (measured by L2 norm with 10K samples), which is reduced by 3.08%.



(a) RM^s (b) RMst

Figure 5: The difference between RM^s and RMst. "S", "T" and "I" refer to source, target and intermediate domains respectively. The red line represents the shortest geodesic path in the embedding space. The I^s (in the blue line) and I^t (in the green line) are intermediate domains formed by source and target mixed recipes respectively.

Why does RMst does not work? As RM can produce source and target mixed recipes, RMst investigates the performance by forming the intermediate domain with the two types of mixed recipes, i.e., the model is trained with recipe mixup loss $\mathcal{L}^{st}_{rm}$ in the Equation 6. Table 1 lists the performance of RMst in yellow zone. Different from RM^s, which outperforms CCFT in all the six RM strategies, RMst constantly degrades the performance of CCFT in terms of MedR and R@K. Figure 5 illustrates the difference between RM^s and RMst. In fact, the source and mixed recipes intuitively lie in different paths between source and target domains, considering all the sections are opposite with each other. Therefore, extra domain shift is introduced by not only source mixed recipes but also target mixed recipes in RMst. The recipe features are obtained by jointly transforming title, ingredients and instructions, while the source and target mixed recipes are symmetric in structure but distinctive in contents. As a consequence, the objectives of RMst forcing source and target mixed recipes to stay on the shortest geodesic path in the embedding space simultaneously are not necessarily aligned. In other words, with two different objectives to fulfill the shortest geodesic path constraint, RMst results in negative transfer.

Qualitative results. As shown in Figure 6, two typical examples are presented showing the top 3 retrieved recipes of RM$^s_4$ and CCFT [53]. Note that the recipes shown in the figure is originally Chinese and translated into English for presentation purpose. Only title and major ingredients are presented for a recipe to save space. In both examples, RM$^s_4$ manages to rank the ground truth (GT) recipes in the first place while the ranks of CCFT are worse at the positions of 14 and 8 respectively. Furthermore, the recipes in the top 3 of RM$^s_4$ also demonstrate more overlapped ingredients with the GT recipes. For instance, in the first example (rows 3-5), the second ("Spicy Chicken") and third ("Mushroom Chicken Soup") retrieved recipes of RM$^s_4$ contain the major ingredients "chicken thigh" or "chicken" and "green onion". In contrast, CCFT only contains one of "green onion" or "chicken breast". The result shows that RM$^s_4$ manages to capture more discriminate features and thus achieves better retrieval performance.

Nevertheless, the overall retrieval performance of RM$^s_4$ is much worse than the traditional retrieval models [36, 44] and domain adaptation models [53] with mono-lingual recipes. Indeed, cross-lingual adaptation for recipe retrieval is an extremely challenging problem. Figure 7 shows two standard failure cases of the top 5

Query Image	RM_4^s				CCFT			
	Top 3 Retrieved Recipes and Paired Images			GT Rank	Top 3 Retrieved Recipes and Paired Images			GT Rank
	**Chenpi Chicken**	Spicy Chicken	Mushroom Chicken Soup	1	Steamed Huangli Fish	Low-Fat Potato Base Pizza	Hoof Jelly	14
	chicken thigh; green onion; chenpi…	chicken thigh; green onion; pepper…	chicken; green onion; mushroom…		fish; ginger; garlic; green onion…	chicken breast; egg; potato; carrots…	pork trotter; green onion; ginger…	
								
	Peach Crisp	Crispy Egg Roll	Cheese Ball	1	Hoof Soup	Honeycomb Cake	Poached Fish with Tomatoes	42
	flour; baking soda; egg; green onion; sugar; pepper...	flour; egg; sugar; butter; milk; black sesame …	cream cheese; egg; sugar; corn starch; cream…		pig trotter; sugar; green onion; bean…	flour; egg; baking soda; milk; honey; sugar…	fish; coriander; ginger; green onion; celery…	
								

Figure 6: Examples showing the top 3 retrieved results of RM_4^s and CCFT. The ground truth (GT) recipe is highlighted with red bounding box. The common ingredients appeared in the GT recipes are marked in red and underlined.

Ground Truth (GT)		Top 5 Retrieved Recipe Titles and Paired Images					GT Rank
Recipe Title	Steamed Chicken with Mushrooms	Rose Soy Sauce Chicken	Curry Chicken Rice	Pork Belly Braised Egg	Rice with Sausage and Mushroom	Pepper Bean Curd Lettuce	65
Query Image							
Recipe Title	Qingming Mugwort Cake	Red Bean Paste Pie	Goldfish Dumplings	Sweet Glutinous Rice Cake	Big Belly Dumpling	Nutrition Fat Reduction Bento	276
Query Image							

Figure 7: Failure cases showing the top 5 retrieved results of RM_4^s.

retrieved recipes by RM_4^s, where the ranks of GT recipes are in relatively large positions. In the first example (rows 2-3), although the GT recipe is ranked in 65, 3 out of the top 5 recipes contain common main ingredients with the GT recipe, for example, "chicken" and "mushroom". In the second example (rows 4-5), the query image ("Qingming Mugwort Cake") is one kind of special food in China which is particularly made in a traditional Chinese Qingming festival, as a result, it is not surprising that the rank for the GT recipe is as bad as 276. The reason is that little common knowledge can be transferred from Recipe 1M where majority of the recipes are quite different western food. Interestingly, although the visual appearance of the top 5 images and the query image is quite different, four of the recipes are made from "flour", which is the same with the GT recipe.

4.3 Impact of Machine Translation

Figure 8 visualizes the Bert features of three recipe sections and ingredient vocabulary in the source (Recipe 1M) and target (Vireo-FoodTransfer) by t-SNE [42] before and after machine translation (MT). As ingredient and instruction sections contains multiple sentences, the visualization results are based on the average of all the Bert features for each section. Different from the recipe sections, ingredient vocabulary refers to the post-processed ingredient labels (each ingredient label includes a few words, e.g, olive oil, white sugar and mushroom) in each dataset, specifically, the 4,102 ingredients in Recipe 1M and 1,635 ingredients in Vireo-FoodTransfer. Observed from the results before machine translation (left side of each sub-figure in Figure 8), all the multilingual Bert features from source and target domains are clearly separated into two distinctive sets, including the ingredient vocabulary. It shows that the domain gap between source and target domains is quite evident. Furthermore, the "inherent" translation by multi-lingual Bert model fails to capture the same semantics across different languages. However, after explicitly translating the Chinese recipes to English (right side of each sub-figure in Figure 8), it can be observed that the distance between source and target domains is shorten, i.e, the domain gap is reduced. In particular, the reduced gap of ingredient vocabulary are much more evident than all the three recipe sections. The result is not surprising because the ingredient vocabulary is

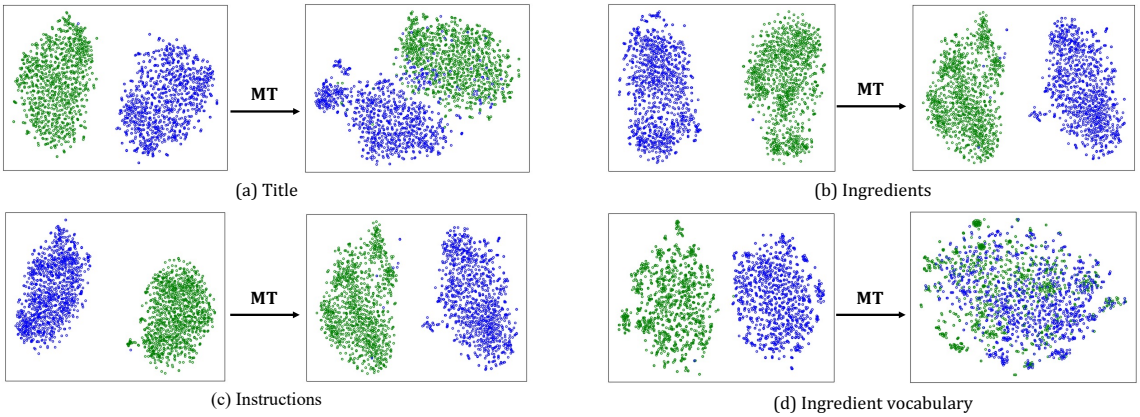


Figure 8: Visualization of recipe Bert features for title, ingredients and instructions and ingredient vocabulary by t-SNE on source (Green dots) and target (Blue dots) domains before and after machine translation (MT).

Table 2: Ablation study based on RM_4^s.

Methods	MedR	R@10	R@50
RM_4^s w/o adv	124.7	12.18	31.70
RM_4^s w/o sem	119.5	13.78	32.65
RM_4^s w/o gen	117.8	13.95	32.92
RM_4^s w/o rm (CCFT [53])	128.0	11.78	30.33
RM_4^s (full)	114.6	14.42	34.16

pre-processed into ingredient labels beforehand, therefore, some of the common ingredients of the Chinese-translated English in Vireo-FoodTransfer can be exactly the same with the original English ingredients in Recipe1M.

4.4 Ablation Study

The significance of each RM_4^s component is assessed. Table 2 lists the performances of the ablation models with one of the loss functions being excluded from training in turn. Adversarial learning (adv) shows a high impact on the performance, where MedR and R@K drop dramatically without employing domain discriminator for feature alignment. In contrast, the semantic regularization (sem) for ingredient recognition and image generation (gen) from recipe embedding have a lower impact on the performance. Without sem, both MedR and R@K degrade a bit. Similarly, the performance also drops slightly without recipe-to-image generation. The result is aligned with CCFT [53], where ingredient recognition from image embedding is more important than image generation for the retrieval performance. Finally and the worst, when RM is taken away, the largest margin of drop is noticed. Since RM_4^s is built upon CCFT with recipe mixup, the performance without RM is the same with CCFT.

5 DISCUSSION

We recapitulate and discuss the limitations of this paper in threefold. First, there is a performance gap between the proposed model and the oracle model which is trained with paired data in the target domain. The performance gap is also larger than the result reported in CCFT [53] on the three Asian cuisines without language gap. Second, the proposed recipe mixup has significantly boosted the performance of traditional model (without domain adaptation) and the recent CCFT (with domain adaptation for the recipes in the same language). Despite these achievements, the physical meaning of using recipe mixup for domain augmentation remains not being fully understood. Similar studies also point out the issue and some progress has been made to explain mixup [46, 49]. Third, we notice that there is an apparent performance gap between using multi-lingual Bert for "inherent" translation and using off-the-shelf language translator for explicit translation. Exploring internal properties of languages (e.g., lexical similarity or structural similarities) [22] and fine-grained alignment (e.g., subword) [2] between different languages could be one possible solution to ease the dependency of machine translation.

6 CONCLUSION

We have presented a novel recipe mixup method for cross-lingual adaptation in the context of image-to-recipe retrieval, which outperforms the baseline models with a large margin. Through the empirical experiments, we have shown that all six kinds of recipe mixup strategies with source mixed recipes constantly achieve better performance than the baseline models. Exchanging both title and ingredient sections between source and target recipes attains the best performance. By using both the source and target mixed recipes as intermediate domain, negative transfer appears. The future work includes exploration of recipe modeling using transformers and machine translation free cross-lingual adaptation.

ACKNOWLEDGMENTS

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11203517), a CityU MF_EXT grant (project no. 9678180), and a Shanghai Pujiang Program (20PJ1401900).

REFERENCES

[1] Oscar Beijbom, Neel Joshi, Dan Morris, Scott Saponas, and Siddharth Khullar. 2015. Menu-match: Restaurant-specific food logging from images. In *2015 IEEE Winter Conference on Applications of Computer Vision*. 844–851.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Proceedings of European Conference on Computer Vision*. 446–461.

[4] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. 2018. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision*. 135–150.

[5] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. 2018. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 35–44.

[6] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM International Conference on Multimedia*. 32–41.

[7] Jingjing Chen, Chong-Wah Ngo, and Tat-Seng Chua. 2017. Cross-modal recipe retrieval with rich food attributes. In *Proceedings of the 25th ACM International Conference on Multimedia*. 1771–1779.

[8] Jingjing Chen, Chong-Wah Ngo, Fu-Li Feng, and Tat-Seng Chua. 2018. Deep understanding of cooking procedure for cross-modal recipe retrieval. In *Proceedings of the 26th ACM International Conference on Multimedia*. 1020–1028.

[9] Jingjing Chen, Bin Zhu, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. 2020. A study of multi-task and region-wise deep learning for food ingredient recognition. *IEEE Transactions on Image Processing* 30 (2020), 1514–1526.

[10] Xilun Chen, Ahmed Hassan, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-Source Cross-Lingual Model Transfer: Learning What to Share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3098–3112.

[11] Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics* 6 (2018), 557–570.

[12] Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. 2021. IDM: An Intermediate Domain Module for Domain Adaptive Person Re-ID. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11864–11874.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.

[14] Han Fu, Rui Wu, Chenghao Liu, and Jianling Sun. 2020. 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*. 14570–14580.

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[16] Xiaoyan Gao, Fuli Feng, Xiangnan He, Heyan Huang, Xinyu Guan, Chong Feng, Zhaoyan Ming, and Tat-Seng Chua. 2019. Hierarchical attention network for visually-aware food recommendation. *IEEE Transactions on Multimedia* 22, 6 (2019), 1647–1659.

[17] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. 2019. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2477–2486.

[18] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. 2013. Unsupervised adaptation across domain shifts by generating intermediate data representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 11 (2013), 2288–2302.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 770–778.

[20] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*. 1989–1998.

[21] Xin Huang, Yuxin Peng, and Mingkuan Yuan. 2018. Mhtn: Modal-adversarial hybrid transfer network for cross-modal retrieval. *IEEE Transactions on Cybernetics* 50, 3 (2018), 1047–1059.

[22] K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *International Conference on Learning Representations*.

[23] Yohei Kikuta, Yuichiro Someya, and Leszek Rybicki. 2017. Approaches to Food/Non-food Image Classification Using Deep Learning in Cookpad. In *Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities in conjunction with The 2017 International Joint Conference on Artificial Intelligence*. 35–38.

[24] Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan Plummer. 2020. Mule: Multimodal universal language embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 11254–11261.

[25] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations*.

[26] Keigo Kitamura, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2009. Foodlog: Capture, analysis and retrieval of personal food images via web. In *Proceedings of the ACM Multimedia 2009 workshop on Multimedia for Cooking and Eating Activities*. 23–30.

[27] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*. 97–105.

[28] Xudong Mao, Yun Ma, Zhenguo Yang, Yangbin Chen, and Qing Li. 2019. Virtual mixup training for unsupervised domain adaptation. *arXiv preprint arXiv:1905.04215* (2019).

[29] Weiqing Min, Shuqiang Jiang, and Ramesh Jain. 2019. Food Recommendation: Framework, Existing Solutions, and Challenges. *IEEE Transactions on Multimedia* 22, 10 (2019), 2659–2671.

[30] Zhao-Yan Ming, Jingjing Chen, Yu Cao, Ciarán Forde, Chong-Wah Ngo, and Tat Seng Chua. 2018. Food photo recognition for dietary tracking: System and experiment. In *International Conference on Multimedia Modeling*. 129–141.

[31] Jian-Yun Nie. 2010. Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies* 3, 1 (2010), 1–125.

[32] Siyuan Pan, Ling Dai, Xuhong Hou, Huating Li, and Bin Sheng. 2020. ChefGAN: Food Image Generation from Recipes. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4244–4252.

[33] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2009), 1345–1359.

[34] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. 2015. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine* 32, 3 (2015), 53–69.

[35] Fan Qi, Xiaoshan Yang, and Changsheng Xu. 2018. A unified framework for multimodal domain adaptation. In *Proceedings of the 26th ACM International Conference on Multimedia*. 429–437.

[36] Amaia Salvador, Erhan Gundogdu, Loris Bazzani, and Michael Donoser. 2021. Revamping Cross-Modal Recipe Retrieval with Hierarchical Transformers and Self-supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[37] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3020–3028.

[38] Yu Sugiyama and Keiji Yanai. 2021. Cross-Modal Recipe Embeddings by Disentangling Recipe Contents and Dish Styles. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2501–2509.

[39] Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation (Volume 1: Research Papers)*. Association for Computational Linguistics.

[40] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7167–7176.

[41] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474* (2014).

[42] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008).

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010.

[44] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven CH Hoi. 2019. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11572–11581.

[45] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. 2020. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6502–6509.

[46] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

[47] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. 2021. Token shift transformer for video classification. In *Proceedings of the 29th ACM International Conference on*

Multimedia. 917–925.

[48] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. 2018. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8156–8164.

[49] Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. 2021. How Does Mixup Help With Robustness and Generalization?. In *International Conference on Learning Representations*.

[50] Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1403–1412.

[51] Bin Zhu and Chong-Wah Ngo. 2020. CookGAN: Causality based Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5519–5527.

[52] Bin Zhu, Chong-Wah Ngo, and Wing-Kwong Chan. 2021. Learning from Web Recipe-image Pairs for Food Recognition: Problem, Baselines and Performance. *IEEE Transactions on Multimedia* (2021).

[53] Bin Zhu, Chong-Wah Ngo, and Jingjing Chen. 2020. Cross-domain Cross-modal Food Transfer. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3762–3770.

[54] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. 2019. R2GAN: Cross-modal recipe retrieval with generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11477–11486.