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
hanwangzhang@ntu.edu.sg
ascymiao@ntu.edu.sg
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 view of "teacher". This leads 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

ACM Reference Format:
Lei Meng, Long Chen, Xun Yang, Dacheng Tao, Hanwang Zhang, Chunyan Miao, and Tat-Seng Chua. 2019. Learning Using Privileged Information for Food Recognition. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350870

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 FoodLog[1], TADA project[2], foodAI[3], and DietLens[4].

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 detected food regions for classification [11, 14, 40]. This typically requires heavy manual efforts 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] proposes 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

MM '19, October 21–25, 2019, Nice, France
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350870

[1] http://www.foodlog.jp/en
[2] http://www.tadaproject.org/
[3] http://foodai.org/
[4] https://www.dietlens.com/

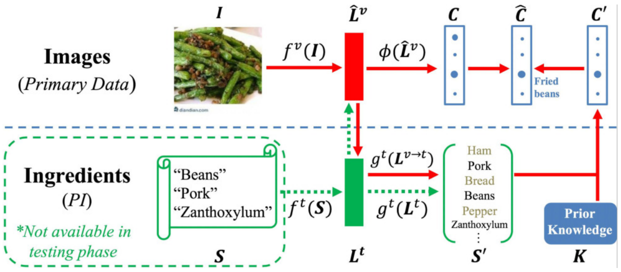
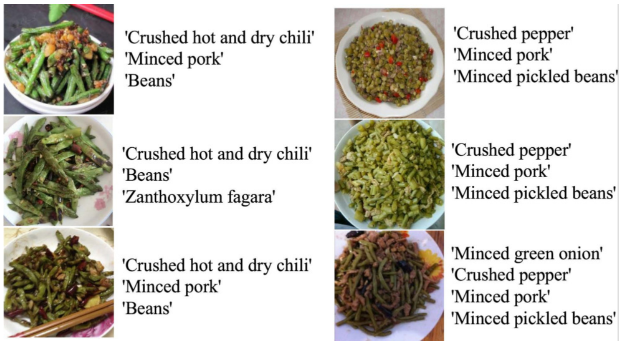


Figure 1: Illustration to ATNet for food recognition under LUPI paradigm. Information passing from images (drawn in red arrows) is allowed for both training and testing, while that from the ingredients (drawn in green dashed arrows), i.e. the privileged information (PI), is only allowed for training. In addition to the conventional pipeline $\mathbf{I} \mapsto f^{v}(\mathbf{I}) \mapsto \mathbf{C}$, ATNet builds the synchronized image and ingredient channels using autoencoders. This enables information passing between visual and semantic channels using two lines of cross-modal mappings: 1) a heterogeneous feature alignment for $f^{v}(\mathbf{I}) \mapsto \hat{\mathbf{L}}^{v} \mapsto \mathbf{C}$ and 2) a cross-modal mapping for $f^{v}(\mathbf{I}) \mapsto \mathbf{L}^{v\mapsto t} \mapsto \mathbf{S}' \mapsto \mathbf{C}'$. The final prediction $\hat{\mathbf{C}}$ is obtained by a fusion of the predictions from both views $\mathbf{C} \oplus \mathbf{C}' \mapsto \hat{\mathbf{C}}$.

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 $\mathbf{I}$ and ingredients $\mathbf{S}$ in the respective channels. This ensures that the learned visual and the semantic embeddings $\mathbf{L}^{v} = f^{v}(\mathbf{I})$ and $\mathbf{L}^{t}$ encode the information of their source data $\mathbf{I}$ and $\mathbf{S}$ as much as possible. **The first interaction** happens in a heterogeneous transfer network, where $\mathbf{L}^{v}$ and $\mathbf{L}^{t}$ are aligned to produce a regularized visual embedding $\hat{\mathbf{L}}^{v}$ using a deep transfer mapping $\theta(\mathbf{L}^{v}, \mathbf{L}^{t})$. Notably, finding a suitable $\theta(\mathbf{L}^{v}, \mathbf{L}^{t})$ is challenging due to the intrinsic heterogeneity of $\mathbf{L}^{v}$ and $\mathbf{L}^{t}$ in feature distributions. Additionally, $\theta(\mathbf{L}^{v}, \mathbf{L}^{t})$ 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 $\mathbf{L}^{v}$ and $\mathbf{L}^{t}$ 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 $\mathbf{C}$ obtained from the image channel $\hat{\mathbf{L}}^{v} \mapsto \mathbf{C}$ benefits from the semantic embeddings in the ingredient channel. **The second interaction** is enabled by a cross-modal mapping $\mathbf{L}^{v} \mapsto \mathbf{L}^{v\mapsto t}$ from visual to the semantic space. This allows the visual embedding $\mathbf{L}^{v}$ to make food prediction in the ingredient channel from the view of “teacher”. The gap between $\mathbf{S}'$ and $\mathbf{C}'$ in the mapping $\mathbf{L}^{v\mapsto t} \mapsto \mathbf{S}' \mapsto \mathbf{C}'$ 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 dependencies $P(s_q|c_j, s_p)$. The final prediction $\hat{\mathbf{C}}$ is obtained by a fusion of the predictions drawn from both the image and ingredient channels, i.e. $\mathbf{C} \oplus \mathbf{C}' \mapsto \hat{\mathbf{C}}$.



(a) Fried beans (b) Beans with minced pork

Figure 2: Illustration to the representation power of images and ingredients using dishes from two similar food classes. Images from both (a) and (b) evidently show the intra-class diversity and inter-class similarity, while ingredients can perfectly distinguish both.

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 using the ingredients of food 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]. However, visual features have limited representation power due to the diverse visual appearances of food (See Figure 2 for an instance). Ingredients have been widely used for food analysis and recipe retrieval [7, 8, 23, 24, 27], 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 non-linear 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 $\mathbf{L}^v \mapsto \mathbf{L}^{v \mapsto t}$ 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 $\mathcal{I} = \{\mathbf{I}_i | i = 1, ..., N\}$ of J classes $C = \{c_j | j = 1, ..., J\}$ and the corresponding ingredients $\mathcal{S} = \{\mathrm{S}_i | i = 1, ..., N\}$, ATNet receives data triplets $(\mathbf{I}, \mathbf{S}, y)$ in the training phase, where y is class label, and learns the pipeline of $\mathbf{I} \mapsto \mathbf{L}^v \mapsto \hat{\mathbf{L}}^v \mapsto \hat{\mathrm{C}}$ via a two-stage interaction with the ingredient channel. It has four main procedures:

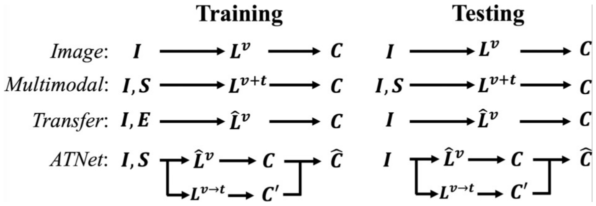

Training
Testing
Image: $I \longrightarrow L^v \longrightarrow C$
$I \longrightarrow L^v \longrightarrow C$
Multimodal: $I,S \longrightarrow L^{v+t} \longrightarrow C$
$I,S \longrightarrow L^{v+t} \longrightarrow C$
Transfer: $I,E \longrightarrow \hat{L}^v \longrightarrow C$
$I \longrightarrow \hat{L}^v \longrightarrow C$
ATNet: I,S $\hat{L}^v \longrightarrow C$ $\hat{C}$; $L^{v\to t} \longrightarrow C'$
I $\hat{L}^v \longrightarrow C$ $\hat{C}$; $L^{v\to t} \longrightarrow C'$

Figure 3: Illustration to different settings for food recognition. *Image*: Conventional image classification; *Multimodal*: Multimodal image classification; *Transfer*: Transfer learning. *ATNet*: our model that learns a two-channel mapping for food recognition. I: images, S: ingredients; E: external data; $\mathbf{L}^v$: visual embeddings; $\mathbf{L}^{v+t}$: a fusion of visual and semantic embeddings; $\hat{\mathbf{L}}^v$: regularized visual embeddings; $\mathbf{L}^{v \mapsto t}$: semantic embeddings mapped from the image channel; C, C′, and Ĉ: food class indicators.

(1) **Embedding modeling using autoencoders:** ATNet independently models the embeddings of I and S using two autoencoders: $\mathbf{I} \mapsto \mathbf{L}^v \mapsto \mathbf{I}'$, and $\mathbf{S} \mapsto \mathbf{L}^t \mapsto \mathbf{S}'$.
(2) **Heterogeneous feature alignment:** Considering the intrinsic heterogeneity of visual and semantic embeddings $\mathbf{L}^v$ and $\mathbf{L}^t$, ATNetfi nds the shared space in two steps: 1) learning two subsets of features $\dot{\mathbf{L}}^v$ and $\dot{\mathbf{L}}^t$ that carry the information describing C and minimizing the KL-divergence $KL(\dot{\mathbf{L}}^t || \dot{\mathbf{L}}^v)$, and 2) aligning $\dot{\mathbf{L}}^v$ and $\dot{\mathbf{L}}^t$ in a shared space to obtain $\dot{\mathbf{L}}^v \mapsto \hat{\mathbf{L}}^v$. Subsequently, the food class indicator C from the image channel can be obtained by $\hat{\mathbf{L}}^v \mapsto \mathrm{C}$.
(3) **Cross-modal food prediction:** Taking advantage of the autoencoder in the ingredient channel, a cross-modal mapping enables $\mathbf{L}^v \mapsto \mathbf{L}^{v \mapsto t} \mapsto \mathbf{S}'$. The food prediction $\mathbf{S}' \mapsto \mathbf{C}'$ is achieved by using the pre-acquired knowledge on food-ingredient associations, including frequencies $P(s_p | c_j)$, co-occurrences $P(s_p, s_q | c_j)$, and sequential dependencies $P(s_q | c_j, s_p)$.
(4) **Multiview prediction fusion:** Thefi nal prediction on food classes $\hat{\mathrm{C}}$ is obtained by a fusion of the indicators obtained from both the image and ingredient channels, i.e. $\mathrm{C} \oplus \mathrm{C}' \mapsto \hat{\mathrm{C}}$.

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 thefi ve 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 $\mathbf{I} \mapsto \mathbf{L}^v \mapsto \mathbf{I}'$ using a CNN encoder $f^v(.)$ and a CNN decoder $g^v(.)$. The encoder can be any state-of-the-art networks, such as VGG [28], ResNet [13], and wide residual network

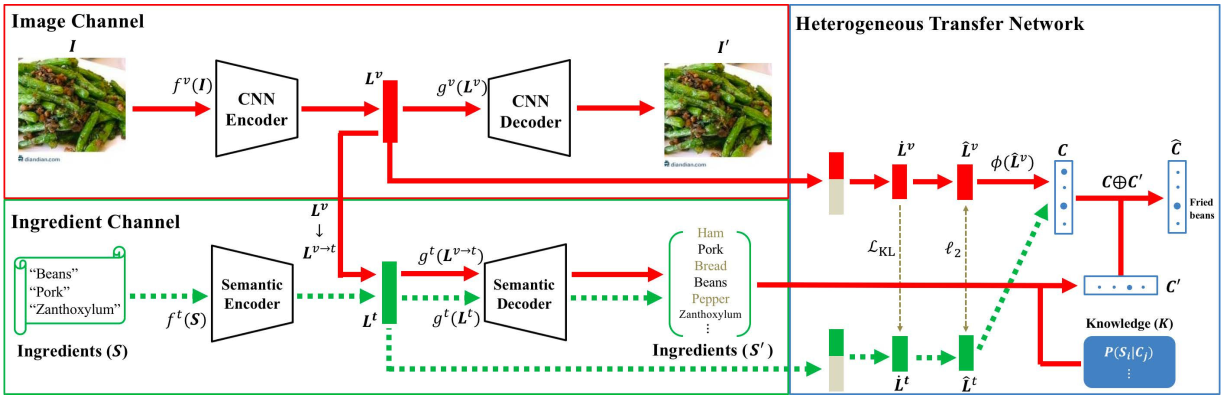


Figure 4: The framework of ATNet. The pipelines for images, i.e. the primary data, and ingredients, i.e. the privileged information, are drawn in red and green dashed arrows, respectively. In the training phase,five lines of mappings are learned: 1) $\mathrm{I} \mapsto \mathrm{L}^v \mapsto \mathrm{I}'$, 2) $\mathrm{S} \mapsto \mathrm{L}^t \mapsto \mathrm{S}'$, 3) $\mathrm{L}^v \mapsto \hat{\mathrm{L}}^v \mapsto \mathrm{C}$, 4) $\mathrm{L}^t \mapsto \hat{\mathrm{L}}^t \mapsto \mathrm{C}$, and 5) $\mathrm{L}^v \mapsto \mathrm{L}^{v\mapsto t} \mapsto \mathrm{S}'$. While in the testing phase, thefinal prediction $\hat{\mathrm{C}}$ is obtained via a fusion of the predictions from both the image and ingredient channels $\mathrm{C} \oplus \mathrm{C}' \mapsto \hat{\mathrm{C}}$, which are obtained via $\mathrm{I} \mapsto \mathrm{L}^v \mapsto \hat{\mathrm{L}}^v \mapsto \mathrm{C}$ and $\mathrm{L}^v \mapsto \mathrm{L}^{v\mapsto t} \mapsto \mathrm{S}' \mapsto \mathrm{C}'$.

(WRN) [21, 38], and the decoder should reverse the operations of the encoder. The autoencoder framework is used to learn $\mathbf{L}^v$ that preserves the information of its input $\mathbf{I}$, using a reconstruction loss:

$$\mathcal{L}_{rec}^v = ||\mathbf{I}' - \mathbf{I}||_F, \tag{1}$$

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 $\mathrm{S} \mapsto \mathbf{L}^t \mapsto \mathrm{S}'$ using a pair of semantic encoder $f^t(.)$ and decoder $g^t(.)$. 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 $f^t(.)$ and $g^t(.)$ leads to a binary "bag-of-word" representation for S, so for $\forall s_m \in \mathrm{S}$, $s_m \in \{0, 1\}$. This makes $\mathbf{L}^t \mapsto \mathbf{S}'$ a multi-label prediction task and requires a sigmoid activation function for $\mathbf{S}'$. This autoencoder is optimized using a reconstruction loss computed by the ℓ_2 norm $||.||_2$:

$$\mathcal{L}_{rec}^t = ||\mathrm{S}' - \mathrm{S}||_2. \tag{2}$$

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 $\mathrm{S} = [\mathbf{s}_1, ..., \mathbf{s}_M]$ where $\mathbf{s}_m \in \Re^{1\times r}$, 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

$$\mathbf{h}_t = \mathrm{LSTM}(\mathbf{s}_t, \mathbf{h}_{t-1}) \tag{3}$$

(2) **Self-Attention for Ingredient Embedding:** The self-attention step fuses the output hidden vectors $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_M]$ to learn a unified embedding for ingredients, defined as

$$\mathbf{L}^t = \mathbf{AH}, \tag{4}$$
$$\mathbf{A} = softmax(\mathbf{W}_2\, tanh(\mathbf{W}_1\mathbf{H})), \tag{5}$$

where $\mathbf{A} \in \Re^{1\times M}$ is a trainable attention vector that evaluates the importance of $\mathbf{h}_m$ from d aspects, $\mathbf{W}_1 \in \Re^{d\times r}$ computes a d-dimensional attention for $\mathbf{H}$, and $\mathbf{W}_2 \in \Re^{1\times d}$ fuses the d types of attention to produce $\mathbf{A}$.

(3) **Decoding for Ingredient Prediction:** Given the semantic embedding $\mathbf{L}^t$, the LSTM decoder follows the conventional image captioning procedures [9] to decode m hidden vectors $\mathbf{H}$, followed by a non-linear mapping to incrementally predict the ingredients $\mathbf{S}' = [\mathbf{s}_1', ..., \mathbf{s}_M']$, defined as

$$\mathbf{s}_t' = softmax(g_2^t(\mathbf{h}_t)), \tag{6}$$
$$\mathbf{h}_t = LSTM(g_1^t([\mathbf{L}^t, \mathbf{h}_{t-1}]), \mathbf{h}_{t-1}) \tag{7}$$

where [.] is the vector concatenation operator, and $g_1^t(.)$ and $g_2^t(.)$ are non-linear mappings.

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

$$\mathcal{L}_{rec}^t = BCE(\mathrm{S}', \mathrm{S}), \tag{8}$$

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

4.3 Heterogeneous Feature Alignment

The encoding-decoding frameworks for the image and ingredient channels independently learn the visual and semantic embeddings $\mathbf{L}^v$ and $\mathbf{L}^t$. 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 thefirst-stage cross-modal interaction of ATNet. Heterogeneous transfer [10, 29] is a commonly-used method for this problem, which aims tofind a subspace where the visual and semantic embeddings overlap, i.e.

$\theta^v(\mathbf{L}^v) = \theta^t(\mathbf{L}^t)$. 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 $\mathbf{L}^v$ and $\mathbf{L}^t$ that carry the shared information on food class. Subsequently, this method finds a shared space for them. It is based on the hypothesis that $\mathbf{L}^v$ and $\mathbf{L}^t$ are described by two types of features, where $\mathbf{L}_1^k$ ($k = \{v, t\}$) contains information on food class label and $\mathbf{L}_2^k$ contains information on their own styles [4]. Therefore, disentangling $\mathbf{L}_1^v$ and $\mathbf{L}_1^t$ from $\mathbf{L}^v$ and $\mathbf{L}^t$ for alignment alleviates the intra-modal diversity and help to discover the latent space to align the heterogeneous features sharing the information on food class.

As shown in Figure 4, the heterogeneous transfer networkfirst masks $\mathbf{L}^v$ and $\mathbf{L}^t$ (shown in grey) to allow only part of them, i.e. $\mathbf{L}_1^v$ and $\mathbf{L}_1^t$, to be aligned for food recognition. Subsequently, two pairs of non-linear mappings $\{\theta_1^k(.), \theta_2^k(.)\}$ ($k = \{v, t\}$) and a shared linear mapping $\phi(.)$ are used to implement the mappings $\mathbf{L}_1^v \mapsto \mathbf{C}$ and $\mathbf{L}_1^t \mapsto \mathbf{C}$. Three loss terms are used to learn the embeddings:

(1) Learning disentangled features $\mathbf{L}_1^k$ ($k = \{v, t\}$) is conditioned on the cross-entropy loss $CE(.)$ of both channels, defined as

$$\mathcal{L}_c^v = CE(softmax(\phi(\hat{\mathbf{L}}^v)), y), \tag{9}$$

$$\mathcal{L}_c^t = CE(softmax(\phi(\hat{\mathbf{L}}^t)), y). \tag{10}$$

(2) The loss of KL-divergence allows for a one-direction alignment of $\dot{\mathbf{L}}^v$ to $\dot{\mathbf{L}}^t$, defined as

$$\mathcal{L}_{KL} = KL(\dot{\mathbf{L}}^t || \dot{\mathbf{L}}^v). \tag{11}$$

It alleviates the cross-modal gap without breaking the distribution of the semantic embeddings.

(3) The ℓ_2 norm aligns $\hat{\mathbf{L}}^v$ and $\hat{\mathbf{L}}^t$ in a shared space, defined as

$$\mathcal{L}_{align} = ||\hat{\mathbf{L}}^v - \hat{\mathbf{L}}^t||_2. \tag{12}$$

In addition, the shared $\phi(.)$ implicitly aligns $\hat{\mathbf{L}}^v$ and $\hat{\mathbf{L}}^t$ 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. $\mathbf{L}^v \mapsto \mathbf{L}^{v \mapsto t} \mapsto \mathrm{S}'$, as shown in Figure 4. Learning such a mapping requires two loss regularizers, defined as

$$\mathcal{L}_{v \mapsto t} = ||\mathbf{L}^v - detach(\mathbf{L}^t)||_2, \tag{13}$$

$$\mathcal{L}_{v \mapsto \mathrm{S}'} = CE(\mathrm{S}', \mathrm{S}), \tag{14}$$

where $detach(.)$ removes the gradient for $\mathbf{L}^t$ from $\mathcal{L}_{v \mapsto t}$, making $\mathbf{L}^t$ from $\mathcal{L}_{v \mapsto t}$ a one-directional alignment from $\mathbf{L}^v$ to $\mathbf{L}^t$.

Note that the prediction S' reveals the image-ingredient association between the ingredients S and the image **I**, i.e. $P(s_p|\mathbf{I})$, $P(s_p, s_q|\mathbf{I})$, and $P(s_q|\mathbf{I}, s_p)$ ($P(s_q|\mathbf{I}, s_p)$ for LSTM decoder only). This enables the mapping $\mathrm{S}' \mapsto \mathrm{C}'$ by using the pre-acquired knowledge on food-ingredient association, including frequencies $P(s_p|c_j)$, co-occurrences $P(s_p, s_q|c_j)$, and sequential dependencies $P(s_q|c_j, s_p)$.

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 $P(s_p|\mathbf{I})$, the probability that **I** belongs to the j-th class of C is defined as

$$P_f(c_j|\mathbf{I}) = \frac{||\mathbf{P}(\mathrm{S}|\mathbf{I}) \wedge \mathbf{P}(\mathrm{S}|c_j)||_1}{||\mathbf{P}(\mathrm{S}|\mathbf{I}) \vee \mathbf{P}(\mathrm{S}|c_j)||_1}, \tag{15}$$

where $\mathbf{P}(\mathrm{S}|\mathbf{I}) = [P(s1|\mathbf{I}), ..., P(s_M|\mathbf{I})]$ and $\mathbf{P}(\mathrm{S}|c_j) = [P(s1|c_j), ..., P(s_M|c_j)]$ are probability distributions. $\wedge$ and $\vee$ are element-wise fuzzy AND and OR operators such that $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. $||.||_1$ is the ℓ_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, $P_{co}(c_j|\mathbf{I})$ and $P_{seq}(c_j|\mathbf{I})$ can be obtained using $P(s_p, s_q|\mathbf{I})$ and $P(s_q|\mathbf{I}, s_p)$, respectively. Therefore, the integrated indicator for food classes $P(c_j|\mathbf{I})$ inferred from ingredient predicts can be computed using a weighted fusion of the three views, defined as

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

where α, β, and γ are hyperparameters and $\alpha + \beta + \gamma = 1$. Note that if S' is predicted using neural network, $P_{seq}(c_j|\mathbf{I})$ is unavailable. In this way, the food class indicator in the ingredient channel $\mathrm{C}' = [P(c_1|\mathbf{I}), ..., P(c_J|\mathbf{I})]$ 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 $\hat{\mathrm{C}}$ using their fused predictions, defined as

$$\hat{\mathrm{C}} = \mathrm{C} \oplus \mathrm{C}', \tag{17}$$

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. $\mathcal{L}_{rec}^v$ and $\mathcal{L}_{rec}^t$, (b) losses for aligning visual and semantic embeddings, i.e. $\mathcal{L}_{KL}$ and $\mathcal{L}_{align}$, (c) losses for food classification, i.e. $\mathcal{L}_c^v$ and $\mathcal{L}_c^t$, and (d) losses for cross-modal ingredient prediction, i.e. $\mathcal{L}_{v \mapsto t}$ and $\mathcal{L}_{v \mapsto \mathrm{S}}$. 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 autoencodersfirst:** 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 $\mathbf{L}^v$ and $\mathbf{L}^t$ 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 $\mathcal{L}_{rec}^v = \mathcal{L}_{rec}^t = 1$, since ATNet uses the pretrained parameters for autoencoders. $\mathcal{L}_c^v$, $\mathcal{L}_c^t$, and $\mathcal{L}_{v \mapsto \mathrm{S}}$, i.e. the losses for the target goals, are weighted to be 10. The losses regularizing the intermediate embeddings, i.e. $\mathcal{L}_{KL}$, $\mathcal{L}_{align}$, and $\mathcal{L}_{v \mapsto t}$ 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:

(1) **vgg19_bn:** Pytorch implementation[5] for 19-layer VGG [28] with each convolutional layer followed by batch normalization.
(2) **resnet50:** Pytorch implementation[6] for 50-layer ResNet [13].
(3) **WRN50-2:** Pytorch implementation[7] for Wide Residule Networks (WRN) [38] using ResNet50 and a wide factor of 2.
(4) **WISeR:** In-house implementation for Wide-Slice Residual Networks (WISeR) [21], which adds a slide branch upon WRN50-2.

The designs in Section 4.2 are used for the ingredient channel:

(1) **NN:** 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. $\mathbf{L}^t$ and $\mathbf{L}^v$ have the same length. The four-layer decoder uses a Sigmoid activation function to map $\mathbf{L}^t$ to the ingredient indicator $\mathbf{S}$.
(2) **LSTM:** The length of the embedding vectors for $\mathbf{S}$ and the hidden vector $\mathbf{h}$ also equals to that of $\mathbf{L}^v$. d in self-attention, as defined for $\mathbf{W}_1$ and $\mathbf{W}_2$ 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 $\mathbf{L}_1^k \mapsto \dot{\mathbf{L}}^k$ and $\dot{\mathbf{L}}^k \mapsto \hat{\mathbf{L}}^k$ ($k = \{v, t\}$) are two-layer networks with each layer followed by a ReLU. The length for these intermediate embeddings is 3/8 of $\mathbf{L}^v$. The network for cross-modal mapping $\mathbf{L}^v \mapsto \mathbf{L}^{v \mapsto t}$ is an eight-layer fully-connected network, each of which has the same length to $\mathbf{L}^v$ and is followed by a ReLU. Regarding cross-modal food prediction, Top-5 and -10 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 $L_1^v \mapsto$ C; Align$_{NN(LSTM)}$: Using $\mathbf{L}^t$ learned by $NN(LSTM)$ for alignment; F: Using multiview prediction fusion.

Components	Base CNN Models			
	vgg19_bn	resnet50	WRN50-2	WISeR
Origin	81.6	80.2	82.5	82.8
+ D	81.7	80.2	82.4	82.6
+ D +Align$_{NN}$	83.7	83.6	85.3	85.2
+ D +Align$_{LSTM}$	84.2	83.9	85.4	85.6
+ D +Align$_{NN}$ + F	84.9	84.8	85.9	**86.2**
+ D +Align$_{LSTM}$ + F	**85.3**	**85.0**	**86.1**	**86.2**

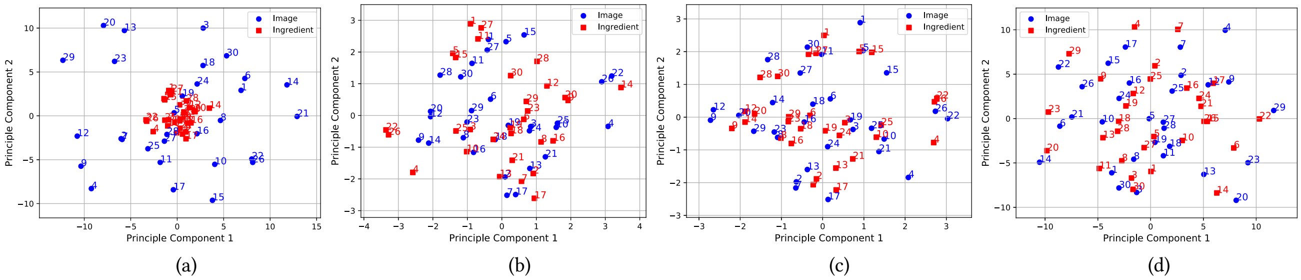
5.3 Ablation Study

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 dependency 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 from autoencoders, i.e. $\mathbf{L}^v$ and $\mathbf{L}^t$, 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 $\mathbf{L}^v$ and $\mathbf{L}^t$ 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] https://github.com/pytorch/vision/blob/master/torchvision/models/vgg.py
[6] https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py
[7] https://github.com/szagoruyko/wide-residual-networks/tree/master/pretrained



(a) (b) (c) (d)

Figure 5: Visualization for the visual and semantic embeddings of 30 randomly-selected testing samples. (a) $\mathbf{L}^v$ and $\mathbf{L}^t$ obtained by autoencoders; (b) $\dot{\mathbf{L}}^v$ and $\dot{\mathbf{L}}^t$ obtained by thefi rst step of ATNet's partial heterogeneous transfer; (c) $\hat{\mathbf{L}}^v$ and $\hat{\mathbf{L}}^t$ obtained by the second step of ATNet's partial heterogeneous transfer; (d) $\hat{\mathbf{L}}^v$ and $\hat{\mathbf{L}}^t$ obtained by aligning $\mathbf{L}^v$ and $\mathbf{L}^t$ using deep Coral.

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. $\mathbf{C}$ and $\mathbf{C}'$, leads to a further improvement upon heterogeneous feature alignment. Note that the prediction $\mathbf{C}'$ is also a fusion of three indicators obtained using different types of food-ingredient associations, i.e. frequencies $P(s_p|c_j)$, co-occurrences $P(s_p, s_q|c_j)$, and sequential dependences $P(s_q|c_j, s_p)$, 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.

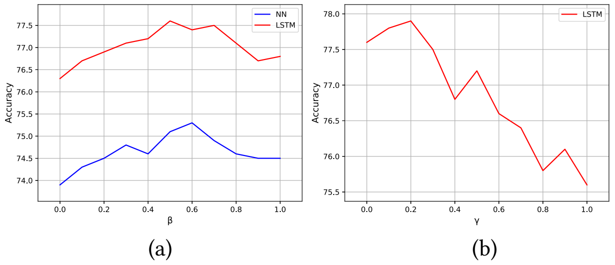
Effects of Weights for Ingredient-to-Food Indicators

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 $P(s_q|c_j, s_p)$ do not apply to the NN-based autoencoder, the importance of $P(s_p|c_j)$ and $P(s_p, s_q|c_j)$ are evaluatedfi rst. As observed in Figure 6(a), using $P(s_p, s_q|c_j)$ solely achieves better performance than using $P(s_p|c_j)$ 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 $P(s_q|c_j, s_p)$ is revealed in Figure 6(b). As observed, the best performance is achieved at $\gamma = 0.2$, 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. $\mathbf{C}$ and $\mathbf{C}'$, 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 $\mathbf{C}$ and $\mathbf{C}'$ 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 $\mathbf{C}$ and $\mathbf{C}'$ 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.



(a) (b)

Figure 6: Illustration to the effects of weights α, β, and γ for $\mathbf{S}' \mapsto \mathbf{C}'$. (a) $\alpha + \beta = 1$ and $\gamma = 0$. (b) $\alpha = \beta = 0.5$, and $\mathbf{C}'$ is normalized by $\mathbf{C}'/(\alpha + \beta + \gamma)$.

Table 2: Performance (in Top-1 accuracy (%)) of ATNet using different fusion operations. $+, \times$, max, and min are element-wise operators on C and C′. $S(.)$ means performing *softmax* to indicators before the operator.

C	C′	+	×	max	min	$S(+)$	$S(\times)$	S(max)	S(min)
84.2	77.9	83.8	84.7	84.4	83.5	84.9	85.1	85.0	**85.3**

5.4 Performance Comparison

This sections 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 vgg19_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

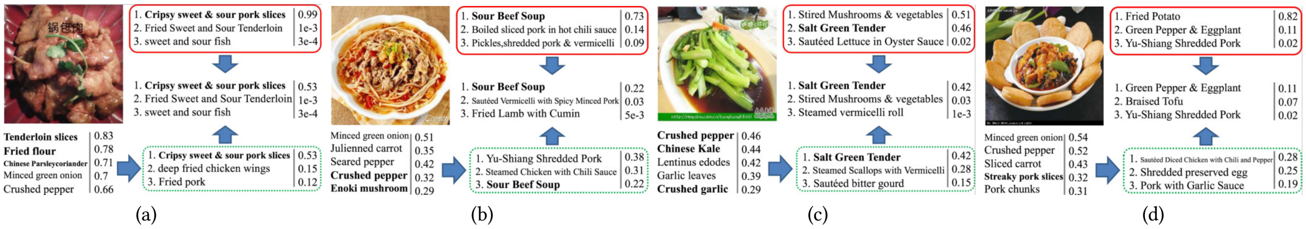

锅包肉
1. Cripsy sweet & sour pork slices 0.99
2. Fried Sweet and Sour Tenderloin 1e-3
3. sweet and sour fish 3e-4
1. Cripsy sweet & sour pork slices 0.53
2. Fried Sweet and Sour Tenderloin 1e-3
3. sweet and sour fish 3e-4
Tenderloin slices 0.83
Fried flour 0.78
Chinese Parsleycoriander 0.71
Minced green onion 0.7
Crushed pepper 0.66
1. Cripsy sweet & sour pork slices 0.53
2. deep fried chicken wings 0.15
3. Fried pork 0.12
(a)
1. Sour Beef Soup 0.73
2. Boiled sliced pork in hot chili sauce 0.14
3. Pickles,shredded pork & vermicelli 0.09
1. Sour Beef Soup 0.22
2. Sautéed Vermicelli with Spicy Minced Pork 0.03
3. Fried Lamb with Cumin 5e-3
Minced green onion 0.51
Julienned carrot 0.35
Seared pepper 0.42
Crushed pepper 0.32
Enoki mushroom 0.29
1. Yu-Shiang Shredded Pork 0.38
2. Steamed Chicken with Chili Sauce 0.31
3. Sour Beef Soup 0.22
(b)
1. Stired Mushrooms & vegetables 0.51
2. Salt Green Tender 0.46
3. Sautéed Lettuce in Oyster Sauce 0.02
1. Salt Green Tender 0.42
2. Stired Mushrooms & vegetables 0.03
3. Steamed vermicelli roll 1e-3
Crushed pepper 0.46
Chinese Kale 0.44
Lentinus edodes 0.42
Garlic leaves 0.39
Crushed garlic 0.29
1. Salt Green Tender 0.42
2. Steamed Scallops with Vermicelli 0.28
3. Sautéed bitter gourd 0.15
(c)
1. Fried Potato 0.82
2. Green Pepper & Eggplant 0.11
3. Yu-Shiang Shredded Pork 0.02
1. Green Pepper & Eggplant 0.11
2. Braised Tofu 0.07
3. Yu-Shiang Shredded Pork 0.02
Minced green onion 0.54
Crushed pepper 0.52
Sliced carrot 0.43
Streaky pork slices 0.32
Pork chunks 0.31
1. Sautéed Diced Chicken with Chili and Pepper 0.28
2. Shredded preserved egg 0.25
3. Pork with Garlic Sauce 0.19
(d)

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 “Cripsy sweet & sour pork slices” made of “Chinese Parsleycoriander”, “Friedfl our”, 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.

Model	VireoFood-172		Ingredient-101	
	Top-1	Top-5	Top-1	Top-5
resnet18	77.1	93.1	79.6	92.9
resnet50	80.2	93.8	82.1	94.3
vgg16_bn	80.4	95.3	80.7	93.4
vgg19_bn	81.6	95.7	81.3	93.7
ARCH-D [7]	82.1	95.9	83.7	96.2
WRN50-2 [38]	82.5	96.1	84.6	96.5
WISeR [21]	82.8	95.6	85.1	96.6
ATNet$_{\text{vgg19\_bn}}$	85.3	96.5	86.4	**96.8**
ATNet$_{\text{resnet50}}$	85.0	96.2	86.7	96.6
ATNet$_{\text{WRN50-2}}$	86.1	**96.6**	**87.3**	96.7
ATNet$_{\text{WISeR}}$	**86.2**	96.4	87.1	96.5

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$_{\text{WRN50-2}}$, 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 withfl at values. In this case, ATNet relies on the image channel tofi lter 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 Funding 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

[1] Ziad Ahmad, Marc Bosch, Nitin Khanna, Deborah A Kerr, Carol J Boushey, Fengqing Zhu, and Edward J Delp. 2016. A Mobile Food Record For Integrated Dietary Assessment. In *Proceedings of International Workshop on Multimedia Assisted Dietary Management*. 53–62.

[2] Kiyoharu Aizawa, Yuto Maruyama, He Li, and Chamin Morikawa. 2013. Food balance estimation by using personal dietary tendencies in a multimedia food log. *IEEE Transactions on multimedia* 15, 8 (2013), 2176–2185.

[3] Yongsheng An, Yu Cao, Jingjing Chen, Chong-Wah Ngo, Jia Jia, Huanbo Luan, and Tat-Seng Chua. 2017. PIC2DISH: A Customized Cooking Assistant System. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 1269–1273.

[4] Ershad Banijamali, Amir-Hossein Karimi, Alexander Wong, and Ali Ghodsi. 2017. JADE: Joint Autoencoders for Dis-Entanglement. *NIPS Workshop on Learning Disentangled Representations: from Perception to Control* (2017).

[5] 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.

[6] 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.

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

[8] 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 *ACM Multimedia Conference on Multimedia Conference*. ACM, 1020–1028.

[9] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6298–6306.

[10] Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. 2018. Unsupervised cross-modal alignment of speech and text embedding spaces. In *Advances in Neural Information Processing Systems*. 7365–7375.

[11] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. 2017. Food recognition: a new dataset, experiments, and results. *IEEE journal of biomedical and health informatics* 21, 3 (2017), 588–598.

[12] Oscar Day and Taghi M Khoshgoftaar. 2017. A survey on heterogeneous transfer learning. *Journal of Big Data* 4, 1 (2017), 29.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*. 770–778.

[14] Ye He, Chang Xu, Nitin Khanna, Carol J Boushey, and Edward J Delp. 2013. Food image analysis: Segmentation, identification and weight estimation. In *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

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

[16] Richang Hong, Lei Li, Junjie Cai, Dapeng Tao, Meng Wang, and Qi Tian. 2017. Coherent semantic-visual indexing for large-scale image retrieval in the cloud. *IEEE Transactions on Image Processing* 26, 9 (2017), 4128–4138.

[17] Hokuto Kagaya, Kiyoharu Aizawa, and Makoto Ogawa. 2014. Food detection and recognition using convolutional neural network. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 1085–1088.

[18] Deborah A Kerr et al. 2016. The connecting health and technology study: a 6-month randomized controlled trial to improve nutrition behaviours using a mobile food record and text messaging support in young adults. *International Journal of Behavioral Nutrition and Physical Activity* 13, 1 (2016).

[19] Chang Liu, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, Ma Yunsheng, Songqing Chen, and Peng Hou. 2018. A new deep learning-based food recognition system for dietary assessment on an edge computing service infrastructure. *IEEE Transactions on Services Computing* 11, 2 (2018), 249–261.

[20] Yong Luo, Tongliang Liu, Yonggang Wen, and Dacheng Tao. 2018. Online Heterogeneous Transfer Metric Learning. In *IJCAI*. 2525–2531.

[21] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. 2018. Wide-slice residual networks for food recognition. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 567–576.

[22] Michele Merler, Hui Wu, Rosario Uceda-Sosa, Quoc-Bao Nguyen, and John R Smith. 2016. Snap, Eat, RepEat: a food recognition engine for dietary logging. In *Proceedings of the 2nd international workshop on multimedia assisted dietary management*. ACM, 31–40.

[23] Weiqing Min, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz. 2017. Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration. *IEEE Transactions on Multimedia* 19, 5 (2017), 1100–1113.

[24] Weiqing Min, Shuqiang Jiang, Shuhui Wang, Jitao Sang, and Shuhuan Mei. 2017. A delicious recipe analysis framework for exploring multi-modal recipes with various attributes. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 402–410.

[25] Zhaoyan 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.

[26] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. 2016. Information bottleneck learning using privileged information for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1496–1505.

[27] 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.

[28] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations* (2015), 1–14.

[29] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*. Springer, 443–450.

[30] Vladimir Vapnik and Rauf Izmailov. 2015. Learning using privileged information: similarity control and knowledge transfer. *Journal of machine learning research* 16, 2023-2049 (2015), 2.

[31] Vladimir Vapnik and Akshay Vashist. 2009. A new learning paradigm: Learning using privileged information. *Neural networks* 22, 5-6 (2009), 544–557.

[32] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. 2015. Recipe recognition with large multimodal food dataset. In *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 1–6.

[33] Yu Wang, Ye He, Fengqing Zhu, Carol Boushey, and Edward Delp. 2015. The use of temporal information in food image analysis. In *ICIAP*. 317–325.

[34] Hao Yang, Joey Tianyi Zhou, Jianfei Cai, and Yew Soon Ong. 2017. Miml-fcn+: Multi-instance multi-label learning via fully convolutional networks with privileged information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1577–1585.

[35] Xun Yang, Meng Wang, Richang Hong, Qi Tian, and Yong Rui. 2017. Enhancing person re-identification in a self-trained subspace. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13, 3 (2017), 27.

[36] Xun Yang, Meng Wang, and Dacheng Tao. 2017. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing* 27, 2 (2017), 791–805.

[37] Xun Yang, Peicheng Zhou, and Meng Wang. 2018. Person reidentification via structural deep metric learning. *IEEE Transactions on Neural Networks and Learning Systems* 99 (2018), 1–12.

[38] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *In British machine vision conference* (2016), 1–12.

[39] Wenbo Zhang, Hongbing Ji, Guisheng Liao, and Yongquan Zhang. 2015. A novel extreme learning machine using privileged information. *Neurocomputing* 168 (2015), 823–828.

[40] Fengqing Zhu et al. 2015. Multiple hypotheses image segmentation and classification with application to dietary assessment. *IEEE journal of biomedical and health informatics* 19, 1 (2015), 377–388.