Emotion Recognition From Multimodal Physiological Signals Using a Regularized Deep Fusion of Kernel Machine

Xiaowei Zhang, *Member, IEEE*, Jinyong Liu, *Member, IEEE*, Jian Shen, *Graduate Student Member, IEEE*, Shaojie Li, Kechen Hou, Bin Hu, *Graduate Student Member, IEEE*, Jin Gao, *Graduate Student Member, IEEE*, Tong Zhang, *Member, IEEE*, and Bin Hu, *Senior Member, IEEE*

***Abstract*—These days, physiological signals have been studied more broadly for emotion recognition to realize emotional intelligence in human–computer interaction. However, due to the complexity of emotions and individual differences in physiological responses, how to design reliable and effective models has become an important issue. In this article, we propose a regularized deep fusion framework for emotion recognition based on multimodal physiological signals. After extracting the effective features from different types of physiological signals, we construct ensemble dense embeddings of multimodal features using kernel matrices, and then utilize a deep network architecture to learn task-specific representations for each kind of physiological signal from these ensemble dense embeddings. Finally, a global fusion layer with a regularization term, which can efficiently explore the correlation and diversity among all of the representations in a synchronous optimization process, is designed to fuse generated representations. Experiments on two benchmark datasets show that this framework can improve the performance of subject-independent emotion recognition compared to single-modal classifiers or other fusion methods. Data visualization also demonstrates that the final fusion representation exhibits higher class-separability power for emotion recognition.**

***Index Terms*—Deep neural network, emotion recognition, kernel machine, multimodal fusion.**

Manuscript received January 21, 2020; accepted April 8, 2020. Date of publication May 13, 2020; date of current version September 8, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFA0706200, in part by the National Natural Science Foundation of China under Grant 61632014 and Grant 61402211, in part by the National Basic Research Program of China (973 Program) under Grant 2014CB744600, and in part by the Program of Beijing Municipal Science and Technology Commission under Grant Z171100000117005. This article was recommended by Associate Editor W. Hu. *(Corresponding author: Bin Hu.)*

Xiaowei Zhang, Jinyong Liu, Jian Shen, Shaojie Li, Kechen Hou, Bin Hu, and Jin Gao are with the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China (e-mail: zhangxw@lzu.edu.cn; liujy2016@lzu.edu.cn; shenj17@lzu.edu.cn; lishj2019@lzu.edu.cn; houkch16@lzu.edu.cn; hub17@lzu.edu.cn; gaoj2018@lzu.edu.cn).

Tong Zhang is with the School of Electronics and Information, South China University of Technology, Guangzhou 510640, China (e-mail: tony@scut.edu.cn).

Bin Hu is with the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China, and also with the CAS Center for Excellence in Brain Science and Institutes for Biological Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China (e-mail: bh@lzu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2020.2987575.

Digital Object Identifier 10.1109/TCYB.2020.2987575

I. Introduction

EMOTION is a high-level cognitive activity of human beings that plays an important role in daily life. In addition to logical intelligence, emotional intelligence is considered an important part of human intelligence [1]. Therefore, improving this cognitive ability of artificial intelligence in human–computer interaction, known as affective computing, has become an increasingly important research topic. The goal of affective computing is the development of artificial intelligence with the ability to perceive, understand, and react to emotions. The identification of emotional states is the first step to achieving the ultimate goal. In this context, emotion recognition has received extensive attention from academia and industry and has been widely used in applications, such as medical care [2], distance education [3], and intelligent robots [4]. Emotions themselves are complex psychological and physiological processes, and they are usually evaluated according to an emotion model that suits the particular application. Existing emotion models fall roughly into the two categories of discrete and continuous emotion models. In discrete emotion models, humans are thought to have an innate set of discrete basic emotions that are cross-culturally recognizable by an individual's facial expressions and biological processes. However, with a limited number of emotions, discrete models are usually too simple to distinguish complex mental states and mixed emotions. Moreover, in continuous emotion models, such as the typical 2-D valence–arousal (V–A) model, most emotional states are described by the dimensions of valence and arousal [5]. In the research of affective computing, the V–A model is widely used to quantitatively describe emotions. The valence dimension represents the positive or negative level of emotion, ranging from unpleasant to pleasant feelings, and the arousal dimension represents the level of excitement or inhibition of emotions, ranging from drowsiness or boredom to extreme excitement. Psychological research has established the interrelationship between these two dimensions [6]. Combining arousal and valence enables

one to represent different emotions, which means continuous emotion models better express moderate emotions (not very intense) and complex real emotions on continuous scales. Therefore, increasing numbers of more researchers have examined continuous emotion models in their studies of emotion recognition.

Next, the selected emotion model would be used to construct a classification scheme that uses different types of data as inputs, such as facial expressions, voice intonation, and body posture, and predicts a user's emotional state as an output [3], [7]. In addition to the above input data, physiological signals that reflect changes in the individuals' nervous and endocrine systems have also been widely studied to more objectively and reliably predict the emotional states. One study explored a method to select suitable subject-specific frequency bands instead of using fixed frequency bands for emotion recognition based on the electroencephalogram (EEG) signals [8]. Other researchers used a discrete wavelet transform (DWT) to extract features from the surface electromyography (EMG) signals followed by a backpropagation neural network for emotion recognition [9]. Hsu *et al.* [10] and Harper and Southern [11] have recently taken advantage of the electrocardiogram (ECG) for emotion recognition by extracting physical ECG features from the time and frequency domains and performing nonlinear analyses of the ECG signals for emotion classification. Li *et al.* [12] proposed a hybrid deep learning model based on multichannel EEG signals to mine the relationship between channels and frequency components to improve the performance of emotion recognition tasks.

Nevertheless, the complexity of emotions and individual differences in physiological responses may lead to declining prediction performance from a classification perspective when using just a single type of physiological signal. Different individuals usually manifest different physiological responses from the same emotional experience. Differences in signals even exist for the same person and conditions at different times, and one's emotional state typically causes simultaneous variation in multiple types of signals. For example, anger shows an increased heart rate, decreased heart rate variability, and increased skin conductivity, while surprise also increases heart rate but without the changes in heart rate variability and skin conductivity. In this case, emotion recognition based on a single type of physiological signal may overlook the impact of individual differences, which decreases the performance of the ultimate classification model. It would be better to integrate multiple physiological signals and construct multimodal fusion strategies to utilize complementarities among different types of signals, so as to gain better performance in both reliability and accuracy.

In this article, we propose an emotion recognition framework based on multimodal physiological signals using a regularized deep fusion of kernel machines (RDFKMs). After extracting the effective features from several types of physiological signals, including EEG, EMG, galvanic skin response (GSR), and respiratory rate (RES) in the DEAP dataset [13], as well as magnetoencephalogram (MEG), EMG, electrooculogram (EOG), and ECG in the DECAF dataset [14], we apply kernel matrices to construct ensemble dense embeddings of multimodal features, to which the representation learning network is applied to learn task-specific representations for each kind of physiological signal from these ensemble dense embeddings. To generate more interactions among multiple modalities and construct better fusion representations, we take advantage of intermediate-level fusion to generate interaction representations between any two modalities. Finally, we use the global fusion layer to fuse all generated representations into a final fusion representation to perform classification tasks. To maximize the capture of the relationships between different representations, we introduce a regularization term to investigate the correlation and diversity among all of the representations in a unified learning process. Experiments conducted on the DEAP and DECAF datasets show that this fusion framework can improve the performance of subject-independent emotion recognition over other fusion methods by efficiently utilizing the correlation and diversity among multimodal physiological signals. The final fusion representation can be observed to exhibit higher class-separability power by using t-distributed stochastic neighbor embedding (t-SNE).

Our framework has several advantages. First, the different levels of the framework are flexible and easily adapt to different tasks. Second, because the initial task-specific representations for each kind of physiological signal are generated from the ensemble dense embeddings using a kernel matrix, our framework is insensitive to synchronization of multimodal signals and can be directly applied to heterogeneous data without time synchronization. Finally, unlike most fusion methods, our framework can explore complex relationships among different types of representations. This characteristic ensures its ability to learn more discriminating fusion representations to efficiently classify emotional states.

The remainder of this article is organized as follows. We review some related work in Section II. In Section III, the proposed ensemble deep kernel machine optimization (eDKMO) for representation learning is proposed. In Section IV, we propose an innovative framework called RDFKM, which extends eDKMO and optimizes its objective function with an alternate strategy. In Section V, several experiments are conducted on the DEAP and DECAF datasets and we compare the results among several state-of-the-art methods. Finally, we discuss our conclusions in Section VI.

II. Related Work

Current methods implement the fusion of multiple physiological signals mainly at three levels: 1) the feature level; 2) the intermediate level; and 3) the decision level.

The feature-level fusion occurs when different types of data are fused before being fed into the model. At this level, the simplest strategy is to concatenate different modalities into a long vector and send it into the model. Multiple kernel learning (MKL) is another popular method of the feature-level fusion. Through the kernel trick, MKL maps the nonlinearly separable features from different modalities to different high-dimensional feature spaces and considers the combination of multiple kernels based on a given criterion to improve classification performance. This method has two advantages.

First, an effective positive semidefinite (PSD) kernel inherently defines a transformation to a reproducing kernel Hilbert space (RKHS) that effectively approximates any spatial transformation function of interest. The RKHS [15] can effectively simplify the empirical risk minimization problem from an infinite-dimensional to a finite-dimensional optimization problem. Second, the fusion of data from multiple modalities is straightforward.

For instance, Guillaumin *et al.* [16] proposed a method using an MKL framework to fuse image and text modalities to improve the object recognition performance of a support vector machine (SVM). Yeh *et al.* [17] proposed an MKL called group lasso regularized MKL (GLMKL) to perform the feature-level fusion, constructing a group of multiple kernels for each modality and using group lasso for each base kernel. Zheng *et al.* [18] used the multimodal deep neural networks to extract features from EEG and eye movements, and predicted emotions based on the learned high-level-shared representation. Shu and Wang [19] proposed a feature-level fusion method using a restricted Boltzmann machine (RBM) to model inherent dependencies among multiple physiological signals, and they validated the effectiveness of multimodal methods on two multimodal benchmark datasets.

The feature-level fusion of multimodal data does not fully utilize the complementary nature of the modalities involved, and it generates large feature vectors containing redundant information. In addition, it is necessary to ensure the synchronism of the multimodal data during feature-level fusion. These shortcomings limit the effectiveness of such methods.

The intermediate-level fusion is an alternative, with the current research extending feature level to the concept of intermediate fusion using deep learning frameworks. Intermediate fusion has good flexibility compared to other fusion approaches, especially in supporting the fusion of representations at different depths. Multimodal-shared representations can be generated directly by a separate fusion layer, or gradually generated by multiple fusion layers at different depths. Intermediate fusion is not without its drawbacks. A large number of network parameters can easily result in overfitting. In addition, the flexibility of intermediate fusion requires a careful design of the architecture, including which modalities to fuse, and how and when to fuse them, making the design process complex.

One example of intermediate fusion is a multiresolution convolutional neural-network architecture [20] that gradually fuses representations learned from video streams across multiple fusion layers during the training process. Compared to feature-level and decision-level fusion, this architecture gives better results in large-scale video classification experiments. A method of multiple expert models using the multimodal data and multiple depth learning methods was proposed [21]. By exploring the correlation among various expert models, the approach combines these expert models into one model to perform a multilabel classification task related to emotion recognition.

The decision-level fusion occurs when different modalities are fused after being fed into the model. This produces multiple independent models for joint decision making. When the input modalities are significantly uncorrelated, the dimensions differ greatly, or different sampling rates are used, it is easier to use decision-level fusion because of the strong independence between modalities. However, the assumption of strong independence also makes it difficult to capture complementary relationships between different types of data.

Shahbe and Hati [22] used a voting scheme (minimization, product, and averaging) to perform decision fusion for infrared (IR) and visible face recognition. Kisku *et al.* [23] proposed to fuse global and local face-matching strategies using the Dempster–Shafer evidence theory for face-recognition tasks, with experimental results demonstrating its effectiveness in the cases of partial occlusion or missing information. Meyer and Mulligan [24] proposed a decision-level fusion method combining speech and visual data that extracted speech features with the MFCC algorithm and visual features from the lip contour portion of a speaker's facial image in a video. Afterward, it used an HMM classifier to obtain independent decisions for each modality, with fusion implemented using the Bayesian inference.

From the above, we see that fusion methods at different levels have their own advantages and disadvantages. How to investigate relationships between different types of signals and fuse them effectively has become an open issue in this important research issue.

III. Ensemble Deep Kernel Machine Optimization

In this section, we first introduce the DKMO model. As a measure of similarity between samples, a kernel matrix constructed using a kernel function is considered to be an embedding of original samples in the DKMO model. Instead of performing representation learning for each embedding using a multilayer fully connected network (FCN) to learn the potential feature space for the corresponding task in original DKMO [25], we propose an eDKMO to perform representation learning for one ensemble embedding. Compared with the original DKMO, the proposed eDKMO can greatly reduce the time and space complexity of computation due to the use of only one multilayer FCN to learn the representation. The original DKMO and proposed eDKMO models are shown in Figs. 1 and 2, respectively. Details of the eDKMO model are as follows.

A. Embedding Layer

Figs. 1 and 2 show that this layer is an important part of the connection between the deep architecture and the kernel machine. We consider the kernel matrix constructed by the kernel as $\mathbf{K} \in \mathbb{R}^{n\times n}$, where $\mathbf{K}_{i,j} = \mathbf{k}(\mathbf{x}_i, \mathbf{x}_j)$ and n represents the number of input samples. Each row $\mathbf{r}_i$ of the kernel matrix models the similarity relation between the sample $\mathbf{x}_i$ and all other samples $\mathbf{x}_j$, which can be considered as an embedding for the sample $\mathbf{x}_i$. In the ideal case, the values of the kernel matrix will be very sparse because samples from the same class have larger values, with samples from different classes having very small values, even close to zero. Furthermore, the number of embedding dimensions increases with the sample size. The combination of sparsity and high dimensionality makes the original embedding unsuitable for inference tasks. It is difficult

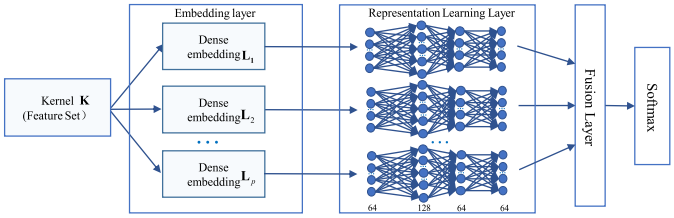


Fig. 1. Original DKMO model. The kernel matrix constructed from the feature data generates multiple dense embeddings using kernel approximation techniques in the embedding layer. Then, in order to learn the potential feature space for the corresponding task, original DKMO performs representation learning for each embedding using a multilayer FCN in the representation learning layer and exploits two kinds of merging strategies (one directly uses concatenation and summation; and the other uses concatenation and summation with kernel dropout regularization) to fuse the learned representations in the fusion layer. Finally, the fused representation is sent to the softmax layer for classification.

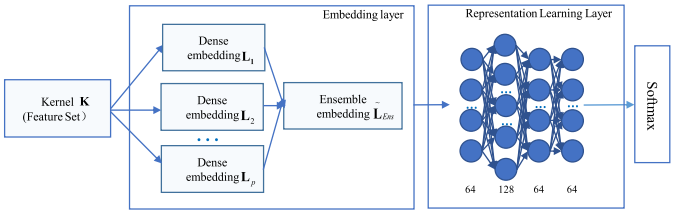


Fig. 2. Proposed eDKMO model. The kernel matrix constructed from the feature data generates an ensemble embedding using kernel approximation techniques and ensemble methods in the embedding layer, which is then sent to the representation learning layer to obtain the potential representation. Finally, the potential representation is sent to the softmax layer to perform the classification task. Note that the proposed eDKMO model has significantly less time and space complexity of computation due to the use of only one multilayer FCN to learn the representation for one ensemble embedding.

to achieve good results by directly building a model on the original embedding.

One simple solution applies the matrix factorization strategies to transform the original embedding into a dense representation with low dimensionality. This is similar to the use of point mutual information (PMI) to construct dense word embeddings in natural language processing [26]. In this article, we use kernel approximation techniques based on the Nyström method [27] to generate dense embeddings directly from the kernel matrix.

For a given kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, we want to find an approximation matrix $\widetilde{\mathbf{K}}_r$ with a rank much smaller than the number of samples. Using the Nyström method, an efficient technique to generate low-rank matrix approximations, we first define matrix $\mathbf{C} \in \mathbb{R}^{n\times s}$ consisting of s columns randomly extracted from the matrix $\mathbf{K}$. These randomly selected s columns can be used to find an approximate kernel map $\mathbf{L} \in \mathbb{R}^{n\times r}$, such that $\mathbf{K} \simeq \mathbf{L}\mathbf{L}^T$, where $s \ll n$ and $r \leq s$. After rearranging the kernel matrix $\mathbf{K}$, we write $\mathbf{C}$ and $\mathbf{K}$ as

$$\mathbf{C} = \begin{bmatrix} \mathbf{W} \\ \mathbf{E} \end{bmatrix} \text{ and } \mathbf{K} = \begin{bmatrix} \mathbf{W} & \mathbf{E}^T \\ \mathbf{E} & \mathbf{F} \end{bmatrix}$$

where $\mathbf{W} \in \mathbb{R}^{s\times s}$ is the matrix containing the intersection of $\mathbf{C}$ and the corresponding s rows of the given kernel matrix $\mathbf{K}$. $\mathbf{E}$ and $\mathbf{F}$ are the remaining parts after rearranging the given kernel matrix $\mathbf{K}$. Since the kernel matrix $\mathbf{K}$ is PSD, $\mathbf{W}$ is also PSD.

For a given $r \leq s$, the Nyström method generates a rank-r approximation

$$\widetilde{\mathbf{K}}_r = \mathbf{C}\widetilde{\mathbf{W}}_r^{+}\mathbf{C}^T \tag{1}$$

where $\widetilde{\mathbf{W}}_r$ is the best-r approximation of $\mathbf{W}$ based on the truncated singular value decomposition (TSVD).

We then obtain the mapping function

$$\mathbf{L} = \mathbf{C}\left(\mathbf{U}_{\widetilde{\mathbf{W}}_r}\mathbf{\Lambda}_{\widetilde{\mathbf{W}}_r}^{-1/2}\right) \tag{2}$$

where $\mathbf{U}_{\widetilde{\mathbf{W}}_r}$ and $\mathbf{\Lambda}_{\widetilde{\mathbf{W}}_r}$ are the top r eigenvalues and eigenvectors of $\mathbf{W}$. To reduce the number of dimensions, we use the approximate mappings $\mathbf{L}$ directly instead of the approximated kernels $\widetilde{\mathbf{K}}_r$. In fact, the mapping functions generated from different samples will generate completely different representations in the RKHS. Thus, we must use different representations to model the characters of different regions in the input space. For this reason, we calculate the different sample subsets obtained in the multiple random sampling process to obtain multiple mapping functions $\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_p$. To learn the potential feature space, the original DKMO model performs representation learning for the p dense embeddings using p multilayer FCNs. Due to performing representation learning for each embedding using a multilayer FCN, the original DKMO can cost a great deal of computational time and space. To solve this problem, we propose an eDKMO model that applies an ensemble method [28] to further improve these p dense embeddings generated by p mapping functions. Thus, the general form of the ensemble dense embedding is

$$\widetilde{\mathbf{L}}_{\mathbf{Ens}} = \sum_{i=1}^{p} \mu_i \mathbf{L}_i. \tag{3}$$

We define the ensemble weight μ_i as the reciprocal of the number of embeddings.

B. Representation Learning Layer

In the representation learning layer, the ensemble dense embedding generated by the embedding layer in the eDKMO model is sent to an FCN to learn the potential feature space for the corresponding task.

In fact, the final representation generated from the fusion layer in DKMO contains almost all of the information (it may lose some information due to dropout) of several learned representations. In other words, it contains almost the entire information of all generated embeddings in the embedding layer. Similarly, the ensemble embedding in eDKMO contains almost the entire information of all of the generated embeddings. The representation generated by DKMO can be expressed as $\mathbf{R_f} = f_1(\mathbf{L}_1) + f_2(\mathbf{L}_2) + \cdots + f_p(\mathbf{L}_p)$, where $\mathbf{R}_f$ is the fused representation and $f_p(\cdot)$ is the pth multilayer FCN in the representation learning layer. The representation learned by eDKMO can be expressed as $\mathbf{R}_e = f_e(\widetilde{\mathbf{L}}_{\mathbf{Ens}})$, where $f_e(\cdot)$ is the multilayer FCN in the representation learning layer. Because the representation learning layers of these two models nearly have the same structure, the losses generated by these layers can be regarded as identical. Therefore, if we ignore the losses generated by the multilayer FCN, the losses generated from the

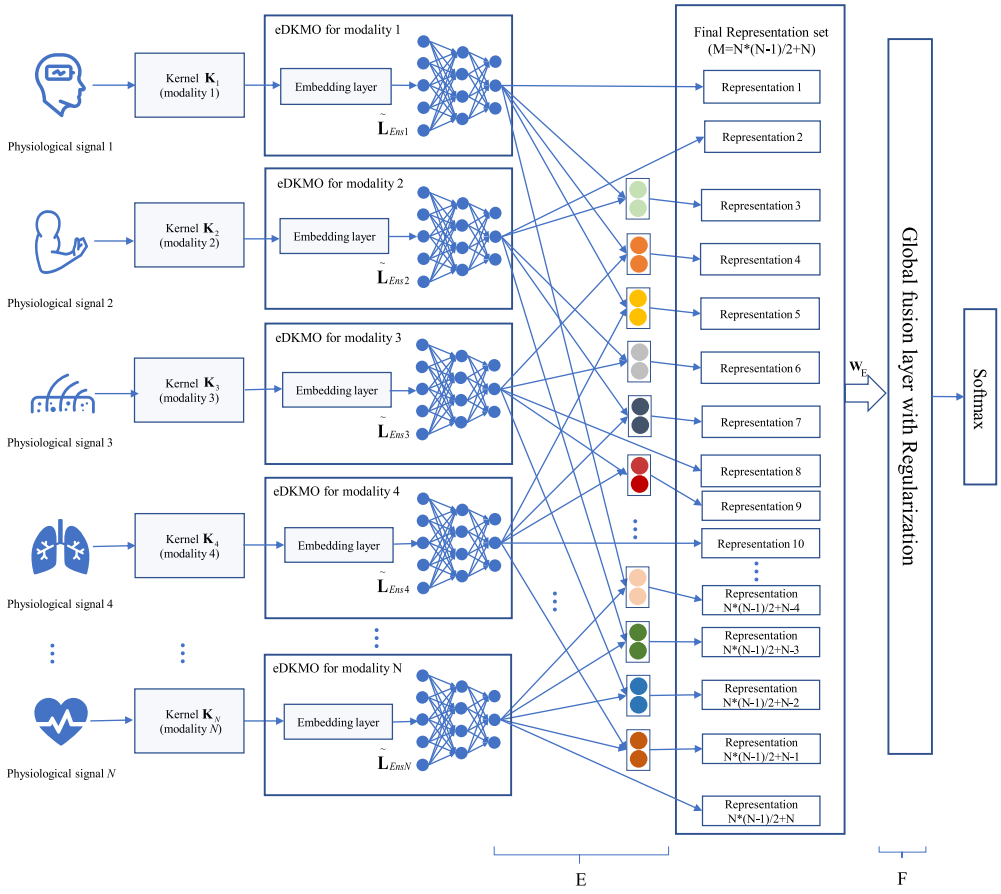


Fig. 3. RDFKM framework. The proposed framework takes advantage of the eDKMO model without the softmax layer to generate potential representations for N physiological signals. These will be further fused at a deeper level to generate intermediate fusion representations between any two modalities in the framework. All of the representations generated by eDKMO models and the intermediate fusion layer are used to construct a fusion representation set, where the intermediate fusion takes advantage of a one-layer FCN, which has the same dimension (64) as the last layer of the representation learning layer in eDKMO. The global fusion layer with regularization fuses representations in the final representation set to generate a final fused representation and sends it to the softmax layer for classification, where $\mathbf{W}_E = [\mathbf{W}_E^1, \ldots, \mathbf{W}_E^m, \ldots, \mathbf{W}_E^M]$ is the weight matrix, which can linearly transform representations in the final representation set ($\mathbf{W}_E^m$ is the weight matrix for the mth representation, where $m = 1, \ldots, M$). Note that the eDKMO models used in this framework are the same as the eDKMO proposed in Section III.

representation layer of DKMO and eDKMO can be expressed as $\|\mathbf{K} - \sum_{i=1}^{p} \mathbf{L_i}\|_F^2$ and $\|\mathbf{K} - \widetilde{\mathbf{L}}_{\mathbf{Ens}}\|_F^2$, respectively, where $\mathbf{K}$ is the kernel matrix and $\|\cdot\|_F^2$ is an upper bound on the norm-2 loss and the Frobenius loss. According to [28], the loss bounds of the representation generated by DKMO and eDKMO are similar in form. However, the bounds of eDKMO are tighter than DKMO. Moreover, eDKMO only costs a fraction of computational time and space in the representation learning layer in comparison with DKMO.

IV. Regularized Deep Fusion of Kernel Machine

To efficiently integrate multimodal physiological signals, we apply the eDKMO model to each modality and fuse them using an innovative framework called the RDFKM framework, as shown in Fig. 3.

A. Multimodal Representation Learning Layer

The data from different sources represent different aspects of the task with complementary information. As shown in Fig. 3, each modality uses an independent kernel function to generate its own kernel matrix. Then, we use a set of eDKMO models without a softmax layer to generate representations from different kernel matrices. In other words, we optimize the multiple kernels by sending each kernel matrix from the different modalities to an eDKMO model to obtain representations for fusion, which is a part of the final representation set. Note that we implement intermediate fusion for representations generated by eDKMO to enable better interaction representations between any two modalities, inspired by Mroueh *et al.* [29]. The neurons (colored dots) in Fig. 3 represent the intermediate fusion for different representations generated by eDKMO. The intermediate fusion takes advantage of the one-layer FCN, which has the same dimension (64) as the last layer of the representation learning layer in eDKMO. The representations acquired by intermediate fusion will also be a part of the final representation set for fusion. We use those representations in the final representation set to generate more effective features by exploiting the relationships between representations in the next global fusion layer.

B. Global Fusion Layer

In the global fusion layer, we fuse all representations in the final representation set.

In fact, different representations input to the fusion layer are both correlated and diverse. The former means that some information may be shared between different representations. The latter means there is information unique to each representation.

This complex relationship implies that a simple fusion strategy may show limited performance improvement because the relationships between multiple representations are ignored. To generate a more robust and efficient fusion representation to improve the performance of the corresponding classification task, we add a regularization term to the objective function to guide the generation of the fusion representation and simultaneously perform fusion and classification.

To facilitate the discussion below, we define some symbols relating to the framework. Considering only the network structure in the framework, it has L layers. To simplify the problem, we assume that the final representations set has M representations of the same size. As shown in Fig. 3, the global fusion layer uses all of the representations in the final representation set to generate a fusion representation as an input to the softmax layer, so the transformation equation for the global fusion layer can be expressed as

$$\mathbf{a}_F = \sigma\left(\sum_{m=1}^{M} \mathbf{W}_E^m \mathbf{a}_E^m + \mathbf{b}_E\right) \tag{4}$$

where E is the index of the layer ahead of the global fusion layer (i.e., the last layer of the representation learning layer in the eDKMO model and the intermediate fusion layer), F is the index of the global fusion layer (i.e., $F = E + 1$), and M is the number of representations in the final representation set ($M = N * (N-1)/2 + N$). Therefore, $\mathbf{a}_E^m$ represents the mth representation of the final representation set. σ is a nonlinear activation function. It can be known from the above equation that a representation among these M representations of the final representation set is first linearly transformed by the weight matrix $\mathbf{W}_E^m$, and then converted to a fusion representation $\mathbf{a}_F$ by nonlinear mapping using an activation function.

Since the weights $\mathbf{W}_E^1, \ldots, \mathbf{W}_E^M$ in the fusion layer convert all of the representations to fusion representations, we can explore relationships between representations based on these weights. We first transform each weight matrix $\mathbf{W}_E^m$, $m = 1, \ldots, M$, to a weight column vector of size $P = |\mathbf{a}_E^m||\mathbf{a}_F|$, and then stack all these vectors together into a matrix $\mathbf{W}_E \in \mathbb{R}^{P \times M}$.

To simultaneously explore the correlation and diversity between modalities, we construct a regularization term in the final objective function

$$\begin{aligned} \min_{\mathbf{W},\boldsymbol{\Psi}} \quad & \mathcal{L} + \frac{\lambda_1}{2}\sum_{l=1}^{L-1}\|\mathbf{W}_l\|_F^2 + \frac{\lambda_2}{2}\mathrm{tr}\left(\mathbf{W}_E\boldsymbol{\Psi}^{-1}\mathbf{W}_E{}^T\right) \\ \text{s.t.} \quad & \boldsymbol{\Psi} \succeq 0. \quad \mathrm{tr}(\boldsymbol{\Psi}) = 1 \end{aligned} \tag{5}$$

where λ_1 and λ_2 are the regularization coefficients, and $\boldsymbol{\Psi} \in \mathbb{R}^{M \times M}$ is a relation matrix modeling the relationship between different columns of $\mathbf{W}_E$. The first term, $\mathcal{L} = \sum_1^N \ell(\hat{(y)}, y)$, is the empirical loss term. The second term is a regularization term to avoid overfitting. The last term helps learn the relationship between representations by forcing the parameters between representations to be similar according to the similarity encoded in the relation matrix inverse. That is to say, through this regularization term, the weight vectors of the correlated representations will have similar values so that they make similar contributions to the fusion representation. Similar regularization terms are often used in multitask learning to explore the relationship between tasks to improve learning performance [30]. After implementing the fusion strategy, we use a softmax layer to obtain the final classification result.

C. Optimization

To optimize (5), we use an alternate optimization strategy to optimize the network weight $\mathbf{W}_l (l = 1, \ldots, L)$ in the entire framework and the relationship matrix $\boldsymbol{\Psi}$.

1) Algorithm: We first optimize $\mathbf{W}_l (l = 1, \ldots, L)$ by making $\boldsymbol{\Psi}$ constant for all the parameters of each layer in the network. With this condition, the original problem becomes the unconstrained optimization problem

$$\min \quad \mathcal{L} + \frac{\lambda_1}{2}\sum_{l=1}^{L-1}\|\mathbf{W}_l\|_F^2 + \frac{\lambda_2}{2}\mathrm{tr}\left(\mathbf{W}_E\boldsymbol{\Psi}^{-1}\mathbf{W}_E^T\right). \tag{6}$$

All of the items in the above loss function are smooth, which means the gradient can be effectively calculated. After calculating the gradient $\mathbf{G}_l$ for $\mathbf{W}_l$, the update equation of $\mathbf{W}_l$ on the kth iteration becomes

$$\mathbf{W}_l(k) = \mathbf{W}_l(k-1) - \alpha\mathbf{G}_l(k) \tag{7}$$

where α is the learning rate of the gradient.

Then, we optimize $\boldsymbol{\Psi}$ while holding the other variables constant. In that case, the optimization problem in (5) becomes

$$\begin{aligned} \min \quad & \mathrm{tr}\left(\mathbf{W}_E\boldsymbol{\Psi}^{-1}\mathbf{W}_E^T\right) \\ \text{s.t.} \quad & \boldsymbol{\Psi} \succeq 0. \quad \mathrm{tr}(\boldsymbol{\Psi}) = 1. \end{aligned} \tag{8}$$

We then obtain

$$\begin{aligned} \mathrm{tr}\left(\boldsymbol{\Psi}^{-1}\mathbf{A}\right) &= \mathrm{tr}\left(\boldsymbol{\Psi}^{-1}\mathbf{A}\right)\mathrm{tr}(\boldsymbol{\Psi}) \\ &= \mathrm{tr}\left(\left(\boldsymbol{\Psi}^{-\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\right)\left(\mathbf{A}^{\frac{1}{2}}\boldsymbol{\Psi}^{-\frac{1}{2}}\right)\right)\mathrm{tr}\left(\boldsymbol{\Psi}^{\frac{1}{2}}\boldsymbol{\Psi}^{\frac{1}{2}}\right) \\ &\geq \left(\mathrm{tr}\left(\boldsymbol{\Psi}^{-\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\boldsymbol{\Psi}^{\frac{1}{2}}\right)\right)^2 = \left(\mathrm{tr}\left(\mathbf{A}^{\frac{1}{2}}\right)\right)^2 \end{aligned} \tag{9}$$

where $\mathbf{A} = \mathbf{W}_E^T\mathbf{W}_E$. Based on the Cauchy–Schwarz inequality, $\mathrm{tr}(\boldsymbol{\Psi}^{-1}\mathbf{A})$ produces the minimum value if and only if $\boldsymbol{\Psi}^{-(1/2)}\mathbf{A}^{(1/2)} = \mathbf{a}\boldsymbol{\Psi}^{(1/2)}$. According to the above analysis, we obtain the analytical solution as

$$\boldsymbol{\Psi} = \frac{\left(\mathbf{W}_E^T\mathbf{W}_E\right)^{\frac{1}{2}}}{\mathrm{tr}\left(\left(\mathbf{W}_E^T\mathbf{W}_E\right)^{\frac{1}{2}}\right)}. \tag{10}$$

Thus, we are able to effectively solve the proposed optimization problem.

In summary, we use the weights in the neural network to estimate the relationship between different representations and then use the relationship matrix to improve classification performance. We note that Jiang *et al.* [31] used a similar solution to model the relationship between features in deep

Algorithm 1 Optimization Process of RDFKM

Process:

1: Random initialization of all network weights $\mathbf{W}_l, l = 1, \ldots, L$ in the entire framework , $\mathbf{\Psi} = \frac{1}{M}\mathbf{I}_M$,where $\mathbf{I}_M$ are identity matrices;

2: **for** *iteration* $k = 1$ to K **do**

3: update the weight matrix $\mathbf{W}_l, l = 1, \ldots, L$ for each layer through backpropagation by calculating

$$\mathbf{W}_l(k) = \mathbf{W}_l(k-1) - \alpha \mathbf{G}_l(k); \tag{15}$$

4: update the representation relation matrix $\mathbf{\Psi}$ by calculating

$$\mathbf{\Psi} = \frac{(\mathbf{W}_E^{\mathbf{T}}\mathbf{W}_E)^{\frac{1}{2}}}{\mathrm{tr}((\mathbf{W}_E^{\mathbf{T}}\mathbf{W}_E)^{\frac{1}{2}})}; \tag{16}$$

5: **end for**

neural networks, but our approach applies structural regularization to complex multimodal deep fusion network architectures to explore the relationships between different representations. Algorithm 1 describes the entire optimization process.

2) Convergence Analysis: We prove convergence of Algorithm 1 as follows.

Theorem 1: The alternating update rules described in Algorithm 1 monotonically decrease the value of the objective function in each iteration until convergence.

Proof: In the iteration process, to update $\mathbf{W}_l, l = 1, \ldots, L$, we fix the relation matrix $\mathbf{\Psi}$ and use (7) to update the parameters of each layer in the network. Then, we have

$$\mathcal{J}\Big((\mathbf{W}_1, \ldots, \mathbf{W}_L)^{t+1}, \mathbf{\Psi}^t\Big) \leq \mathcal{J}\big((\mathbf{W}_1, \ldots, \mathbf{W}_L)^t, \mathbf{\Psi}^t\big). \tag{11}$$

To update the relational matrix $\mathbf{\Psi}$, we fix the weight matrix $\mathbf{W}_l, l = 1, \ldots, L$, in the network and use (10) to update the relational matrix $\mathbf{\Psi}$. According to the analytical process of the analytical solution, we have

$$\mathrm{tr}\Big(\mathbf{W}_E\mathbf{\Psi}^{-1}\mathbf{W}_E^T\Big) \geq \left(\mathrm{tr}\left(\left(\mathbf{W}_E^T\mathbf{W}_E\right)^{\frac{1}{2}}\right)\right)^2. \tag{12}$$

The analytical solution (10) is constructed under the condition that the equal sign holds and the relation matrix $\mathbf{\Psi}$ is updated according to it. Then, we have

$$\mathcal{J}\Big((\mathbf{W}_1, \ldots, \mathbf{W}_L)^t, \mathbf{\Psi}^{t+1}\Big) \leq \mathcal{J}\big((\mathbf{W}_1, \ldots, \mathbf{W}_L)^t, \mathbf{\Psi}^t\big). \tag{13}$$

Based on (11) and (13), we obtain

$$\begin{aligned}\mathcal{J}\Big((\mathbf{W}_1, \ldots, \mathbf{W}_L)^{t+1}, \mathbf{\Psi}^{t+1}\Big) &\leq \mathcal{J}\Big((\mathbf{W}_1, \ldots, \mathbf{W}_L)^t, \mathbf{\Psi}^{t+1}\Big) \\ &\leq \mathcal{J}\big((\mathbf{W}_1, \ldots, \mathbf{W}_L)^t, \mathbf{\Psi}^t\big).\end{aligned} \tag{14}$$

Therefore, the overall objective function monotonically decreases using Algorithm 1 and Theorem 1 is proved. ■

3) Time Complexity Analysis: In the training phase, we must input N kernel matrices generated from N modalities. Thus, the complexity for calculating the kernel matrices is $\mathcal{O}(N * n^2)$. The time complexity in the embedding layer for N modalities is $\mathcal{O}(N * p * n * r)$. Then, we alternately update $\mathbf{W}$ and $\mathbf{\Psi}$. For each iteration of the algorithm, the variables $\{\mathbf{W}, \mathbf{\Psi}\}$ are updated with the time complexity of $\mathcal{O}(N*n*w_1*r + N*n*w_1*w_2 + N*n*w_2*w_3 + N*n*w_3*w_4 + N*n*w_4*w_5 + M*n*w_5*w_6)$ and $\mathcal{O}(M^2)$, respectively, where w_1, w_2, w_3, and w_4 are the number of nodes (width) in each layer of the FCN, w_5 is the number of nodes in the intermediate fusion layer, w_6 is the number of nodes in the global fusion layer, and the complexity of intermediate fusion is $\mathcal{O}(N*n*w_4*w_5)$. In practice, the values of N, r, w_1, w_2, w_3, w_4, w_5, and w_6 are much smaller than n. Hence, the computationally expensive part of the framework is due to computing the kernel matrices. However, this particular computation is independent of update rules in the iterations, and we conduct it only once in the initialization phase of the algorithm, which considerably accelerates the convergence speed.

V. Experiments

A. Datasets

To evaluate the effectiveness of our method, we conducted several experiments on the DEAP [13] and DECAF datasets [14].

The DEAP and DECAF databases are two multimodal datasets for analyzing emotional states. The DEAP dataset consists of different types of data from an experiment, in which 32 participants watched 40 video clips while their participant ratings, physiological responses, and facial expressions were recorded. Similarly, the DECAF dataset provides different types of data, in which 30 participants watched 36 video clips while their participant ratings in the V–A model, physiological responses, and near-infrared (NIR) facial video signals were recorded. In our experiments, we concentrated on the multimodal physiological data in these datasets.

The physiological data in the DEAP dataset contain EEG signals, which are random signals containing highly complex information [32], and peripheral physiological responses, including EMG, GSR, and RES. Similarly, the DECAF dataset contains MEG signals and peripheral physiological responses, including EMG, EOG, and ECG.

Because arousal and valence can effectively represent various aspects of emotion, which are widely used in affective computing, we primarily consider these two dimensions during the experiments. We also use the same setting as the datasets providers and map each dimension into positive and negative. At this point, we obtain two classification tasks (high/low valence and arousal) to judge the effectiveness of our method.

B. Data Preprocessing and Feature Extraction

1) Data Preprocessing: In our experiments, we used the preprocessing physiological data recordings from the DEAP and DECAF datasets. Preprocessing on the DEAP dataset consisted of downsampling all of the data to 128 Hz, averaging EEG data to the common reference, removing EOGs artifacts, and implementing a bandpass frequency filter between 4.0 and 45.0 Hz. Preprocessing on the DECAF datasets consisted of trial segmentation, frequency-domain filtering, including downsampling the MEG signal to 300 Hz, low-pass and high-pass filtering with cutoff frequencies of 95 and 1 Hz,

TABLE I
FEATURE SET USED IN EXPERIMENT

Dataset	Cha	Feature	Description
DEAP	EEG	Power Spectral Density (PSD)	$log(P_x(f))$ in different bands: delta,theta,slow alpha, alpha, beta, and gamma.
	EMG	Power	$log(P_x(f)), f \in [20, f_s/2]Hz$
		Statistical moments	Mean, SD, skewness, and kurtosis
	GSR	Number of peaks	Number of peaks in resistance exceeding 100 Ω
		Amplitude of peaks	GSR peak amplitude from the saddle point preceding the peak
		Rise time	The time it takes GSR to reach its peak from the saddle point in seconds
		Statistical moments	Mean,SD
	RES	Main frequency	Frequency at which the power spectrum reaches its maximum value($f \in [0.16, 0.6]Hz$)
		PSD	$log(P_x(f)), f \in [0, 0.1], [0.1, 0.2], [0.2, 0.3], [0.3, 0.4], [0.4, 0.5], [0.5, 0.6], [0.6, 0.7], [0.7, 0.8], [0.8, 0.9], [0.9, 1.0]Hz$
		Statistical moments	Mean, SD, skewness, and kurtosis
DECAF	MEG	Discrete Cosine Transform (DCT)	The first two DCT coefficients from the spatial, temporal, and spectral dimensions
	EMG	PSD	$log(P_x(f)), f \in [0, 0.5], [0.5, 1.5], [1.5, 2.5], [2.5, 3.5], [3.5, 5.0], [5.0, 10], [10, 15], [15, 25], [25, 45], [55, 95], [105, 145]Hz$
	EOG	PSD	$log(P_x(f)), f \in [0, 0.1], [0.1, 0.2], [0.2, 0.3], [0.3, 0.4], [0.4, 0.6], [0.6, 1.0], [1.0, 1.5], [1.5, 2.0], [105, 115], [115, 130], [130, 145]Hz$
	ECG	PSD	$log(P_x(f)), f \in [0.1, 0.2], [0.2, 0.3], [0.3, 0.4], [0.4, 0.5], [0.5, 0.6], [0.6, 1.0], [1.0, 1.5], [1.5, 2.0], [2.0, 2.5], [2.5, 5.0]Hz$
		Inter-Beat Intervals	Statistical measurements
		Heart Rate	Statistical measurements
		Heart Rate Variability	Statistical measurements

respectively, to remove low-frequency ambient noise and high-frequency artifacts, channel correction, and time–frequency analysis for frequency smoothing. Due to noise, three trials in the DECAF dataset were removed.

2) Feature Extraction: Other researchers have tended to cut the data into a number of short time periods to better describe the information contained in the signal, but this method may not be applicable to the DEAP and DECAF datasets. During DEAP and DECAF data collection, subjects were asked to watch videos for a period of time, which means they were in the process of watching the video and not always in a state of high stimulation. Thus, many time segments contain only useless information [33]. Such simple data segmentation does not meet expectations for improved performance, so we extract features from the signals of the entire time period. In other words, we regard each trial as a sample, meaning that there are 40 trials for each of the 32 participants in the DEAP dataset, and 33 trials for each of the 30 participants in the DECAF dataset.

For all of the physiological signals in the DEAP dataset, we used the TEAP toolbox [34] to extract features from four types of signals: 1) EEG; 2) EMG; 3) GSR; and 4) RES. In the DECAF dataset, we used the features that were extracted by the dataset providers [14]. Table I lists the features extracted from the data.

C. Experimental Setup

After data preprocessing and feature extraction, we obtained two datasets with 1280 samples (32 subjects $\times$ 40 trials) in DEAP, and 990 samples (30 subjects $\times$ 33 trials) in DECAF, and then conducted single-modal classification and multimodal fusion classification experiments on these two datasets. The performance of our method is compared with that of several fusion methods at different levels to verify the effectiveness of our method. In all of the experiments, we compared classification performance both on the valence and arousal dimensions.

According to the way of dividing a given dataset into training and testing sets, we can carry out subject-dependent or subject-independent classification. In the former case, the model is trained and tested based on the data of an individual subject. In the latter case, the model is trained from data of various subjects and then tested on data of new subjects (not included those in the training set). In fact, in real application scenarios, an emotional recognition system often faces the latter scenario, so it is more valuable to establish a good subject-independent emotional recognition model. For these reasons, we used leave-one-subject-out (LOSO) cross-validation to evaluate the performance of several methods for subject-independent emotion recognition [35]. For example, in the DEAP dataset, for subject 1, we trained a model using the data from subjects 2 to 32 (training set) and recorded the evaluation score for subject 1 (testing set). Similarly, for subject 2, we trained with data from subject 1 and subjects 3 to 32 (training set) and recorded the evaluation score for subject 2 (testing set). This was done for all 32 subjects. The training set was further divided into training data and a validation set with the same strategy. We evaluated our method according to the average of the 32 scores. To explore the effect of the kernel function in the proposed RDFKM framework, we conducted a series of experiments on several commonly used kernel functions, including RBF, linear, and polynomial kernel with degree $d = 2$ and 3. We provided the best experimental results acquired from the selected kernel function in this article. Moreover, the kernel parameters were optimized through LOSO cross-validation.

The following paragraphs detail the models and the corresponding parameters we compared. In single-modal experiments, we used three algorithms to evaluate the classification performance: 1) SVM; 2) decision tree (DCT); and 3) naive Bayes (NB). In multimodal experiments, we compared our method with six fusion methods at different levels, including feature-level fusion with SVM (SVM-FLF), AverageMKL, SimpleMKL, EasyMKL, FCN, and decision-level fusion with SVM (SVM-DLF).

1) *SVM:* In the single modality experiments, we searched the appropriate value of C in $[10^{-3}, 10^{-2}, 10^{-1}, 10^{-0}, 10^{1}]$.
2) *DCT:* DCT is an instance-based inductive learning method whose model presents a tree structure. We used the Gini index to measure the split quality.
3) *NB:* NB is a classification method based on Bayes' theorem and the conditional independence hypothesis between features.
4) *SVM-FLF:* In multimodal experiments, SVM-FLF linearly combines features from different modalities at the feature level and uses an SVM for classification. We attempted to find an optimal value of the SVM parameter C from the list $[10^{-3}, 10^{-2}, 10^{-1}, 10^{-0}, 10^{1}]$.
5) *AverageMKL:* This method of averaging all base kernels to construct the target kernel has become a strong baseline for MKL.
6) *SimpleMKL:* We compared this classic method that uses a general MKL algorithm with semi-infinite linear programming [36]. We obtained the optimal value of the parameter C from $[10^{-3}, 10^{-2}, 10^{-1}, 10^{-0}, 10^{1}]$.
7) *EasyMKL:* We also compared the state-of-the-art MKL algorithm [37], an MKL method that can quickly process large-scale kernels. We searched the parameter space $[10^{-3}, 10^{-2}, 10^{-1}, 10^{-0}, 10^{1}]$ to determine the parameter C.
8) *FCN:* In the multimodal experiments, we used a fusion layer to merge the results of the hidden layer for each modality. During the training process, we set the dropout parameter to 0.5. All of the networks used the Adam optimizer with the learning rate set to 0.001 to optimize network parameters.
9) *SVM-DLF:* We trained an SVM classifier using features from different modalities independently. The outputs of all of the classifiers were fused based on a hard voting strategy. We determined the parameter C from $[10^{-3}, 10^{-2}, 10^{-1}, 10^{-0}, 10^{1}]$.
10) *RDFKM (Our Method):* In the eDKMO part of the multimodal representation layer, we set $s = 3\%n$, $r = 30$, and $p = 10$, according to [28]. We sent each ensemble embedding to the independent eDKMO representation learning layer with sizes 64, 128, 64, and 64 empirically. In the subsequent fusion layer, each intermediate layer of size 64 further fused the representations from different modalities. Finally, the global fusion layer with size 128 fused all of the representations. Other parameters of the network were the same as the FCN settings. The regularization parameters were set to 0.01. We implemented the entire network architectures using TensorFlow and conducted the experiments using a single Tesla M60 GPU.

D. Results

In the experiments, we defined two binary classification tasks. We implemented two commonly used metrics (accuracy and F1 score) to evaluate the results of all of the algorithms.

To alleviate individual differences, we performed standardization by scaling the features of each subject to a value between 0 and 1 to reduce interparticipant variability, according to the following equation:

$$Z_k = \frac{X_k - U_k}{S_k} \tag{17}$$

where X_k, U_k, S_k, and Z_k are the features of the kth subject without standardization, the mean of X_k, standard deviation of X_k, and the standardized features of the kth subject, respectively.

First, we compared the classification performance between single modality and combinations of different modalities on the DEAP and DECAF datasets, as shown in Tables II and III. The experimental results on single-modal data are the best obtained from SVM, DCT, or NB. All of the results on combinations of different modalities are achieved using our RDFKM framework. We can see that the best experimental results are obtained when fusing all of the modalities both on arousal and valence, irrespective of the dataset. We also show the best experimental results acquired among the kernels and the corresponding kernel in Tables II and III. As illustrated in Tables II and III, most experimental results (best experimental results acquired from several kernel functions) are obtained using the RBF kernel function.

In addition, we observe that the combination of different modalities may not necessarily improve the performance compared to a single modality. For example, in the DEAP dataset, the classification performance of EEG+EMG was inferior to EEG or EMG on the valence dimension, and in the DECAF dataset, both MEG and ECG showed superior performance on the arousal dimension compared to MEG+ECG. According to the definition of the relation matrix $\mathbf{\Psi}$ and the regularization term in our RDFKM framework, the weight vectors of the correlated representations will have similar values so that they make similar contributions to the fusion representation. If there are only two modalities with lower correlation, it will be difficult to determine the weights using the similarity defined in the relation matrix. The final global fusion representation will exhibit lower class-separability power under the influence of interaction representation of these two uncorrelated modalities. Following the increase of modalities, the RDFKM framework will more efficiently investigate the correlation and diversity among all of the representations. By increasing weights of similar representations, our framework could gain consistency among different representations and further reduce the influences of uncorrelated representations, ultimately obtaining a more efficient global fusion representation to improve generalization in emotion recognition.

Second, we analyzed the experimental results of multimodal fusion methods. As shown in Tables IV and V, the RDFKM method achieved the best performance in all of the comparison methods. On the DEAP dataset, the RDFKM method achieved an improvement of approximately 5%–7% in both accuracy and F1 score on the valence dimension. It improved accuracy by approximately 10% and F1 scores by approximately 8%, in the arousal classification task. Similarly, on the DECAF dataset, the RDFKM

TABLE II
EXPERIMENTAL RESULTS OBTAINED BY ENABLING DIFFERENT MODAL COMBINATIONS ON THE DEAP DATASET

Modalities	Valence				Arousal			
	Classifier	Kernel	ACC	F1	Classifier	Kernel	ACC	F1
EEG	SVM	Poly2	0.589	0.673	SVM	RBF	0.563	0.699
EMG		RBF	0.613	0.687			0.568	0.705
GSR		RBF	0.563	0.696			0.583	0.725
RES		Poly3	0.562	0.703			0.580	0.720
EEG+EMG	RDFKM	RBF	0.578	0.579	RDFKM	RBF	0.572	0.641
EEG+GSR			0.573	0.602			0.606	0.672
EEG+RES			0.566	0.572			0.570	0.630
EMG+GSR			0.589	0.651			0.564	0.637
EMG+RES			0.601	0.657			0.557	0.621
GSR+RES			0.566	0.712			0.559	0.667
EEG+EMG+GSR	RDFKM	RBF	0.595	0.631	RDFKM	RBF	0.590	0.662
EEG+EMG+RES			0.584	0.626			0.566	0.642
EEG+GSR+RES			0.580	0.647			0.606	0.615
EMG+GSR+RES			0.581	0.638			0.584	0.623
EEG+EMG+GSR+RES	RDFKM	RBF	**0.645**	**0.696**	RDFKM	RBF	**0.631**	**0.701**

TABLE III
EXPERIMENTAL RESULTS OBTAINED BY ENABLING DIFFERENT MODAL COMBINATIONS ON THE DECAF DATASET

Modalities	Valence				Arousal			
	Classifier	Kernel	ACC	F1	Classifier	Kernel	ACC	F1
MEG	SVM	Linear	0.612	0.687	NB	-	0.560	0.614
EMG		RBF	0.585	0.714	DCT	-	0.523	0.531
EOG		RBF	0.629	0.701	SVM	Ploy2	0.556	0.534
ECG		Poly3	0.589	0.709		RBF	0.520	0.531
MEG+EMG	RDFKM	RBF	0.639	0.669	RDFKM	RBF	0.533	0.518
MEG+EOG			0.637	0.632			0.537	0.561
MEG+ECG			0.652	0.680			0.565	0.537
EMG+EOG			0.594	0.623			0.519	0.547
EMG+ECG			0.580	0.639			0.548	0.489
EOG+ECG			0.663	0.700			0.522	0.577
MEG+EMG+EOG	RDFKM	RBF	0.642	0.694	RDFKM	RBF	0.554	0.589
MEG+EMG+ECG			0.656	0.668			0.533	0.497
MEG+EOG+ECG			0.647	0.688			0.548	0.535
EMG+EOG+ECG			0.694	0.709			0.538	0.510
MEG+EMG+EOG+ECG	RDFKM	RBF	**0.712**	**0.719**	RDFKM	RBF	**0.571**	**0.605**

achieved an improvement of approximately 6%–11% in accuracy and 0.1%–4% in F1 score on the valence dimension. It improved accuracy by approximately 0.6%–4% and F1 scores by approximately 6%, in the arousal classification task.

We used the Friedman test [38] and two-tailed Nemenyi test [39] to determine the statistical significance. In the DEAP dataset, we calculated $F(6, 186) = 4.62$ and $F(6, 186) = 4.55$ in accuracy and F1 score on arousal, as well as $F(6, 186) = 3.28$ and $F(6, 186) = 3.14$ in accuracy and F1 score on valence, which showed $p < 0.05$ and CD $= 1.59$. So our method was significantly ($p < 0.05$) better than SVM-FLF, AverageMKL, SimpleMKL, EasyMKL, FCN, and SVM-DLF in accuracy on arousal. Similar results were obtained for the F1 score. Moreover, our method was significantly ($p < 0.05$) better than SVM-FLF and EasyMKL in accuracy on valence, which can be also observed in the F1 score. In the DECAF dataset, we calculated $F(6, 174) = 7.32$ in the F1 score on arousal, as well as $F(6, 174) = 3.45$ in accuracy on valence, which showed $p < 0.05$ and CD $= 1.65$. Thus, our method was significantly ($p < 0.05$) better than AverageMKL, FCN, and SVM-DLF in the F1 score on arousal. Our method was significantly ($p < 0.05$) better than FCN and SVM-DLF in accuracy on valence.

We compared RDFKM with five state-of-the-art studies that used the same or similar features on the DEAP and DECAF datasets, and the results are shown in Table VI. The cross-validation strategies include subject-dependent leave-one-trial-out (SDLOTO), subject-dependent three-fold (three-fold), and subject-independent LOSO. According to Table VI, RDFKM performs better than most of the state-of-the-art studies. In fact, it is more difficult to obtain good results with a subject-independent cross-validation strategy because individual differences will degrade the generalization performance of models. Under these conditions, our method still improves the performance of emotion recognition on these two datasets compared to other state-of-the-art methods.

To more intuitively present the representations generated by different methods, we also used the t-SNE algorithm to reduce dimensions and visualize the generated representations in the classification experiments. Fig. 4 shows the representations generated by the fusion method in the valence classification task on the DEAP dataset, except for decision-level fusion. We can see that the separability of fusion representations generated by MKL methods is not good enough, which makes the classification performance worse. For FCN and RDFKM, we extracted the output of the fusion layer and visualized it. It can be clearly seen that the fusion representation generated

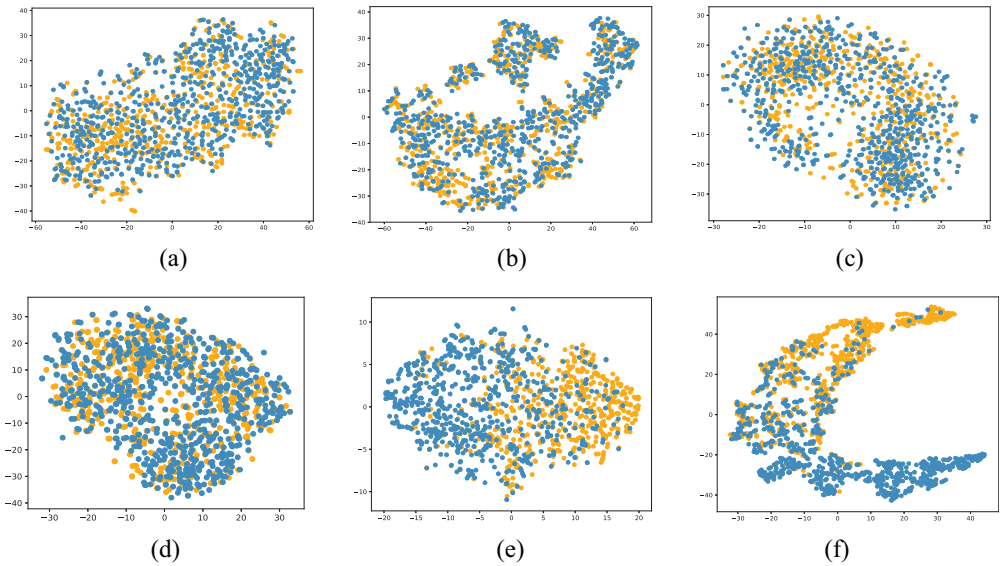


Fig. 4. Visualization of the representation in valence classification experiments on the DEAP dataset. Data from the 17th subject were used for testing (scoring), with data from subjects 1–16 and 18–32 used for training. (a) AvgMKL. (b) SimpleMKL. (c) easyMKL. (d) SVM-ELF. (e) FCN. (f) RDFKM.

TABLE IV
EXPERIMENTAL RESULTS FOR MULTIMODAL SETTINGS ON THE DEAP DATASET

Method	Evaluation	Valence	Arousal
SVM-FLF	ACC	0.573	0.534
	F1	0.631	0.608
AverageMKL	ACC	0.580	0.528
	F1	0.630	0.599
SimpleMKL	ACC	0.588	0.539
	F1	0.644	0.625
EasyMKL	ACC	0.570	0.532
	F1	0.622	0.601
FCN	ACC	0.595	0.559
	F1	0.650	0.639
SVM-DLF	ACC	0.598	0.533
	F1	0.668	0.647
RDFKM	ACC	**0.645**	**0.631**
	F1	**0.696**	**0.701**

TABLE V
EXPERIMENTAL RESULTS FOR MULTIMODAL SETTINGS ON THE DECAF DATASET

Method	Evaluation	Valence	Arousal
SVM-FLF	ACC	0.632	0.565
	F1	0.713	0.547
AverageMKL	ACC	0.642	0.528
	F1	0.709	0.529
SimpleMKL	ACC	0.636	0.547
	F1	0.716	0.540
EasyMKL	ACC	0.643	0.546
	F1	0.713	0.545
FCN	ACC	0.594	0.562
	F1	0.673	0.518
SVM-DLF	ACC	0.619	0.531
	F1	0.718	0.418
RDFKM	ACC	**0.712**	**0.571**
	F1	**0.719**	**0.605**

by RDFKM has better separability than the other methods. These results demonstrate the effectiveness of our approach in improving fusion performance.

TABLE VI
COMPARISON WITH STATE-OF-THE-ART STUDIES

Method	Dataset	Strategy	Evaluation	Valence	Arousal
Tzelepis et al. [40]	DEAP	SDLOTO 3-fold	ACC	0.659	0.650
			F1	0.551	0.609
Soleymani et al. [34]	DEAP	SDLOTO	ACC	0.586	0.559
			F1	0.570	0.523
Romeo et al. [41]	DEAP	SDLOTO	ACC	0.636	0.611
			F1	0.612	0.546
Kandemir et al. [42]	DEAP	LOSO	ACC	0.600	0.580
			F1	0.560	0.530
SAbadi et al. [14]	DECAF	SDLOTO	ACC	0.600	0.560
			F1	0.590	0.550
RDFKM	DEAP	LOSO	ACC	0.645	0.631
			F1	0.696	0.701
RDFKM	DECAF	LOSO	ACC	0.712	0.571
			F1	0.719	0.605

TABLE VII
DATA DIMENSION OUTPUT FROM EACH MODULE OF THE RDFKM FRAMEWORK FOR EACH MODALITY DURING THE TRAINING PHASE ON THE DEAP AND DECAF DATASETS

Datasets	DEAP	DECAF
Modalities	EEG, EMG, GSR, RES	MEG, EMG, EOG, ECG
Kernel	1200×1200	924×924
Embedding Layer	1200×100	924×100
FCN	64×1200	64×924
Intermediate Fusion Layer	64×1200	64×924
Global Fusion Layer	128×1200	128×924

To further illustrate the RDFKM framework, we depict the data dimension of each module in the RDFKM framework on two datasets. As shown in Table VII, the data dimension from the embedding layer decreases significantly by using kernel approximation techniques and ensemble methods, which can reduce the computational complexity compared to directly using the kernel matrix. The data output from the FCN and the

intermediate fusion layer have the same dimension to facilitate the global fusion. The data output from the global fusion layer will be sent to the softmax layer for classification.

E. Discussion

As can be seen from the experimental results, our method efficiently integrated multimodal physiological signals and gained more effective fusion representation. Before proceeding, we note that many previous studies do not use the LOSO verification method [43], [44]. In practical application, new participants are completely unknown individuals. Cross-validation with random segmentation may incorporate part of the samples of test subjects in both the training and testing sets, resulting in overfitting and degrading generalization of classification models. Therefore, it is more reasonable to use the LOSO method for performance validation. In fact, it is more difficult to obtain good results with the LOSO validation method because the influence of individual differences makes it difficult to build a good general model for all of the subjects. Under these conditions, our method still improves the accuracy in the valence and arousal dimensions compared to other benchmark methods.

It can be seen that our proposed method is closely related to the MKL method. In fact, the fusion method based on MKL strongly relies on the constructed optimization problem, and the goal of this optimization problem is generally to find a linear combination of multiple kernels. Compared to the traditional kernel method, although the MKL method shows the advantages of the automated kernel combination, the multikernel method does not always exceed the performance of the single-kernel method. The main reason is that the target kernel domain is not rich enough, so the performance of the multikernel method is not necessarily stronger than that of the single-kernel method. In other words, the MKL method is essentially a shallow learning method. The insufficiency of the target kernel domain directly affects the performance of the fusion method based on MKL, so that the model cannot effectively improve performance when facing the problems of complex relationships, such as the fusion of multimodal physiological signals. In our method, the kernel matrix representing different modalities will construct an ensemble representation from multiple linear subspaces of the RKHS. This ensemble representation will serve as the native space of a predefined kernel, which contains information from multiple linear subspaces of the RKHS. Subsequent multilayer network architectures will convert the ensemble representation to task-specific representations driven by tasks, which means that the different modalities of the native space of a predefined kernel are transformed by the task-specific feature space.

The computational expense of most MKL methods owes to computing the kernel matrices. Therefore, although the proposed RDFKM framework takes advantage of several useful modules and tricks to improve fusion performance, the RDFKM framework does not consume much more computational time than other MKL methods. The following global fusion layer fuses the transformed feature space into an implicit multikernel combination. Compared to conventional MKL methods, our method utilizes the superior performance of the multilayer deep network architecture in representation learning and flexibility of the shared representation fusion method, and it has stronger fusion capability.

In addition, our method models the relationships between weights of different representations in the fusion layer by constructing the relation matrix $\mathbf{\Psi}$, which is essentially the covariance matrix between weight column vectors constructed based on the matrix-variate distribution [45]. Intuitively, representations making similar contributions to the final global fusion representation should have similar weights. We explore correlation and diversity between different representations through the optimization process. After convergence, larger values of nondiagonal terms represent larger similarities, and the weights of corresponding representations are guaranteed to be similar by minimizing the trace norm. The regularization term in our method can guide the exploration of correlation and diversity between different representations and improve the quality of global fusion representation.

VI. Conclusion

In this article, we proposed an innovative multimodal fusion framework with regularization based on a new perspective of the kernel matrix and a deep network architecture. We used the superior performance of the deep network architecture in representation learning to transform the native space of a predefined kernel to a task-specific feature space and adopted a shared presentation layer to learn the fusion representation, which means implicitly combining multiple kernels. At the same time, a new regularization term was introduced in the loss function to model relationships between representations to improve the performance of multimodal fusion.

The experimental results show that our method efficiently captures relationships between multimodal physiological signals through the kernel function representation and gains better fusion performance.

References

[1] M. Ptaszynski, P. Dybala, W. Shi, R. Rzepka, and K. Araki, "Towards context aware emotional intelligence in machines: Computing contextual appropriateness of affective states," in *Proc. Int. Joint Conf. Artif. Intell.*, 2009, pp. 1469–1474.

[2] J. Shen, X. Zhang, B. Hu, G. Wang, Z. Ding, and B. Hu, "An improved empirical mode decomposition of electroencephalogram signals for depression detection," *IEEE Trans. Affective Comput.*, early access, Aug. 14, 2019, doi: 10.1109/TAFFC.2019.2934412.

[3] L. Qi and H. Tan, "Facial and speech recognition emotion in distance education system," in *Proc. Int. Conf. Intell. Pervasive Comput.*, 2007, pp. 483–486.

[4] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, "Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots," *Int. J. Soc. Robot.*, vol. 11, no. 4, pp. 1–19, 2019.

[5] J. Ibáñez, "Emotional sea: Showing valence and arousal through the sharpness and movement of digital cartoonish sea waves," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 4, pp. 901–910, Jun. 2013.

[6] R. D. Lane, P. M. Chua, and R. J. Dolan, "Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures," *Neuropsychologia*, vol. 37, no. 9, pp. 989–997, 1999.

[7] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Trans. Affective Comput.*, vol. 10, no. 2, pp. 223–236, Apr. 2019.

[8] J. Pan, Y. Li, and J. Wang, "An EEG-based brain–computer interface for emotion recognition," in *Proc. Int. Joint Conf. Neural Netw.*, 2016, pp. 2063–2067.

[9] B. Cheng and G. Y. Liu, "Emotion recognition from surface EMG signal using wavelet transform and neural network," *J. Comput. Appl.*, vol. 28, no. 2, pp. 1363–1366, 2008.

[10] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung, "Automatic ECG-based emotion recognition in music listening," *IEEE Trans. Affect. Comput.*, vol. 11, no. 1, pp. 85–99, Jan.–Mar. 2017.

[11] R. Harper and J. Southern, "A Bayesian deep learning framework for end-to-end prediction of emotion from heartbeat," 2019. [Online]. Available: arXiv:1902.03043.

[12] X. Li, D. Song, P. Zhang, Y. Hou, and B. Hu, "Deep fusion of multi-channel neurophysiological signal for emotion recognition and monitoring," *Int. J. Data Min. Bioinformat.*, vol. 18, no. 1, p. 1, 2017.

[13] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan.–Mar. 2012.

[14] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 209–222, Jul.–Sep. 2015.

[15] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, no. 3, pp. 337–404, 1950.

[16] M. Guillaumin, J. J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. Comput. Vis. Pattern Recognit.*, 2010, pp. 902–909.

[17] Y.-R. Yeh, T.-C. Lin, Y.-Y. Chung, and Y.-C. F. Wang, "A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 563–574, Jun. 2012.

[18] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.

[19] Y. Shu and S. Wang, "Emotion recognition through integrating EEG and peripheral signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017, pp. 2871–2875.

[20] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1725–1732.

[21] S. E. Kahou *et al.*, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[22] M. Shahbe and S. Hati, "Decision fusion based on voting scheme for IR and visible face recognition," in *Proc. IEEE Comput. Graph. Imag. Visual. (CGIV)*, 2007, pp. 358–364.

[23] D. R. Kisku, M. Tistarelli, J. K. Sing, and P. Gupta, "Face recognition by fusion of local and global matching scores using DS theory: An evaluation with uni-classifier and multi-classifier paradigm," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR Workshops)*, 2009, pp. 60–65.

[24] G. Meyer and J. Mulligan, "Continuous audio-visual digit recognition using decision fusion," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (CASSP)*, 2002, p. 305.

[25] H. Song, J. J. Thiagarajan, P. Sattigeri, and A. Spanias, "Optimizing kernel machines using deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5528–5540, Nov. 2018.

[26] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2177–2185.

[27] S. Kumar, M. Mohri, and A. Talwalkar, "Sampling methods for the Nyström method," *J. Mach. Learn. Res.*, vol. 13, pp. 981–1006, Apr. 2012.

[28] S. Kumar, M. Mohri, and A. Talwalkar, "Ensemble Nyström method," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1060–1068.

[29] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015, pp. 2130–2134.

[30] Y. Zhang and Q. Yang, "Learning sparse task relations in multi-task learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 2914–2920.

[31] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 352–364, Feb. 2018.

[32] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 417–429, Jul. 2019.

[33] L. Piho and T. Tjahjadi, "A mutual information based adaptive windowing of informative EEG for emotion recognition," *IEEE Trans. Affect. Comput.*, early access, May 28, 2018, doi: 10.1109/TAFFC.2018.2840973.

[34] M. Soleymani, F. Villaro-Dixon, T. Pun, and G. Chanel, "Toolbox for emotional feature extraction from physiological Signals (TEAP)," *Front. ICT*, vol. 4, p. 1, Feb. 2017.

[35] S. Walter, J. Kim, D. Hrabal, S. C. Crawcour, H. Kessler, and H. C. Traue, "Transsituational individual-specific biopsychological classification of emotions," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 4, pp. 988–995, Jul. 2013.

[36] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, Nov. 2008.

[37] F. Aiolli and M. Donini, "EasyMKL: A scalable multiple kernel learning algorithm," *Neurocomputing*, vol. 169, pp. 215–224, Dec. 2015.

[38] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assoc.*, vol. 32, no. 200, pp. 675–701, 1937.

[39] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 1–30, 2006.

[40] C. Tzelepis, V. Mezaris, and I. Patras, "Linear maximum margin classifier for learning from uncertain data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2948–2962, Dec. 2018.

[41] L. Romeo, A. Cavallo, L. Pepa, N. Berthouze, and M. Pontil, "Multiple instance learning for emotion recognition using physiological signals," *IEEE Trans. Affective Comput.*, early access, Nov. 19, 2019, doi: 10.1109/TAFFC.2019.2954118.

[42] M. Kandemir, A. Vetek, M. Goenen, A. Klami, and S. Kaski, "Multi-task and multi-view learning of user state," *Neurocomputing*, vol. 139, pp. 97–106, Sep. 2014.

[43] J. Li, S. Qiu, Y. Shen, C. Liu, and H. He, "Multisource transfer learning for cross-subject EEG emotion recognition," *IEEE Trans. Cybern.*, early access, Mar. 27, 2019, doi: 10.1109/TCYB.2019.2904052.

[44] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, Jul.–Sep. 2014.

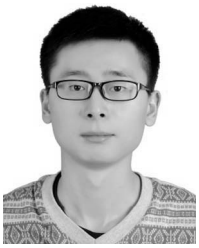
[45] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions*. New York, NY, USA: Chapman & Hall, 2018.

Xiaowei Zhang (Member, IEEE) received the Ph.D. degree in computer application technology from Lanzhou University, Lanzhou, China.

He is currently an Associate Professor with the School of Information Science and Engineering, Lanzhou University. He is interested in research fields about affective computing, multimodal fusion, and machine learning.

Jinyong Liu (Member, IEEE) received the B.Sc. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2016. He is currently pursuing the master's degree with the School of Information Science and Engineering, Lanzhou University, Lanzhou, China.

His research interests include multimodal fusion and affective computing.



Jian Shen (Graduate Student Member, IEEE) received the master's degree from the College of Computer Science and Engineering, Northwest Normal University, Lanzhou, China, in 2016. He is currently pursuing the Ph.D. degree with the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou.

His research interests include data mining, affective computing, and pervasive computing.

Shaojie Li received the bachelor's degree from the Department of Energy and Power Engineering, Northeast Electric Power University, Jilin City, China, in 2016. He is currently pursuing the master's degree with the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou, China.

His research interests include multimodal fusion and affective computing.

Kechen Hou received the first undergraduation degree from the Department of Information Security, School of Information Science and Engineering, Lanzhou University, Lanzhou, China, in 2016, and the second undergraduation degree from the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, in 2019.

His research interests include deep learning, affective computing, and multimodal fusion.

Bin Hu (Graduate Student Member, IEEE) received the bachelor's degree from the Department of Nuclear Science and Technology, Lanzhou University, Lanzhou, China, in 2017, where he is currently pursuing the master's degree with the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering.

His research interests include multimodal fusion and affective computing.

Jin Gao (Graduate Student Member, IEEE) received the bachelor's degree from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, in 2018. She is currently pursuing the master's degree with the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou, China.

Tong Zhang (Member, IEEE) received the B.S. degree in software engineering from Sun Yat-sen University, Guangzhou, China, in 2009, and the M.S. degree in applied mathematics and the Ph.D. degree in software engineering from the University of Macau, Macau, China, in 2011 and 2016, respectively.

He is currently an Assistant Professor with the School of Electronics and Information, South China University of Technology, Guangzhou. He has been working in publication matters for many IEEE conferences. His research interests include affective computing, evolutionary computation, neural network, and other machine learning techniques and their applications.

Bin Hu (Senior Member, IEEE) was born in Beijing, China, in 1965. He received the M.S. degree in computer science from the Beijing University of Technology, Beijing, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Science, Beijing.

From 2007 to 2009, he was a Reader with the Head of Context Aware Computing Research Group, School of CTN, Birmingham City University, Birmingham, U.K. Since 2009, he has been the Dean of the School of Information Science and Engineering, Lanzhou University, Lanzhou, China. He is also a Guest Professor with the CAS Center for Excellence, Brain Science and Institutes for Biological Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China. His current research interests include pervasive computing, cognitive computing, and mental healthcare.

Dr. Hu served as an Editor for *IET Communications*, *Cluster Computing*, *Wireless Communications and Mobile Computing*, the *Journal of Internet Technology*, *Security and Communication Networks* (Wiley), and *Brain Informatics* and an Associate Editor of some peer-reviewed journals in computer science, such as the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and the IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS. He is also an IET Fellow, an IET Fellow Assessment Panel Member (China Committee), a Member-at-Large of ACM China, the Director of Web Intelligence Consortium (China Committee), and a Board Member of the International Society for Social Neuroscience (China Committee).