Two-Stage Spatial Mapping for Multimodal Data Fusion in Mobile Crowd Sensing

JIANCUN ZHOU¹,², TAO XU², SHENG REN², AND KEHUA GUO², (Member, IEEE)
¹All-Solid-State Energy Storage Materials and Devices Key Laboratory of Hunan Province, Hunan City University, Yiyang 413000, China
²School of Computer Science and Engineering, Central South University, Changsha 410083, China

Corresponding author: Kehua Guo (guokehua@csu.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61672535, Grant 61502540, and Grant 61977062, in part by the National Science Foundation of Hunan Province, China, under Grant 2019JJ20025 and Grant 2019JJ40406, in part by the Earmarked Fund for China Agriculture Research System, in part by the Opening Foundation of All-Solid-State Energy Storage Materials and Devices Key Laboratory of Hunan Province under Grant 2017TP1024, and in part by the 111 Project under Grant B18059.

ABSTRACT Human-driven Edge Computing (HEC) integrates the elements of humans, devices, Internet and information, and mobile crowd sensing become an important means of data collection. In HEC, the data collected from large-scale sensing usually includes a variety of modalities. These different modality data contain unique information and attributes, which can be complementary. Combining data from many different modalities will get more information. However, current deep learning is usually only for bimodal data. In order for artificial intelligence to make further breakthroughs in understanding our real world, it needs to be able to process data in different modalities together. The key step is to be able to map these different modalities data into the same space. In order to process multimodal data better, we propose a fusion and classification method for multimodal data. First, a multimodal data space is constructed, and data of different modalities are mapped into the multimodal data space to obtain a unified representation of different modalities data. Then, through bilinear pooling, the representations of different modality are fused, and the fused vectors are used in the classification task. Through the experimental verification on the multi-modal data set, it proves that the multi-modal fusion representation is effective, and the classification effect is more accurate than the single-modal data.

INDEX TERMS Multimodal data, unified representation, mobile crowd sensing, human-driven edge computing.

I. INTRODUCTION

The rapid development of communication technology realizes the information collection and dissemination from various mobile crowd sensing (MCS) services in Human-driven Edge Computing (HEC) environment [1]. However, the data collected from large-scale sensing usually includes a variety of modalities [2]. The environment in which we live is multiple, and our perception to this world also includes multiple factors which are jointly accomplished through language, vision, sound, action, and touch. Humans form meaningful perceptual experience by integrating information from different sensory modalities such as sight, sound, touch, smell and taste into coherent representation. The information from different sensory modalities can complement each other and provide richer information, so multimodal data analysis is getting more and more attention. In order for artificial intelligence to better understand our world, it needs an ability to understand and reason about multimodal data [3]. However, there is a semantic gap between different modality data, which brings great difficulties to processing multimodal data [4].

Taking the deep neural network, a representative technology of artificial intelligence deep neural network, as an example, although the current deep neural network has been widely used in various fields, the data of neural network usually consists of bimodal pairs, such as images and labels, so it cannot process multimodal data [5].

The first step in processing multimodal data is how to uniformly represent the multi-modal data [6]. The purpose of unified representation is to find a latent space, and map data of many different modalities into the space. The mapping process can retain the unique characteristics of each modal data and can also indicate the relationship between different modal data [7]. Because of the complementarity and redundancy between multimodal data, the unified representation
needs to combine the unique information of different modal data and remove the redundant information of all modal data. At the same time, the heterogeneous nature of multimodal data makes uniform characterization more challenging [8]. In addition, the data of different modalities are very different when they represent the same object. For example, for the object of bicycle, the image representation is a series of irregular pixels, and the corresponding label is bicycles. Differences also need to be taken into account. These differences make learning multimodal data more difficult.

There are also some related studies on current multimodal data representation. These studies mainly focus on some common modal data, and analyze the relationships and differences between these data, such as images and text [9], video and audio. The methods for certain modality are not applicable to other modality data. Therefore, we propose a two-stage method for the fusion of multimodal data. First, different modality data is encoded through a specific encoding structure. During the encoding process, each modality data is coded according to a specific structure, for example: using convolutional neural network (CNN) for images [10], and Word2Vec for text [11]. After encoding, the data is represented as a vector, but this vector representation is not the suitable multimodal representation. Because it does not consider the relationship between different modalities, it is only a single modality data representation. Secondly, a mapping relationship is learned through neural networks to map different modal data into latent space. The goal is to map different modal representations into space and make their correlation as original as possible. By elaborately designing the loss function, the same objects are as close as possible in the space, so as to maximize their correlation, and different objects are as far away as possible. After getting the representation of each modal in space, the representation of each modal in space can be used for cross-modal retrieval. Multimodal representations are constructed based on the embedded features of each modality, and compact bilinear pooling is used to generate a fusion representation of multimodal data. Based on the fusion representation, it can be used for the classification of multimodal data, and at the same time, the entire network is back-propagated according to the label, so that the label information is also incorporated. Through this two-stage method, on the one hand, the matching module can cope with the gap between different modal data and map it into a unified space; on the other hand, the labels can supplement some information, making the sneaking into the space more robust.

The innovations of this paper are as follows: 1) A latent space is proposed, which can map data of various modalities to the space. They can retain the previous correlation as much as possible, and the representation of the same objects in the space will be similar. 2) The compact bimodal pooling is used to fuse the multimodal representations to obtain a unified representation in the latent space. The representation can be used for classification and other purposes. 3) In theory, multimodal fusion does not limit the type of data. Any type of data can be mapped into space after the data has been specifically encoded.

The remainder of the paper is structured as follows. Section II provides background by discussing related work. Section III describes the work process of our method. The key details are introduced including data encode process, the mapping and fusion processes in Section IV. Section V provides the experimental results. Finally, Section VI presents conclusions and describes future work.

II. RELATED WORK

In many areas, data is presented in various modalities. We hope that machines can like humans, fuse different aspects of information when making cognition and decision-making. Because it is beneficial to fuse multiple aspects of information, different modal data can complement each other [12], [13]. This requires a model that can fuse different modal data to make decisions. According to reference [7], multimodal tasks are divided into 5 different categories, which are representation, translation, alignment, fusion, and co-learning. In recent years, research in the field of multimodality has attracted a large number of researchers, resulting in a large number of related studies, including cross-modal retrieval [14], video ranking recommendation [15], message authentication [16], and so on. For example, in the field of images and texts, Niu and others have proposed a hierarchical multi-modal visual semantic embedding method that can model the hierarchical relationship between local areas in an image and text [17].

Many current studies have shown that the effect of multi-modal fusion is better than that of single-modal data. For example: in the field of facial expression recognition, reference [18] proposed a multi-modal learning method for facial expression recognition. In order to combine data of different modalities, structure regularization was introduced, which demonstrated the superiority of the proposed method. Based on the deep Boltzmann machine, Pang established a joint density model of multimodal input space, including visual, auditory and text modes [19]. It can learn the joint representation of multimodal and can also cope with the situation of certain modality data. In the medical field, the restricted Boltzmann machine-based deep network is used to fuse images from Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), and the fusion can further improve the accuracy of diagnosis of Alzheimer’s Disease and its precursor stage [20]. In terms of neuroimaging technology, reference [21] proposed a method of integrating individual modalities, which improved the problem that each modality can only provide a limited field of view. Based on many achievements in the field, a table was proposed for analysis, the need of priors, dimension reduction strategies and input data type. In the field of recommendation, reference [22] proposes a recommendation in the field of video, which represents the correlation among videos as the correlation combination of textual, visual, and aural. The above proves that multimodal fusion can achieve better results in different fields.
Unifying the multimodal data into a representation is an important step in multi-modal fusion. The fused representation should contain the information of the original data and the relationship between them. One of the more popular areas is the fusion of images and text. In reference [23], the non-negative matrix decomposition is used to integrate the visual and text information. The results of experiment have achieved good results in various tasks of image index and search. In reference [24], the identity is identified by combining vision and sound. The powerful performance of the residual network is also applied to the multi-modal field. With the powerful advantages of the residual network, the multimodal residual network (MRN) is proposed to solve the visual question answering [25]. MRN inherently uses shortcuts and residual mapping for multimodality, and effectively learns multimodal representations by combining residuals. However, these methods are aimed at specific two or more modal data types, so they cannot deal with general multi-modal fusion problems.

Information can be complementary between multimodal data to achieve more accurate and closer to human effect. For example, based on SVM, the situation of the elderly in a solid home is achieved by combining multiple different sensors in the smart home, and the situation of the elderly is divided into different 7 states [26]. By fusing different biological characteristics, such as: face, fingerprints, and palm prints, using SVM classification after PCA fusion can achieve higher accuracy identification [26]. In pathological diagnosis, different histological images are fused to make pathological diagnosis. A framework based on the novel and robust Collateral Representative Subspace Projection Modeling is used to fuse different histological images [28]. Finally, the classification of the image is determined by multimodal late fusion algorithm with a weighted majority voting strategy.

However, the general multimodal representation method is only applicable to the data of certain common modality data, and there are also restrictions on the common modality data. When the data types are different, it cannot be used. To solve the above problem, we propose a method of multimodal data representation, which can map different modality data to the same potential space and retain their original semantics and corresponding relations.

III. NETWORK STRUCTURE
Multimodal data is often a transmission medium for multiple information. For example, a video has image and sound information. These different modality data contain its unique features and attributes, which can complement each other. Combining data from many different modalities will get more information. Due to the large differences between different modal data, it is difficult to apply deep learning techniques to multi-modal data. The key of eliminating the differences between these data is how to map these different modality data into the same space. The characterization of single-modal data is to express the data as a high-dimensional vector through certain calculations and abstractions. The representation of multimodal data is different from that of single-modal data. In addition to considering the representation of single-modal data, the complementarity between multimodal data also needs to be used to abstract better multimodal representation.

We propose a two-stage spatial mapping network for the fusion and classification of multimodal data, which is mainly divided into two stages:

(1) Data mapping stage. In order to map data of different modalities to the same space at the same time, and maintain the original relation and semantics, data mapping is required. First, the multimodal data are encoded by different encoding methods. Then using multi-branch network to train the data of different modalities separately, the purpose is to make the different modal representations of the same object still maintain the original relation in space, that is, the different modal representations of the same object are as close as possible. After data mapping, it is equivalent to successfully mapping data of different modalities into the same space at the same time and maintaining the original relation and semantics.

(2) Data fusion stage. Through bilinear pooling technology, the data representations of different modalities are fused into a multimodal representation in the space, and the fused representations are classified by a classifier. The results are compared with the real labels to obtain errors. Backpropagation to adjust the classifier network and space mapping network simultaneously. This training method makes full use of the information of multimodal data including labels, making the results more accurate and robust.

The specific process is shown in Fig. 1.

Fig. 1 shows the flow chart of the multimodal data fusion and classification method. The flow of this method mainly includes the following parts: (1) Data Encoding. First, the data of different modalities are encoded by different structures. The data of different modalities are encoded by different structures, which is convenient for representing the data of different modalities more accurately; (2) Data Mapping. The encoded data are mapped through the fully connected (FC) network into the same space; (3) Data Fusion. Multimodal features are fused in pairs through bilinear pooling in the same space, and then all fusions are spliced together to obtain multimodal fusion representation; (4) Classification. The multimodal fusion representation passes through the fully connected network and then through the softmax layer to obtain the classification results of the corresponding multi-modal data.

IV. MULTIMODAL FUSION AND CLASSIFICATION METHODS
This section will introduce the method we used in detail, from the following three aspects: data encoding, spatial mapping, fusion classification process.

A. DATA ENCODING
Data encoding is the first step. Since different modalities data has different representations and cannot be processed in

VOLUME 8, 2020

96729
a unified manner. For different modalities data, if a unified coding method is adopted, it may not be able to obtain good results. The same encoding structure may not be able to adapt to data of many different modalities. So, different modalities data should be encoded by different structures. On the one hand, the different modalities data contains different information in different ways. For example: the image is in the form of pixels arranged in a certain way, and the text is in the form of characters which have a specific meaning. On the other hand, the different modalities data describe the same object, and they are related to each other. So, they need to be processed and extracted separately. Therefore, we need use different encoding methods for different modalities data. Each modality is encoded using the structure that is good at processing to encode the modality data. In theory, there is no limit to the structure of the encoder, and any modal corresponding encoder structure can be applied. For example, for common image data, you can use Inception or ResNet [29, 30] structure trained on ImageNet. For text data, you can use recurrent neural network (RNN) structure processing, pre-training through a lot of expectations. The word vector expression of the text can be obtained from RNN.

In order to better describe the details and quantitative analysis, the following defines some symbolic representations that need to be used.

There are \( N \) pairs of data in the dataset, \( C \) categories, and the modality number of each object is \( m \). \((x_i, y_i)\) represents a pair of data. Each pair of data is a set that consist of representation of different modal data of an object.

Where \( x_i \) is a set of representations of different modality data, \( x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}] \) and \( y_i \) is the label of the object. The encoder structure is defined as \( E = [E_1, E_2, \ldots, E_m] \) for each modality. After the data of each modality is encoded by the encoder structure, the vector is expressed as \( (E(x_i), y_i) \). The representation after encoding is as follows:

\[
F = [(f_1, y_1), (f_2, y_2), \ldots, (f_N, y_N)]
\]  \hspace{1cm} (1)

where

\[
f_i = [E_1(x_{i,1}), E_2(x_{i,2}), \ldots, E_m(x_{i,m})]
\]  \hspace{1cm} (2)

### B. SPATIAL MAPPING

After the multimodal data is encoded, it is converted into a vector representation, and then the vector is converted and mapped into the constructed space. The purpose of this process is to learn the relation between different multimodal data.

Specifically, if different multimodal data represent the same object, then there may be a common part or a potential relation between the different modalities. At the same time, each modality data also carries some other unique information. Our goal is to find the characteristics that exist in all modality data. And a mapping relationship is learned to transform the representation vectors of different modality data and retain the characteristics of each modal data itself. The relation between the different modality is learned, so that the unified representation can retain the original feature information as much as possible and eliminate redundant information. This part is mainly elaborated from the following aspects.

The first is the construction of space. Assuming that \( F \) is the space after encoding of multimodal data. A mapping relation is defined \( G : F \rightarrow T \) which is used to map the multimodal data represented to the multimodal representation space. In this space, all modalities data of same object in a set of data are relatively close, and the distance between different objects of data is relatively long.

This mapping process is complicated and difficult to express in a general form, so a neural network is used to fit this mapping process. A large number of studies have shown that multilayer neural networks can fit various transformation relationships [31]. As long as the back propagation is done by setting the appropriate structure and loss, the neural network can fit the required mapping.

The mapping network is composed of multiple different branches, and each branch processes data of one modality data. Different modality data are separately mapped into the constructed space. In each branch, four fully connected layers are set up to map features into the constructed space. The four fully connected layers are FC1, FC2, FC3 and FC4. The activation function between fully connected layers is ReLU.
The specific form of parameter setting for each layer of network is shown in Table 1 below:

| Layer | Input | Output | BN | Activation |
|-------|-------|--------|----|------------|
| FC1   | ?     | 512    | YES| ReLu       |
| FC2   | 512   | 512    | YES| ReLu       |
| FC3   | 512   | 512    | NO | ReLu       |
| FC4   | 512   | 512    | NO | ReLu       |

In Table 1, the input size of FC1 depends on the encoded vector dimension. In order to meet the requirements of mapping, a loss function needs to be specially set. The ultimate goal of mapping is to make different multimodal data representing the same object as close as possible in space and keep different objects as far away as possible.

First, in the constructed space, the definition of the similarity between modalities refers to the cosine similarity. The cosine similarity range is \([-1, 1]\], the larger the value is, the closer the two objects are to each other. To facilitate the use of back propagation, normalize its value to between 0 and 1, and take the opposite. The definition of similarity \(D\) is shown in Equation (3):

\[
D(x, y) = -0.5 + 0.5 \cdot \frac{x \cdot y}{\|x\| \cdot \|y\|} \tag{3}
\]

where \(x, y\) are two different vector representations in space, and \(\|\cdot\|\) is the 2-Norm of \(\cdot\). When \(D\) is larger, it means that the distance between \(x\) and \(y\) in the space is closer, that is, \(x\) and \(y\) are similar. The smaller \(D\) is, the greater the distance between \(x\) and \(y\) in the space, and the less similar \(x\) and \(y\) are.

After network mapping, the best mapping relationship is that if the similarity between any two modality representations of the same object is small enough, and the similarity between any two modality representations of different objects is large enough. This mapping relationship is considered sufficient. In order to quantify the similarity between the same object and different objects, it is necessary to process not only different modality data of the same object but also the same modality data of different objects. To express, some basic concepts are defined as follows.

**Definition 1:** Matching multimodal data pair \((x, y)\), unmatching multimodal data pair \((x \setminus i, y)\), where:

\[
x = \{x_1, x_2, \ldots, x_m\} \tag{4}
\]

\[
x \setminus i = \{x_1, x_2, \ldots, \hat{x_i}, \ldots, x_m\} \tag{5}
\]

where \(y\) is the label of the matching data pair, \(x\) is the composition of all modality data, which is a representation of the unified object, \(x \setminus i\) is the \(i\)-th modality data is replaced with the most dissimilar data among all the data.

In each calculation, select the top \(k\) multimodal data in the most unmatched data. If the difference exceeds a certain value, then the modal mapping relationship is considered good enough at this time. Therefore, the definition of the loss function is shown in (5):

\[
\mathcal{L} = \sum_{x} \sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} \left[ \alpha \cdot \text{Trun} \left( D \left( x_i, x_j \right), D \left( x_i, \bar{x}_{j,k} \right) \right) + \beta \cdot \text{Trun} \left( D \left( x_i, x_j \right), D \left( x_{j,k}, x_j \right) \right) \right] \tag{6}
\]

where \(\alpha\) and \(\beta\) are balance factors, which are used to balance the weights of different modalities. \(x_{j,k}\) is the top \(k\) least matched data pair, and \(\text{Trun}\) is the truncation function, used to control the similarity between matched data pair and the unmatched data pairs. \(\text{Trun}\) is defined as shown in (7):

\[
\text{Trun} \left( a, b \right) = \max \left( b - a + m, 0 \right) \tag{7}
\]

where, the value of \(m\) is an allowable difference, that is, when the difference between the two-modality data is greater than \(m\), they are considered to be sufficiently different in space. At this time, their value is set to 0. The purpose of this process is to avoid the impact of some data on other data pairs with too large differences. Intuitively, the function of the loss function is to reduce the similarity of the matched data pair, that is, the similarity of \(x_i\) and \(x_j\); and increase the similarity of the unmatched data pair, that is, \(x_i, x_{j,k}\) and \(x_{j,k}, x_j\).

**C. FUSION AND CLASSIFICATION PROCESS**

After getting the spatial representation of each modal, it is necessary to merge the multimodal representations to generate a multimodal fusion representation that can be used for classification. In this part, the multimodal data fusion process and classification process are introduced in detail.

Multimodal fusion representation is implemented with Bilinear Pooling [32]. Bilinear Pooling is mainly used to merge features. Its main form is:

\[
B(x, y) = x^T y \tag{8}
\]

where \(x\) and \(y\) are two \(1 \times m\) feature vectors. After the operation, an \(m \times m\) matrix will be obtained. This will generate a large number of parameters, which is not convenient for subsequent calculations. Therefore, in this paper, its improved scheme, compact bilinear pooling (cbp) that can greatly reduce the number of parameters and the effect can remain unchanged be adopted [33]. A mapping function and fast Fourier transform, multi-modal data is mapped to a low-dimensional space.

Assuming that the two features of multimodal data are represented as \(X\) and \(Y\), then:

\[
\langle B(X), B(Y) \rangle = \langle \sum_{s \in S} \sum_{u \in U} x_s y_u^T, \sum_{s \in S} \sum_{u \in U} y_u x_s^T \rangle
\]

\[
= \sum_{s \in S} \sum_{u \in U} \langle x_s y_u^T, y_u x_s^T \rangle
\]

\[
= \sum_{s \in S} \sum_{u \in U} \langle x_s, y_u \rangle \tag{9}
\]
From the above formula, it is obvious that Bilinear Pooling calculates the features of two positions in an image, and the result of the operation is a second-order polynomial kernel. Bilinear Pooling makes the linear classifier have second-order discriminative power. Therefore, we need to find a projection function \( \phi(x) \) that satisfies \( \phi(x) \in R^p \), \( p \ll c^2 \), and project the original dimension into a lower-dimensional space, so there are:

\[
(B(X), B(Y)) = \sum_{s \in S} \sum_{u \in U} \langle x_s, y_u \rangle^2 \\
\approx \sum_{s \in S} \sum_{u \in U} \langle \phi(x), \phi(y) \rangle \\
= (C(X), C(Y))
\] (10)

where

\[
C(x) = \sum_{s \in S} \phi(x_s)
\] (11)

The sketch function is defined and used to control the computational complexity in the projection process, which can often provide better mapping effects. The sketch function is defined as follows:

\[
\Psi(x, h, s) = \{(Qx)_1, \ldots, (Qx)_p \}
\] (12)

where \( (Qx)_l = \sum_{t: h(l) = s(t)} s(t)x_t \). Both \( h \) and \( s \) are vectors of length \( p \). Each element in \( h \) is randomly generated from \( \{1, 2, \ldots, p\} \), and the probability of each element is the same. Each element in \( s \) is selected from \( +1 \) and \( -1 \), and the probability of selecting \( +1 \) or \( -1 \) each time is equal. Therefore, the mapping function is:

\[
\psi_T(x) = FFT^{-1}(FFT(\Psi(x, h_1, s_1)) \cdot FFT(\Psi(x, h_2, s_2)))
\] (13)

where \( FFT(*) \) is the fast Fourier transform. \( FFT^{-1}(*) \) is the inverse fast Fourier transform. \( \cdot \) is the outer product operation of the vector.

This structure is suitable for parameter optimization using back propagation to achieve end-to-end optimization. Suppose \( \ell \) is the loss function, \( s \) is the position feature information, and \( i \) is the index of the training sample. For the mapping function \( \phi(x) \), the process of back propagation is:

\[
\frac{\partial \ell}{\partial x_s} = \sum_p \frac{\partial \ell}{\partial y_p} \sum_k T_p^k(x_s^i) \cdot s_k
\] (14)

\[
\frac{\partial \ell}{\partial s_k} = \sum_l \frac{\partial \ell}{\partial y_p} \sum_s T_p^k(x_s^i) \cdot x_s^i
\] (15)

where \( T_p^k(x) = \Psi(x, h_k, s_k) \), \( k = \{1, 2\} \), \( \bar{k} = \{2, 1\} \). Through the above back propagation, parameter updates can be achieved. The multimodal data fusion process will be explained below.

First, after normalizing the different modality representations in the space, cbp operation is performed between any two-modality data to get the fusion vector. By splicing the fusion vector, a multimodal fusion representation is obtained. The detailed process is shown in Algorithm 1:

**Algorithm 1 Multimodal Fusion Algorithm**

| Input: | Multimodal dataset: \( D = \{(X_1, y_1), (X_2, y_2), \ldots, (X_N, y_N)\} \) |
| Output: | Multimodal fusion representation: \( D_F = \{(F_1, y_1), (F_2, y_2), \ldots, (F_N, y_N)\} \) |
| 1. | \( i = 0 \) |
| 2. | for \( X \) in \( D \):
| 3. | for \( x \) in \( X \):
| 4. | calculate \( B_c \) of any two components of \( x \) by equation 13 |
| 5. | \( F_i \leftarrow \text{concat} \ B_c \) |
| 6. | end for |
| 7. | \( i \leftarrow i + 1 \) |
| 8. | end for |
| 9. | return \( D_F \) |

1) **SINGLE LABEL**

\[
\mathcal{L} = \frac{1}{N} \sum_i \sum_c y_{ic} \log p_c
\] (16)

where \( y_{ic} \) is the indicator variable, when the category is the same as sample \( i \), it is 1, otherwise it is 0. \( p_c \) is the result which is the output \( Z \) of the classification network after softmax normalization, that is:

\[
p_c = \frac{e^{z_c}}{\sum_c e^{z_c}}
\] (17)

2) **MULTI LABEL**

Similar to the single label, the softmax function is replaced by sigmoid function. The output of each node is calculated by dichotomy and cross-entropy, and the calculation formula is as follows:

\[
\mathcal{L} = -\frac{1}{N} \sum_i \sum_c y_{ic} \log p_c + (1 - y_{ic}) \log(1 - p_c)
\] (18)

where \( y_{ic} \) is the indicator variable. If the label actually exists in the prediction result, it is 1, otherwise it is 0. \( p_c \) is the output \( Z \) of the classification network after sigmoid calculation. That is:

\[
p_c = \frac{1}{1 + e^{-z_c}}
\] (19)
By training the mapping network first, and then training the classification network. The mapping network is also adjusted while training the classification network, so that the mapping network can help the needs of subsequent classification tasks while mapping, and the label information is integrated into the mapping network.

D. FORWARD PROPAGATION AND TRAINING PROCESS
In the above subsections, the details of each part of the multimodal data fusion and classification method are introduced. The details of the network training and forward propagation of the method will be described below.

The forward propagation process can be divided into two parts, namely multi-modal data matching task and multi-modal classification task.

Multimodal data matching task. For a piece of data, the goal is to search for data related to the data from other modal data. For example: for an image, you need to query the text description corresponding to the image. In our method, multimodal representations are obtained through corresponding branches from multimodal data in space. By calculating the similarity $D$ from the sample to be queried, and then using the K-nearest neighbor algorithm, $k$ similar objects with the smallest similarity $D$ are selected from all modality data, that is, other modality data corresponding to the sample to be queried.

Multimodal classification task. The matched data pairs to be classified are input into the network, and the probability values belonging to each category are calculated. If it is a single label, the largest probability value corresponds to the predicted label. If it is a multilabel task, the label with a probability value greater than 0.5 corresponds to the label contained in the data.

The following describes the training process of multimodal data fusion and classification methods. The training process of the multimodal fusion and classification method proposed in this paper is divided into two stages. The first stage is the construction of multimodal space and the training of the mapping process, training the mapping function of the network. The second stage is the training of fusion and classification networks. In the process of back propagation, the parameters of the multimodal mapping network are also adjusted. The detailed training Algorithm 2 is as follows:

By the above algorithm, the mapping network is trained firstly to make multimodal data mapped in space. The multimodal fusion representation is obtained by the vector completed mapping through the multi-modal fusion algorithm. Then, the classification network is trained by the multimodal fusion representation and the data label, and the parameters of the mapping network are also adjusted. The purpose of the classification network is to predict the multimodal fusion representation of multimodal data, and adjust the parameters according to the labels, so that the classification network can cope with the classification task of this label of data. Since the input data of the classification task is the output of the mapping network, the output results also affect the performance of the classification network. Therefore, when the classification network is trained, the mapping network is also trained. The data label itself also contains certain information and integrating it into the training process of the entire network, which will make the overall performance of the network better.

V. EXPERIMENTS
In this section, corresponding experiments will be designed to verify the effectiveness of the proposed method on multimodal data. First, we will introduce the dataset used in the experiments, evaluation indicators, and experimental environment. Then some experiments are designed to evaluate the multimodal data matching performance and multimodal data classification performance of the proposed method, and we analyze the experimental results.

The experimental data is a series of data pairs composed of images and text descriptions from the Pascal dataset. The dataset includes 20 categories, and there are 50 images in each category, a total of 1000 images. For each category, randomly 5 images are selected as test images. The remaining 45 images are used as training images. Each image has a corresponding text describing its content, including five texts describing its content. Some of the data are shown in Fig. 2.
An image and a text description are combined into a sample, so that the data set contains 4500 pairs of data in the training set.

In the data encoding process of the experiments, for image data, ResNet is used as a feature extraction network to obtain the feature vector of the image. ResNet is pre-trained on ImageNet. For text data, fastText is used as a word processing framework to obtain sentence vector expression [34].

These experiments are implemented using the PyTorch deep learning framework. To accelerate the training of deep models, one NVIDIA GTX TITAN XP GPU with 12GB memory is employed. The verification experiments of this article will be introduced below.

A. VALIDATION EXPERIMENT OF MULTIMODAL DATA MAPPING AND MATCHING

Firstly, the spatial mapping function of multimodal data is verified. The goal of mapping is to make different modality data of the same object as close as possible, and modality data of different objects as far away as possible. If the mapping effect is good, there should be several corresponding text descriptions around an image representation. In order to verify the effectiveness of the mapping function, cross-modality retrieval between data is adopted. The experiment uses the k-nearest neighbor algorithm to retrieve the k data closest to each modality data. If there are other modality data that are paired with the data, the retrieval is considered to be accurate.

This experiment is divided into two parts, namely image to text description retrieval and text description to image retrieval. First, the image data and text data in the test data set are mapped into a common space through mapping network. For each sample, the k other modality data samples nearby are found in turn. The values of k are 1, 5, 10. The specific experimental results are shown in Table 2:

| TABLE 2. Retrieval accuracy. |
|-------------------------------|
| Retrieval     | k=1 | k=5 | k=10 |
|----------------|-----|-----|------|
| Image→Text    | 46.0| 82.0| 88.0 |
| Text→Image    | 32.4| 65.4| 76.8 |

As can be seen from Table 2, the accuracy of image retrieval text descriptions is always more accurate than text description retrieval images. Because an image has 5 text descriptions corresponding to it, and each text description has only one image corresponding to it, so the hit rate of image retrieval text descriptions will be higher. On the other hand, the image can provide more information than the text description. Although the text description can also provide corresponding information, the text description is sometimes ambiguous and uncertain, resulting in a low retrieval accuracy. It can also be seen that as the value k increases, the search range also increases, and the hit rate of the search also improves.

B. MULTIMODAL DATA FUSION AND CLASSIFICATION EFFECT

In this experiment, the fusion and classification effects of the proposed method will be verified. After the training process is completed, the data of the test set is input into the multimodal fusion and classification network. The multimodal data will first be mapped into a space, and then each pair of data will be fused into a multimodal representation. In order to more intuitively show the effect of the fusion, the t-SNE algorithm is used to reduce the multimodal data to two-dimensional space and display it visually. The distribution of multi-modal representation after fusion is shown in Fig. 3.

In Fig. 3, different colors represent different categories, and each point is a fusion representation of a pair of multimodal data in space. As can be seen from Fig. 3, the data for each category is basically concentrated in a range. But some categories are distributed more than one cluster, and some categories are still far apart from each cluster. The reason may be that when the image is fused with different text descriptions, the representation of an image and its corresponding 5 text descriptions after fusion is similar in space. The representations of different images and their corresponding textual descriptions present a more scattered phenomenon in the space, which also leads to a number of different small clusters in a category.

Effective fusion is a prerequisite for effective classification. In order to verify the effectiveness of the fusion
representation, the following experiments verify the accuracy of multimodal classification and compare it with the accuracy of single-modality classification. In this experiment, ResNet is used to classify images, and fastText is used to classify text descriptions. The accuracy of single-modal data classification and multimodal data classification are compared and analyzed. Among them, ResNet is pre-trained on the ImageNet data set. Through fine tuning, the fully connected layer with an output dimension of 20 is trained on the last layer of the network for classification. At the same time compared with the scheme mentioned in the reference [27], the accuracy rate of single-modal classification and the accuracy rate of multimodal classification are shown in Table 3.

### TABLE 3. Classification accuracy results.

| Scheme          | Accuracy (%) |
|-----------------|--------------|
| Images          | 64.0         |
| Text descriptions | 57.2       |
| Sharma[27]      | 66.6         |
| Multimodal Fusion | 69.2       |

It can be seen from Table 3 that the classification accuracy of multimodal data is higher than that of single-modality data classification. When using images or text descriptions for classification alone, we can only obtain features from images or text descriptions. The multimodal classification method can obtain more feature information from multimodal data, so it can also get higher classification accuracy. Compared with the using image classification alone, the accuracy of classification has been improved after adding text description. Compared with the using text description classification alone, the accuracy of classification is also greatly improved after adding image information. In contrast, it can be seen that images can provide richer feature information than text descriptions. Compared with the current multimodal data classification scheme, the proposed scheme has higher classification accuracy.

### C. IMPACT OF DIFFERENT MAPPING VECTOR DIMENSIONS ON PERFORMANCE

In the multi-modal data mapping process, the above experiments mapped various modal data to 512 dimensions. In this experiment, the accuracy of multimodal data retrieval under different mapping dimensions is compared to verify the effect of different dimensions on the retrieval accuracy, and to find a better dimension value. The dimension values are set to 256, 512, 1024, 2048, 4096, 8192, and 10000 respectively. That is, the output size of the mapping network is adjusted, then the model is trained, and then the test data are used to test the accuracy of the model. When \( k \) is 1, 5, 10 respectively, the retrieval accuracy of multi-modal data under different mapping dimensions is verify. The specific results are shown in Fig. 4.

![FIGURE 4. Retrieval accuracy in different dimensions.](image)

In theory, the higher the mapping dimension, the more information it contains and the higher the retrieval accuracy. But this is not the case. As can be seen from Fig. 4, when the dimension is low, the accuracy rate will increase as the dimension increases. When the dimension increases to a certain degree, then increasing the dimension value will cause the accuracy to decline. When the maximum dimension set in this experiment is 10000, the accuracy rate is basically at a low level. It can be seen from Fig. 4-a and 4-d that when \( k \) is 1, both the image retrieval text description and the text description retrieval image obtain the highest accuracy rate at the 2048 dimension. As can be seen from Fig. 4-b, c, e, f, after the search range is increased, the best accuracy rate can basically be obtained when the dimension is 1024.

The reason for the above result is that, when the dimension is low, the included information is insufficient, so the accuracy rate will increase as the dimension increases. As the dimension increases, the information included will also increase, but there will also be more irrelevant information. Therefore, too high dimensions will lead to a decrease in accuracy. And the increase of dimension will also lead to double the amount of calculation. Therefore, in the actual use process, the appropriate dimension should be selected.

### D. IMPACT OF DIFFERENT FUSION VECTOR LENGTHS ON CLASSIFICATION PERFORMANCE

In the second experiment, the image and text were fused into 512 dimensions before classification, and the resulting
classification effect was better than that of using images or text descriptions alone for classification. This experiment will verify the effect of different modal fusion length on classification accuracy. Different fusion vector lengths are set to 256, 512, 1024, 2048, 4096, 8192, and 10000, respectively. Then the fusion vector is classified through the classification network. The classification accuracy is compared on different fusion vector lengths when value $k$ are 1 and 5. The result is shown in Fig. 5.

As can be seen from Fig. 5, no matter when $k = 1$ or $k = 5$, the accuracy rate increases first and then decreases as the dimension increases. The results are similar to the multimodal retrieval results under different mapping dimensions in Experiment 3. At $k = 1$, when the fusion vector length is increased from 256 to 512, the accuracy rate is improved by nearly 25%. Regardless of $k = 1$ or $k = 5$, when the length of the fusion vector is increased from 256 to 512, the accuracy improves the most. This is because when the length of the fusion vector is 256, it is not enough to provide sufficient information for classification. Therefore, when the length of the fusion vector is doubled, the classification accuracy is greatly improved. However, when the dimension is increased again, the classification accuracy rate is rarely improved. This is because when the fusion vector length is 512, sufficient information has been provided. When the $k = 5$, the highest accuracy rate is obtained when the fusion vector length is 1024. However, when the fusion vector length is increased from 512 to 1024, the accuracy rate is improved by less than 2%, but the increase in the amount of calculation is doubled. Subsequently, as the length of the fusion vector increases, the accuracy rate basically shows a downward trend. Therefore, in actual use, a balance must be made between performance and computation.

VI. CONCLUSION

In HEC environment, the data collected from large-scale sensing usually includes a variety of modalities. Aiming at the problem that common deep learning techniques are difficult to deal with multi-modal data, a method for fusing multi-modal data is proposed in this paper. This method is used to solve the problem of deep learning training multimodal data and obtain an effective model for identifying multimodal data. First the data of different modalities are encoded, and then map the encoded data into a unified space through the mapping network. The goal is to map the different modal representations into the space and retain their original relations as much as possible. Bilinear pooling is used to generate a fusion representation of multimodal data. Through this method, on the one hand, the mapping network can cope with the gap between different modal data and map it into a unified space; on the other hand, during the training process, you can add some information through the label to make the mapped space is more robust.

In the future, tests will be carried out in the field where multimodal data identification is needed to verify the practicability and effectiveness of the method with the actual results and improve the advantages of the method.

ACKNOWLEDGMENT

The authors declare that they have no conflicts of interests.

REFERENCES

[1] L. Shao, C. Wang, L. Liu, and C. Jiang, “RTS: Road topology-based scheme for traffic condition estimation via vehicular crowdsensing,” Concurrency Comput. Pract. Exp., vol. 29, no. 3, p. e3778, Feb. 2017.
[2] R. Zhu, X. Zhang, X. Liu, W. Shu, T. Mao, and B. Jalaian, “ERDT: Energy-efficient reliable decision transmission for intelligent cooperative spectrum sensing in industrial IoT,” IEEE Access, vol. 3, pp. 2366–2378, 2015.
[3] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, “Multimodal deep learning,” in Proc. 28th Int. Conf. Mach. Learn., 2011, pp. 689–696.
[4] J. Zhou, G. Ding, and Y. Guo, “Latent semantic sparse hashing for cross-modal similarity search,” in Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 2014, pp. 415–424.
[5] Y. Zhang, L. Yuan, Y. Guo, Z. He, I.-A. Huang, and H. Lee, “Discriminative bimodal networks for visual localization and detection with natural language queries,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 557–566.
[6] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, Feb. 2019.
[7] Y. Gao, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, “Pairwise relationship guided deep hashing for cross-modal retrieval,” in Proc. 31st AAAI Conf. Artif. Intell., 2017, pp. 1–9.
[8] J. Guo, B. Song, P. Zhang, M. Ma, W. Luo, and J. Lv, “Affective video content analysis based on multimodal data fusion in heterogeneous networks,” Inf. Fusion, vol. 51, pp. 224–232, Nov. 2019.
[9] J.-W. Ha, B.-H. Kim, B. Lee, and B.-T. Zhang, “Layered hypernetwork models for cross-modal associative text and image keyword generation in multimodal information retrieval,” in Proc. Pacific Rim Int. Conf. Artif. Intell., 2010, pp. 76–87.
[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2012, pp. 1097–1105.
[11] X. Rong, “Word2vec parameter learning explained,” 2014, arXiv:1411.2738 [Online]. Available: http://arxiv.org/abs/1411.2738.
[12] X. Liu, R. Zhu, B. Jalaian, and Y. Sun, “Dynamic spectrum access algorithm based on game theory in cognitive radio networks,” Mobile Netw. Appl., vol. 20, no. 6, pp. 817–827, Dec. 2015.
[13] J. Zhang, J. Cui, H. Zhong, Z. Chen, and L. Liu, “PA-CRT: Chinese remainder theorem based conditional privacy-preserving authentication scheme in vehicular ad-hoc networks,” IEEE Trans. Depend. Sec. Comput., early access, Mar. 11, 2019, doi: 10.1109/TDSC.2019.2904274.
[14] F. Feng, X. Wang, and R. Li, “Cross-modal retrieval with correspondence autoencoder,” in Proc. ACM Int. Conf. Multimedia (MM), 2014, pp. 7–16.
[15] S. Wei, Y. Zhao, Z. Zhu, and N. Liu, “Multimodal fusion for video search reranking,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 8, pp. 1191–1199, Aug. 2010.

[16] J. Cui, X. Zhang, H. Zhong, Z. Ying, and L. Liu, “RSMA: Reputation system-based lightweight message authentication framework and protocol for 5G-enabled vehicular networks,” IEEE Internet Things J., vol. 6, no. 4, pp. 6417–6428, Aug. 2019, doi: 10.1109/JIOT.2019.2895136.

[17] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, “Hierarchical multimodal LSTM for dense visual-semantic embedding,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1881–1889.

[18] W. Zhang, Y. Zhang, L. Ma, J. Guan, and S. Gong, “Multimodal learning for facial expression recognition,” Pattern Recognit., vol. 48, no. 10, pp. 3191–3202, Oct. 2015.

[19] L. Pang, S. Zhu, and C.-W. Ngo, “Deep multimodal learning for affective analysis and retrieval,” IEEE Trans. Multimedia, vol. 17, no. 11, pp. 2008–2020, Nov. 2015.

[20] Y. Ohno, H. Koyama, H. Y. Lee, T. Yoshikawa, and K. Sugimura, “Magnetic resonance imaging (MRI) and positron emission tomography (PET)/MRI for lung cancer staging,” J. Thoracic Imag., vol. 31, no. 4, pp. 215–227, Jul. 2016.

[21] J. Sui, T. Adali, Q. Yu, J. Chen, and V. D. Calhoun, “A review of multivariate methods for multimodal fusion of brain imaging data,” J. Neurosci. Methods, vol. 204, no. 1, pp. 68–81, Feb. 2012.

[22] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li, “Online video recommendation based on multimodal fusion and relevance feedback,” in Proc. 6th ACM Int. Conf. Image Video Retr. (CIVR), 2007, pp. 73–80.

[23] J. C. Caicedo, J. BenAbdallah, F. A. González, and O. Nasraoui, “Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization,” Neurocomputing, vol. 76, no. 1, pp. 50–60, Jan. 2012.

[24] S. Anzellotti and A. Caramazza, “Multimodal representations of person identity individuated with fMRI,” Cortex, vol. 89, pp. 85–97, Apr. 2017.

[25] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, “Multimodal residual learning for visual QA,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 361–369.

[26] A. Fleury, M. Vacher, and N. Noury, “SVM-based multimodal classification of activities of daily living in health smart homes: Sensors, algorithms, and first experimental results,” IEEE Trans. Inf. Technol. Biomed., vol. 14, no. 2, pp. 274–283, Mar. 2010.

[27] P. Sharma and M. Kaur, “Multimodal classification using feature level fusion and SVM,” Int. J. Comput. Appl., vol. 76, no. 4, pp. 26–32, Aug. 2013.

[28] T. Meng, L. Lin, M.-L. Shyu, and S.-C. Chen, “Histology image classification using supervised classification and multimodal fusion,” in Proc. IEEE Int. Symp. Multimedia, Dec. 2010, pp. 145–152.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2818–2826.

[30] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[31] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” Neural Netw., vol. 2, no. 5, pp. 359–366, Jan. 1989.

[32] T.-Y. Lin, A. RoyChowdhury, and S. Maji, “Bilinear CNN models for fine-grained visual recognition,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 1449–1457.

[33] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, “Compact bilinear pooling,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 317–326.

[34] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” Trans. Assoc. Comput. Linguistics, vol. 5, pp. 135–146, Dec. 2017.

JIANCUN ZHOU received the master’s degree in applied mathematics from Guizhou University, Guiyang, China, in 2006. He is currently a Professor with Hunan City University, Yiyang, China. His research interests include edge computing, big data, information fusion.

TAO XU received the bachelor’s degree in information and computing science from Ludong University, in 2017. He is currently a Graduate Student with the School of Computer Science and Engineering, Central South University. His research interests include deep learning and multimodal data.

SHENG REN received the M.S. degree in software engineering from Central South University, Changsha, China, in 2010, where he is currently pursuing the Ph.D. degree in computer science and engineering. His research interests include medical big data, image, and video super-resolution.

KEHUA GUO (Member, IEEE) received the Ph.D. degree in computer science and technology from the Nanjing University of Science and Technology, Nanjing, China, in 2008. He is currently a Professor with Central South University, Changsha, China. He has served as a Guest Editor, the Workshop Chair, the Publicity Chair, a Technical Program Committee Member, and a Reviewer of international journals/conference proceedings. His research interests include social computing, ubiquitous computing, big data, and image retrieval.

***