CMRDF: A Real-Time Food Alerting System Based on Multimodal Data

Pengfei Zhou, Cong Bai, *Member, IEEE*, Jie Xia, and Shengyong Chen, *Senior Member, IEEE*

***Abstract*—A healthy diet is a major concern for everyone, especially for those with specific diseases, such as diabetes. Meanwhile, with the rapid development of new technologies, it is feasible for us to detect the deep latent relationship between daily meals and wellbeing. Advanced Internet of Things devices, such as smart bracelets and wearable cameras, make it possible for people to know how food is related to health at any time. However, it is still arduous for individuals to memorize all the health information and utilize them to regulate their diet. To deal with such problems, we propose a novel system called cross-modal retrieval on diabetogenic food (CMRDF) which realizes a real-time dietary notice based on multimodal data captured from wearable devices. In this system, we propose a new graph-based cross-modal retrieval method named graph correlation analysis with ranking loss that finds the latent information in multimodal data. We use graph convolutional networks to dig the deep latent information in modalities and represent the data in finer granularity. It uses visual and physiological information to estimate whether the food that a user tries to obtain is diabetogenic or not, and feeds back the reasons in detail. Extensive experiments on the MSCOCO data set and the new proposed multimodal diabetogenic food database real-life diabetogenic show that the proposed cross-modal retrieval method outperforms state-of-the-art methods and CMRDF can achieve reliable results on preventing diabetic patients from inappropriate food.**

***Index Terms*—Cross-modal retrieval, graph convolutional network (GCN), multimodal data, personal health, wearable devices.**

I. Introduction

FOOD is one of the most important necessities in people's lives, which has a great impact on health and wellbeing. Nowadays, new agricultural technologies ensure us with sufficient varieties of food to be selected. However, the number of diabetics is also growing rapidly [1], [2]. It is necessary for people to have a healthier diet in order to keep away from diabetes [3]. For those who are already suffering from diabetes, the issues that are related to food may be more serious. Diabetic patients can find it difficult to arrange and monitor their daily meals [4], [5] since there are so many dietary taboos. It is impossible for individuals to memorize all medical knowledge of food, and the frequency of diet makes it bothersome to keep looking for information on their smartphones all the time [6], [7]. Therefore, it is worthwhile to develop a real-time alerting system that can track the characters of food and push the medical advice to users to help them establish the balance between dietary health and eased life [8].

On the other hand, with the rapid development of the Internet of Things (IoT), a great number of new technologies that can sense, record, and analyze different resources of data are produced to bring more convenience to people [9]–[12]. Especially with the help of various sensors and wearable devices, physiological information data are ready to be captured and combined with more personal information in the form of digital diaries. Individuals are now allowed to use digital technology to track the details of their own daily activities, such as diet, exercising, and sleeping [13]–[17]. Such kinds of data collected by various devices have been summarized as lifelogs [18]–[22]. Those lifelogs include all kinds of data generated in the interaction between individuals and smart devices, such as visual images and biometric data, as shown in Fig. 1. These various formats of data have been utilized to help personal lives in real-world scenarios [23]–[28].

There are already some research efforts dedicated to using these data in personalized food-related area [29]. For example, Nag *et al.* [30] presented a novel food recommendation engine, which analyzes different sources of data to give users advice about where and what to eat outdoors. Yang *et al.* [31] developed a visual system called PlateClick that could feed food composition and nutrient details back to users. Dehais *et al.* [32] used the images to estimate the volume of food. Jiang *et al.* [33] recognized food images using the multiview deep feature. Salvador *et al.* [34] used cross-modal embeddings to retrieve recipes via food images. Min *et al.* [35] retrieved recipes of food based on ingredients with the help of the deep belief network. It is also noticeable that Kawano and Yanai [36] developed a real-time system that can achieve real-time food recognition on smartphones.

Furthermore, there are also several researchers who are working on the field of physiological data trying to help people with a healthy diet. Pouladzadeh *et al.* [37] built a recognizing system that can help users calculate the energy intake of

Manuscript received December 1, 2019; revised March 15, 2020; accepted May 17, 2020. Date of publication May 20, 2020; date of current version April 25, 2022. This work was supported in part by the National Key Research and Development Program under Grant 2018YFB1305200, and in part by the Natural Science Foundation of China under Grant U1908210 and Grant 61976192. *(Corresponding author: Cong Bai.)*

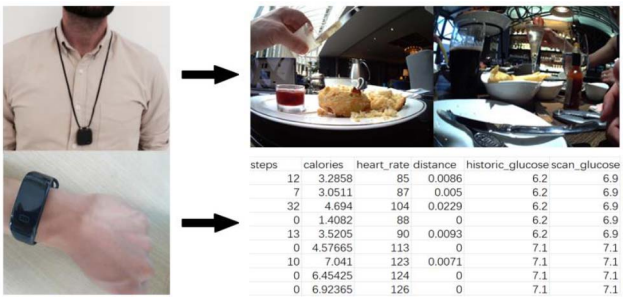
Pengfei Zhou is with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: pengfeizhou@zjut.edu.cn).

Cong Bai is with the College of Computer Science and Technology and Key Laboratory of Visual Media Intelligent Processing Technology of Zhejiang Province, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: congbai@zjut.edu.cn).

Jie Xia is with the College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: jiexia@zjut.edu.cn).

Shengyong Chen is with the College of Computer Engineering, Tianjin University of Technology, Tianjin 300384, China (e-mail: sy@ieee.org).

Digital Object Identifier 10.1109/JIOT.2020.2996009



steps	calories	heart_rate	distance	historic_glucose	scan_glucose
12	3.2858	85	0.0086	6.2	6.9
7	3.0511	87	0.005	6.2	6.9
32	4.694	104	0.0229	6.2	6.9
0	1.4082	88	0	6.2	6.9
13	3.5205	90	0.0093	6.2	6.9
0	4.57665	113	0	7.1	7.1
10	7.041	123	0.0071	7.1	7.1
0	6.45425	124	0	7.1	7.1
0	6.92365	126	0	7.1	7.1

Fig. 1. Lifelog data. Images are captured from wearable camera constantly and the instant biodata are captured from the smart bracelet.

their daily food. Lee *et al.* [38] proposed a system based on a textual database. This system helps diabetic patients to learn what kind of food is safe for them. Anthimopoulos *et al.* [5] used image recognition to tell patients about the categories of food. Rivas *et al.* [39] analyzed the electrocardiograph signal to detect allergy reactions after children taking food in real time. Tusor *et al.* [40] managed to establish a general system that could help individuals make decisions about choosing food. It takes different health conditions (diabetes, celiac disease, food allergies, etc.) into consideration based on medical knowledge.

However, few methods use various novel sensors, such as mobile blood glucose meter [41] and wearable sensor of blood pressure [42]. Many of these sensors have been embedded in digital devices, such as smart watches and smart bracelets. All of these sensors can be used not only for basic self-care but also for developing a practical food monitor system which is effective enough to give personalized advice about what to eat.

Besides the insufficiency of sensors, current frameworks also have these issues discussed as follows.

First, the current food analyzing and recommendation systems are not real time enough, which means that they did not take the time emergency into consideration. It can be grueling for users to pull out their smartphones, take pictures of food, and wait for some time to get the feedback that retrieved from the Internet every time they try to enjoy a great meal. Still, it could be dangerous for diabetic patients to eat unclear food without notification outdoors, in the situation that they cannot remember the taboos of food all the time. Hence, it is necessary to develop an automatic monitoring system that could alert people whenever they are trying to eat any inappropriate food. The retrieval techniques in this system have to be quick responsive and reliable enough.

Second, it is unpractical to put the previous systems into healthcare use because of insufficient data coverage. They did not include enough information to serve the special groups. One reason is that these systems retrieve information based on data in a single modality. For example, the vision-based system basically uses data in the image modality and dietary assistants only work on the textual environment. The data set of a single modality is limited by the number of labels (it is unrealistic to label all data captured from the real world by reason of its variety) and it is impossible to cover different vital information via these data. All of these factors lead to deficient retrieval results. So, the cross-modal retrieval method based on multimodal data is supposed to be employed to improve the practicability of the retrieval system.

Third, although several systems are able to match data from the modalities of images and texts, the effectiveness of them is limited. Once the cross-modal retrieval system is expanded to employ multimodal data, the data set could be extremely complex, which means that it is difficult to find out the causal relationship and internal rules in such various kinds of data. It makes the retrieval results unsatisfactory, and sometimes misleading. Therefore, how to find out the latent feature in each modality and the relationship between these modalities is another technical problem that needs to be solved.

Finally, the existing food-related databases are not suitable for evaluating the real-world food alerting system. It is not only because of insufficient accuracy but also the vital information missed. The current food data sets cannot provide enough diabetes-related information. Moreover, these data sets have images and texts only, whereas biological data of food composition are needed for diabetogenic alerting in particular. Therefore, an effective data set has to be established to boost the development of the diabetogenic food alerting system.

In this article, we propose a new system, cross-modal retrieval on diabetogenic food (CMRDF), which can not only retrieve multimodal data efficiently but also help people learn more knowledge about how diet is related to illness. This smart alerting system, which is developed for diabetes specifically, is able to send the user a timely alert whenever the user tries to eat inappropriate food, and the user is able to get the details of reason retrieved by the system.

To achieve the above-mentioned requirements, the contributions of this article can be summarized as follows.

1) We introduce the data of other modalities into our system. To deal with the real-time and real-world issues, the input of instant lifelog data is used to detect food in the system proposed. Also, we implement the food composition data to supply the coverage of output. Such a supplement could improve the performance of cross-modal retrieval and thus food alerting. The retrieved results of alerting notification and pathogenic details are directly produced and then sent to users automatically.
2) We propose a new cross-modal retrieval approach called graph correlation analysis with ranking loss (GCARL) to narrow the semantic gap between modalities caused by the introduction of the multimodal data, which could achieve accurate cross-modal retrieval results. It uses graph convolutional networks (GCNs) to learn the graph representation in the food image modality and description text modality, respectively. The GCN implemented in this system can help to handle the latent connection of features in each modality and find the deep correlation between pathogenic factors and biological data (such as: doughnut-> high-oil food -> calorie spike -> diabetogenic -> cannot be eaten).
3) We made efforts to establish a new food database called real-life diabetogenic (RLD) to measure the performance of diabetogenic food alerting in the real-world environment. It is built based on various authoritative databases,

such as USDA Food Composition Database[1] and abundant lifelog data set [19]. In addition to food images and nutrition descriptions, it also contains the healthcare-related biodata. To the best of our knowledge, it is the first database that includes so much multimodal data.

4) We evaluate the proposed CMRDF on various databases including RLD from different perspectives. We compare the results of our methods with several state-of-the-art approaches. The result indicates that our approach has better performance in cross-modal information retrieval as well as in diabetogenic food alerting. In the meantime, this system can be used to help analyze the correlation between diabetogenic food composition and individual physical condition.

The remainder of this article is organized as follows. Related work is discussed in Section II, followed by the details of the proposed system given in Section III. We describe the establishment of the new data set in Section IV. Then, we present our experiment and discuss the evaluation results in Section V. Eventually, we give the conclusion and perspectives in Section VI.

II. Related Work

A. Cross-Modal Retrieval

Cross-modal retrieval is an effective solution to query in one modality and return the retrieval results in other different modalities, which is widely used to match the image and text data. For example, in a traditional image to text cross-modal retrieval task, the model retrieves a most similar text as the output, whether the text is covered in the training set or not. In such an end-to-end cross-modal retrieval task, the effective cross-modal retrieval method has to retrieve data bridging the semantic gap between modalities. So the cross-modal retrieval is developed to use multimodal data to retrieve directly with no dependency on labels. Moreover, it can improve the efficiency of retrieval by using complementary data of different modalities. In recent years, cross-modal retrieval is becoming one hotspot in multimedia processing research area.

In the field of cross-modal retrieval, canonical correlation analysis (CCA) is first proposed by Hotelling [46], and Rasiwasia *et al.* [47] applied it to the cross-modal retrieval task to match texts and images, which has attracted extensive attention. Those large numbers of researches related to CCA are known as cross-modal retrieval technology based on subspace [48]–[50]. In addition, there are some other exploring attempts in cross-modal retrieval tasks, such as hashing-based cross-modal retrieval [51]–[53] and topic model methods [54], [55]. Some researchers combine cross-modal retrieval with deep learning. Ngiam *et al.* [56] proposed a deep network-based cross-modal learning method. Andrew *et al.* [57] proposed the deep CCA, which maximizes the common expression correlation by learning complex nonlinear projection through a multilayer deep neural network.

While their methods combine simply the cross-modal mathematical algorithm with deep neural networks, they cannot guarantee the accuracy of its retrieval. So, there are still numerous problems in specific applications. Furthermore, most of the previous studies, such as dimension reduction by selecting several features of images, do not extract the deep latent information in multimodal data as much as possible to satisfy the demands of retrieval. How to use different kinds of related multimodal data as efficient as possible is a technical problem to be solved. So, in our proposal, we adopt the latest advanced neural network technology based on graph learning, which not only extracts deep features more effectively but also mines the latent relevance of features in modalities.

[1]https://ndb.nal.usda.gov/ndb/search/list

B. Graph Learning

The graph neural network (GNN) is first proposed by Gori *et al.* [58] and modified by Scarselli *et al.* [59] to improve the performance of graph learning. Then, GCNs are proposed by Kipf and Welling [60], which is a landmark contribution in the graph learning field. Then, GCN is identified as a promising technology in classical computer vision researching areas, such as object detection [61], image captioning [62], semantic segmentation [63], etc. The core idea of graph learning is to learn a mapping function, through which the node in a graph can aggregate its own feature and its neighbor's feature to generate a new representation of the node so that GCN can extend the convolution operation from the traditional data to graph data. The GCN can understand the semantic relationship between objects both in the visual scene and the natural language. Applying GCN to end-to-end subspace cross-modal retrieval is breaking the ice. For example, the GCN-based hashing method for cross-modal retrieval is first proposed in [80].

The proposed CMRDF in this article adopts GCN to discover the deep semantic correlations in each modality. To be more specific, GCNs are used as feature extractors in our proposal due to its strength in handling the latent relevance of features, which could be extremely complex in real-world data. These associations are represented by graphs to help match the data in different modalities.

C. Lifelog Analysis

Wearable devices provide more personal lifelog data such as the images and biodata captured during daily lives. These data can be applied to personal health information analysis. However, the modalities of lifelog data are various, which makes the overall analysis of multimodal data challenging. To deal with such real-world issues, there are many lifelog-related competitions held all over the world every year, such as the lifelog image retrieval contest organized by ImageCLEF [64]. As many of the researchers have made breakthroughs and contributions constantly, the practical value of this research direction is also proved. Several lifelog retrieval systems have been relatively modified and put into use, such as LIFER [65], an interactive life record retrieval system developed by ImageCLEF [64], which is based on MyLifeBits [66] life record database. It can effectively interact with users according to different requirements. LifeXplore [67], a Lifelog

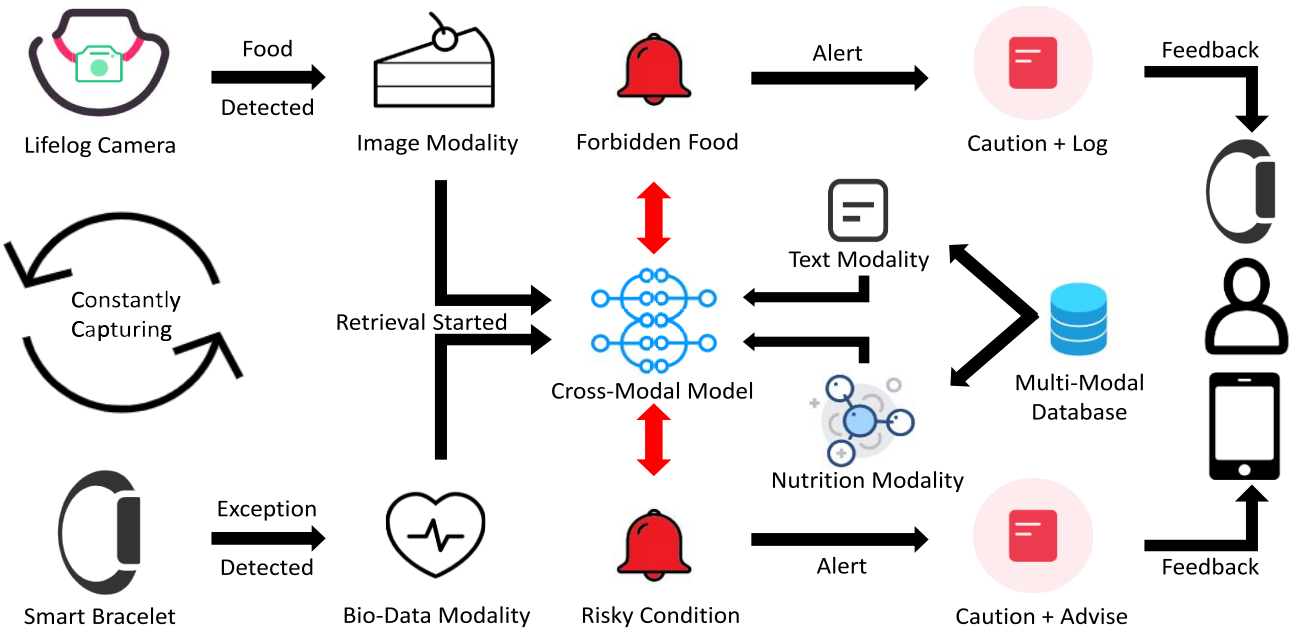


Fig. 2. Overview of the CMRDF system. Once the intake of diabetogenic food or the change of physiological condition is detected, the user will be warned by the wearable bracelet. Meanwhile, the key information, such as food analysis and matters which need to be aware of, will be fed back to the user via smartphone.

video search system developed by Bernd, has been successfully applied in several multimedia retrieval competitions and achieves outstanding results.

There are also some researches on proposing different strategies and models to explore the inherent law of lifelog. Nguyen *et al.* [68] analyzed the deep correlation of lifelog biodata with Adaboost, Rand Tree, and correlation matrix, which makes the implementation of lifelog easier. At the same time, there are also researchers who contributed various models to provide deep learning solutions for the retrieval methods of personal lifelog data. For instance, Kavallieratou *et al.* [69] studied the feature selection using a convolutional neural network (CNN).

Unlike the previous system, we choose GCN to deal with the lifelog metadata set. In our proposal, we replace the traditional CNN with GCN to learn the deep latent information in data.

III. Food Alerting by Cross-Modal Retrieval for Diabetogenic Food

In this section, we give the details of the proposed system and the core algorithms.

A. Overview

The purpose of CMRDF is to develop a collaboration mechanism based on the daily needs of diabetic patients and implement precise real-time retrieval on multimodal lifelog data. As shown in Fig. 2, corresponding to the user's wearable devices [13] (wearable camera and smart bracelet) capturing data every minute, the real-time images obtained from the visual sensor are extracted into semantic features constantly. Once the food is detected in images by a pretrained classifier, this system gets ready to activate a cross-modal retrieval sector, which is able to bring a timely alert. As described in [19], the biometric sensors in smart bracelets can capture data on a second-by-second basis, whereas wearable cameras can capture five images per minute. Since the sampling interval of lifelog devices (wearable camera and smart bracelet) is adjustable, the collection of data can be conducted less than 1 min. So data collection is timely enough to start the alerting mechanism before the food is consumed.

The goal of cross-modal retrieval is to retrieve the diabetogenic information via food images directly by a cross-modal retrieval model, which is named as GCARL. It uses the fusion of real-time images and biodata as its multimodal input. As the model is trained in a multimodal data set RLD, which is established through a variety of scientific methods, it could match the different modalities of information, such as food description and nutrition composition effectively. If there are any inappropriate factors in retrieval results, the system starts the alerting mechanism. Through the vibration and flicker of his/her wearable bracelet, the user realizes that he/she is facing the risks of taking improper food. Meanwhile, the user is able to find detailed reasons and learn to eat a healthier diet by the information pushed on their smartphones. This information, including food description (text modality) and nutrition composition (biodata modality), is retrieved by GCARL and fed back to the user. In the meantime, if the system finds the abnormal data captured by the bracelet, such as the quick increase in blood glucose, it also alerts. The key feedback can help the user know the actual situation. Furthermore, the health logs are made at the end of the day, which is generated by the retrieval results based on the user's diet, steps, blood glucose management, and other conditions in the daytime. So the user can

make the balance of scientific diet and easy life with the help of CMRDF. Furthermore, the user is able to retrieve desired information via the smartphone at any time.

B. Graph Correlation Analysis With Ranking Loss

As the proposed system is a food-related alerting system, it needs cross-modal information retrieval with high accuracy. The diversification of results should also be taken into account, which means that the results of retrieval should cover different relevant information as much as possible. Furthermore, our system analyzes the data automatically so as to give alerts or feedback, which should have the ability to retrieve useful information as relevant as possible.

The above-mentioned requirements demand a high-precision cross-modal retrieval method. To improve the accuracy in the specific retrieval task, we propose a novel approach of cross-modal retrieval based on graph learning called GCARL. We design it to narrow the semantic gap between the data in different modalities and meet the demand for accuracy. When GCARL is used for cross-modal retrieval, the input of the network is the pair of image–texts (images and texts are paired with two sets of biodata when food alerting). Then, GCN is implemented to process the branches of the feature node in each modality of data. The fusion of biodata is another procedure developed for food alerting in particular (this procedure can be removed when GCARL is used for a two-modal retrieval). The trace norm with ranking loss increases the relations of two networks between different modalities.

The general idea of our cross-modal method is presented as follows.

Assuming there is a training set $X = \{x_i = [i_i, t_i, c_i], i = 1, 2, \ldots, n\}$, the images set is $I = [i_1, i_2 \cdots i_n] \in \mathbf{R}^{d_i \times n}$, and the texts set is $T = [t_1, t_2 \cdots t_n] \in \mathbf{R}^{d_t \times n}$. Let $I - T$ represents a pair of image and text. There is a common label set to mark the $I - T$ pairs into different classes, which is $L = [l_1, l_2 \cdots l_n] \in \mathbf{R}^{l \times n}$. So, $\{I_i, T_i\}_{i=1}^{n}$ represents the semantic relation of images and texts. Let the test set be $Y = \{y_j = [i_j, t_j], j = 1, 2, \ldots, m\}$. The target of cross-modal retrieval is to learn a similarity measurement $F(\cdot)$. For a specific query $i_q \in I$ or $t_q \in T$, it returns the most similar sample in the other modality: $t_q = \min_j F(i_q - t_j)$ or $i_q = \min_j F(t_q - i_j)$.

First, we adopt GCN to learn the graph embedding of images and texts. We combine the GCN with the trace norm objective and ranking loss to solve the semantic gap between modalities.

To be more specific, we use two new proposed feature extractors, Res-GCN and BERT-GCN to perform the graph embedding in the modality of image and text, respectively. Then, we extract the positive semantic vectors in the graph to generate a feature matrix. After that, the matrices of different modalities are put into a spatial pyramid pooling (SPP) layer to be pooled into vectors, which are input into the fully connected (FC) layer and the trace norm with the ranking loss for cross-modal mapping.

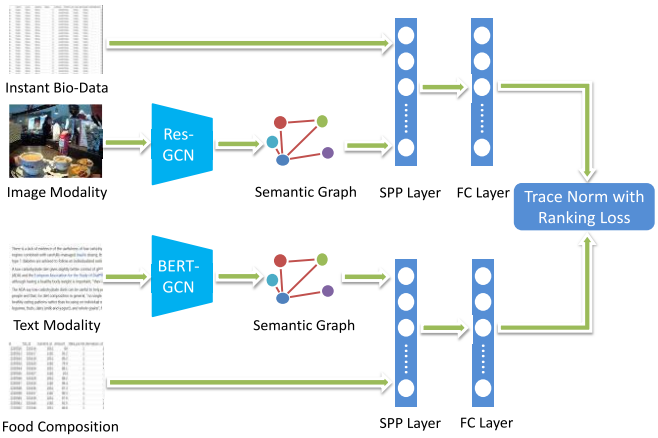


Fig. 3. GCARL. The GCNs are used as feature extractor and SPP is used to solve the input size problem caused by a variant number of semantic vectors. The trace norm with ranking loss is used for matching the data from different modalities.

Finally, we get an end-to-end cross-modal retrieval model GCARL which could be used to analyze the relationship of data cross different modalities directly. The architecture of the GCARL is presented in Fig. 3.

C. Graph-Based Feature Extractor

Given an image, the graph conventional network can detect, identify objects, and predict the semantic relationships between concepts [81]. In the text modality, natural language can also be parsed into semantic graphs [82], where each word represents a concept. It is a promising method of aligning a given text description with an image bridging the semantic gaps.

When the algorithm of the GCN runs at the node level, the pooling module which connects to the graph convolutional layer can transfer the graph into a high-level substructure roughly. This architectural design can be used to extract the various levels of representation of the graph. In our approach, the pretrained GCN is implemented to extract semantic features on account of its good performance and the ability of making use of the latent relevance in various modalities. We build our two GCNs based on the proposal of [60] and [71] to improve the accuracy of retrieval. To be more specific, we build a ResNet [78]-based GCN called Res-GCN and a BERT [74]-based GCN called BERT-GCN. One major difference between them is the composition of the backbone. The former is based on the backbone of ResNet-101 and the latter is based on BERT. The following processing procedures are developed based on image processing and text processing, respectively. In Res-GCN, we employ ResNet-101 and global max pooling to obtain the image-level feature as the input for GCN. As for BERT-GCN, it uses the processing factors, such as tokenizer and transformer to map the text into text-level feature directly. However, the outputs of two GCNs are both semantic features, which could be used for classification as well as for cross-modal retrieval. Thus, if it is not specified whether it is Res-GCN or Bert-GCN, we define the GCN as a generalized case which is composed of the backbone model

and graph convolutional layers in the illustrations below. We pretrain them in a multilabel image data set and multilabel text data set, respectively to get the practical semantic feature extractors.

1) Overview: In GCARL, two GCNs are pretrained on multilabel classification tasks and then used as the feature extractor to improve the performance of retrieval. Specifically, we use a graph to construct the interdependencies between semantic features in each modality, which is an effective way to capture the topological structure in a semantic space. To capture and explore the deep latent correlations between semantic features, we represent each node of the graph as the word embedding of a concept produced by the backbone model. The GCN is supposed to map these word embeddings into a set of interdependent semantic features, which can be directly applied to cross-modal retrieval. The entire cross-modal retrieval model is supposed to be trained together at last to fine-tune the multimodal networks. We introduce the theory of GCNs used in this system as follows.

The goal of GCN in GCARL is to learn a function $H(\cdot)$ on graph representation. Let the inputs be described as feature $A^i \in \mathbf{R}^{n\times d}$ and the corresponding correlation matrix $C \in \mathbf{R}^{n\times n}$, where n represents the number of nodes and d represents the dimension of node features. After the procedure of graph convolution, we then get the semantic matrix $S \in \mathbf{R}^{m\times e}$ through the mapping $S = H(G(\widehat{C} \cdot A^i \cdot W^i))$, where m is the number of new nodes and e is the dimensionality of semantic vectors.

In the image modality, we employ ResNet-101 [78] which is pretrained on ImageNet [79] to learn the initial feature representation $A = x$ of an image. Thus, we input image I in size of 448×448 and obtain $2048 \times 14 \times 14$ feature maps from the conv5 layer. Then, we apply global max pooling to obtain the image-level feature $x = F_{\mathrm{GMP}}(F_{\mathrm{CNN}}(I))$. Similarly, in the text modality, we use a pretrained BERT [74] to learn encoder representations from texts and obtain a word embedding feature $A = t$.

However, the representation methods of graph nodes and edges are expected to be determined. We design the concepts of graph nodes corresponding to the labels in a multilabel data set and thus construct the correlation matrix.

2) Graph Preparing: The connection of edges represents a structural relation between nodes. The graph convolution focuses on the topological structure and properties of the network, which means that GCN works by propagating information between nodes based on a correlation matrix. Therefore, how to establish the correlation matrix C is the key problem. The correlation matrix C is predetermined by analyzing the association between labels in multilabel data sets. We determine the correlation between labels by mining the co-occurrence pattern of labels in the data set.

We use the form of conditional probability to describe the correlation of labels. In order to construct the correlation matrix, we first calculate the number of times when the label pairs occur in the whole data set and get the matrix $M \in \mathbf{R}^{C\times C}$. To be more specific, C is the number of label classes and $M_{i,j}$ is the number of simultaneous occurrences of the pair of L_i and L_j. By using the label co-occurrence matrix, we then can obtain the conditional probability matrix $P_i = M_i/T_i$, where T_i denotes the occurrence times of L_i in the training set and $P(L_i, L_j)$ represents the probability of label i and label j appears together in data set.

However, this simple correlation has the drawback of long-tail distribution, which could bring the noise of the rare co-occurrence. To solve this problem, we binarize the correlation P via a threshold τ to filter noisy edges, where $C_{i,j}$ in correlation matrix C can be presented as

$$C_{i,j} = \begin{cases} 1, & \text{if } P_{i,j} = \tau \\ 0, & \text{if } P_{i,j} < \tau. \end{cases} \tag{1}$$

3) Graph Convolution: The idea of graph convolution is to exchange the information between the nodes that have strong associations, which is able to make the connection degree between samples with different attributes as sparse as possible and make the connection degree between samples with same attributes as close as possible so that the relationship between vectors can share the associated information as well as maintain the structural relationship between nodes in the graph structure.

Let the mapping function of GCN be $G(\cdot)$, the inputs be described as feature $A^i \in \mathbf{R}^{n\times d}$, and the corresponding correlation matrix be $C \in \mathbf{R}^{n\times n}$ (where n represents the number of nodes and d represents the dimension of the node feature). The graph convolutional layer updates the node features as $A^{i+1} \in \mathbf{R}^{n\times d'}$ via the convolution operation $A^{i+1} = G(A^i, C)$. To be explicit, it can be rewritten into $A^{i+1} = G(\widehat{C} \cdot A^i \cdot W^i)$, where $W^i \in \mathbf{R}^{d\times d'}$ is a transformation matrix, $\widehat{C}$ is the normalized version of the correlation matrix, and $G(\cdot)$ represents a nonlinear operation operated by LeakyReLU [45] in our model. Therefore, the procedure of graph convolution in our proposal can be regarded as the process of graph embedding. Then, we are able to model the complex interrelationships between feature nodes using graph conventional layers. After the graph convolution, it is able to map graph embeddings into relevant semantic feature vectors.

4) Res-GCN: Since the pretrained backbone model placed in GCN for image processing is ResNet-101 [78], we name the GCN-based feature extractor in the image modality as Res-GCN.

In this image representation learning task, the final learning result is a set of semantic feature vectors $V = [v_1, v_2 \cdots v_n]$.

To do this, we learn the graph embedding-dependent concepts from label representation through a stacked GCN with two graph convolutional layers first. For the first graph convolutional layer, the input is $W \in R^{C\times D}$, where D represents the dimensionality of the image representation. For the second layer, the output is a matrix $Z \in R^{C\times d}$, where d is the dimension of the label word feature in graph embedding.

The whole GCN module is pretrained in a multilabel classification task. By applying the pretrained classifier to the image representation, we can get the predicted confidence score $\hat{y} = Wx$. Let the ground truth be $y \in Y$, where $y_i = \{0, 1\}$ denotes whether label i appears in the samples or not, and $\sigma(\cdot)$ is the sigmoid function.

The loss function used in this network is defined as

$$L = \sum_{c=1}^{C} y^c \log\left(\sigma\left(\hat{y}^c\right)\right) + \left(1 - y^c\right)\log\left(1-\sigma\left(\hat{y}^c\right)\right). \quad (2)$$

Therefore, in order to obtain better attribute prediction results, it is also necessary to further reduce the error between the prediction attributes and the real attributes of the training set samples. The error value is described by the Euclidean distance

$$\Delta = \sum_{i=1}^{m} \|f_i - y_i\|^2 = \|F - Y\|^2 \quad (3)$$

where the matrix $Y = [y_1, y_2 \cdots y_n]$ is the semantic matrix of the training set samples. It is able to determine that the label matrix can be constructed through the correlation matrix of the attribute graph, $Y = 2C-1$. If the constraint item is added to the function of the AP algorithm, the following can be obtained:

$$\hat{F} = \underset{F}{\operatorname{argmin}}\{L + \lambda\Delta\}. \quad (4)$$

In the above equation, λ is a positive parameter used to get the balance between graph embedding and class prediction. By solving the above equation, we can obtain a set of feature vectors that can be used for the mapping function.

Then, we perform the graph-based mapping function and obtain

$$S = \{s_i\}_{i=1}^{E} \quad (5)$$

where E represents the number of positive vectors.

5) BERT-GCN: A GCN-based feature extractor is also developed for the text modality and named as BERT-GCN due to the integration of BERT and GCN.

We use the feature approach proposed in [74] to get the text representation as feature vectors of $t \in R^{C\times R}$, where R is the size of the final hidden layer of the transformer in BERT and is set as 1024 in our model. This text representation is used as the input of GCN in BERT-GCN.

Then, we prepare graph embedding through graph convolution, which is the same as Res-GCN, and then, we employ the graph-based mapping function and get $Q = \{t_i\}_{i=1}^{C}$, where C represents the number of positive vectors. We establish the module of GCN using the same method as mentioned above. The output of BERT-GCN is the matrix $Q \in R^{C\times e}$, where e is the dimension of the label word embedding.

The result of graph embedding in the text modality is a set of semantic feature vectors $Z = [z_1, z_2 \cdots z_n]$.

The details of graph-based mapping are presented as follows.

6) Graph-Based Mapping: Taking the text modality as an example, we use the pretrained GCN which is trained on the multilabel classification task using the loss function in (2) and obtain the predicted confidence score $\hat{j} = Bt$ from t. Let the ground truth be $j \in J$, where $y_j = \{0, 1\}$ denotes whether label j appears in the samples or not.

As the goal of graph-based mapping is to map feature vectors to the semantic space, the problem of solving graph regularized class predictors can be transformed to the problem of mapping a sample eigenspace X^T to the graph embedded in the semantic space X. Assuming that the mapping matrix corresponding to the sample is B, the corresponding mapping process is as follows: $X = X^T B$.

At the same time, like the role of a positive parameter λ, η is the parameter used to control the constraint of punishments for graph embedding. B is presented as

$$\begin{aligned} B &= \left(XL_C X^T + \lambda X^T X + \eta I\right)^{-1}(\lambda XY) \\ &= \left(XL_C X^T + \lambda X^T X + \eta I\right)^{-1}(\lambda X(2C-1)). \end{aligned} \quad (6)$$

L_C is the Laplace matrix which is symmetric and semipositive definite. Therefore, the solution of the above loss function can be converted into the solution of a regularized least square method problem, and the partial derivative of the projection matrix can be solved. The solution process is as follows:

$$B^T XL_H X^T B + \lambda\left(X^T X - XY\right) + \eta B = 0. \quad (7)$$

For the input sample z with unknown attributes, its attribute value can be obtained by projecting it from the feature space to the attribute space through projection matrix B

$$P = \operatorname{sign}\left(z^T B\right) \quad (8)$$

$$\operatorname{sign}(x) = \begin{cases} 1, & \text{if } x = \tau \\ 0, & \text{if } x < \tau \end{cases} \quad (9)$$

where τ is a threshold set artificially, and we set $\tau = 0.4$ by finding that it is the optimal value for our system. Then, we use P to extract semantic features.

The final semantic feature matrix can be $\{z_i\}_{i=1}^{N}$, where $P_i = 1$ and $N \leq C$.

Finally, the GCN could integrate the pivotal features in graphs and output these related vectors as the semantic feature matrix.

D. Cross-Modal Matching

1) Spatial Pyramid Pooling: As we refer the GCNs pretrained on multilabel classification tasks as feature extractors, we basically select the semantic feature vectors with respect to a threshold to form the feature matrices. So, the dimensionalities of matrices obtained from the extractor could be uncertain, which means that the number of key semantic vectors can be various for different images or texts. It is obvious that the matrices with uncertain dimensions cannot be input to the FC layer directly, as the FC layer needs the input of fixed size. To deal with this problem, we choose the SPP model, which is proposed by He *et al.* [72] to map the various dimensionalities of multimodal data into the same size. The implement of SPP not only aligns the matrices in various dimensionality effectively but also helps the procedure of dimensionality reduction in cross-modal retrieval.

As shown in Fig. 4, there are two graphs with a different number of vectors generated by GCN in the image modality. We use SPP here to carry out pooling operations on each feature matrix. For each SPP, the largest bin represents the partition of the original matrix, the second one divides the matrix into four partitions, and the last one divides the feature matrix into 16 parts. Then, we apply the global max pooling

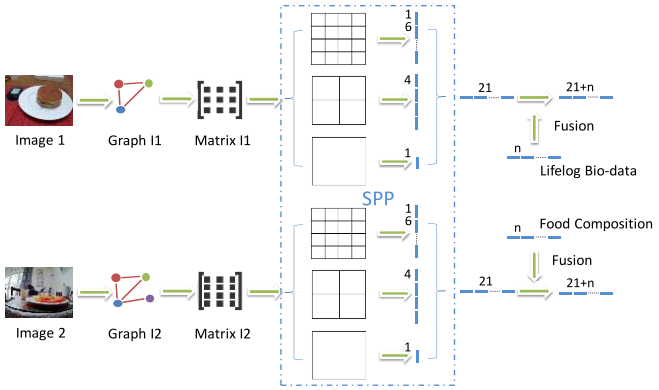


Fig. 4. Illustration of SPP. The graph refers to the semantic features output by GCNs. SPP carries out pooling operations on the semantic matrices and output feature vectors in the dimension of 21, then feature vector and biodata are fused on each modality, respectively.

on each partition and obtain a feature vector of dimension in $16 + 4 + 1 = 21$ per matrix. So, the SPP solves the problem of different feature matrix sizes in each modality at last.

2) Trace Norm With Ranking Loss: To deal with the real-world diabetogenic food problem, we must expand the coverage of data applied in the system. It means that we use the multimodal data in both the image and text branches in order to take advantage of the instant biodata and food composition data. We make the fusion of data in the image and text modalities, and the dimensions of vectors become $21 + m$ and $21 + n$, respectively. This change brings the problem that not only the dimension difference reoccurs but also the semantic gap of input and output expands.

To handle such a problem, we propose the trace norm with ranking loss inspired by the subspace learning in CCA. We map the data into the same subspace to reduce the gap between different modalities.

We use the trace norm objective to transform the multidimensional x and y into same-dimensional x' and y' by linear transformation. Then, the correlation is analyzed by a correlation coefficient. Therefore, the trace norm objective is looking for two linear transformations corresponding to two groups of variables ω_x and ω_y (equal to the dimension of x and y, respectively), so that correlation coefficient between the two combination variables after the linear transformation can be maximized. The algorithm process is presented as follows.

Assuming the two sets of random variables have N samples, linear transformation is conducted on all N samples and two sets of data obtained

$$s_x w_x = \left(w_x^T x_1, \ldots, w_x^T x_N\right) \tag{10}$$

$$s_y w_y = \left(w_y^T x_1, \ldots, w_y^T x_N\right). \tag{11}$$

The maximum correlation is calculated between the two sets of data, which can be expressed as the following equation:

$$\rho = \max_{w_x, w_y} \operatorname{corr}\left(S_x w_x, S_y W_y\right) = \max_{W_X, W_Y} \frac{\left\langle S_x W_x, S_y w_y \right\rangle}{||S_x w_x|| \left|\left|S_y W_y\right|\right|}. \tag{12}$$

After that, we can obtain the following formula according to the detailed solution in [70]:

$$w_y = \frac{C_{yy}^{-1} C_{yx} w_x}{\lambda} \tag{13}$$

$$C_{xy} C_{yy}^{-1} C_{yx} w_x = \lambda^2 C_{xx} w_x. \tag{14}$$

Because the covariance matrices C_{xx} and C_{yy} are symmetric and positively definite, a complete Cholesky decomposition can be performed

$$C_{xx} = R_{xx} \cdot R'_{xx}. \tag{15}$$

Let $\mu_x = R'_{xx} \cdot W_x$.

Substitute C_{xx} in (15) into (14), and we can get

$$R_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} R_{xx}^{-1} u_x = \mu^2 u_x. \tag{16}$$

This is an eigenvalue problem, we can get the eigenvalue u_x and final eigenvector W_x by solving the function.

So that we can solve $s_x w_x$ and $s_y w_y$ by the above results, and got $\rho = \operatorname{corr}(X, Y)$ at last. We can use ρ to map data from different modalities to the embedding space, which is called the subspace.

For paired input samples, the characteristics learned by two FC layers should be as similar as possible, which can be realized through a similarity loss L. So, a loss function that is best relevant to subspace embedding projection is needed for backpropagation. The ranking loss presented in [84] is well suited for nearest neighbor retrieval in the embedded space, as it analyzes the optimal embedding weights by ranking correlated trace norm embedding projections. So, we promote the pairwise ranking loss based on the cosine distance in our approach, and we place it between the trace norm objective layer and the optimization target.

Let us optimize a pairwise ranking loss

$$L_{\text{rank}} = \sum_x \sum_k \max\{0, \alpha - S(x, y) + S(x, y_k)\} \tag{17}$$

where x is the embedded sample of the first modality, y is the correlated embedded sample of the second modality, y_k is the mismatched sample of the second modality, and α defines the margin of the loss function.

The cosine distance is used as the scoring function in ranking loss

$$S(x, y) = \cos(x, y). \tag{18}$$

By optimizing the loss, making the mapping matrix change in the ideal direction and mapping the characteristic data of different modalities into the same subspace, the cosine similarity measurement can be used to get the correlation between different modalities and reduce the difference of features in multimodal. We implement the pairwise ranking loss to coordinate the trace norm objective and improve the retrieval performance of our cross-modal retrieval model. Finally, the cross-modal feature matching is realized.

IV. Real-Life Diabetogenic Database

In this section, we explain the procedures for establishing a new database, the RLD database. We process data by data mining, filtering, and clustering to establish a cross-modal food data set for better evaluating the food alerting system. The pipeline of this article is presented in Fig. 5. The database is formed into multimodal pairs at last, while each pair contains an instant food image paired with a line of instant lifelog biodata and a textual food description paired with a line of food composition biodata.

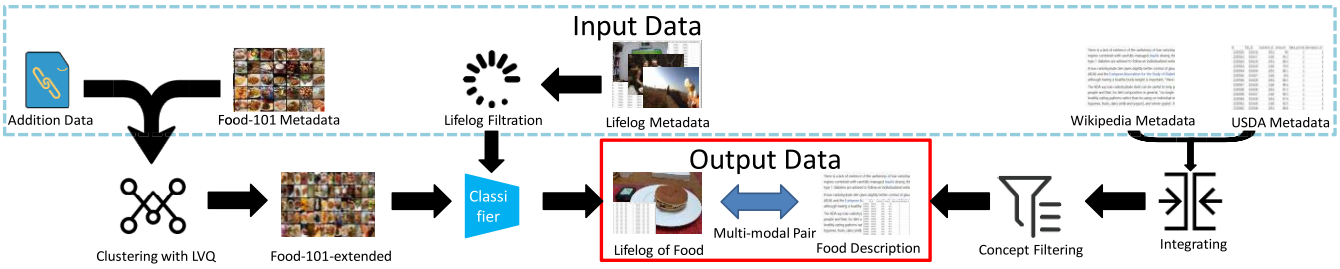


Fig. 5. Pipeline of establishing the RLD database. The input of metadata is encircled by the blue dashed outline. We then process the data with several methods, and build the new data set in the form of multimodal pairs, which are encircled by the solid red frame.

A. Food-101 Processing

Food-101 [43] is a data set of food images. It includes 101 types of western food, and each type has about 1000 food images. We choose the Food-101 as one of our basic data sets with respect to its huge amounts of images and the abundant categories of food. However, it has these drawbacks.

First, although Food-101 has 101 000 images, it is not big enough to cover all features of the food item since we want to apply this system to a real-world retrieval task. If we want to train a classifier that could recognize the real-world food images as much as possible, the amount of its images needs to be extended.

Second, the images in this database contain not only food images but also other kinds of images, for example, the images of chairs or sticks, as this data set is downloaded from a social network site[2] and classified roughly by random forests. Therefore, the data set has not been cleaned carefully and there are also many images divided into wrong classes, which means that it is mixed with many noises. These noises must be eliminated so as to improve the performance of training.

Third, the coverage of its categories is too narrow to detect all images shown in lifelog. It must have the ability to cover the forbidden food list as much as possible. So, the categories of Food-101 must be extended.

To solve the first two problems, we replenish the images with google API. Then, we implement a practical cluster method, which is learning vector quantization (LVQ) to clean the data. We extend the Food-101 [43] with multilabels to solve the last problem. Details are explained in the following.

1) Wiping Off Noise: In our framework, we choose a cluster to get rid of noises in the metadata. The LVQ algorithm shown in Algorithm 1 is used for clustering. We extract the feature vectors by the pretrained ResNet-101 [78] and then use the cluster to calculate the distance between images and the class sample point. Eventually, we give the right label to each picture and remove the pictures whose feature distance is too far from the sample point. The main steps of the algorithm include initialize the prototype vector, iterative optimization, and update the prototype vector. The details of the algorithm are introduced follows: we assume that a set of prototype vectors $\{p_1, p_2, \ldots, p_q\}$ is initialized by selecting a sample labeled t_q from positive samples first.

2) Food-101 Extending: To extend the original Food-101 data set, we merge it with the data crawled from the Internet previously at first. Then, we implement the LVQ cluster to clean the data and filter the noises. Then, we carefully put these 101 classes into more categories by analyzing the category relationship of food. For example, an image of filet steak can be labeled by several classes, such as animal protein and meat. All of these categories are selected from the USDA database to ensure the authority of classification. We choose the classes in USDA because it is also one of the metadata sets we use to get the nutrition composition data of food to verify whether a food is proper for diabetes exactly. Besides, these data are the key information that we are supposed to feedback to users.

The extended food data set, named Food-101-Extended, is able to ensure that the least improper foods could steer by this alerting system. It includes ten highest categories, 40 middle classes, and 101 detailed classes at last.

Algorithm 1 LVQ Algorithm

Repeat
  Pick a random sample
  Calculate the Euclidean distance from the sample to each prototype vector
  Find the shortest distance
  If $y_j = t_i$ (same class)
    $p' = p_j + a(x_j - p_j)$ (reduce the distance so that p become closer to the sample point)
  Else
    $p' = p_j - a(x_j - p_j)$ (increase the distance so that p become farther from the sample point)
  $p_j = p'$(update)
Until the condition to end is met.

B. Lifelog Filtration

The lifelog data are collected all day around by several users with no given limitation [19], so the quality of these data cannot be guaranteed. Thus, the metadata needs to be preprocessed.

First, there are many futile images, such as blur images or images that are covered with black entirely. The existence of these data would influence the effectiveness of the whole pipeline. We apply three filtering metrics over the data set to filter them. Two of them filter the blur images and the other one filter images that are covered by any objects. Then, we use

[2]http://www.foodspotting.com/

the concept filter to extract the food pictures with biodata in the lifelog metadata set.

1) Blur Filtering: The first filter is a Laplacian filter (3×3 kernel) calculating the blur as the variance of the convolution result. After filtering, it is necessary to implement a more refined metric to improve the quality of the data set. Then, a fast Fourier transform is applied to images. Once this step is completed, the average value in the transformed image is obtained and then scaled according to the size of the images to compensate for the tearing effect. The average value is then used as a threshold to filter the images, in which the image with a larger value represents the focused image and that of with a lower value represents the blurred image.

2) Coverage Filtering: To detect if an image is covered by something or facing the ceiling or wall, we use detectors with a maximum connected area calculator to calculate the proportion of subjects in the image. Then, we remove the images that have more than 90% of the major subject area. The method used is described as follows.

Step 1: Converting the images into grayscale images.
Step 2: Converting grayscale images into binary images.
Step 3: Converting the binary images into matrices.
Step 4: Finding the largest pattern in matrices and calculating the proportion of it.
Step 5: Removing the images according to the result of matrix calculation.

3) Concept Filtering: With the contribution of our previous methods on lifelog preprocessing in ImageCLEFlifelog2019 competition [22], we filter the metadata into a high-precision food data set.

This idea is implemented by two pretrained classifiers to achieve the target. We use a pretrained classifier to extract the concepts of images first, and then, we maintain images related to several topics and remove the others. We arrange a two-classes training set by selecting samples from Food-101-Extended and MSCOCO [44]. After that, we use the pretrained ResNet-101 [78] to fine-tune a two-class classifier on this two-class training set. Finally, we use this fine-tuned classifier to detect whether an image is a food image, and if it is not, then we delete it.

In the above procedures, all of the biodata matched the specific images are filtered together. Finally, we get a set of food images paired with instant biodata.

C. Real-Life Diabetogenic Database

The RLD database is the new data set that we build based on various metadata sets, which are Lifelog [19], Wikipedia, USDA, and Food-101-Extended. After the procedures of processing, the RLD has 4500 multimodal pairs. Each multimodal pair contains a food image (filtered from the lifelog metadata set) paired with a line of instant lifelog biodata (such as blood glucose) and a textual description of food (obtained from Wikipedia) paired with a line of food composition biodata (such as fat content), as shown by the example of our database presented in Fig. 6.

1) Datamining the Diabetic Information: We analyze several authoritative diabetic healthcare websites and get

Fig. 6. Examples of RLD. The bar on the bottom presents the instant biodata that paired with the image. And the bar on the top right shows the food composition biodata paired with the text.

suggestions from medical experts to acquire key information on diabetic maintenance. The key information is summarized into a csv file as the metadata of our database.

We also preprocess the public data set USDA to integrate the useful data into our database. We reserve the glucose, fat, carbohydrate, and calories of foods in USDA because these indicators are proximate factors that can cause or aggravate diabetes.

2) Concept Filtering: We choose a wide variety of paragraphs in Wikipedia and use it as the metadata set of the text modality in our system. Since the data must be related to food and diabetic illness, we also train a concept extractor based on the pretrained BERT [74].

Then, we maintain useful texts which are related to diabetogenic food topics and filter all other texts. We double-check the texts to ensure the quality of data. Afterward, we get a number of paragraphs that are related to food and diabetes. We label these texts with the chosen classes in USDA and Food-101-Extended. This corpus includes food descriptions and multicategory labels. We also use this food-description data set of the text modality to pretrain BERT-GCN.

Finally, we integrate the texts with the food nutrition composition data to get the pairs of two description modalities (food description texts and food composition biodata). Then, we affiliate the two description modalities with two lifelog modalities by matching the picture to text to shape the final form of the RLD. We call it image–text pair informally since the data set has three modalities in fact.

After all procedures above, the four kinds of data in RLD are prepared completely.

V. Experiment

To evaluate the capability of the proposed system CMRDF in the cross-modal diabetogenic food alerting task, we first test the reliability of GCN-based feature extractors on multimodal data, Food-101-Extended data set, and food description data set, respectively. Then, we use the MSCOCO [44] data set to test the performance of the proposed cross-modal retrieval

TABLE I
COMPARATION OF DIFFERENT IMAGE FEATURE EXTRACTION METHODS ON THE FOOD-101-EXTENDED DATA SET

Methods	mAP	CP	CR	CF1
Inception [76]	76.3	79.6	63.4	70.6
SRN [75]	81.4	86.1	68.2	76.1
ResNet-101 [78]	82.9	85.4	69.3	76.5
Res-GCN	**87.5**	**91.1**	**77.6**	**83.8**

TABLE II
COMPARATION OF TEXT FEATURE EXTRACTION METHODS ON THE FOOD-DESCRIPTION DATA SET

Methods	CP	CR	CF1
Word2Vec [73]	49.15	37.62	42.62
BERT [74]	96.75	92.15	92.63
BERT-GCN	**98.38**	**94.12**	**96.20**

approach GCARL. Finally, we evaluate CMRDF on our new proposed data set RLD with the retrieval results. The details of the result are discussed in this section.

A. Evaluation Metrics

The average per-class precision (CP), average per-class recall (CR), and average per-class F1 (CF1) following the settings of [75] are used for evaluation. We also report the mean average precision (mAP). The definition of mAP is presented as follows.

The average accuracy (AP) of a single topic is the average of the accuracy of each related retrieval result. The AP is defined as

$$AP = \frac{1}{L}\sum_{i=1}^{R} \mathrm{Pre}(i) \cdot \mathrm{Rel}(i) \tag{19}$$

where L is the number of relevant texts in the retrieval set, and $\mathrm{Pre}(i)$ is the percentage of relevant texts in the top i retrieved text. In the order of prediction, $\mathrm{Rel}(i)$ is an index function. If the item in the prediction ranking j is relevant to query, then $\mathrm{Rel}(i) = 1$, otherwise $\mathrm{Rel}(i) = 0$. R is the number of retrieved texts to be examined. The mAP is the average of the AP of all categories, which shows not only the precision of results but also the coverage of relevant results.

Except specified, we refer to all rank in evaluation metrics, such as mAP@all and CR@all.

B. Feature Extractor Evaluation

In this experiment, we select 87 000 images, 12 000 images, and 12 000 images as the training set, validation set, and test set, respectively. We train the Res-GCN on the Food-101-Extended data set for classification tasks, then we use the pretrained GCN as the feature extractor in GCARL on the RDL data set for the cross-modal retrieval task. We present the results compared with other classic multilabel classification models on the Food-101-Extended data set in Table I.

TABLE III
PERFORMANCE COMPARISON IN TERMS OF MAP@R(%) ON THE MSCOCO DATA SET

Methods	R=50			R=All		
	T q I	I q T	average	T q I	I q T	average
PL-ranking [83]	43.2	40.2	41.7	33.7	34.1	33.9
LDSP [84]	50.1	44.3	47.2	34.7	35.3	35.0
MNiL [77]	53.4	46.4	49.9	35.6	38.8	37.2
GCARL-without RL	55.5	47.2	51.3	36.1	39.0	37.5
GCARL-with RL	**56.9**	**48.3**	**52.1**	**36.4**	**39.8**	**38.1**

From this table, we can see that Res-GCN achieves the best performance in all kinds of evaluation metrics.

We pretrain BERT-GCN in the food-description data set, and then, we evaluate the features extracted by BERT-GCN and compare it with other classical feature extraction methods, such as Word2Vec [73] and BERT [74] on the text classification task.

Since it is a classification task in the text modality, we only use CP, CR, and CF1 as evaluation metrics. All results are presented in Table II. The results show that BERT-GCN gets appreciable performance.

From Tables I and II, we can conclude that Res-GCN and BERT-GCN outperform other methods in classification tasks, thus we can say that Res-GCN and BERT-GCN are promising feature extractor for image and text, respectively.

C. Cross-Modal Retrieval Evaluation

MSCOCO [44] is selected to evaluate the proposed GCARL for large-scale cross-modal retrieval. Due to the limitation of data in MSCOCO, the modalities of this experiment only involve image and text. We split train–validation–test set following the setting of [77]. We also calculate mAP by dividing the image–text pairs that share at least one semantic label into the same category on the MSCOCO data set.

We compare our proposed GCARL model with three state-of-the-art cross-modal retrieval methods: 1) PL-ranking [84]; 2) LDSP [83]; and 3) MNiL [77]. The results are shown in Table III, where the *T q I* means use text query to get matched image, and *I q T* means *vice versa*. From this table, we can conclude that the proposed GCARL outperforms other subspace learning methods.

Furthermore, we also execute the ablation experiment of ranking loss in GCARL. The results are also shown in Table III. The GCARL-without RL means the GCARL methods without ranking loss. We can see that ranking loss does improve retrieval performance.

D. Food Alerting Evaluation

Finally, we evaluate the proposed CMRDF on the food alerting scene. Fig. 7 shows two examples. We randomly choose 4000 multimodal pairs in RLD as the training set and leave

Fig. 7. Food alerting results. The images captured with instant biodata are presented in the left of the blue dotted line. The five top-ranked output retrieved (from top to bottom) and the ground truth (in black frame) is shown in the right. The keywords on the far right that control whether alerting of not are determined by GCARL based on the category of retrieved multimodal pair.

TABLE IV
PERFORMANCE COMPARISON IN TERMS OF MAP@R(%) AND TIME(S) ON THE RLD DATA SET

Methods	mAP		Time
	R=50	R=All	
DCCA [57]	57.1	43.5	0.067
MNiL [77]	60.2	48.9	0.078
DSCMR [85]	63.7	50.6	0.096
CMRDF-2M	64.3	50.4	0.091
CMRDF-3M	**75.9**	**59.7**	**0.093**

500 pairs as a test set. We divide the data into three categories (forbidden, healthy, and other) for the evaluation of mAP. Once the pair is located in the range of forbidden class, the food alerting starts. The alerting time is also taken into consideration. The classical cross-modal retrieval method DCCA [57] is chosen as the baseline method. In addition, two state-of-the-art methods MNiL [77] and DSCMR [85] are chosen for comparison. We test our CMRDF method using only two modalities of data and three modalities of data, respectively, and all the results are shown in Table IV. Because we perform CMRDF on a cross-modal retrieval task to track down details related to diabetogenic, mAP@R is used to measure the alerting performance at the fixed number of retrieved samples. $R = 50$ means calculating mAP by top 50 retrieved samples and "all" for all retrieved samples.

In Table IV, CMRDF-2M means that two-modal data, images and texts, are used for retrieval and CMRDF-3M means that three-modal data (images, texts, and two kinds of biodata) are used for retrieval, we can see that the proposed CMRDF outperforms DCCA and MNiL obviously. Also, the comparison shows that the fusion of biodata makes it easier for diabetogenic food alerting as well as detail retrieval. In the meantime, the alerting time per query of CMRDF-3M is 0.093 s, which is a little shorter than that of DSCMR, we can say that CMRDF is not only precise but real-time enough to handle the food alerting problem.

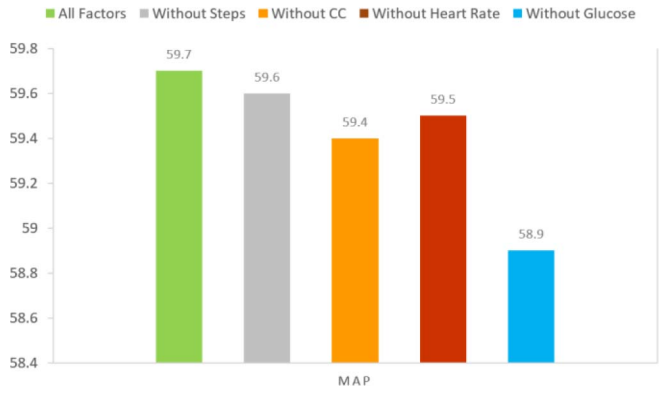


Fig. 8. Performance comparison of mAP when each kind of lifelog biodata is removed. The CC means the calories consumption in lifelog biodata.

E. Discussion

We select the latest GCN model to make it easier to learn the correlation between internal data structure and features of each modality, and take fine-grained semantic space as a common space to realize cross-modal retrieval. Finally, our proposed approach shows the most advanced performance among the different classical or state-of-the-art methods.

We can conclude from Tables I and II that our feature extractors have the best performance in the multilabel classification task. One major reason is that it uses the correlation between labels to learn the latent semantic information. It can explain why the accuracy of our cross-modal retrieval is better than other state-of-the-art models. It is the GCN mines the deep relations and represents them in semantic structures in the graph embedding space, exaggerating the associations between correlated semantic features. In two multilabel classification tasks, besides the procedure of graph convolution, the label correlation matrix employed also helps the GCN to make the most of the implicit information about correlated nodes such as the semantic connections. The above information helps the model to learn a better representation of images or text. So the two feature extractors are able to achieve the best performance in these two tasks.

In Table III, we can confirm that GCARL is more precise than other methods in cross-modal retrieval task. The graph convolutions in each modality contribute to cross-modal retrieval since it embeds the features in different modalities into a similar semantic space. Then, the semantic features with sufficient information are used for cross-modal matching, and the SPP is able to maintain such correlation information in the semantic feature matrix. Finally, with the help of trace norm with ranking loss, two feature vectors are paired in common space precisely. So our model GCARL achieves the best performance in the traditional cross-modal retrieval task.

For the results of the food alerting task shown in Table IV, we can say that with the help of data in the third modality, the proposed CMRDF system can obtain even better performance than it can gain with two modalities. We can also conclude

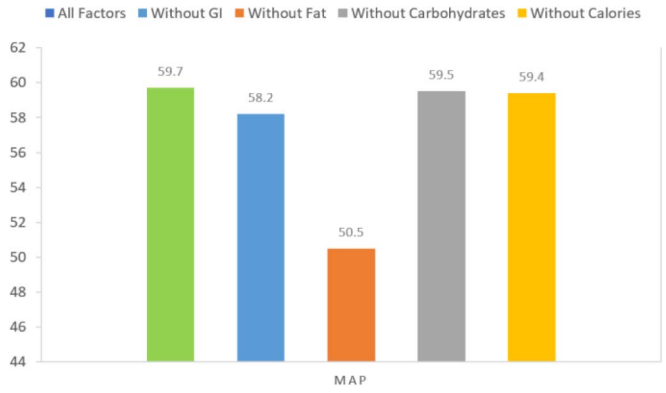


Fig. 9. Performance comparison of mAP when each kind of food composition data is removed. The GI means the glycemic index in food composition data.

that the introduction of the third modality alleviates semantic divergence between modalities in our model, which indeed promotes the cross-modal matching. Also, the results verify the relevance between the food composition and the change of physiological condition during the diet. So the proposed CMRDF achieves the best accuracy in the food alerting task.

In the proposed CMRDF, whenever the food detected, the retrieval model is able to tell the user whether the food is diabetogenic or not through analyzing the retrieved data automatically. We can also learn the correlation between diabetes and food intake as well as analyze the influence of food on physiological characteristics (glucose, etc.) with the help of our system.

We further analyze the influence factor by removing a certain set of data from the training set and test set and comparing the changes of mAP in retrieval results. The comparisons are shown in Figs. 8 and 9. Each column represents the retrieval performance when we remove this kind of data. We can see it from Fig. 8 that the glucose has the greatest positive impact on retrieval performance. It means that the fluctuation of glucose is the most relevant to food, which is corresponding to common sense. Fig. 9 shows that the retrieval performance drops when the GI or fat is removed from the data set. As the fat composition has the biggest influence on the retrieval information, we are able to say that the fat content of food is the most related factor to diabetogenic food in our system.

VI. Conclusion

This article presented CMRDF, a diabetogenic food alerting system for health and wellbeing issues, which can conduct real-time alerting when it detects risky food for diabetic patients. Also, it is able to give important notification to patients by retrieving food details using real-time lifelog images and instant biodata. As the data used in this system are multimodal, including image, text, and numerical value, a new graph-based cross-modal retrieval method, GCARL was proposed. Furthermore, we created a novel diabetogenic food data set, namely, RLD. The experimental results showed that our proposal is of high precision and time effectiveness compared with other state-of-the-art cross-modal retrieval methods.

As for future work, we should establish a larger data set for food information retrieval or diabetogenic food alerting. Such kind of work could improve the performance of real-world food information retrieval. Moreover, we are supposed to establish a more diversified food data set that includes not only diabetogenic information but also other food-related illness, such as allergy.

Meanwhile, the core retrieval model is deployed on a powerful server in our experiments, which means that the cloud services are needed for connecting the lifelog devices with the server to serve the common users. The wireless data transmission can bring uncertain factors and time delays. So, in the future implementation, we are planning to develop a knowledge distillation system to extract a student model from the current system. This kind of student model is lightweight enough to be run on mobile devices. The tiny-size model is able to retrieve information offline with a local database in removable devices, which solves the efficiency of retrieval and the problem of privacy. While the latter is the major problem of our application, a local database and mobile retrieval system may solve it for good.

References

[1] J. MacLeod *et al.*, "Academy of nutrition and dietetics nutrition practice guideline for type 1 and type 2 diabetes in adults: Nutrition intervention evidence reviews and recommendations," *J. Acad. Nutrition Dietetics*, vol. 117, no. 10, pp. 1637–1658, Oct. 2016.

[2] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," *ACM Comput. Surveys*, vol. 52, no. 5, pp. 1–36, Sep. 2019.

[3] W. Min, S. Jiang, and R. Jain, "Food recommendation: Framework, existing solutions and challenges," *IEEE Trans. Multimedia*, early access, Dec. 9, 2019, doi: 10.1109/TMM.2019.2958761.

[4] I. D. Santaren *et al.*, "Serum pentadecanoic acid (15:0), a short-term marker of dairy food intake, is inversely associated with incident type 2 diabetes and its underlying disorders," *Amer. J. Clin. Nutrition*, vol. 100, no. 6, pp. 1532–1540, Dec. 2014.

[5] M. M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou, "A food recognition system for diabetic patients based on an optimized bag-of-features model," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 4, pp. 1261–1271, Jul. 2014.

[6] F. Aguiree *et al.*, *Diabetes Atlas*. Brussels, Belgium: IDF, 2013.

[7] K. Grifantini, "Knowing what you eat: Researchers are looking for ways to help people cope with food allergies," *IEEE Pulse*, vol. 7, no. 5, pp. 31–34, Sep./Oct. 2016.

[8] B. D. Malhotra, R. Singhal, A. Chaubey, S. K. Sharmac, and A. Kumarc, "Recent trends in biosensors," *Current Appl. Phys.*, vol. 5, no. 2, pp. 92–97, 2005.

[9] R. Atat, L. Liu, J. Wu, G. Li, C. Ye, and Y. Yang, "Big data meet cyber-physical systems: A panoramic survey," *IEEE Access*, vol. 6, pp. 73603–73636, 2018.

[10] J. Wu, S. Guo, J. Li, and D. Zeng, "Big data meet green challenges: Big data toward green applications," *IEEE Syst. J.*, vol. 10, no. 3, pp. 888–900, Sep. 2016.

[11] J. Wu, S. Guo, H. Huang, W. Liu, and Y. Xiang, "Information and communications technologies for sustainable development goals: State-of-the-art, needs and perspectives," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2389–2406, 3rd Quart., 2018.

[12] M. S. Mahmud, H. Wang, and H. Fang, "Sensoring: An integrated wearable system for continuous measurement of physiological biomarkers," in *Proc. IEEE ICC*, Kansas City, MO, USA, 2018, pp. 1–7.

[13] W. Min, B.-K. Bao, S. Mei, Y. Zhu, Y. Rui, and S. Jiang, "You are what you eat: Exploring rich recipe information for cross-region food analysis," *IEEE Trans. Multimedia.*, vol. 20, no. 4, pp. 950–964, Apr. 2018.

[14] J. Lee, P. Paudyal, A. Banerjee, and S. K. S. Gupta, "FIT-EVE&ADAM: Estimation of velocity & energy for automated diet activity monitoring," in *Proc. IEEE ICMLA*, Cancun, Mexico, 2017, pp. 1071–1074.

[15] M. Farooq and E. Sazonov, "Accelerometer-based detection of food intake in free-living individuals," *IEEE Sensors J.*, vol. 18, no. 9, pp. 3752–3758, May 2018.

[16] Y. Nishiyama, T. Okoshi, T. Yonezawa, J. Nakazawa, K. Takashio, and H. Tokuda, "Toward health exercise behavior change for teams using lifelog sharing models," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 3, pp. 775–786, May 2016.

[17] F. Li, J. Clemente, and W. Song, "Non-intrusive and non-contact sleep monitoring with seismometer," in *Proc. IEEE GlobalSIP*, Anaheim, CA, USA, 2018, pp. 449–453.

[18] C. Gurrin, A. F. Smeaton, and A. R. Doherty, "Lifelogging: Personal big data," *Found. Trends Inf. Retrieval*, vol. 8, no. 1, pp. 1–125, 2014.

[19] D. T. D. Nguyen, L. Zhou, R. Gupta, M. Riegler, and C. Gurrin, "Building a disclosed lifelog dataset: Challenges, principles and processes," in *Proc. CBMI Workshops*, 2017, pp. 1–6.

[20] R. Gupta and C. Gurrin, "Considering manual annotations in dynamic segmentation of multimodal lifelog data," in *Proc. PerCom Workshops*, Kyoto, Japan, 2019, pp. 34–39.

[21] A. R. Doherty, A. F. Smeaton, K. Lee, and D. P. W. Ellis, "Multimodal segmentation of lifelog data," in *Proc. Large Scale Semantic Access Content (RIAO) Workshops*, Pittsburgh, PA, USA, 2007, pp. 21–38.

[22] D. T. Dang-Nguyen *et al.*, "Overview of ImageCLEFlifelog 2019: Solve my life puzzle and lifelog moment retrieval," in *Proc. Working Notes CLEF Workshops*, Lugano, Switzerland, Sep. 2019, pp. 1–25.

[23] B. Jiang, J. Yang, Z. Lv, and H. Song, "Wearable vision assistance system based on binocular sensors for visually impaired users," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1375–1383, Apr. 2019.

[24] D. Giansanti, G. Maccioni, S. Cesinaro, F. Benvenuti, and V. Macellari, "Assessment of fall-risk by means of a neural network based on parameters assessed by a wearable device during posturography," *Med. Eng. Phys.*, vol. 30, no. 3, pp. 367–372, 2008.

[25] A. Mihailidis, B. Carmichael, and J. Boger, "The use of computer vision in an intelligent environment to support aging-in-place, safety, and independence in the home," *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 3, pp. 238–247, Sep. 2004.

[26] M. He, W. Gu, Y. Kong, L. Zhang, C. J. Spanos, and K. M. Mosalam, "CausalBG: Causal recurrent neural network for the blood glucose inference with IoT platform," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 598–610, Jan. 2020.

[27] J. Lokoè, T. Souèek, and G. Kovalèík, "Using an interactive video retrieval tool for lifelog data," in *Proc. ACM Workshop Lifelog Search Challenge*, 2018, pp. 15–19.

[28] M. Bolanos, M. Dimiccoli, and P. Radeva, "Toward storytelling from visual lifelogging: An overview," *IEEE Trans. Human–Mach. Syst.*, vol. 47, no. 1, pp. 77–90, Feb. 2017.

[29] R. Xu, L. Herranz, S. Jiang, S. Wang, X. Song, and R. Jain, "Geolocalized modeling for dish recognition," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1187–1199, Aug. 2015.

[30] N. Nag, V. Pandey, and R. Jain, "Live personalized nutrition recommendation engine," in *Proc. 2nd Int. Workshop Multimedia Pers. Health Health Care*, 2017, pp. 61–68.

[31] L. Yang, Y. Cui, F. Zhang, J. P. Pollak, S. Belongie, and D. Estrin, "Plateclick: Bootstrapping food preferences through an adaptive visual interface," in *Proc. ACM Int. Conf. Inf. Knowl. Manag.*, 2015, pp. 183–192.

[32] J. Dehais, M. Anthimopoulos, S. Shevchik, and S. Mougiakakou, "Two-view 3D reconstruction for food volume estimation," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1090–1099, May 2017.

[33] S. Jiang, W. Min, L. Liu and Z. Luo, "Multi-scale multi-view deep feature aggregation for food recognition," *IEEE Trans. Image Process.*, vol. 29, no. 7, pp. 265–276, Jul. 2019.

[34] A. Salvador *et al.*, "Learning cross-modal embeddings for cooking recipes and food images," in *Proc. CVPR*, 2017, pp. 3020–3028.

[35] W. Min, S. Jiang, J. Sang, H. Wang, X. Liu, and L. Herranz, "Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1100–1113, May 2017.

[36] Y. Kawano and K. Yanai, "FoodCam-256: A large-scale real-time mobile food recognition system employing high-dimensional features and compression of classifier weights," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 761–762.

[37] P. Pouladzadeh, S. Shirmohammadi, and R. Al-Maghrabi, "Measuring calorie and nutrition from food image," *IEEE Instrum. Meas. Mag.*, vol. 63, no. 8, pp. 1947–1956, Aug. 2014.

[38] C. S. Lee, M. H. Wang, H. C. Li, and W. H. Chen, "Intelligent ontological agent for diabetic food recommendation," in *Proc. IEEE Int. Conf. Fuzzy Syst. World Congr. Comput. Intell.*, Hong Kong, 2008, pp. 1803–1810.

[39] R. G. Rivas, J. J. G. Domínguez, W. P. Marnane, N. Twomey, and A. Temko, "Real-time allergy detection," in *Proc. IEEE 8th Int. Symp. Intell. Signal Process.*, 2013, pp. 21–26.

[40] B. Tusor, G. Simon-Nagy, J. T. Tóth, and A. R. Várkonyi-Kóczy, "Personalized dietary assistant—An intelligent space application," in *Proc. IEEE INES*, 2017, pp. 27–32.

[41] G. Wang, M. D. Poscente, S. S. Park, C. N. Andrews, O. Yadid-Pecht, and M. P. Mintchev, "Wearable microsystem for minimally invasive, pseudo-continuous blood glucose monitoring: The *e*-mosquito," *IEEE Trans. Biomed. Circuits Syst.*, vol. 11, no. 5, pp. 979–987, Oct. 2017.

[42] W. Chonghe *et al.*, "Monitoring of the central blood pressure waveform via a conformal ultrasonic device," *Nat. Biomed. Eng.*, vol. 2, no. 9, pp. 687–695, Sep. 2018.

[43] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101—Mining discriminative components with random forests," in *Proc. ECCV*, 2014, pp. 446–461.

[44] T. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, Apr. 2014, pp740–755.

[45] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013, pp. 1–6.

[46] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.

[47] W. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM MM*, Florence, Italy, 2010, pp. 251–260.

[48] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE CVPR*, Boston, MA, USA, 2015, pp. 3441–3450.

[49] Z. Li, W. Lu, E. Bao, and W. W. Xing, "Learning a semantic space by deep network for cross-media retrieval," in *Proc. DMS*, 2015, pp. 199–203.

[50] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Mach. Learn.*, 2015, pp. 1083–1092.

[51] Y. Liu, Y. Yuan, Q. Huang, and Z. Huang, "Hashing for cross-modal similarity retrieval," in *Proc. Int. Conf. Semantics Knowl. Grids (SKG)*, Beijing, China, 2015, pp. 1–8.

[52] G. Wang, Q. Hu, J. Cheng, and Z. Hou, "Semi-supervised generative adversarial hashing for image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 491–507.

[53] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 4242–4251.

[54] Y.Cong, Z. Qin, J. Yu, and T. Wan, "Cross-modal information retrieval—A case study on Chinese Wikipedia," in *Proc. ADMA*, 2012, pp. 15–26.

[55] Z. Qin, J. Yu, Y. Cong, and T. Wan, "Topic correlation model for cross-modal multimedia information retrieval," *Pattern Anal. Appl.*, vol. 19, no. 4, pp. 1007–1022, 2016.

[56] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, Washington, CA, USA, 2011, pp. 689–696.

[57] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, Atlanta, CA, USA, 2013, pp. 1247–1255.

[58] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 2, 2005, pp. 729–734.

[59] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[60] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, 2017, pp. 1–14.

[61] Z. Liu, Z. Jiang, and W. Feng, "OD-GCN: Object detection by knowledge graph with GCN," in *Proc. Comput. Vis. Pattern Recognit.*, Nov. 2019, pp. 1–6.

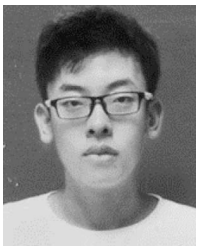
[62] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10685–10694.

[63] Y. Lu, Y. Chen, D. Zhao, and J. Chen, "Graph-FCN for image semantic segmentation," in *Proc. Adv. Neural Netw.*, 2019, pp. 97–105.

[64] B. Ionescu *et al.*, "Overview of ImageCLEF 2017: Information extraction from images," in *Proc. Int. Conf. Cross Lang. Eval. Forum Eur. Lang.*, 2017, pp. 315–337.

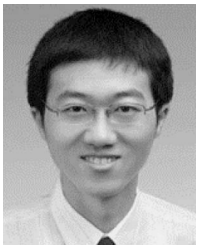
[65] L. Zhou, Z. Hinbarji, D. D. Nguyen, and C. Gurrin, "LIFER: An interactive lifelog retrieval system," in *Proc. ICMR*, 2018, pp. 9–14.

[66] J. Gemmell, A. Aris, and R. Lueder, "Telling stories with mylifebits," in *Proc. IEEE Int. Conf. Multimedia Expo*, Amsterdam, The Netherlands, 2005, pp. 1536–1539.
[67] B. Münzer, A. Leibetseder, S. Kletz, M. Primus, and K. Schoeffmann, "Lifexplore at the lifelog search challenge 2018," in *Proc. ACM Workshop Lifelog Search Challenge*, 2018, pp. 3–8.
[68] I. V. K. Nguyen *et al.*, "THUIR at the NTCIR-14 lifelog-3 task: How does lifelog help the user's status recognition," in *Proc. NTCIR*, 2019, pp. 27–39.
[69] E. Kavallieratou, C. R. del-Blanco, C. Cuevas, and N. Garcia, "Retrieving events in life logging," in *Proc. CLEF Workshops*, Sep. 2018, pp. 235–238.
[70] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–3664, 2004.
[71] Z. M. Chen X. S. Wei, P. Wang, and Y. W. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. Comput. Vis. Pattern Recognit.*, Apr. 2019, pp. 5177–5186.
[72] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 346–361.
[73] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," in *Proc. Int. Conf. Mach. Learn.*, Nov. 2014, pp. 1–32.
[74] J. Devlin, M. W. Chang, K. Lee, and T. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Oct. 2018, pp. 4171–4186.
[75] F. Zhu, H. S. Li, W. L. Ouyang, N. H. Yu, and X. G. Wang, "Learning spatial regularization with image level supervisions for multi-label image classification," in *Proc. CVPR*, 2017, pp. 5513–5522.
[76] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 1–9.
[77] L. Zhang, B. Ma, G. Li, and Q. M. Huang, "Multi-networks joint learning for large-scale cross-modal retrieval," in *Proc. ACM Multimedia*, 2017, pp. 907–915.
[78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.
[79] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 248–255.
[80] R. Q. Xu, C. Li, J. C. Yan, D. Cheng, and X. L. Liu, "Graph convolutional network hashing for cross-modal retrieval," in *Proc. IJCAI*, 2019, pp. 10–16.
[81] T. Yao, Y. W. Pan, Y. H. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. ECCV*, 2018, pp. 684–699.
[82] A. Narayanan, M. Chandramohan, R. Venkatesan, and L. H. Chen, "Graph2vec: Learning distributed representations of graphs," 2017. [Online]. Available: arxiv.abs/1707.05005
[83] L. Zhang, B. P. Ma, G. R. Li, and Q. Tian, "PL-ranking: A novel ranking method for cross-modal retrieval," in *Proc. ACM Multimedia*, 2016, pp. 1355–1364.
[84] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 5005–5013.
[85] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 10386–10395.



Pengfei Zhou was born in Jinan, China, in 1998. He is currently pursuing the bachelor's degree with the College of Computer Science, Zhejiang University of Technology, Hangzhou, China.

His researching interests include computer vision and cross-modal retrieval.



Cong Bai (Member, IEEE) received the B.E. degree from Shandong University, Jinan, China, in 2003, the M.E. degree from Shanghai University, Shanghai, China, in 2009, and the Ph.D. degree from the National Institute of Applied Science of Rennes (INSA de Rennes), Rennes, France, in 2013.

He was with the School of Information Science and Engineering, Shandong Agricultural University, Tai'an, China, from 2003 to 2006. Since 2013, he has been the faculty of the College of Computer Science, Zhejiang University of Technology, Hangzhou, China. His research interests include computer vision and multimedia retrieval/understanding.

Jie Xia was born in Hangzhou, China, in 1999. She is currently pursuing the bachelor's degree with the College of Information Engineering, Zhejiang University of Technology, Hangzhou.

Her research interests include image processing and cross-modal retrieval.

Shengyong Chen (Senior Member, IEEE) received the Ph.D. degree in computer vision from the City University of Hong Kong, Hong Kong, in 2003.

He was with the University of Hamburg, Hamburg, Germany, from 2006 to 2007. He is currently a Professor with the Tianjin University of Technology, Tianjin, China, and also with the Zhejiang University of Technology, Hangzhou, China. He has authored over 100 scientific papers in international journals. His research interests include computer vision, robotics, and image analysis.

Prof. Chen received the National Outstanding Youth Foundation Award of China in 2013. He received the Fellowship from the Alexander von Humboldt Foundation of Germany. He is a Fellow of IET and a Senior Member of CCF.