Cross-domain Cross-modal Food Transfer

Bin Zhu
City University of Hong Kong
binzhu4-c@my.cityu.edu.hk

Chong-Wah Ngo
City University of Hong Kong
cscwngo@cityu.edu.hk

Jing-jing Chen
Fudan University
chenjingjing@fudan.edu.cn

ABSTRACT

The recent works in cross-modal image-to-recipe retrieval pave a new way to scale up food recognition. By learning the joint space between food images and recipes, food recognition is boiled down as a retrieval problem by evaluating the similarity of embedded features. The major drawback, nevertheless, is the difficulty in applying an already-trained model to recognize different cuisines of dishes unknown to the model. In general, model updating with new training examples, in the form of image-recipe pairs, is required to adapt a model to new cooking styles in a cuisine. Nevertheless, in practice, acquiring sufficient number of image-recipe pairs for model transfer can be time-consuming. This paper addresses the challenge of resource scarcity in the scenario that only partial data instead of a complete view of data is accessible for model transfer. Partial data refers to missing information such as absence of image modality or cooking instructions from an image-recipe pair. To cope with partial data, a novel generic model, equipped with various loss functions including cross-modal metric learning, recipe residual loss, semantic regularization and adversarial learning, is proposed for cross-domain transfer learning. Experiments are conducted on three different cuisines (Chuan, Yue and Washoku) to provide insights on scaling up food recognition across domains with limited training resources.

CCS CONCEPTS

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

KEYWORDS

food recognition; cross-modal food retrieval; cross-domain transfer

ACM Reference Format:
Bin Zhu, Chong-Wah Ngo, and Jing-jing Chen. 2020. Cross-domain Cross-modal Food Transfer. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413809

1 INTRODUCTION

The demand for huge number of training examples is known to be a problem for supervised learning of neural networks. In the domain of food recognition, the problem is prevalent as image annotation requires not only category-level labels [3, 6, 22] but also their attributes such as ingredients [2, 6, 23], cooking and cutting methods [8]. In practice, the number of categories can easily go beyond a thousand for a city-scale food dataset [24]. Acquiring sufficient number of positive examples for all food categories poses a huge challenge, not mentioning the required efforts for data cleaning and labeling. Despite these daunting demands, transferring of an already-trained model to a new domain (e.g., a different city or cuisine) still requires intensive effort in crawling and annotating new food images.

Food recognition is recently posted as a problem of cross-modal retrieval [7, 29]. Specifically, instead of predicting food categories and attributes [6, 8], the recipes of query images are retrieved. As a recipe typically contains three sections, i.e., title, ingredients and cooking steps, food recognition is achieved by extracting recipe titles and relevant attributes. Instead of training a neural network for food classification, this new paradigm learns to embed image and recipe features in a common space for similarity comparison. The annotation effort is keep to minimum as the required training examples are simply image-recipe pairs that can be acquired from cooking sharing websites [7, 29]. Due to its potential in scaling up food recognition, this cross-modal retrieval paradigm has spurted various research interests, including ingredient recognition [2, 6], cooking causality analysis [35], recipe and food image synthesis [28, 35].

Based upon these prior works [4, 7, 9, 29, 33, 36], this paper extends from cross-modal to cross-domain food retrieval. Leveraging on image-recipe pairs in a source domain, we consider the problem of food transfer as recognizing food in a target domain with new food categories and attributes. The training resources in a target domain is assumed incomplete. Specifically, only partial data of dishes, for instance, titles and ingredients instead of their recipes, are available. Food images may be available but are possibly not annotated or linked to recipes. The consideration of resource incompleteness can be mapped to the following real scenarios when transferring food recognition to an unknown domain. First, considering that only a sparse list of dish names written in text are available in a target domain. The list may be crawled from the menus of local restaurants in the new domain and is incomplete. In this case, food transfer is to leverage the sparse list of dishes to train a model for recognizing dishes not only new to the source domain but also unseen in the provided list of dish titles. Second, it is not strange to see that the restaurant menus provide both dish titles and their ingredients for reference. In this case, the model can be trained with additional food attribute information. Third, when images, whether being annotated with food names, are also available, the model can be further trained with multiple modalities. Finally, if pairing information between image and text are also available for some of the data, the model can be trained in cross-modal manner [7, 29] to embed text and image features in a common space.

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413809

Table 1: Partial data in a target domain, summarized from the scarcest resource (Case 1) to near sufficiency (Case 10). Note that “pair" refers to the cases that the image-text pairing relations are available for some resources.

Methods	Cases	Recipe			Image	Pair
		Title	Ingredients	Instructions		
Unsupervised	Case 1	✓				
	Case 2	✓	✓			
	Case 3	✓	✓	✓		
	Case 4				✓	
	Case 5	✓			✓	
	Case 6	✓	✓		✓	
	Case 7	✓	✓	✓	✓	
Semi-supervised	Case 8	✓			✓	✓
	Case 9	✓	✓		✓	✓
	Case 10	✓	✓	✓	✓	✓

In this paper, we refer the situation of incomplete data as resource scarcity in a target domain. The aim of food transfer is to utilize the incomplete data, together with complete data in a source domain, to train a cross-domain cross-modal neural network for food retrieval. We generalize the issue of resource scarcity into ten different cases listed in Table 1. Our main contribution is proposal of a food transfer model that is generic to deal with various situations when data is incomplete. While cross-domain transfer is not a new problem [5, 17, 27], framing the problem from the perspective of resource scarcity has not been addressed. Particularly, food transfer is considered in the setting of cross-modal learning, where model adaptation across domains is required to align within and across modalities under the situation of resource scarcity. Note that our work is different from the recent studies in cross-modal transfer [5, 17] which define different modalities as domains. Instead, domains refer to different cuisines of dishes. A domain is composed of multimedia resources describing dish preparation. To expedite food transfer, only limited resources are acquired in the target domain for training. To the best of our knowledge, this is the first work that addresses the problem of domain transfer for the topic of food recognition.

2 RELATED WORK

Domain adaptation is not a new problem and has been intensively studied in [12, 20, 21, 25, 26, 31, 32]. The goal is to transfer the knowledge learnt in a source domain to annotate data in a target domain. The challenges of transfer attribute to various reasons, such as different sources where data are acquired [12, 21], limited amount of training data in the target domain [16, 37] and different modalities between source and target domains [5, 17]. A general solution to this problem is by aligning the feature distribution of two domains. The representative approaches include minimizing data gap by maximum mean discrepancy (MMD) [21, 32] and learning domain invariant features by adversarial learning [12, 31]. In the literature, most works are dedicated to single modality transfer, for examples, image-to-image [12, 21] and text-to-text [1, 37] between two domains.

Recently, multi-modal [27] and cross-modal [17] transfers are proposed. In [27], the data in both domains are composed of visual-audio pairs. Transfer is carried out by learning intra-modality and inter-modality domain invariant features using generative adversarial network (GAN). Inter-modality transfer refers to the generation of attention-layer and fusion-layer features that are indistinguishable by the domain discriminator in GAN. The learning is unsupervised as class labels are assumed not available for target data. In [17], the source and target domains are allowed to have data in different modalities. The goal is to learn modality invariant shared representation, such that knowledge learnt in a modality is transferable to a new modality unseen by the source domain. To achieve this, the pairing information between different modalities and their category labels are required to be known in the target domain. In this case, not only domain discriminator as in [27] but also category-level regularization are employed for learning shared representation.

While similar in spirit as [17, 27], our work is more generic in dealing with various situations that may arise in a target domain. First, different from [17, 27], we deal with variants of representation within a single modality. Specifically, within the text modality, there are recipe title (i.e., a phrase or short sentence in few words), ingredient list (i.e., a sparse set of categories describing dish content) and cooking steps (i.e., a stepwise procedural description of cooking methods). Under the scenario of resource scarcity, the data within a modality can be incomplete. For example, only recipe title is available but not ingredients and cooking steps. In other words, the proposed food transfer needs to handle incomplete data within a single modality. Second, different from [17, 27], pairing information across different modalities may not be available. In other words, while the model in source domain is trained with paired modalities, in contrast to [27], the target domain may contain only one of the modalities. Furthermore, different from [17], neither pairing nor category-level information may be available for training. Among the ten cases consider in this paper (see Table 1), only cases 7 and 10 are the situations where [27] and [17] deal with respectively. Finally, as our model is specifically design for food recognition, a rich set of domain discriminators and regularizers are considered as loss functions. These include multi-label classification of ingredients as semantic regularizer as in [33], image and text domain discriminators as in [12], and reconstruction of images from recipes for shared representation learning [33, 36].

While food recognition has recently captured numerous research attentions [3, 6, 8, 11, 18, 24], there are yet to have any studies in

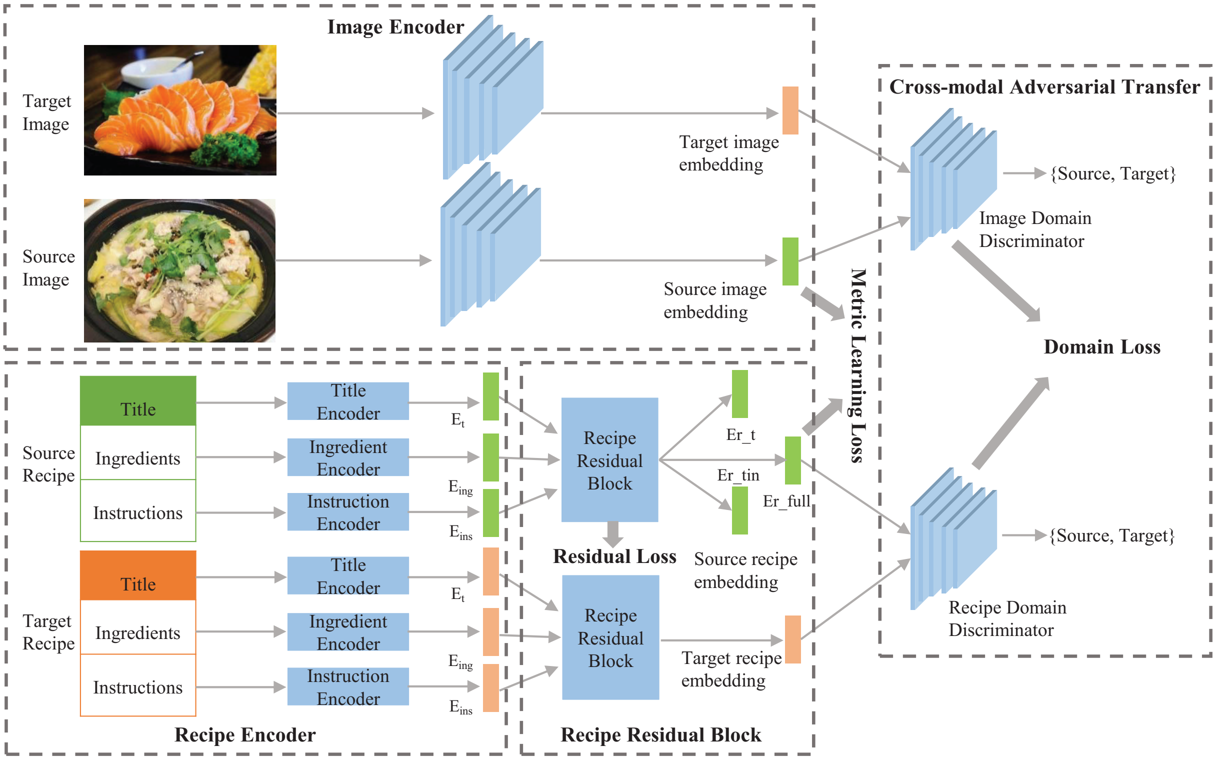


Figure 1: The architecture of cross-domain food transfer.

cross-domain transfer learning for food recognition. The works in [18] and [34] pose few-shot food recognition as a problem of meta-learning using relation and prototypical networks respectively. These works are formulated under C-way K-shot classification, where C is generally a small value, for example, referring to 20 food categories in [18] and 5 categories in [34]. Extending C to over hundreds of dish categories as in this paper is not a trivial problem. Finally, these works [18, 34] assume that a complete view of data is accessible for learning, which is fundamentally different from our work to address learning from partial data view. In [11], knowledge graphs, which model the relations between different cooking attributes, are leveraged to annotate ingredients unseen during training. The relations include ingredient hierarchy, co-occurrence and the association with cooking and cutting methods. These relations, nevertheless, are expected to vary across different cuisines due to variations in taste and food favour which will result in different use of ingredients and cooking techniques.

3 CROSS-DOMAIN FOOD TRANSFER

The aim is to perform food recognition on a target domain with incomplete training data. It is assumed that there exists a source domain with complete training data. Hence, the problem is defined as annotating images in a target domain with partial resources by leveraging the complete resources available in a source domain. The resource types include image V, recipe R, food title T, title and ingredients T_{in}, and recipe-image pair (V, R). The complete training data refers to the set of recipe-image pairs for fully supervised model training [10, 33, 36]. In the remaining sections, we abbreviate the source and target domains with the superscripts s and t respectively. Let the resources in a source domain as $D^s = \{(r_i^s, v_i^s)\}_{i=1}^{N^s}$, with N^s recipe-image pairs.

There are ten possible cases to describe resource scarcity in a target domain, as listed in Table 1. For example, Case-3 contains recipes and hence the full set of textual information (i.e., title, ingredients and cooking instructions) is available. In contrast, Case-1 only has the dish titles while Case-2 has both titles and ingredients but no cooking steps. Images are available for cases 4 to 7, but the pairing between images and other resources are unknown. Due to missing of pairing information, only unsupervised training can be conducted on the target domain for cases 1 to 7. In practice, pairing relations among the available sources may exist. For example, a restaurant menu comes with dish titles along with their images or even ingredients. These situations are included in cases 8 and 9. Finally, Case-10 represents the situation that recipe-image pairs can be downloaded, for example, from the cooking sharing websites. As pairing information usually only exists for some popular dishes, not all resources in a target domain can be linked. In the experiment, we assume that there are at most 5,000 pairs available for training. Therefore, model training under the cases with pairing relations being partially observed is characterized by semi-supervised learning.

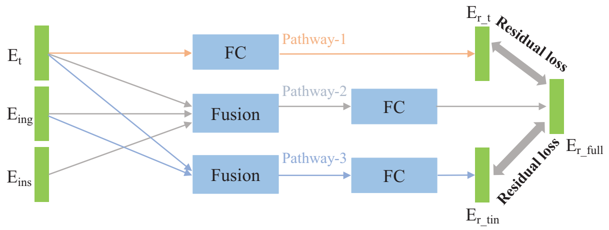


Figure 2: The architecture of Recipe Residual Block.

3.1 Model Overview

Figure 1 depicts the design of network architecture. As other networks [17, 27], the design includes encoders for supervised feature learning and discriminators for adversarial learning. The novelties are the proposals of recipe encoder and recipe residue block (RRB). The former selectively extracts various cooking data depending on resource availability. The latter deals with partial data learning. RRB is generic to enable either unsupervised or semi-supervised learning for all the ten cases in Table 1.

The training starts by encoding the source recipes and images into latent features by their respective encoders. As these information are paired, metric learning, such as cosine similarity [29] and triplet ranking loss [4, 10, 33, 36], can be employed to ensure cross-modal feature closeness in the latent space. Different from other architecture designs [4, 10, 29, 33, 36], the proposed recipe encoder learns three separate features corresponding to different sections (title, ingredient, cooking steps) of a recipe. In case where only food titles are available in the target domain (i.e., Case-1 in Table 1), the recipe features are generated based on titles only. RRB is the module that, leveraging on complete recipe information in the source domain, learns to enrich the recipe features transformed from partial data. The various embeddings generated by RRB are then forwarded to discriminators for cross-modal transfer learning.

The objective of unsupervised learning is:

$$L = L^s_{ml} + \lambda L_{residual} + \gamma L_d + \sigma L_{reg}, \tag{1}$$

where L^s_{ml}, $L_{residual}$, L_d and L_{reg} represent metric learning loss for source data, residual loss (Section 3.2), domain loss (Section 3.3) and regularization loss respectively. The losses are linearly combined with trade-off hyper-parameters λ, γ and σ. Note that the L^s_{ml} and L_{reg} terms are similar as other models [4, 10, 33, 36]. The former, L^s_{ml}, is learnt with triplet ranking loss [30]. The latter, L_{reg}, is a regularizer composed of two parts: multi-labeling of ingredients for an image embedding [33] and reconstruction of food image from a recipe embedding [33, 36].

When pairing information are available for some of the target data (Cases 8-10), metric learning can also be conducted for the target pair data. In this case, semi-supervised learning is carried out with the objective function:

$$L_{semi} = L + \mu L^t_{ml}, \tag{2}$$

where μ balances the relative importance of the two parts.

3.2 Recipe Residual Block (RRB)

The spirit of RRB is to interpolate missing information when recipe feature is generated from partial data such as title-only (e.g., Case-1) or title and ingredients only (e.g., Case-2). The motivation comes from the fact that there exists correlation between ingredients and cooking methods. Specifically, given a set of ingredients as a prior knowledge, the likelihood of cooking steps can be predicted. In other words, the ways of cooking are not in random but restricted by the available ingredients. In general, dish title is expected to be informative to include main ingredients and cooking methods. In some cases, the title even gives clue to the visual appearance of a dish. Hence, even in the case when only title is known but not ingredients, it is possible to imagine how a dish will possibly be prepared and its visual appearance. RRB leverages the full recipe information available in source domain to ensure consistency among features extracted from different parts of recipes. For example, the feature extracted from a title should resemble to the feature encoded by the sequence of cooking steps. Furthermore, with the paired information between recipes and images, the learned features are also enforced to be consistent with image features through metric learning. In brief, RRB aims to synchronize various information extracted from recipes as well as their paired images, and then transfer to target domain with partial data.

Denote E_t, E_{ing}, E_{ins} as features extracted by the recipe encoder, corresponding to title, ingredients and cooking steps respectively. RRB routes these features through three pathways as depicted in Figure 2. Along a pathway includes an optional fusion layer for feature integration and a fully connected (FC) layer for feature embedding. The first pathway transforms E_t to E_{r_t}, while the third pathway integrates both E_t and E_{ing} before transforming to E_{r_tin}. Similarly for the second pathway that fuses all the three features to produce E_{r_full}. The embedded features, having the same dimensionality, are trained to resemble of each other in RBB. Specifically, supervised by E_{r_full} that encapsulates the entire information of a recipe, E_{r_t}, E_{r_tin} learn to interpolate missing data through residual loss, as following:

$$L_{residual} = \left\|E_{r\_full} - E_{r\_t}\right\|_2^2 + \left\|E_{r\_full} - E_{r\_tin}\right\|_2^2. \tag{3}$$

Note that the three pathways are activated simultaneously during training in the source domain. When only partial data is available in target domain, the pathways are selectively in function. In the extreme case when only title is available, RBB simply performs feature transformation based on knowledge learnt in source domain.

3.3 Cross-modal Adversarial Transfer

Inspired by [12, 13], adversarial learning is employed to reduce the gap between source and target domains. A discriminator is trained to predict whether an embedding feature originates from source or target domain. The discriminator and RRB play a min-max game to align the feature distributions in both domains. Additional discriminator to distinguish image embeddings is also trained, when images are available in the target domain (Cases 4-10). The domain

Table 2: Dataset overview. The last three rows show the average number of ingredients, cooking steps and images per recipe.

Cuisines	Chuan	Yue	Washoku
Num. of recipes	42,797	27,256	9,626
Num. of images	155,750	119,758	48,485
Avg. num. of ingredients	6.9	6.1	6.2
Avg. num. of instructions	7.2	7.8	8.5
Avg. num. of images	3.6	4.4	5.0

loss is formalized as:

$$\begin{aligned} L_d = & \underbrace{\mathbb{E}_{r^s \sim p_r^s}[\log D_1(r^s)] + \mathbb{E}_{r^t \sim p_r^t}[\log (1 - D_1(r^t)]}_{\text{recipe discriminator}} + \\ & \underbrace{\mathbb{E}_{v^s \sim p_v^s}[\log D_2(v^s)] + \mathbb{E}_{v^t \sim p_v^t}[\log (1 - D_2(v^t))],}_{\text{image discriminator}} \end{aligned} \tag{4}$$

where D_1 and D_2 are recipe and image discriminators respectively to predict the domain of an embedding. Note that only D_1 is considered in Cases 1-3 due to absence of image, and similarly only D_2 in Case 4 due to absence of recipe.

4 EXPERIMENT SETTING

Dataset. Three datasets, corresponding to Sichuan ("Chuan"), Cantonese ("Yue") and Washoku ("Japanese") cuisines, are constructed. The datasets are composed of image-recipe pairs crawled from one of the most popular Chinese recipe sharing websites "xiachufang" [1]. A list of ingredients common to all the three cuisines are compiled from the recipes. A total of 1,635 ingredients are kept after excluding those found in less than 20 recipes. Note that some ingredients are common in one cuisine but rare in other cuisines. For example, "Sichuan pepper" is frequent in Chuan (found in 5,642 recipes) but rare in Washoku (only in 154 recipes). Among the 1,635 ingredients, "egg" are popular in all the three cuisines.

The dataset statistics is listed in Table 2. Chuan has a relatively larger number of ingredients per recipe on average, while Washoku has the most number of cooking steps per recipe. We split each cuisine into three subsets, 70% for training, 10% for validation and 20% for testing. Please refer to the supplementary document for details about the dataset, including sample images and recipes, word cloud of each cuisine and general discussion about the difference of these three cuisines.

Performance Evaluation. The experiment is conducted in the setting of cross-modal food retrieval, similar as [10, 33, 36]. Specifically, given a query image, the task is to retrieve either the dish title or recipe of the image from a dataset. The availability of resources in a dataset follows the ten cases listed in Table 1. In cases 3, 7 and 10, recipes will be retrieved and ranked based on their similarity to a query image. For other cases, dish titles are ranked. The performance will be mainly evaluated by median rank (MedR), which is the median rank of ground-truth titles or recipes for all the testing queries. Recall at top-K (R@K) is also reported when necessary. During testing, depending on the number of data in a target domain, we randomly sample either 1,000 or 5,000 recipes from testing data for experiment. The corresponding image of each recipe is then issued as testing query. For fair comparison, we repeat the experiment for 10 times and report the average MedR. The setting closely follows [10, 33, 36], where each test involves at least 1,000 different dishes for recognition. The proposed approach can also be applied for searching images when a query is title or recipe, as in [10, 33, 36]. Due to space limitation, we only present the result of image search in the supplementary document.

Implementation details. The dimensions of recipe and image embeddings are set to 1024. The backbone of image encoder is ResNet-50 [14] pre-trained on ImageNet, by replacing the last fully-connected layer with 1024 neurons. Bidirectional LSTM [15] is employed by title and ingredient encoders for feature transformation. The text features are undergone word2vec embedding and then projected as vectors of 300 dimensions. As cooking steps are much lengthy, hierarchical LSTM [29] is adopted by instructor encoder. Words are embedded sequentially and then followed by sentences into a vector of 1024 dimensions. Both recipe and image discriminators are implemented as a differentiable three-layer perceptron for domain classification. The model is trained end-to-end from scratch using Adam optimizer [19] with batch size of 32 in all experiments. We set the balance hyperparameters to be $\lambda = 1$, $\gamma = 0.01$ and $\sigma = 0.002$ in Equation 1. For semi-supervised learning, we set $\mu = 0.1$ in Equation 2. Additional 5,000 pairing information, i.e., title-image pairs for cases 8-9 and recipe-image pairs for Case-10, are randomly sampled from a target domain for model training.

5 RESULT ANALYSIS

5.1 Impact of Resource Scarcity

We first study the impact of resources on model training based on the 10 cases defined in Table 1. The performances are then benchmarked against two oracle runs that assume all resources are available for training. Both runs, Oracle-1 and Oracle-2, are fully supervised learnt based on the architecture proposed in Figure 2. Oracle-1 leverages all the resources (i.e., recipe-image pairs) in source and target domains for training. Oracle-2 does not have knowledge of source domain and the model training involves only resources in the target domain.

As Yue and Washoku are relatively smaller than Chuan in dataset size, we set them as target domains in this experiment. Figure 3 shows the result of three different transfers on testing data size of 1,000. The general trend is that recipe is an essential resource that guarantees better MedR. When the domain gap is relatively small (e.g., Chuan→Yue), recipe-only attains MedR=4.2 which is fairly close to MedR=2.7 of Oracle-1. In other words, having a subset of recipes in a target domain can sufficiently train a model with decent recognition performance. When comparing image-only (Case-4) to recipe-only (Case-3), the improvement introduced by image modality is less. Particularly in Chuan→Washoku, where the domain gap is expected to be the largest among the three transfers, the MedR of recipe-only is 63.2% better than image-only. The result basically verifies that recipes, which are textually resourceful, are more effective than images in domain transfer. When pairing information

[1] https://www.xiachufang.com/

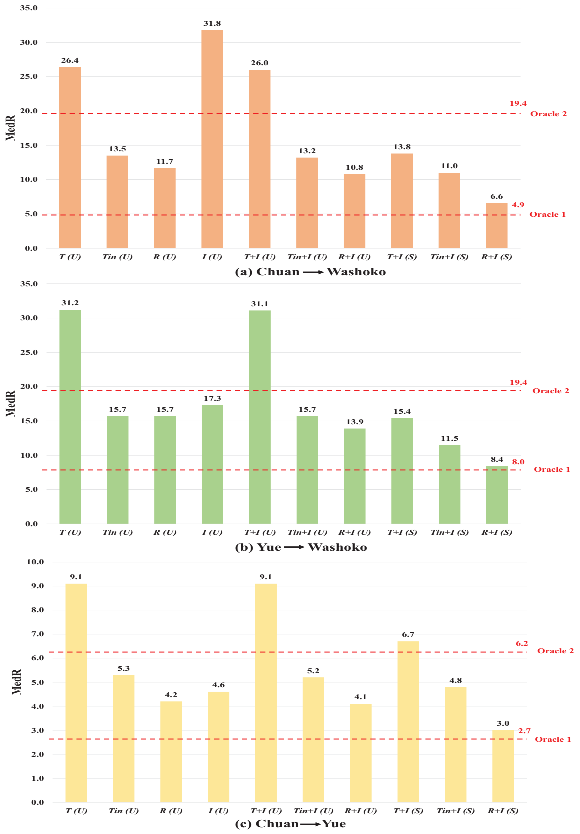


Figure 3: MedR performance (y-axis) varies across ten different cases of resource scarcity (x-axis) given in Table 1. T: title-only, Tin: title and ingredient only, R: recipe, I: image-only. The first seven bars refer to cases 1-7 under unsupervised learning (U). The last three bars refer to cases 8-10 under semi-supervised learning (S). The two oracles (dotted line) show the results of fully supervised models.

between recipes and images are available for semi-supervised learning, MedR values approach the performance of Oracle-1 for all the three transfers. In the case of transferring to Washoku, where the gap is expected to be larger, pairing information improves MedR almost by double from Case-3 to Case-10.

In the most resource scarcity situation (Case-1) where only title is available for unsupervised learning, the performance in terms of MedR is suboptimal. Despite this, the MedR of title-only is still 17.0% better than image-only in Chuan⟶Washoku. The result potentially implies that, when the gap is large, title is more informative than image modality for domain transfer. Surprisingly, when both titles and images are available but without pairing information (Case-5), the improvement compared to title-only is insignificant for all the transfers. Improvement becomes significant only if title-image pairing information (Case-8) are available. Compared to image modality, ingredient appears to be more complementary to title. In all the transfers, using both title and ingredient (Case-2) result in performance leap compared to using title-only. The result is even close to that of using recipe-only (Case-3). In other words, among the three text-based resources (title, ingredients, cooking instructions), ingredients appear to be essential to bridge the domain gap. Figure 4 shows three typical examples illustrating how different information in recipes are helpful for transfer. In Figure 4(a), the recipe title already captures all the visible ingredients in the dish. Hence, dish title can be retrieved at the top-1 position with title-only information for transfer. If a title only partially captures some ingredients, such as in 4(b), further using ingredients for transfer is required to rank the recipe at top position. For popular Washoku dishes such as "rich ball lunch" in 4(c), there could exist several similar recipes. In this case, cooking steps which give clue to the ingredient shapes, for example "chopped shrimp" and "shredded carrot", are required to distinguish different versions of similar dishes for retrieval.

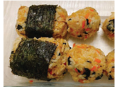
	Query	Recipe	Case-1	Case-2	Case-3
(a)		**Steamed Japanese Tofu with Shrimp** Japanese tofu, shrimp, salt, flour, sugar. 1. Peel the shrimp and draw the string. 2. Cut the Japanese tofu into pieces, don't break it…	1	1	1
(b)		**Schnauzer Rice Ball** cheese, rice, sesame salt, seaweed… 1. Put the sesame salt into the beating bowl… 2. 2/3 of the rice is mixed with sesame salt, and the remaining 1/3...	23	1	1
(c)		**Rice Ball Lunch** shrimp, seaweed, carrot, rice, butter, Sesame oil, ginger… 1. Shrimps are soaked and chopped… 2. Take out the ginger slices, shred the carrots, then chop them...	8	3	1

Figure 4: The retrieved dish title for a query image by using different information for transfer: title-only (Case-1), title and ingredient (Case-2), recipe (Case-3).

In all the transfers, additional use of image modality without pairing information with text modality only manages to introduce slight improvement. We believe that this is due to large visual variation between different cuisines. For example, Chuan dishes are mostly spicy and, as a result, the presentations are apparently different from other two cuisines in terms of color and texture. The large difference causes difficulty for image encoder to fool discriminator in adversarial learning. When pairing information is available, nevertheless, improvements in MedR are consistently observed. This is mainly because pairing information enables metric learning such that text and image embeddings are learnt to be compatible of each other. To provide further insights, Figure 5 shows the consistent improvement in MedR with the increase number of pairs. The pairing information is particularly useful if only recipe titles are available for transfer. When ingredients are also available for transfer, nevertheless, the impact of having pairing information is less obvious. For example, the improvement in MedR is less than 2 ranks by increasing the number of pairs from 1K to 5K. On the other hand, when the entire recipes are available, constant gradual improvement is noticeable. This is due to the fact that the pairing information enables not only metric learning, but also the reconstruction of food images from recipes for adversarial transfer learning. In addition, fine-grained cooking and cutting methods can be more effectively leveraged to disambiguate recipes of similar ingredients composition.

Using resources in target domain only (i.e., Oracle-2) is consistently outperformed by Oracle-1, which uses all resources in source

Table 3: MedR performances for different cuisine transfers. The size of testing dataset (1K or 5K) is indicated in the parenthesis. The baselines are trained using all the training examples in the source domain and then applied to answer queries by retrieving from the partial resources available on the target domain. For ease of visualization, the runs that using the same type of resources for retrieval are marked with the same color. For example, the light cyan color indicates the runs that their recipe resources are formed by dish titles only.

Methods	Resources	C⟶Y (1K)	C⟶Y (5K)	C⟶W (1K)	Y⟶C (1K)	Y⟶C (5K)	Y⟶W (1K)
Baseline (source-only)	Title	10.0	44.9	26.7	19.1	87.6	31.4
	Title+Ingredients	5.4	22.5	13.8	10.4	45.6	16.4
	Recipe	4.8	19.6	33.6	12.5	54.7	19.2
Unsupervised	Case-1	9.1	42.8	26.4	17.6	80.5	31.2
	Case-2	5.3	20.4	13.5	9.9	42.7	15.7
	Case-3	4.2	16.6	11.7	8.8	38.6	15.7
	Case-4	4.6	17.6	31.8	11.6	52.2	17.3
	Case-5	9.1	40.6	26.0	17.3	79.7	31.1
	Case-6	5.2	20.3	13.2	9.8	42.5	15.7
	Case-7	4.1	15.9	10.8	8.3	37.4	13.9
Semi-supervised	Case-8	6.7	28.6	13.8	11.3	49.4	15.4
	Case-9	4.8	18.7	11.0	7.2	31.1	11.5
	Case-10	3.0	11.9	6.6	6.0	25.5	8.4
Fully-supervised	Oracle-2	6.2	25.3	19.4	3.6	13.6	19.4
	Oracle-1	2.7	8.7	4.9	2.0	6.7	8.0

and target domains. The result indeed implies that food recognition is highly transferable even when the domain gap is large (e.g., Chuan⟶Washoku). Under the aid of source domain, using only ingredients and their titles in a target domain for unsupervised learning already surpasses Oracle-2 in all the transfers. Indeed, Oracle-2 basically only performs better than the most scarcity case where only titles are available for training.

In summary, we can characterize the impact of resources on performance as followings. When only dish titles are known, recognition performance are likely to be suboptimal. However, the performance is expected to boost significantly and surpass Oracle-2 if the ingredients of dishes are involved in training. Images are not necessary useful unless if pairing information is available. Under this situation, even if pairings are only exists for some data (5,000 pairs in the experiment), cross-modal embedding of features can introduce noticeable improvement in MedR. The performance is close to Oracle-1 when recipe-image pairs are available for semi-supervised learning.

5.2 Impacts of Domain Gap and Datasize

In practice, the difficulty of transfer depends on the domain gap. For example, transferring from Chuan to Washoku is potentially more challenging than Chuan to Yue. Table 3 lists the details of performances for transferring between different cuisines. Note that the baselines are the fully supervised models trained with the training data in a source domain only. When being employed for food recognition, the same baseline could exhibit different MedR performances depending on the resource available in a target domain. Using the transfer from Chuan to Yue (C⟶Y) as example, the baseline achieves better MedR when the testing dataset is composed of recipes rather than dish titles. In general, when domain gap is smaller, such as C⟶Y, constant improvement in MedR is observed for unsupervised and semi-supervised learnings when more resources are available for training and retrieval. Comparing C⟶W to C⟶Y, the domain gap is larger as observed from the higher MedR value. An interesting fact is that, when the gap is larger, the baseline can perform the worst (MedR=33.6) when the target dataset is composed of recipes. In C⟶W, the larger gap attributes to different cooking techniques in two cuisines. As a result, food recognition by retrieving recipes is not necessarily an ideal solution. Instead, applying the baseline to retrieve dish titles alone can attain better performance. Nevertheless, when the recipes of a target domain are available for training, the performance boost is significant. For example, the MedR is boosted to 11.7 (Case-3) under unsupervised learning and to 6.6 (Case-10) under semi-supervised learning, closely following the performance of Oracle-1 (MedR=4.9).

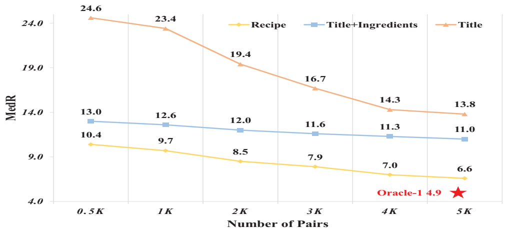


Figure 5: Performance trend with increasing number of pairing information available for linking text and image modalities (Chuan⟶Washoku). The performance of Oracle-1 is indicated with asterisk for reference.

Note that domain gap is not symmetric, for example, the performance in C⟶Y is not necessarily similar to Y⟶C. Instead, the performance depends largely on the number of training examples available in a source domain. As the number of image-recipe pairs

Table 4: Ablation study conducted on C→Y transfer for Case-7. RRB: recipe residual block; Adv: Adversary learning with recipe discriminator (Adv_R) and image discriminator (Adv_V); Reg: semantic regularization with multi-label ingredient recognition (Reg_R) and recipe-to-image generation (Reg_V).

Models	MedR	R@10	R@50
Full model w/o RRB	4.1	63.81	83.18
Full model w/o Adv	4.4	63.51	82.73
Full model w/o Adv_R	4.3	63.52	82.81
Full model w/o Adv_V	4.2	63.58	82.84
Full model w/o Reg	4.8	62.47	80.89
Full model w/o Reg_R	4.4	63.40	82.20
Full model w/o Reg_V	4.2	63.62	83.18
Full model	4.1	63.86	83.40

in Yue cuisine is only about half of that in Chuan cuisine, Y→C is a harder transfer than C→Y as seen in Table 3. Nevertheless, the performance gain by transferring a source model using partial data in the target domain is almost the same. For example, through semi-supervised learning with titles and images, the rank is improved by 40.8% from MedR=19.1 (baseline) to 11.3 (Case-8). Similar margin of improvement is noticed, i.e., 33.0% from MedR=10.0 to 6.7 in C→Y. It is also worth noticing that, the performance difference between two source models does not just depend on the number of training data. For example, in terms of cooking techniques, Washoku is more similar to Yue than Chuan. Hence, despite that the source model in Chuan cuisine is trained with more examples, its performance for retrieving Washoku recipes is worse than the source model trained in Yue cuisine. Furthermore, due to relative large difference in visual appearance of Chuan and Washoku dishes, using only the images in Washoku cuisine for unsupervised learning (Case-4) does not perform better than directly using the source model trained in Yue cuisine.

Increasing the size of testing dataset from 1K to 5K also impacts the performance considerably. Nevertheless, constant improvement is also noticeable when leveraging partial resource in a target domain for model transfer. The degree of performance gains across different cases are indeed bigger if comparing to the dataset of small size. For instance, in C→Y, using titles for unsupervised learning (Case-1) elevates the MedR by 2.1 ranks versus 0.9 rank in 1K-size dataset. Larger degree of gain is obtained in Y→C, where the MedR is elevated by 7.1 ranks versus 1.5 ranks in 1K dataset. When pairing information is available, semi-supervised training using title-only manages to improve MedR by 16.3 ranks (C→Y) and 38.2 ranks (Y→C).

5.3 Ablation Study

As this is the first work that address transfer learning with partial data in target domain, we are not able to draw any relevant prior work for direct comparison. Instead, ablation study assessing impact of different network components is conducted. Our network architecture is similar in spirit as [17] for Case-7. The major difference from [17] is the use of recipe residual block (RRB) to deal with problem of data incompleteness, in addition to loss functions such as recipe-to-image generation and ingredient recognition specific to the application. Table 4 lists the performances when one out of three components is taken away from the proposed model. Note that each of the adversary learning and semantic regularization has two sub-components. In Case-7, since recipes are available, RRB can be omitted and replaced by concatenating the three embeddings from recipe encoder. This is a common setting adopted by cross-modal recipe retrieval [33, 36]. As shown in Table 4, RRB does not impact the model negatively, showing almost the same performances as the ordinary feature concatenation. Similar as other reported works in the literature [12, 17, 27], adversarial learning (Adv) is essential for learning domain-invariant embedding. Without Adv, the MedR drops by 0.3 rank. Recipe discriminator shows slightly greater contribution than image discriminator when one of them is considered by the model. We also replace Adv with the commonly adopted maximum mean discrepancy (MMD) measure in [21]. However, no improvement is noted. Comparatively, semantic regularization shows larger margin of degradation when being taken away from the model. The result is consistent with the conclusion made in [36]. When either ingredient prediction (Reg_R) or image generation (Reg_V) is incorporated, noticeable improvement in MedR is observed.

6 CONCLUSION

We have presented a new perspective of scaling up food recognition across different cuisines by dealing with partial data view, either due to missing of multi-modality information or incomplete textual resource. The network architecture, with the proposal of recipe residual block (RRB), can deal with missing data in a generic manner. The empirical studies, under the ten different scenarios of resource scarcity, reveals the feasibility and impact of different resources in performing domain adaptation. Even when using dish title-only for model transfer, incremental improvement can be noticed. Ingredients play a vital role in model transfer. In all the experiments, ingredients contribute more significantly than image modality if pairing information between them do not exist. Cooking instruction, while proven as important for within-cuisine food recognition [10], may hurt performance if domain gap is large and source model is applied without adaptation. In this case, model transfer using recipe-only resources can lead to significant performance boost. Cross-modal pairing information introduces consistent improvement. The degree of improvement is proportional to the increase number of pairings and the type of resources. Specifically, dish titles and recipes are benefited more from pairing information than ingredients. Finally, the improvement due to model transfer is also shown to be more evident in larger than smaller testing dataset. Currently, our work can only deal with recipes of the same language. The future work includes extension to cross-lingual and cross-domain food transfer.

ACKNOWLEDGMENTS

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11203517).

REFERENCES

[1] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*. 440–447.

[2] Marc Bolaños, Aina Ferrà, and Petia Radeva. 2017. Food ingredients recognition through multi-label learning. In *International Conference on Image Analysis and Processing*. Springer, 394–402.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *European conference on computer vision*. Springer, 446–461.

[4] 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.

[5] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2940–2949.

[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, Lei Pang, and Chong-Wah Ngo. 2017. Cross-modal recipe retrieval: How to cook this dish?. In *International Conference on Multimedia Modeling*. Springer, 588–600.

[8] Jing-jing 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.

[9] Jing-Jing 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.

[10] Jing-Jing 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.

[11] Jing-Jing Chen, Liangming Pan, Zhipeng Wei, Xiang Wang, Chong-Wah Ngo, and Tat-Seng Chua. 2020. Zero-shot Ingredient Recognition by Multi-Relational Graph Convolutional Network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*.

[12] 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.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[16] Ehsan Hosseini-Asl, Yingbo Zhou, Caiming Xiong, and Richard Socher. 2019. Augmented Cyclic Adversarial Learning for Low Resource Domain Adaptation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

[17] Xin Huang, Yuxin Peng, and Mingkuan Yuan. 2018. Mhtn: Modal-adversarial hybrid transfer network for cross-modal retrieval. *IEEE transactions on cybernetics* (2018).

[18] Shuqiang Jiang, Weiqing Min, Yongqiang Lyu, and Linhu Liu. 2020. Few-Shot Food Recognition via Multi-View Representation Learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2020).

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

[20] Shuang Li, Chi Harold Liu, Binhui Xie, Limin Su, Zhengming Ding, and Gao Huang. 2019. Joint Adversarial Domain Adaptation. In *Proceedings of the 27th ACM International Conference on Multimedia*. 729–737.

[21] 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.

[22] Yuji Matsuda, Hajime Hoashi, and Keiji Yanai. 2012. Recognition of multiple-food images by detecting candidate regions. In *2012 IEEE International Conference on Multimedia and Expo*. IEEE, 25–30.

[23] Weiqing Min, Linhu Liu, Zhengdong Luo, and Shuqiang Jiang. 2019. Ingredient-Guided Cascaded Multi-Attention Network for Food Recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1331–1339.

[24] 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*. Springer, 129–141.

[25] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[26] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. 2019. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2239–2247.

[27] 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.

[28] Amaia Salvador, Michal Drozdzal, Xavier Giro-i Nieto, and Adriana Romero. 2019. Inverse cooking: Recipe generation from food images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10453–10462.

[29] 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 conference on computer vision and pattern recognition*. 3020–3028.

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

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

[32] 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).

[33] 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 Conference on Computer Vision and Pattern Recognition*. 11572–11581.

[34] Heng Zhao, Kim-Hui Yap, Alex C Kot, Lingyu Duan, and Ngai-Man Cheung. 2019. Few-shot and Many-shot Fusion Learning in Mobile Visual Food Recognition. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1–5.

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

[36] 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 Conference on Computer Vision and Pattern Recognition*. 11477–11486.

[37] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 1568–1575.