Learning Using Privileged Information for Food Recognition

Lei Meng
National University of Singapore
lmeng@nus.edu.sg

Long Chen
Zhejiang University
longc@zju.edu.cn

Xun Yang
National University of Singapore
xunyang@nus.edu.sg

Dacheng Tao
The University of Sydney
dacheng.tao@sydney.edu.au

Hanwang Zhang
Chunyan Miao
Nanyang Technological University

Tat-Seng Chua
National University of Singapore
dcscts@nus.edu.sg

Abstract

Food recognition for user-uploaded images is crucial in visual diet tracking, an emerging application linking multimedia and healthcare domains. However, it is challenging due to the various visual appearances of food images. This is caused by different conditions when taking the photos, such as angles, distances, light conditions, food containers, and background scenes. To alleviate such a semantic gap, this paper presents a cross-modal alignment and transfer network (ATNet), which is motivated by the paradigm of learning using privileged information (LUPI). It additionally utilizes the ingredients in food images as an "intelligent teacher" in the training stage to facilitate cross-modal information passing. Specifically, ATNet first uses a pair of synchronized autoencoders to build the base image and ingredient channels for information flow. Subsequently, the information passing is enabled through a two-stage cross-modal interaction. The first stage of interaction adopts a two-step method, called partial heterogeneous transfer, to 1) alleviate the intrinsic heterogeneity between images and ingredients and 2) align them in a shared space to make their carried information about food classes interact. In the second stage, ATNet learns to map the visual embeddings of images to the ingredient channel for food recognition from the viewpoint of "teacher". This leads to a refined recognition by a multi-view fusion. Experiments on two real-world datasets show that ATNet can be incorporated with any state-of-the-art CNN models to consistently improve their performance.

CCS Concepts

• Computing methodologies → Classification and regression trees; Learning latent representations; Neural networks.

KEYWORDS

Food recognition; Learning using privileged information; Heterogeneous feature alignment; Cross-modal fusion

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:


1 INTRODUCTION

Visual diet tracking [18, 22, 25] is an emerging AI-powered application in healthcare domain. It collects the users’ uploaded food photos and incorporates food recognition algorithms to understand their eating habits. Comparing with text-based approaches for diet management, such as MyFitnessPal, visual diet tracking is more convenient for users to record their daily intake. This has triggered the launch of several research projects and startup companies, such as FoodLog1, TADA project2, foodAI3, and DietLens4.

Food recognition for user-uploaded images is the key to visual diet tracking, which, in an ideal case, should be able to accurately recognize all the food dishes in the uploaded photos. However, it is an open problem [2, 17, 25, 33, 40] due to the intra-class diversity in visual appearances and the complexity of background scenes. This motivates the exploration of auxiliary information about the food images to assist in the recognition process. One direction is to use the additional features to label data and limits such approaches to images taken in the lab settings. Another line of research explores the descriptive data to food images, such as ingredients. Chen et al. [3, 7] propose a multitask learning algorithm based on VGG [28] for the simultaneous food recognition and ingredient prediction. The visual and semantic features are therefore aligned by the shared convolutional layers of VGG. However, this algorithm uses ingredients as labels. Therefore, it does not explore the semantic representation of food to take full advantage of the multimodal associations between images and ingredients.

To address the aforementioned issues, this paper presents a cross-modal alignment and transfer network (ATNet) to explore the semantic counterpart of food images, i.e., ingredients, for effective image-based food recognition. It follows a learning paradigm called learning using privileged information (LUPI), which was first introduced by Vapnik et al. [31] to utilize the descriptive information

1http://www.foodlog.jp/en
2http://www.tadaproject.org/
3http://foodai.org/
4https://www.dietlens.com/
of primary data as an “intelligent teacher” to guide the learning process of SVM+; i.e. the “student”, in the training stage. Such information is called privileged information (PI) since it is not allowed in the testing phase. LUPI regularizes the learning process in two ways, i.e. similarity control and knowledge transfer [30]. Similarity control models PI as a regularizer to guide the mapping from the feature to the label space, while knowledge transfer allows the “student” to learn and infer in the view of the “teacher”. LUPI has been extended to other learning algorithms, such as ELM [39], information bottleneck theory [26] and CNN [34], which have been applied to image recognition and other related tasks [26, 34].

Following this paradigm, ATNet uses the ingredients of food images as PI, i.e. the “intelligent teacher”, and implements the functions for similarity control and knowledge transfer of LUPI through a two-stage cross-modal interaction. As illustrated in Figure 1, ATNet uses a pair of synchronized autoencoders to model the representations of images I and ingredients S in the respective channels. This ensures that the learned visual and the semantic embeddings \( \hat{L}^v = f^v(I) \) and \( \hat{L}^s = f^s(S) \) encode the information of their source data I and S as much as possible. The first interaction happens in a heterogeneous transfer network, where \( L^v \) and \( L^s \) are aligned to produce a regularized visual embedding \( \hat{L}^v \) using a deep transfer mapping \( \theta(L^v, L^s) \). Notably, finding a suitable \( \theta(L^v, L^s) \) is challenging due to the intrinsic heterogeneity of \( L^v \) and \( L^s \) in feature distributions. Additionally, \( \theta(L^v, L^s) \) needs to align the paired image and ingredient features, rather than a match of distributions in conventional transfer learning. To address this issue, a partial heterogeneous transfer method is proposed. It first finds a subset of features in \( L^v \) and \( L^s \) that carries the shared information on food classes to alleviate the problem of heterogeneity. Subsequently, it aligns these features in a shared space, making the food class prediction \( C \) obtained from the image channel \( \hat{L}^v \) benefits from the semantic embeddings in the ingredient channel. The second interaction is enabled by a cross-modal mapping \( \hat{L}^v \mapsto \hat{L}^s \) from visual to the semantic space. This allows the visual embedding \( \hat{L}^v \) to make food prediction in the ingredient channel from the view of “teacher”. The gap between \( S \) and \( C \)’ in the mapping \( \hat{L}^s \mapsto S \mapsto C \) is bridged by the pre-acquired knowledge extracted from the training corpus. It contains the statistics on the ingredient distributions \( P(s_i) \) over the food classes \( c_j \), including frequencies \( P(s_p | c_j) \), co-occurrences \( P(s_p, s_q | c_j) \), and the sequential dependences \( P(s_q | s_p) \). The final prediction \( C \) is obtained by a fusion of the predictions drawn from both the image and ingredient channels, i.e. \( C \oplus C' \mapsto \hat{C} \).

Experiments were conducted on the VireoFood172 [7] and the Ingredient101 [5] datasets. We conducted ablation studies to evaluate the effectiveness of each component of ATNet, compared ATNet with the state-of-the-art food recognition algorithms, and used case studies to illustrate the behaviors of ATNet in different successful and failure cases. The results show that the proposed partial heterogeneous transfer significantly improves the recognition performance of different base CNN models. This indicates that ATNet is agnostic to backbone CNN models. Besides, the multiview prediction fusion further increased the robustness of ATNet.

In summary, we make three contributions in this paper:

1. ATNet is the first work that addresses the food recognition problem using the LUPI paradigm. In the training phase, ATNet explicitly take advantage of the multimodal representations of images and ingredients for food recognition. While in the testing phase, only images are needed as input.

2. We propose a two-step method, called partial heterogeneous transfer, to alleviate the intrinsic heterogeneity of the image and ingredient embeddings. It shows promising experimental performance (See Figure 5) and can be incorporated to other tasks involving heterogeneous feature alignment.

3. We explore the mapping between ingredient distribution and food classes using three statistical measures, including ingredient frequencies, co-occurrences, and sequential dependencies. It enables the use of visual embeddings to make food prediction in the ingredient channel. This prediction serves as an additional view to refine the final food recognition performance.
2 RELATED WORK
This paper investigates the use of ingredients as privileged information (PI) to regularize the learning process of the image-based food recognition. Related work lies in the following directions:

- **Food recognition**: Early efforts on visual food recognition rely on image processing, ranging from handcrafted features [2], food region detection [33, 40], to deep learning features [17, 21, 32], and have shown superior performance when incorporated with images for food recognition [7, 32]. However, Wang et al. [32] show that, when using concatenated features of images and ingredients, the performance highly depends on the ingredient features. Chen et al. [7] propose a multitask framework that uses image features from a shared VGG to perform both food class and ingredient prediction.

- **Learning using privileged information**: The LUPI paradigm [31] assumes a teacher-student learning scenario, where a teacher can provide descriptive information (privileged information) about a course (primary data) to assist a student (model) to learn through the guidance of similarity control and knowledge transfer [30]. It is distinct to multimodal analysis in that such privileged information is not available in the testing phase. Despite the use of different machine learning algorithms, it has been widely-used to refer to studies that utilize auxiliary information as a regularizer to enhance the learning of primary data, such as using bounding box features to help image classification [26], using image captions to assist in multi-object detection [34], and using auxiliary knowledge for person re-identification [35–37].

- **Heterogeneous Feature Alignment**: Different from multimodal fusion [16], aligning features in heterogeneous domains usually requires transfer learning [12, 20], which learns linear or nonlinear mappings to align data from both domains in a latent space. Recent studies usually use a shared neural network to learn the high-level features for heterogeneous data and put constraint on their similarity, such as the widely-used KL-divergence, the covariance matrix of feature distributions [29], and the loss from generative adversarial network (GAN) [10]. However, these methods do not address the intrinsic heterogeneity of the data from different domains.

3 PROBLEM FORMULATION
This paper investigates the use of ingredients to enhance image-based food recognition under the LUPI paradigm. As shown in Figure 3, a significant difference between ATNet and the conventional settings is the mapping \( L^v \mapsto L^{v+I} \) that enables food prediction in the ingredient channel. This makes ATNet explicitly take advantage of the semantic embeddings to infer food classes in the semantic space. Additionally, in contrast to multimodal food recognition, ATNet uses the ingredients \( S \) only in the training phase, so it requires only images as input in the testing phase.

As shown in Figure 3, given a dataset including food images \( I = \{i_{ij} \mid j = 1, \ldots, N \} \) of \( I \) classes \( C = \{c_i \mid i = 1, \ldots, I\} \) and the corresponding ingredients \( S = \{s_{ij} \mid j = 1, \ldots, N \} \), ATNet receives data triplets \( (I, S, y) \) in the training phase, where \( y \) is class label, and learns the pipeline of \( I \mapsto L^v \mapsto \hat{L}^v \mapsto \hat{C} \) via a two-stage interaction with the ingredient channel. It has four main procedures:

1. **Embedding modeling using autoencoders**: ATNet independently models the embeddings of \( I \) and \( S \) using two autoencoders: \( I \mapsto L^v \mapsto \hat{I} \), and \( S \mapsto \hat{I} \mapsto S' \).

2. **Heterogeneous feature alignment**: Considering the intrinsic heterogeneity of visual and semantic embeddings \( L^v \) and \( L^{v+I} \), ATNet finds the shared space in two steps: 1) learning two subsets of features \( L^v \) and \( L^{v+I} \) that carry the information describing \( C \) and minimizing the KL-divergence \( KL(L^v || L^{v+I}) \); and 2) aligning \( L^v \) and \( L^{v+I} \) in a shared space to obtain \( L^v \mapsto \hat{L}^v \). Subsequently, the food class indicator \( C \) from the image channel can be obtained by \( \hat{L}^v \mapsto \hat{C} \).

3. **Cross-modal food prediction**: Taking advantage of the autoencoder in the ingredient channel, a cross-modal mapping enables \( L^v \mapsto L^{v+I} \mapsto S' \). The food prediction \( S' \mapsto \hat{C} \) is achieved by using the pre-acquired knowledge on food-ingredient associations, including frequencies \( P(s_p | c_j) \), co-occurrences \( P(s_q | c_i, s_p) \), and sequential dependencies \( P(s_q | c_j, s_p) \).

4. **Multiview prediction fusion**: The final prediction on food classes \( C \) is obtained by a fusion of the indicators obtained from both the image and ingredient channels, i.e. \( C \odot C' \mapsto \hat{C} \).

4 CROSS-MODAL ALIGNMENT AND TRANSFER NETWORK
The cross-modal alignment and transfer network (ATNet), as depicted in Figure 4, has three main modules, i.e. a pair of synchronized autoencoders for the image and ingredient channels and the heterogeneous transfer network. This section illustrates the five key procedures of ATNet, including the autoencoders for the image and ingredient channels, the heterogeneous feature alignment, the cross-modal food prediction, and the multiview prediction fusion.

4.1 Autoencoder for Image Channel
As shown in Figure 4, the autoencoder for the image channel learns the mapping \( I \mapsto L^v \mapsto \hat{I} \) using a CNN encoder \( f^v(\cdot) \) and a CNN decoder \( g^v(\cdot) \). The encoder can be any state-of-the-art networks, such as VGG [28], ResNet [13], and wide residual network.
where (WRN) [21, 38], and the decoder should reverse the operations of the encoder. The autoencoder framework is used to learn $L^I$ that preserves the information of its input $I$, using a reconstruction loss:

$$L^I_{rec} = \|I' - I\|_F.$$  

where $\|.\|_F$ is the Frobenius norm.

### 4.2 Autoencoder for Ingredient Channel

Similar to image channel, the autoencoder for the ingredient channel learns the mapping $S \mapsto L' \mapsto S'$ using a pair of semantic encoder $f^S(.)$ and decoder $g^S(.)$. Two designs for the semantic encoder and decoder have been investigated, as illustrated below.

#### 4.2.1 Feed-Forward Neural Network

Using a feed-forward neural network (NN) to implement $f^S(.)$ and $g^S(.)$ leads to a binary “bag-of-word” representation for $S$, so for $S_m \in S, s_m \in \{0, 1\}$. This makes $L' \mapsto S'$ a multi-label prediction task and requires a sigmoid activation function for $S'$. This autoencoder is optimized using a reconstruction loss computed by the $\ell_2$ norm $\|.|\|_2$:

$$L^S_{rec} = \|S' - S\|_2.$$  

#### 4.2.2 Long-Short Term Memory

The second design uses long-short term memory (LSTM) [15] as a building block for the semantic encoder and decoder, in order to captures the semantics and dependences between ingredients. It has three key steps:

1. **Sequential encoding of ingredients**: Having the word embeddings of ingredients $S = [s_1, ..., s_M]$ where $s_m \in \mathbb{R}^{1 \times r}$, the LSTM encoder incrementally processes the ingredient vectors and output the same number of hidden vectors encoding the past inputs. At time $t$, the LSTM encoder is defined as

   $$h_t = \text{LSTM}(s_t, h_{t-1})$$  

2. **Self-Attention for Ingredient Embedding**: The self-attention step fuses the output hidden vectors $H = [h_1, ..., h_M]$ to learn a unified embedding for ingredients, defined as

   $$L' = AH_1$$  

where $A = \text{softmax}((W_2 \text{tanh}(W_1H)))$,  

(4) $A$ is a trainable attention vector that encodes the importance of $h_m$ from $d$ aspects, $W_1 \in \mathbb{R}^{d \times r}$ computes a $d$-dimensional attention for $H$, and $W_2 \in \mathbb{R}^{1 \times d}$ fuses the $d$ types of attention to produce $A$.

3. **Decoding for Ingredient Prediction**: Given the semantic embedding $L'$, the LSTM decoder follows the conventional image captioning procedures [9] to decode $M$ hidden vectors $H$, followed by a non-linear mapping to incrementally predict the ingredients $S' = [s'_1, ..., s'_M]$, defined as

   $$s'_t = \text{softmax}(g^S_2(h_t)), \quad h_t = \text{LSTM}(g^S_1((L', h_{t-1})), h_{t-1})$$  

where $\{,\}$ is the vector concatenation operator, and $g^S_1(.)$ and $g^S_2(.)$ are non-linear mappings.

The autoencoder for the ingredient channel is optimized using a binary cross-entropy loss for ingredient prediction, defined as

$$L^S_{rec} = \text{BCE}(S', S),$$  

(8) where $\text{BCE}(.)$ essentially measures the reconstruction error between the predicted ingredients $S'$ and the true ingredients $S$.

### 4.3 Heterogeneous Feature Alignment

The encoding-decoding frameworks for the image and ingredient channels independently learn the visual and semantic embeddings $L'$ and $L'$. Considering the much stronger discriminative power of the semantic embedding for food recognition, it is straightforward to align the distributions of visual embeddings to their semantic counterparts for improved performance, leading to the first-stage cross-modal interaction of ATNet. Heterogeneous transfer [10, 29] is a commonly-used method for this problem, which aims to find a subspace where the visual and semantic embeddings overlap, i.e.
\( \theta^l(L^v) = \theta^l(L^l) \). However, the intrinsic heterogeneity of image and ingredient data makes it an intractable task.

To alleviate this problem, we propose a two-step method, called partial heterogeneous transfer. It first aligns the features in \( L^v \) and \( L^l \) that carry the shared information on food class. Subsequently, this method finds a shared space for them. It is based on the hypothesis that \( L^v \) and \( L^l \) are described by two types of features, where \( L^v_k \) (\( k = (v, t) \)) contains information on food class label and \( L^l_k \) contains information on their own styles [4]. Therefore, disentangling \( L^v_k \) and \( L^l_k \) from \( L^v \) and \( L^l \) for alignment alleviates the intra-modal diversity and helps to discover the latent space to align the heterogenous features sharing the information on food class.

As shown in Figure 4, the heterogeneous transfer network first adjusts \( L^v \) and \( L^l \) (shown in grey) to allow only part of them, i.e. \( L^v_k \) and \( L^l_k \), to be aligned for food recognition. Subsequently, two pairs of non-linear mappings \( \{ \theta^v_k(\cdot), \theta^l_k(\cdot) \} (k = (v, t)) \) and a shared linear mapping \( \phi(\cdot) \) are used to implement the mappings \( L^v_k \mapsto C \) and \( L^l_k \mapsto C \). Three loss terms are used to learn the embeddings:

1. Learning disentangled features \( L^v_k \) (\( k = (v, t) \)) is conditioned on the cross-entropy loss \( CE(\cdot) \) of both channels, defined as
   \[
   L^v_k = CE(\text{softmax}(\phi(L^v_k)), y),
   \]
   \[
   L^l_k = CE(\text{softmax}(\phi(L^l_k)), y). \tag{9}
   \]
2. The loss of KL-divergence allows for a one-direction alignment of \( L^v \) to \( L^l \), defined as
   \[
   L_{KL} = KL(L^l || L^v). \tag{11}
   \]
   It alleviates the cross-modal gap without breaking the distribution of the semantic embeddings.
3. The \( \ell_2 \) norm aligns \( L^v \) and \( L^l \) in a shared space, defined as
   \[
   L_{align} = ||L^v - L^l||_2. \tag{12}
   \]
   In addition, the shared \( \phi(\cdot) \) implicitly aligns \( L^v \) and \( L^l \) by the shared food class indicator.

### 4.4 Cross-Modal Food Prediction

In addition to the cross-modal feature alignment, the second stage of interaction is enabled by the synchronized autoencoders, which makes it possible to learn a cross-channel mapping for ingredient prediction, i.e. \( L^v \mapsto L^{v \rightarrow t} \mapsto S' \), as shown in Figure 4. Learning such a mapping requires two loss regularizers, defined as

\[
L_{v \rightarrow t} = ||L^{v \rightarrow t} - \text{detach}(L^t)||_2, \tag{13}
\]

\[
L_{v \rightarrow s'} = CE(S', S), \tag{14}
\]

where \( \text{detach}(\cdot) \) removes the gradient for \( L^l \) from \( L_{v \rightarrow t} \), making \( L^l \) from \( L_{v \rightarrow t} \) a one-directional alignment from \( L^v \) to \( L^l \).

Note that the prediction \( S' \) reveals the image-ingredient association between the ingredients \( S \) and the image \( I \), i.e. \( P(s|p[I]), P(s|s_q[I], p[I], s_p) \). This enables the mapping \( S' \mapsto C' \) by using the pre-acquired knowledge on food-ingredient association, including frequencies \( P(s_p|c_j) \), co-occurrences \( P(s_q|s_q|c_j) \), and sequential dependencies \( P(s_q|s_q|c_j). \)

Specifically, the similarity between the image \( I \) and a food class \( c_j \) can be measured by a histogram matching. For example, given the ingredient frequencies \( P(s_p|I) \), the probability that \( I \) belongs to the \( j \)-th class of \( C \) is defined as

\[
P_f(c_j|I) = \frac{||P(S|I) \land P(S|c_j)||}{||P(S|I) \lor P(S|c_j)||}, \tag{15}
\]

where \( P(S|I) = [P(s_1|I), \ldots, P(s_M|I)] \) and \( P(S|c_j) = [P(s_1|c_j), \ldots, P(s_M|c_j)] \) are probability distributions. \( \land \) and \( \lor \) are element-wise fuzzy AND and OR operators such that \( a \land b = \min(a, b) \) and \( a \lor b = \max(a, b) \), \( ||.|.|| \) is the \( \ell_1 \) norm.

Note that, we typically use the Top-10 food classes and the Top-\( n \) ingredients for the matching, where \( n \) is data-dependent. Moreover, to take advantage of the sequential prediction of LSTM, the Top-3 predictions for each word in a sequence are used. In a similar way, \( P_{co}(c_j|I) \) and \( P_{seq}(c_j|I) \) can be obtained using \( P(s_q|s_q[I], p[I], s_p) \), respectively. Therefore, the integrated indicator for food classes \( P(c_j|I) \) inferred from ingredient predictions can be computed using a weighted fusion of the three views, defined as

\[
P(c_j|I) = \alpha P_f(c_j|I) + \beta P_{co}(c_j|I) + \gamma P_{seq}(c_j|I), \tag{16}
\]

where \( \alpha, \beta, \text{ and } \gamma \) are hyperparameters and \( \alpha + \beta + \gamma = 1 \). Note that if \( S' \) is predicted using neural network, \( P_{seq}(c_j|I) \) is unavailable. In this way, the food class indicator in the ingredient channel \( C' = [P(c_1|I), \ldots, P(c_j|I)] \) is obtained.

### 4.5 Multiview Prediction Fusion

Having the predictions on food classes from both the image and ingredient channels, i.e. \( C \) and \( C' \), ATNet computes the final prediction \( \hat{C} \) using their fused predictions, defined as

\[
\hat{C} = C \oplus C', \tag{17}
\]

where \( \oplus \) can be any of the commonly used vector operators, such as plus, multiplication, max-pooling, and min-pooling.

### 4.6 Training Strategies

ATNet is optimized using four groups of loss terms, including (a) losses for the autoencoders of the image and ingredient channels, i.e. \( L_{v \rightarrow rec} \) and \( L_{l \rightarrow rec} \), (b) losses for aligning visual and semantic embeddings, i.e. \( L_{KL} \) and \( L_{align} \), (c) losses for food classification, i.e. \( L_{v \rightarrow c} \) and \( L_{l \rightarrow c} \), and (d) losses for cross-modal ingredient prediction, i.e. \( L_{v \rightarrow s'} \) and \( L_{l \rightarrow s'} \). So balancing their weights for optimization is important and non-trivial. Therefore, we use two strategies to guarantee a smooth training:

1. **Independently training the autoencoders first**: That is, before training ATNet, we first train the autoencoders for the image and ingredient channels independently. The pretrained parameters are used to initialize the autoencoders of ATNet. This ensures that \( L^v \) and \( L^l \) are trained to carry meaningful information of image and ingredients.
2. **Weighting loss terms to match a ratio**: The eight loss terms are weighted to match a ratio. For example, the autoencoder losses are weighted to be \( L_{v \rightarrow rec} = L_{l \rightarrow rec} = 1 \), since ATNet uses the pretrained parameters for autoencoders. \( L_{v \rightarrow c}, L_{l \rightarrow c}, \) and \( L_{v \rightarrow s'} \), i.e. the losses for the target goals, are weighted to be 10. The losses regularizing the intermediate embeddings, i.e. \( L_{KL}, L_{align}, \) and \( L_{v \rightarrow t} \) are weighted to be 5.
5 EXPERIMENTS

5.1 Datasets
The performance of ATNet was evaluated on:

- **VireoFood-172 dataset [7]**: It has 110,241 Chinese food images from 172 classes, which are of size 256×256 and manually annotated with a vocabulary of 353 ingredient terms (three per image on average). It is notable that the ingredients are made by visual tagging, which contain mainly visually appeared ingredients in the images. In the experiments, we followed the original paper’s experimental setup [7] to use 66,071, 11,016, and 33,154 images for training, validation, and testing, respectively.

- **Ingredient-101 dataset [5]**: It uses the images of the Food-101 dataset [6] and annotates them with nine ingredient terms per image on average. This dataset has 1,000 western food images of size 256×256 for each of the 101 classes and 446 ingredient terms. The data split includes 68,175 images for training, 7,575 for validation, and 25,250 for testing.

5.2 Model Details
We investigated several base CNN networks for image channel:

- **vgg19_bn**: Pytorch implementation² for 19-layer VGG [28] with each convolutional layer followed by batch normalization.
- **resnet50**: Pytorch implementation³ for 50-layer ResNet [13].
- **WRN50-2**: Pytorch implementation² for Wide Residue Networks (WRN) [38] using ResNet50 and a wide factor of 2.
- **WiSeR**: In-house implementation of a four-layer fully-connected network, each of which has the number of neuron equal to the number of ingredients and is followed by a ReLU activation function. L^v and L^h have the same length. The four-layer decoder uses a Sigmoid activation function to map L^t to the ingredient indicator S.
- **LSTM**: The length of the embedding vectors for S and the hidden vector h also equals to that of L^v, d in self-attention, as defined for W₁ and W₂ in Equation (5), is set to 5.

In experiments, ATNet consistently uses the batch size of 64 and the learning rate of 1e-4 for vgg19_bn, 1e-3 for resnet50, and 5e-5 for other networks (decay by 0.1 for every five epochs) using Adam optimizer. Since all the base models are pretrained on ImageNet, input images are normalized using the published ImageNet means and standard deviations. Random horizontal flip is also used. In the heterogeneous transfer network, the mappings for L^k_i → L^k and L^k → L̂^k (k = {v, t}) are two-layer networks with each layer followed by a ReLU. The length for these intermediate embeddings is 3/8 of L^v. The network for cross-modal mapping L^v → L^t is an eight-layer fully-connected network, each of which has the same length to L^v and is followed by a ReLU. Regarding cross-modal food prediction, Top-5 and -10 and ingredients are used for the VireoFood-172 (three ingredients in average) and the Ingredient-101 (nine ingredients in average) datasets, respectively.

## Table 1: Classification performance (in Top-1 accuracy (%)) of ATNet with different combinations of components on VireoFood-172 dataset. D: food classification using L^t \rightarrow C; AlignNN(LSTM): Using L^t learned by NN(LSTM) for alignment; F: Using multiview prediction fusion.

<table>
<thead>
<tr>
<th>Components</th>
<th>Baseline Models</th>
<th>Base CNN Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Origin</td>
<td>vgg19_bn</td>
<td>86.1</td>
</tr>
<tr>
<td>+ D</td>
<td>resnet50</td>
<td>80.2</td>
</tr>
<tr>
<td>+ AlignNN</td>
<td>WRN50-2</td>
<td>82.5</td>
</tr>
<tr>
<td>+ AlignNN</td>
<td>WiSeR</td>
<td>82.8</td>
</tr>
<tr>
<td>+ AlignNN + F</td>
<td>LSTM</td>
<td>83.7</td>
</tr>
<tr>
<td>+ AlignNN + F</td>
<td>LSTM</td>
<td>83.9</td>
</tr>
<tr>
<td>+ AlignNN + F</td>
<td>LSTM + AlignNN</td>
<td>85.4</td>
</tr>
<tr>
<td>+ AlignNN + F</td>
<td>LSTM + AlignNN</td>
<td>85.6</td>
</tr>
<tr>
<td>+ AlignNN + F</td>
<td>LSTM + AlignNN</td>
<td>85.5</td>
</tr>
</tbody>
</table>

5.3 Ablation Study

5.3.1 Evaluation on Base Models. We first evaluate the performance of the base models. As observed in the row “Origin” of Table 1, all four base models achieve comparably good performance. Especially, vgg19_bn outperforms resnet50 at the cost of efficiency. Its network size is nearly six times larger than that of resnet50. Additionally, dense filters significantly improve the visual representation. WRN50-2 and WiSeR double the filters of resnet50. This leads to a much better performance.

5.3.2 Evaluation on Heterogeneous Feature Alignment. As shown in the row “+D” of Table 1, using part of features for classification does not lead to an obvious drop in performance. This demonstrates that the base models can learn to compress the information on food class into a subset of features. Adding the component of heterogeneous feature alignment significantly improves the performance of all base models. Notably, the performance using LSTM as semantic encoder consistently outperforms that using NN, demonstrating the effectiveness of learning the semantics dependence of ingredients. Interestingly, resnet50 achieves a comparable performance to vgg19_bn after the alignment. This demonstrates the effectiveness of ATNet in bridging the semantic gap of the visual embeddings.

To investigate the reason, we visualize the visual and semantic embeddings that are obtained in different steps of ATNet with vgg19_bn and LSTM. PCA was chosen instead of t-SNE to map the embeddings to 2D space, since a linear transformation can preserve the domain-specific characteristics, such as scales, of the heterogeneous embeddings. As shown in Figure 5(a), the visual and semantic embeddings learned by autoencoders, i.e., L^v and L^t, have a significantly different distributions in the 2D space, caused by domain shift. The first step of partial heterogeneous transfer, as defined in Equations (11), successfully maps the visual embedding to align with its semantic counterpart in the same scale, as shown in Figure 5(b). The final aligned embeddings using ATNet are shown in Figure 5(c), where all of the paired embeddings are better aligned. For further comparison, Figure 5(d) depicts the aligned embeddings using the deep coral method [29]. It shows that directly aligning L^v and L^t changes their original scales and therefore does not preserve their distributions in the respective channels, leading to a downgraded performance. The above analysis demonstrates the effectiveness of the proposed partial heterogeneous transfer method for heterogeneous feature alignment.
5.3.3 Evaluation on Multiview Prediction Fusion. The last two rows of Table 1 show that fusing the food predictions made from the image and ingredient channels, i.e. C and C’, leads to a further improvement upon heterogeneous feature alignment. Note that the prediction C’ is also a fusion of three indicators obtained using different types of food-ingredient associations, i.e. frequencies P(sp|ci), co-occurrences P(sp,qi|ci), and sequential dependences P(qi|ci,sp), as defined in Equation (16). Therefore, this section, using ATNet with vgg19_bn and LSTM as an example, evaluates the effects of different fusion operators for them.

Effects of Weights for Ingredient-to-Food Indicators

First, we evaluate the importance of the three types of food-ingredient associations in food recognition. It is achieved by measuring their influences on the classification performance of the fused food indicators. Since sequential dependences P(qi|ci,sp) do not apply to the NN-based autoencoder, the importance of P(sp|ci) and P(qi|ci,sp) are evaluated first. As observed in Figure 6(a), using P(sp,qi|ci) solely achieves better performance than using P(sp|ci) for both NN- and LSTM-based autoencoders, indicating the stronger discriminative power of ingredient co-occurrences. Additionally, the best performance is usually obtained when α and β are equally weighted, demonstrating their information complementarity. The importance of P(qi|ci,sp) is revealed in Figure 6(b). As observed, the best performance is achieved at γ = 0.2, followed by a significant drop. This means that the sequential dependencies of ingredients carry helpful information, but it solely is not sufficient.

Effects of Operators for Multiview Prediction Fusion

Similarly, this section illustrates the effects of different operators for the fusion of food class indicators from the visual and ingredient channels, i.e. C and C’, as defined in Equation (17). In the experiments, we selected the indicator of Top-10 classes for refinement, since the visual channel of Alignment can capture the true labels of over 96% of images in the Top-5 predicted classes.

As shown in Table 2, four operators have been investigated to fuse the class indicators C and C’ obtained from the image and ingredient channels, respectively. It is observed that operating on raw values may lead to a drop in performance, which is caused by domain shift. As such, the softmax function is used to make C and C’ in the same scale, leading to a consistent improvement for all operators. The best performance is achieved by the S(min) operator, which works by decreasing the inconsistent predictions.

Figure 5: Visualization for the visual and semantic embeddings of 30 randomly-selected testing samples. (a) L_v and L_t obtained by autoencoders; (b) L_v and L_t obtained by the first step of ATNet’s partial heterogeneous transfer; (c) L_v and L_t obtained by the second step of ATNet’s partial heterogeneous transfer; (d) L_v and L_t obtained by aligning L_v and L_t using deep Coral.

Figure 6: Illustration to the effects of weights α, β, and γ for S’ = C’. (a) α + β = 1 and γ = 0. (b) α = β = 0.5, and C’ is normalized by C’/(α + β + γ).

Table 2: Performance (in Top-1 accuracy (%)) of ATNet using different fusion operations. +, ×, max, and min are elementwise operators on C and C’, S(.) means performing softmax to indicators before the operator.

<table>
<thead>
<tr>
<th>α</th>
<th>β</th>
<th>γ</th>
<th>+</th>
<th>×</th>
<th>max</th>
<th>min</th>
<th>S(+)</th>
<th>S(×)</th>
<th>S(max)</th>
<th>S(min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>84.2</td>
<td>77.9</td>
<td>83.8</td>
<td>84.7</td>
<td>85.4</td>
<td>85.0</td>
<td>85.0</td>
<td>85.3</td>
<td>85.3</td>
<td>85.3</td>
<td>85.3</td>
</tr>
</tbody>
</table>

5.4 Performance Comparison

This section presents the performance comparison on food recognition between ATNet and the state-of-the-art methods, including four CNN networks (resnet18, resnet50, vgg16_bn, and vgg19_bn) that are commonly used for food recognition [1, 19, 32], ARCH-D [7], WRN50-2 [38], and WISER [21]. The implementations of all algorithms except ARCH-D are illustrated in Section 5.2. ARCH-D is an in-house implementation based on Pytorch’s vgg16_bn implementation, so its setting follows that of vgg16_bn in Section 5.2. Notably, it is the only algorithm that uses ingredients to aid the image-based food recognition, using a multitask learning approach. ATNet uses LSTM for the ingredient channel and uses min(·) with softmax as the fusion operator.

As reported in Table 3, ATNet consistently improves its base models on both datasets. This demonstrates the effectiveness of the proposed LUPI framework for food recognition. Comparing with the performance of base models, ATNet alleviates their semantic gap and makes them achieve comparable performance on both datasets. This indicates the successful alignment of the visual embeddings to the semantic space of the ingredient channel. Notably, the improvement over WRN50-2 and WISER is less than their base model, i.e. resnet50. Besides, ATNet does not bring a significant
Figure 7: Case Study on ATNet in successful and failure cases. Results from image and ingredient channels are shown in red and green dashed boxes, respectively. Correct predictions are in boldface. (a) Prediction on “Crispy sweet & sour pork slices” made of “Chinese Parsley coriander”, “Fried flour”, and “Tenderloin slices”. Both channels make correct prediction. (b) Prediction on “Sour Beef Soup” made of “Crushed pepper”, “Sliced Fatty Beef”, “Water”, and “Enoki mushroom”. Only image channel makes the correct prediction. (c) Prediction on “Salt Green Tender” made of “Crushed pepper”, “Crushed garlic”, and “Chinese Kale”. Only ingredient channel makes correct prediction. (d) Prediction on “Sauteed Spicy Pork” made of “Black fungus”, “Streaky pork slices”, “Steamed bread”, and “Garlic leaves”. Both channels make the wrong predictions.

Table 3: Performance comparison of food recognition algorithms (in Top-1 and Top-5 accuracy (%)) on VireoFood-172 and Ingredient-101 datasets.

<table>
<thead>
<tr>
<th>Model</th>
<th>VireoFood-172</th>
<th>Ingredient-101</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-5</td>
</tr>
<tr>
<td>resnet18</td>
<td>77.1</td>
<td>93.1</td>
</tr>
<tr>
<td>resnet50</td>
<td>80.2</td>
<td>93.8</td>
</tr>
<tr>
<td>vgg16_bn</td>
<td>80.4</td>
<td>95.3</td>
</tr>
<tr>
<td>vgg19_bn</td>
<td>81.6</td>
<td>95.7</td>
</tr>
<tr>
<td>ARCH-D [7]</td>
<td>82.1</td>
<td>95.9</td>
</tr>
<tr>
<td>WRN50-2 [38]</td>
<td>82.5</td>
<td>96.1</td>
</tr>
<tr>
<td>WISer [21]</td>
<td>82.8</td>
<td>95.6</td>
</tr>
<tr>
<td>ATNet_vgg19_bn</td>
<td>85.3</td>
<td>96.5</td>
</tr>
<tr>
<td>ATNet_resnet50</td>
<td>85.0</td>
<td>96.2</td>
</tr>
<tr>
<td>ATNet_WRN50-2</td>
<td>86.1</td>
<td>96.6</td>
</tr>
<tr>
<td>ATNet_WISer</td>
<td>86.2</td>
<td>96.4</td>
</tr>
</tbody>
</table>

improvement to both their Top-5 accuracies. This is likely due to the low distinguishing power of the visual embeddings produced by the CNN encoder, which may hinder both its alignment with the semantic embedding and its mapping to the ingredient channel. We will provide an in-depth analysis in the following section, and an evidence is revealed in Figure 7(d).

5.5 Case Study

This section provides an in-depth analysis on the behaviors of ATNet_WRN50-2, as discussed in Section 5.4, in various successful and failure cases. The samples are randomly selected from the testing set of the VireoFood-172 dataset.

As observed in Figure 7(a), a perfect prediction in the image channel also leads to a correct prediction of ingredients and food class in the ingredient channel. These enable a robust decision fusion. On the other hand, when the ingredients in an image are difficult to identify, the ingredient channel may make incorrect predictions with flat values. In this case, ATNet relies on the image channel to filter the classes with incorrect ingredients, as shown in Figure 7(b). In the case that the ingredients are clearly visible while the predictions of the image channel are uncertain (See Figure 7(c)), the prediction of food class of the ingredient channel can help to depress the wrong classes. Lastly, Figure 7(d) depicts the case when the visual embedding is not representative for the correct class. This leads to failure in food recognition, and the prediction values, either from the ingredient channel or the fused prediction, are low. The above analysis reveals the behaviors of ATNet using cross-modal mapping to help food recognition in the image channel. ATNet is effective especially when the ingredients are visible in the images. Wrong predictions in ingredient channel is not harmful unless it also happens in the image channel. This demonstrates the effectiveness of ATNet on food recognition.

6 CONCLUSIONS

This paper presents a cross-modal alignment and transfer network (ATNet) under the paradigm of learning using privileged information (LUPI) to assist in the image-based food recognition. It uses food ingredients as PI and implements the similarity control and knowledge transfer of LUPI using a two-stage cross-modal interaction. The interaction for similarity control aligns the visual and semantic embeddings encoded from images and ingredients. To alleviate their intrinsic heterogeneity, we propose a partial heterogeneous transfer, which learns to align the features in both embeddings that carry information about their shared food classes, rather than those carrying information on their own styles. The interaction for knowledge transfer is achieved by mapping the visual embedding to the ingredient channel for food class prediction.

Despite the achievements of ATNet, future work can be further explored in two directions. First, stronger transfer learning techniques that better align the visual embeddings to the semantic space can significantly improve the performance. Second, a well-defined knowledge graph on food-ingredient associations can further enhance food prediction in the ingredient channel.

ACKNOWLEDGMENTS

This NEXT+ research is supported by the National Research Foundation, Prime Minister’s Office, Singapore under its IRC@SG Fund- ing Initiative; in part, by the National Research Foundation, Prime Minister’s Office, Singapore under its IDM Futures Funding Initiative; in part, by the Australian Research Council Project FL-170100117; and in part, by the 2018 Zhejiang University Academic Award for Outstanding Doctoral Candidates.
REFERENCES


