Fusion of Granular-Ball Visual Spatial Representations for Enhanced Facial Expression Recognition

Shuaiyu Liu[1], Qiyao Shen[1], Yunxi Wang[1], Yazhou Ren[1,2*], Guoyin Wang[3]
[1]School of Computer Science and Engineer, University of Electronic Science and Technology of China
[2]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
[3]Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications
silencemobai@gmail.com, 18272214732@163.com, cyanwang234@gmail.com,
yazhou.ren@uestc.edu.cn, wanggy@cqupt.edu.cn

Abstract

Facial Expression Recognition (FER) is a fundamental problem in computer vision. Despite recent advances, significant challenges remain. Current methods primarily focus on extracting visual representations while overlooking other valuable information. To address this limitation, we propose a novel method called Component Separation and Granular-ball Space Bootstrap Fusion (CS-GBSBF), which leverages granular balls to transform visual images to spatial graphs, thereby enlarging the spatial information embedded in images. Our method separates the face into different components and utilizes the spatial information to bootstrap the fusion. More specifically, CS-GBSBF mainly consists of three crucial networks: Represent Extraction Network (REN), Represent Separation Network (RSN) and Represent Fusion Network (RFN). First, granular balls are used to represent expression images as graphs, which are fed into REN along with images. Then, RSN separates basic visual/spatial representations extracted from REN into a set of component visual/spatial representations. Next, RFN utilizes spatial representations to bootstrap component visual integration. A significant challenge in two-stream models is feature alignment, for which we have developed Attention Guidance Module (AGM) and Bootstrap Alignment Loss ($\mathcal{L}_{BA}$) in REN and RFN, respectively. Results of experiment on eight databases show that CS-GBSBF consistently achieves higher recognition accuracy than several state-of-the-art methods. The code is available at https://github.com/Lsy235/CS-GBSBF.

1 Introduction

As artificial intelligence technologies advance at a rapid pace, understanding human emotions through facial expressions has become paramount for creating intelligent systems that can interact naturally with humans. Facial expression is one way for us to express our emotions without words, and if machines can read these "silent signals", they can engage with us more effectively. According to researches [Taheri *et al.*, 2013], the variation in facial expressions is caused by the brain transmitting emotional signals, and the resulting facial muscle movements lead to morphological changes in the organs of the facial component (hereafter referred to as components), thus forming different facial expressions. As shown in Fig. 1, when someone is angry, his eyes will obviously widen, pupils will become smaller, eyebrows will be depressed, etc. These forms of expression are not only intuitive visual expressions, but also have internal spatial structural differences, such as changes in relative positions between different components, changes in the shape of individual components, etc. However, the vast majority of researches ignore the exploration of spatial information or fails to use spatial information effectively for facial expression recognition tasks.

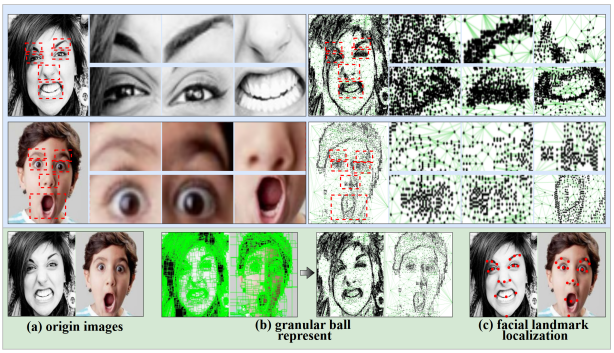


Figure 1: Component analysis of facial expressions. (b) is facial graphs gained by using granular ball to represent expressions. The top part is separation of the expression components.

Many proposed deep learning-based expression recognition methods [Psaroudakis and Kollias, 2022; Roy and Etemad, 2023; Li *et al.*, 2022] have achieved good results. However, most current recognition methods focus entirely on extracting the visual representation of expression images to achieve expression recognition tasks [Wu *et al.*, 2023b]. This type of method focuses solely on extracting visual representations and directly discards the spatial structure information of facial expressions. After the emergence of graph neural network (GNN), a considerable number of researchers applied

*Corresponding author.

it to the field of expression recognition, hoping to use the advantages of graph neural network to extract spatial structure information of visual representation [Li *et al.*, 2023]. However, the inputs transmitted to GNN by these methods are hidden features of the same image passing through different network layers, or hidden features of different patches passing through the same network layer. What these methods have in common is that the features processed by GNN are non-graph structured data or non-source data that have been extracted from images through the backbone of CNN-based or Transformer-based. As mentioned earlier, image itself has very limited spatial information, and a lot of spatial information has been lost after backbone.

With the advent of facial landmark localization, as shown in Fig. 1c, more researchers try to use landmark information to extract spatial structure information in expressions [Zhao *et al.*, 2024; Wu and Cui, 2023]. However, accurate facial landmarks require not only labor-intensive labeling to obtain but also provide limited spatial information that is highly correlated with the quality of label. As shown in Fig. 1b, we propose a novel and more effective way to represent spatial structure information of expressions by importing granular balls method [Xia *et al.*, 2023], which represents expression as graphs with rich spatial information.

Compared to images composed of tens of thousands of pixel points, graphs consist of a finite number of edges and points. In addition, the edge-point structure makes graphs richer in spatial information while the number of data points is reduced. Granular balls representation is a data representation method that reduces redundant data and improves model generalization. Applying it to 2D image data to process pixel points [Xia *et al.*, 2023], we are able to give more spatial structure to expressions while ensuring the quality of expression images. In order to solve the problem of how to effectively utilize spatial information to improve the accuracy of FER, we propose a novel FER method based on Component Separation and Space Bootstrap Fusion (CS-GBSBF), which mainly consists of Represent Extraction Network (REN), Represent Separation Network (RSN) and Represent Fusion Network (RFN).

Specifically, input data are divided into image and graph streams, then the data are first fed into the two backbone of the REN respectively. For images and graphs, the backbone is used to extract basic visual and spatial representation features. Then, two types of features are input to Vision-represent Separation Network (VRSN) and Space-represent Separation Network (SRSN) in RSN, which effectively separate the basic features into component representation features. Next, component features are input into RFN, which uses component spatial representation features to bootstrap the fusion of component visual representation features. In particular, bootstrap alignment loss is developed to solve the non-alignment problem that occurs when guide fusion of component features. Finally, a classify head network is employed for expression classification.

Our main contributions can be summarized as follows.

- A novel CS-GBSBF method is proposed to perform FER. In CS-GBSBF, granular balls representation is innovatively applied to expression images and imports a new spatially structured representation for the 2D visual expression domain.
- We innovatively propose a novel FER paradigm based on component representation, where the overall facial expression is separated into different components. RSN is developed to extract the visual representation features of components along with the spatial representation features. In RFN, bootstrap alignment loss is developed to solve the problem of unaligned fusion between components caused by RSN.
- Our method is evaluated on multiple popular real and wild databases compared with a variety of state-of-the-art methods in recent years and achieves superior results. In particular, CS-GBSBF achieves recognition accuracy of 97. 34% and 96. 74% on the databases Oulu-CASIA and SAMM, respectively.

2 Related Works

2.1 Facial Expression Recognition

Facial expression recognition (FER) is a complex and practical CV task that has gained significant attention recently.

With the substantial advancements over the past decades, deep learning has shown extraordinary performance in extracting visual representational features of expressions, making the application of deep learning models to FER tasks increasingly popular. Since then, an increasing number of researchers have used deep learning to extract more effective visual representational features [Psaroudakis and Kollias, 2022; Roy and Etemad, 2023]. Sun [Sun *et al.*, 2023a] propose a novel Feature Decomposition and Reconstruction Learning method for FER. In the process of research, successive researchers have discovered that there is spatial structure information between facial components. The graph neural network model was imported to extract the spatial information, and different graph structure data fed into GNN were designed [Liu *et al.*, 2024; Li *et al.*, 2023]. Kim [Kim *et al.*, 2023] propose that a face graph is constructed by combining the attention map with face patches and then is fed to GCN.

2.2 Granular-ball Computing

Granular ball is a method for data representation, which is inspired by the "large scale first" cognitive mechanism. Granular balls represent the data using multiple scales granular balls, which greatly reduce the amount of data while preserving the quality of the original dataset. In recent years, granular balls has made significant strides in many fields, such as clustering [Cheng *et al.*, 2023], classification [Xie *et al.*, 2024], graph generation [Xia *et al.*, 2023], etc. Zhang *et al.* [Zhang *et al.*, 2023] proposed GBRS based on incremental granularity computation for better classification tasks. Quadir *et al.* [Quadir and Tanveer, 2024] proposed a granular ball twin support vector machine to deal with significant challenges in the TSVM field.

In this paper, we import the method of granular balls representing images [Xia *et al.*, 2023] to the field of FER by structuring expression images as graphs. Image is one kind of high-density point-integrated data, while graph is the kind of

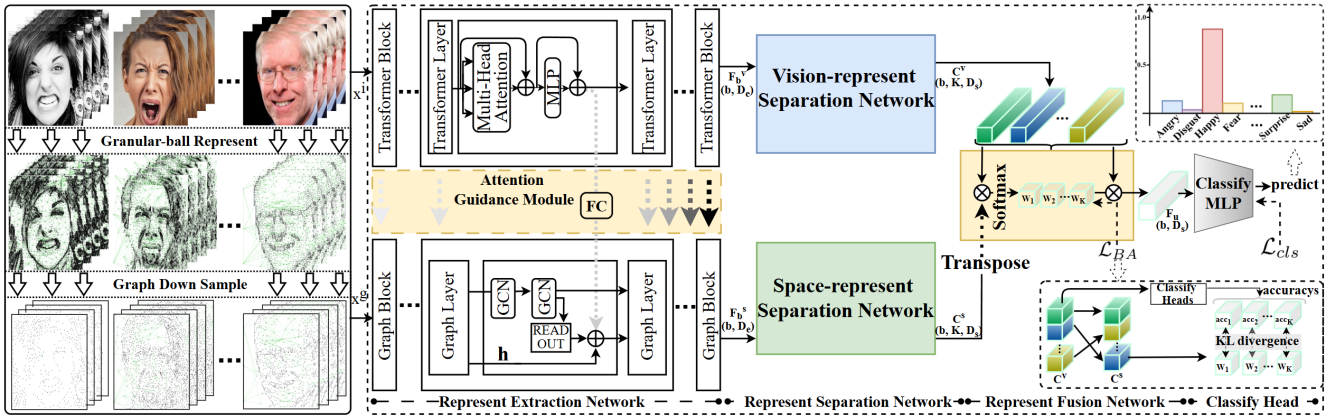


Figure 2: Overview of our proposed CS-GBSBF method. VREN (Top of REN) and SREN (Bottom of REN) form the main structure of REN. Each block contains several layers. Vision-represent Separation Network (Fig. 3a) and Space-represent Separation Network (Fig. 3b) form the main structure of RSN.

sequence-structured edge-point data. Compared with image, graph has richer spatial information, such as relative position, spatial topology, angular direction and so on. Particularly, the conversion from image to graph is highly consistent with the principles and advantages of granular balls representing data.

3 Proposed Method

3.1 Overview

The proposed CS-GBSBF method consists of a Represent Extraction Network (REN), a Represent Separation Network (RSN), a Represent Fusion Network (RFN) and a Classify Head Network (CHN). An overview is shown in Fig. 2.

Given a batch of facial expression images, we firstly use granular balls representation to generate the corresponding graphs of the images. For the too large graphs that are difficult to be processed, the model will use Graph Down Sample (GDS) to downsample graphs to obtain the graphs that are in appropriate size. Then, images are input into Vision-represent Extraction Network (VREN) in REN to extract basic visual representation features, while graphs are input into Space-represent Extraction Network (SREN) in REN to extract basic spatial representation features. In order to align the regions of features extracted by VREN and SREN, we set up an attention steering module that employs the attentional output of VREN to correctly guide the attention regions of SREN. Next, basic visual and spatial representation features will be fed to VRSN and SRSN, respectively, in order to separate and enhance the basic visual/spatial features and then gain the component visual/spatial representation features. Then, the two types of component features will be fed into RFN, which uses component spatial features to guide the fusion of the component visual features. Finally, CHN, equipped with a categorical linear layer, performs FER on the fused features.

3.2 Granular-ball Represent

With the advent of GNN, there is a growing number of methods for representing non-graph data as graph-structured data. In the case of expression images, basically all the representations follow the same paradigm, which first extracting the features of the same sized region block using visual network model, and then constructs graph-structured data to expect to enrich spatial information of data. Unlike previous methods, we propose to utilize granular balls [Xia *et al.*, 2023] to represent expression images x_j^i, that is:

$$x_j^g = GBR(x_j^i), \tag{1}$$

where GBR is the process of granular balls representation images and x_j^g is the graph after representation.

Compared to previous methods, our method preserves rather than produces spatial information from the source of images rather than the intermediate data, and extracts spatial information such as the shape and density of granular balls, as well as the relative positions and relative distances between granular balls, in a multi-granularity rather than a single-size patch manner. The basic idea of GBR algorithm is to divide the image into matrix regions of different sizes with the criterion of ensuring that the pixel values of all positions within each matrix are as similar as possible, as shown in Fig. 1b. Each of these matrices is treated as one point in graph, and the region intersecting matrices are treated as if there is one edge between the matrices.

The number of nodes of graph represented by granular balls method is highly dependent on the resolution and color complexity. An image with resolution 384×384 may extract nearly 10,000 nodes, some of which are located in the hair, clothing, and other non-expression related areas, which will import noise nodes. We use downsample to merge nodes to reduce noisy nodes while keeping the effective spatial information. Unlike the edge-focused downsample method proposed in [Xia *et al.*, 2023], we improve the downsample method with the basic idea that if a node and another node with minimal distance is merged for the downsample operation. More specific details can be viewed in Appendix[1] A.

[1]The appendix material version is available at https://github.com/Lsy235/CS-GBSBF/tree/main/supplyMaterials.

3.3 Represent Extraction Network

Given the *j-th* expression image x_j^i and the *j-th* corresponding graph x_j^g, VREN will extract the basic visual representational features $F_b^v \in \mathbb{R}^{1\times D_e}$ in x_j^i. As well as SREN will extract the basic spatial representational features $F_b^s \in \mathbb{R}^{1\times D_e}$ in x_j^g, which can be formulated as:

$$F_b^v = L_1^v(...(L_l^v(...(L_N^v(x_j^i)))) \ \ for \ \ l = 1, 2, ..., N, \quad (2)$$

$$F_b^s = L_1^s(...(L_l^s(...(L_N^s(x_j^g)))) \ \ for \ \ l = 1, 2, ..., N, \quad (3)$$

where L_l^v denotes the *l-th* network layer of the VREN and L_l^s denotes the *l-th* network layer of the SREN. N indicates the number of layers in the backbone network.

Since REN contains dual data processing streams for VREN and SREN, there are bound to be unaligned problems in the feature regions where VREN and SREN focus on. To solve this problem, we set Attention Guidance Module (AGM) in REN, which is expected to be used for the attention output of VREN to guide the attention of SREN to the corresponding region. Specially, the attention output of the *l-th* layer of VREN is summed with the output of the *l-th* layer of SREN after a layer of fully connected layer transformation, which can be formulated as:

$$out_l^s = L_l^s(out_{l-1}^s) \oplus FC(L_l^v(out_{l-1}^v)) \ \ for \ \ l = 1, ..., N, \quad (4)$$

where $out_l^s \in \mathbb{R}^{1\times D_e}$ and $out_l^v \in \mathbb{R}^{1\times D_e}$ denote the output of the *l-th* network layer of SREN and VREN, respectively. *FC* indicates a fully connected layer.

Images have better location information, while graphs have more information of relative location that can reflect spatiality. We use VREN to guide SREN. When l takes a later value, the attention of VREN is more obvious and the effect of guidance is better. In the realization, the value of l is set to N.

3.4 Represent Separation Network

F_b^v and F_b^s extracted by REN which are broadly focused on all the components of the face but not specific to one, as shown in the attention heat map in Fig. 3. As mentioned in the opening section, facial expression consists of a combination of variations of components. Therefore, in order to extract the component visual representation features $C_k^v \in \mathbb{R}^{1\times D_s}$ with the component spatial representation features $C_k^s \in \mathbb{R}^{1\times D_s}$, we propose RSN to separate F_b^v and F_b^s, which can be formulated as:

$$\begin{cases} C_k^v = sigmoid_k(FC_k(GeLU_k(FC_k(F_b^v)))), \\ C_k^s = sigmoid_k(FC_k(F_b^s)) \ \ for \ \ k = 1, ..., K, \end{cases} \quad (5)$$

where *K* denotes the number of components separated by the overall face. $sigmoid_k$ indicates the sigmoid activation function for the *k-th* component and $GeLU_k$ indicates the GeLU activation function for the *k-th* component. In addition, each component has its own network layers.

We separate the overall face into K components for the expectation that the representational features of each component correspond to the representational features of an unoverlapped region of the face. As shown in the attention heat map in Fig. 3, F_b^v and F_b^s mainly focus on the overall face. While C_k^v and C_k^s in Eq. (5) appear to split the overall attention and gradually focus on different face regions. Next, we employ sigmoid function to augment the attention. sigmoid often plays an important role in anomaly detection and logistic regression binary classification, while the meaning of C_k^v and C_k^s are whether or not to pay attention to one region of the face, which belongs to the same binary classification task. In the process of continuous supervised training, C_k^v and C_k^s tend to focus more and more on the regions where the differences between samples are large. The most obvious regions where the differences between expressions are at the region of facial component, such as mouth, eyes, eyebrows and so on. After the model has fully converged, the regions of attention for both C_k^v and C_k^s are distributed near the facial components.

3.5 Represent Fusion Network

For the task of FER, not all C_k^v and C_k^s are beneficial for recognition. At the same time, expression recognition does not require observation of the full components [Barros and Sciutti, 2021]. As shown in Fig. 1, based on the changes in the eyebrows, eyes and mouth alone, we can fully deduce that the one is in an angry spirit. Only the components that undergo change are useful, the more change, the more useful. However, due to the differences between samples, the degree of usefulness of component change for expression recognition can be inconsistent across samples. Then, to learn the fusion feature $F_u \in \mathbb{R}^{1\times D_s}$ with high generalization, adaptive weights reflecting the degree of usefulness must be designed to guide feature fusion. C_k^s extracted by SRSN is fully compatible. The spatial structure information can well reflect the degree of change that has occurred in component. When one component has a relatively large spatial structure transition, then that component must be an important representation of that individual's expression. As shown in Fig. 2, we first multiply the transposed C_k^s with C_k^v in vector counterparts, and the result obtained is passed through softmax function. Then, the weight $W \in \mathbb{R}^{1\times K}$, which reflect the degree of component change, are obtained, that is:

$$W = softmax([C_k^v \otimes C_k^{sT}]) \ \ for \ \ k = 1, ..., K, \quad (6)$$

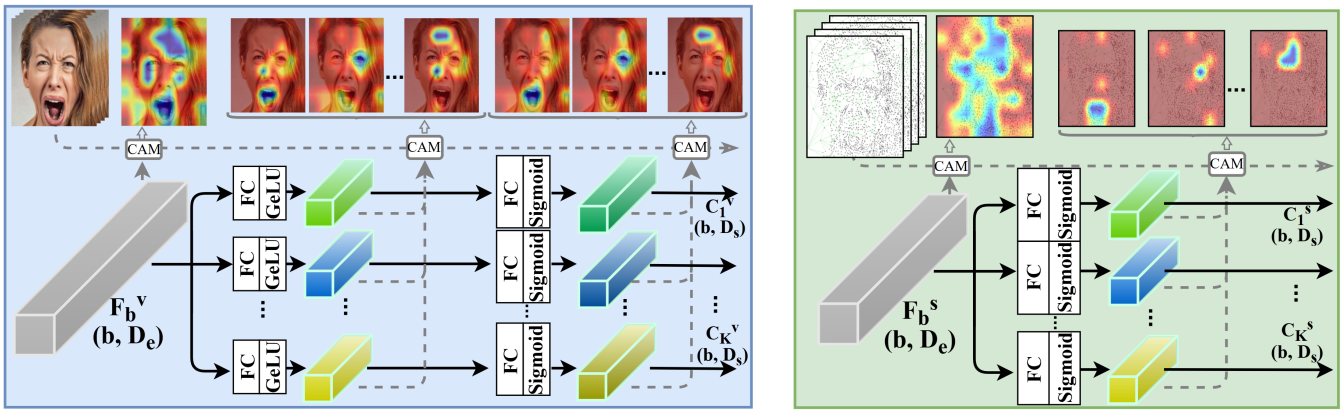
where $softmax$ denotes the softmax activation function.

C_k^s with C_k^v are learned by REN and RSN based on samples. The differences between the samples make C_k^s with C_k^v also have variations. W in Eq. (6) will adaptively modify the weight values for different samples, which makes F_u have higher generalization. Eventually, W is used to guide the fusion of visual representation features to obtain F_u. Specially, matrix multiplication of $W_k \in \mathbb{R}^{1\times 1}$ with C_k^v leads to F_u, which can be formulated as:

$$F_u = \sum_{k=1}^{K} W_k \otimes C_k^v, \quad (7)$$

where F_u will be fed to CHN for the task of FER.

Bootstrap Alignment Loss. *W* is particularly important in the whole process of feature fusion in RFN. As well as in Eq. (6), although we have given an explicit inference formula, we have neglected the question of whether the components indicated by C_k^v and C_k^s are the same component. In other words,



(a) Vision-represent Separation Network

(b) Space-represent Separation Network

Figure 3: The proposed (a) Vision-represent Separation Network (VRSN) and (b) Space-represent Separation Network (SRSN). More details of VRSN and SRSN can be found in Sec. 3.4.

although REN is designed with AGM to guide attention region alignment, whether C_k^v and C_k^s after RSN separation can still be aligned. It is obvious that RSN does not understand alignment. When the RSN separates F_b^v and F_b^s of different samples, there is a high probability that the resulting C_k^v and C_k^s cannot be aligned. To solve the problem of representation alignment, we propose Bootstrap Alignment Loss $\mathcal{L}_{BA}$. The ideal meaning of W is the degree of component change. In addition, the larger the degree of component change is, the more effective the corresponding C_k^v is for expression recognition. Then, we directly perform facial expression recognition based only on each C_k^v. Next, we can obtain K recognition accuracies, which are denoted as $ACCs \in \mathbb{R}^{1\times K}$. It can be formulated as:

$$ACCs = [acc(CHN(C_k^v), y)] \quad for \quad k = 1, ..., K, \tag{8}$$

where *acc* denotes the computational function of recognition accuracy and *y* denotes the label of expression images.

At this point, *ACCs* are also expressed as the extent to which each C_k^v contributes to the recognition of that expression. In addition, the values of *ACCs* and W are at the same order of magnitude, ranging from 0 to 1. Then, we calculate the KL divergence of *ACCs* and W to see if the distributions are consistent. If there is agreement, then it is numerically proven that W has the same distribution as *ACCs*. The meaning of *ACCs* for FER task does not alter and then the only thing that can be proven is that W reaches the desired meaning during the process of training. Mathematically, $\mathcal{L}_{BA}$ is formulated as:

$$\mathcal{L}_{BA} = KL(ACCs, W), \tag{9}$$

where *KL(..)* indicates the function of KL divergence.

3.6 Joint Loss Function

In the proposed CS-GBSBF, REN, RSN, RFN and CHN are jointly trained in an end-to-end manner. The whole network minimizes the following joint loss function:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{BA}, \tag{10}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{BA}$ represent the classification loss and bootstrap alignment loss. In this paper, we use the cross-entropy loss as the classification loss. λ denotes the regularization parameter. By minimizing the joint loss, CS-GBSBF is able to extract discriminative represent features for FER.

4 Experiments

4.1 Databases and Evaluation Setups

The databases involved in the experiments include **AffectNet-8**, **CAER-S**, **RAF-DB**, **Oulu-CASIA**, **CK+**, **SFEW 2.0**, **FER-2013** and **SAMM**, as shown in Table 1.The evaluation metric of Accuracy and UF1 are selected. UF1, named unweighted f1-score, is one metric that is more effective in evaluating on multicategorical unbalanced databases, and is commonly used in the field of microexpression.

Databases	Size	Resolution	Class	Type
AffectNet-8 [2017]	37,303	100 × 100	8-cla	wild
CAER-S [2019]	70,000	712 × 400	7-cla	wild
RAF-DB [2017]	29,672	100 × 100	7-cla	wild
Oulu-CASIA [2011]	64,913	320 × 240	6-cla	lab
CK+ [2010]	774	48 × 48	7-cla	lab
SFEW [2011]	1,766	143 × 181	7-cla	wild
FER-2013 [2013]	35,886	48 × 48	9-cla	wild
SAMM [2016]	11,816	960 × 650	7-cla	lab

Table 1: Detailed information about all databases. The type of wild indicates that this database belongs to the in-the-wild database, and the type of lab indicates in-the-lab.

For the database divided by the original test database, we keep it constant. And for the undivided database, we follow the rules most methods [Gan *et al.*, 2019] adopt for classification.

4.2 Experimental Settings

For each database, all the facial images need to be represented by granular balls, followed by graph downsample operation

to obtain the input graphs. During the process of training, images are further resized to 224×224 and then a random crop is applied for data augmentation. During the test process, images are resized to the size of 224×224 and then fed into the trained model along with graphs. The CS-GBSBF method is implemented with the Pytorch toolbox, employing swinT-base [Liu *et al.*, 2021] as the backbone of VREN and utilizing GCN to build the backbone of SREN, where the swinT-base is pre-trained on the ImageNet-1K database.

The dimensions of both F_b^s and F_b^s are $\mathbb{R}^{b\times 1024}$ and C_k^v and C_k^s have dimensions of $\mathbb{R}^{b\times 128}$. Based on extensive experiments in the validation set, the value of λ in Eq. (10) is empirically set to 0.1, while the value of the hyperparameter K is set to 9. We train CS-GBSBF in an end-to-end manner with one single NVIDIA GeForce RTX 4080 SUPER for 40 epochs, and the batch size for all databases is set to 16. Then our model is trained using the Adam algorithm with an initial learning rate of 0.0001, weight decay = 0.01.

#1	#2	#3	$Acc\uparrow /UF1\uparrow /UAR\uparrow$		
			CAER-S	CK+	FER-2013
✗	✗	✗	89.4/88.6/88.5	98.3/98.0/97.7	82.2/80.1/79.3
✗	✓	✗	89.0/88.1/88.1	98.1/97.8/97.4	81.9/79.9/78.7
✓	✓	✗	90.1/89.8/89.7	99.0/98.5/98.5	82.5/81.2/80.6
✗	✓	✓	92.3/91.7/91.7	99.1/98.9/98.8	83.3/81.6/81.2
✓	✓	✓	**93.1/92.7/92.6**	**100./100./100.**	**85.0/82.1/81.9**

Table 2: The ablation study of the proposed important modules on the validation set of the three databases. #1 denotes AGM module, #2 denotes SRSN module and #3 denotes $\mathcal{L}_{BA}$ module.

Database	Backbone				
	ResNet -34	ResNet -50	VGG -16	swinT -base	swinT -large
CAER-S	91.82	91.94	92.57	**93.12**	92.73
FER-2013	82.04	82.81	81.72	**84.98**	83.69
RAF-DB	87.91	88.23	85.72	90.57	**90.71**

Table 3: The ablation study of backbone in VREN. Note that the evaluation metric remains *Accuracy (%)*.

4.3 Ablation Studies

The ablation studies in this section are recorded in the validation set. The ablation studies set up in this subsection focus on the important modules of the proposed design (AGM, SRSN, $\mathcal{L}_{BA}$), variation in the value of the component number K, and the weight value λ. The effect of each module in the proposed method is explored through ablation studies on the validation set. Ablation studies for the other model parameters are shown in Appendix B.

Influence of the key modules. When only the SRSN module is introduced without any bootstrap alignment methods, C^v do not contribute to improving the recognition accuracy compared to when SRSN is not included. This also shows that the direct introduction of spatial representations is not suitable for FER. When only the AGM module is added, which only ensures that C_k^v and C_k^s are aligned at the stage of REN, the accuracy of our method on the three validation database drops by an average of 1.93%. When both SRSN and $\mathcal{L}_{BA}$ modules are added, the extent of the decline is attenuated. The two ablation experimental schemes differ in model design in that the former tends to realize bootstrap alignment in REN and the latter tends to realize bootstrap alignment in RFN. The experiments show that the alignment design models for the front and rear modules work, with the latter being somewhat more effective.

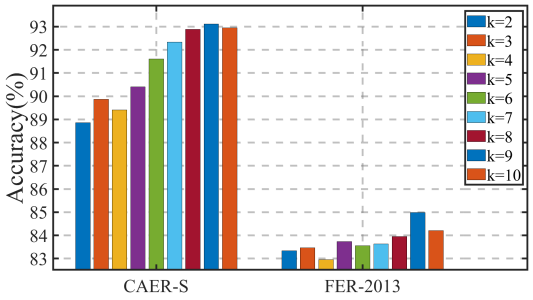
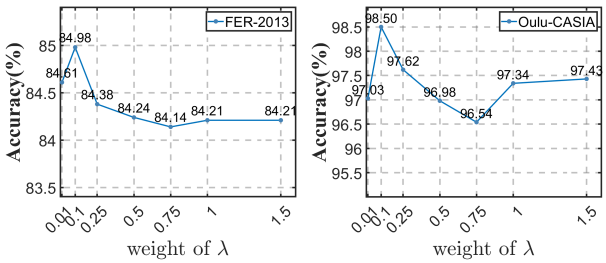


Figure 4: Ablation studies for the different value of K on the validation set of CAER-S and FER-2013 databases.



(a) Influence in FER-2013 (b) Influence in Oulu-CASIA

Figure 5: Ablation studies for the different value of λ on the validation set of FER-2013 and Oulu-CASIA databases.

Influence of the *K* value. We evaluate the recognition performance of our method with the different values of K in Eq. (5), as shown in Fig. 4.

In particular, we can observe that our method achieves the best recognition accuracy when value of K is set to 9. When separated into too many components, the basic representations are separated too finely, weakening the representation in someone component. However, too few results in merging of components, which makes it impossible to obtain spatial information between the merging components and results in difficult to bootstrap better feature fusion.

Influence of the λ weight. As shown in Fig. 5, we can see that CS-GBSBF achieves the best performance when the value of λ is set to 0.1. When a large weight of $\mathcal{L}_{BA}$ is set, the model will ignore the loss of cross-entropy, resulting in a worse classification effect. However, when the value of λ is set to 0, the lack of consideration of feature alignment in RFN leads to worse fusion.

Influence of backbone in VREN. We show our CS-

CAER-S		Oulu-CASIA		RAF-DB		AffectNet-8	
Methods	Acc.↑	Methods	Acc.↑	Methods	Acc.↑	Methods	Acc.↑
MA-net [2021]	88.42	FDRL [2021]	88.26	RUL [2021]	88.98	RUL [2021]	60.66
Poster [2023]	92.73	GEPm [2022]	89.05	FDRL [2021]	89.47	Ada-CM [2022]	57.42
SMResNet [2023]	88.52	SPNDL [2023a]	90.14	Ada-CM [2022]	84.42	CDB [2022]	64.23
Poster++ [2024]	93.00	SSF-ViT [2023]	88.06	RUL-C [2024]	89.51	Poster [2023]	63.34
HAM [2024]	92.86	im-cGAN [2023b]	93.34	Poster++ [2024]	**91.09**	Poster++ [2024]	63.77
CS-GBSBF (ours)	**93.06**	CS-GBSBF (ours)	**97.34**	CS-GBSBF (ours)	90.10	CS-GBSBF (ours)	**64.60**
FER-2013		CK+		SFEW		SAMM	
Methods	Acc.↑	Methods	Acc.↑	Methods	Acc.↑	Methods	Acc.↑/UF1↑
RUL [2021]	73.75	FDRL [2021]	99.54	MA-net [2021]	59.40	MESTD [2021]	91.90/89.60
ResMask [2021]	76.82	FERS [2022]	97.83	FDRL [2021]	62.16	LR-GAC [2021]	88.24/–
Ad-Corre [2022]	72.03	SPNDL [2023a]	99.69	FERS [2022]	35.91	SparseT [2022]	80.15/–
SSF-ViT [2023]	74.95	HCCL [2023]	93.00	Ada-CM [2022]	52.43	μ-BERT [2023]	87.35/83.36
RUL-C [2024]	71.83	SSF-ViT [2023]	98.96	SSF-ViT [2023]	63.69	HTNet [2024]	86.67/81.31
CS-GBSBF (ours)	**83.99**	CS-GBSBF (ours)	**100.00**	CS-GBSBF (ours)	**65.56**	CS-GBSBF (ours)	**96.74/95.70**

Table 4: Performance comparisons among different methods on the test set of several public FER databases. The best results are boldfaced amd the second results are underlined.

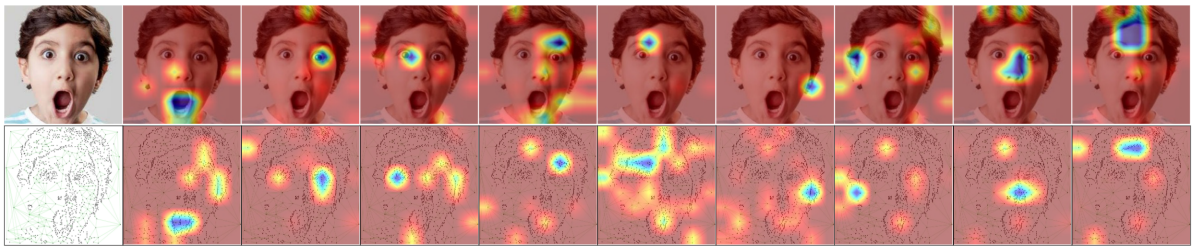


Figure 6: Visualization of component representation features C^v and C^s by CAM. More visualization can be viewed in Appendix B.

GBSBF with different backbones in Table 3, and give the accuracy of CS-GBSBF as a metric for comparison. CS-GBSBF with different backbones can work well. Results prove it that swinT-base [Liu *et al.*, 2021] is able to extract F_b^v that are more suitable for CS-GBSBF.

4.4 Comparisons with State-of-the-Art Methods

Table 4 shows the comparison results between the proposed CS-GBSBF method and several state-of-the-art FER methods on all databases shown in Table 1. Only a different portion of the databases in Table 1 is generally selected in the article for each comparison method, and not all comparison methods have open source code. Therefore, for each database, we try to find the state-of-the-art approach to compare with our proposed method.

As shown in Table 4, the proposed CS-GBSBF method basically improves the recognition accuracy on all databases with an average improvement of 2.48%. This is significantly improved by 2.71% and 7.17% in Oulu-CASIA and FER-2013, respectively. Even on the micro-expression database, SAMM, CS-GBSBF far outperforms the compared methods on Acc. and UF1. It also improves on all other databases except RAF-DB. These comparison experiments with the state-of-the-art method clearly demonstrate our advantage in FER as well as the existence of the proposed CS-GBSBF method with high recognition generalization on different databases.

4.5 Visualization

Attention visualization. We use CAM [Zhou *et al.*, 2016] to draw the attention heat maps based on C^v and C^s respectively, as shown in Fig. 6. We can observe that after passing through RSN, the attentional attention regions of each C^v and C^s are different but the attention regions of both correspond to each other. All the regions are distributed near the different component organs of face, which is one of the reasons why we call proposed method component separation. Similarly in exploring the attention regions of C^s on graphs, the graph is visualized and plotted, and then the regions of attention obtained by CAM are appended to the visualized graphs in the same way as images.

5 Conclusion

In this paper, we propose a novel CS-GBSBF method consisting of REN, RSN, and RFN. To address the problem that current researches focus too much on visual information and ignore other valid information, we import granular balls to represent images and extract visual and spatial information from the source of data. In our method, to address the arising problem of features alignment, we propose AGM and $\mathcal{L}_{BA}$ in the stages of REN and RFN, respectively. We conducted ablation studies to demonstrate the effectiveness of our proposed modules. The extensive experimental results on multiple databases demonstrate the superiority of our method for facial expression recognition.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Nos. 62221005, 62476052, 62450043, 62222601), Sichuan Science and Technology Program (No. 2024NSFSC1473), and Central Guidance for Local Science and Technology Development Fund Projects (No. 2024ZYD0268).

References

[Barros and Sciutti, 2021] Pablo Barros and Alessandra Sciutti. I only have eyes for you: The impact of masks on convolutional-based facial expression recognition. In *CVPR*, pages 1226–1231, 2021.

[Bisogni *et al.*, 2022] Carmen Bisogni, Aniello Castiglione, Sanoar Hossain, Fabio Narducci, and Saiyed Umer. Impact of deep learning approaches on facial expression recognition in healthcare industries. *IEEE TII*, 18(8):5619–5627, 2022.

[Chen *et al.*, 2023] Xuanchi Chen, Xiangwei Zheng, Kai Sun, Weilong Liu, and Yuang Zhang. Self-supervised vision transformer-based few-shot learning for facial expression recognition. *INFORM SCIENCES*, 634:206–226, 2023.

[Chen *et al.*, 2024] Ning Chen, Ven Jyn Kok, and Chee Seng Chan. Enhancing facial expression recognition under data uncertainty based on embedding proximity. *IEEE Access*, 2024.

[Cheng *et al.*, 2023] Dongdong Cheng, Ya Li, Shuyin Xia, Guoyin Wang, Jinlong Huang, and Sulan Zhang. A fast granular-ball-based density peaks clustering algorithm for large-scale data. *IEEE TNNLS*, 2023.

[Davison *et al.*, 2016] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. Samm: A spontaneous micro-facial movement dataset. *IEEE T-AFFC*, 9(1):116–129, 2016.

[Dhall *et al.*, 2011] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *ICCV*, pages 2106–2112. IEEE, 2011.

[Fard and Mahoor, 2022] Ali Pourramezan Fard and Mohammad H Mahoor. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, 10:26756–26768, 2022.

[Gan *et al.*, 2019] Yee Siang Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, and Lit-Ken Tan. Off-apexnet on micro-expression recognition system. *SIGNAL PROCESS-IMAGE*, 74:129–139, 2019.

[Goodfellow *et al.*, 2013] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *ICONIP 2013, Daegu, Korea, Nov. 3-7, 2013, Proc., Part III*, pages 117–124. Springer, 2013.

[Kim *et al.*, 2023] Hyeongjin Kim, Jong-Ha Lee, and Byoung Chul Ko. Facial expression recognition in the wild using face graph and attention. *IEEE Access*, 11:59774–59787, 2023.

[Kumar and Bhanu, 2021] Ankith Jain Rakesh Kumar and Bir Bhanu. Micro-expression classification based on landmark relations with graph attention convolutional network. In *CVPR*, pages 1511–1520, 2021.

[Lee *et al.*, 2019] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *ICCV*, pages 10143–10152, 2019.

[Li *et al.*, 2017] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, pages 2852–2861, 2017.

[Li *et al.*, 2022] Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, and Xinbo Gao. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In *CVPR*, pages 4166–4175, 2022.

[Li *et al.*, 2023] Chunlei Li, Xiao Li, Xueping Wang, Di Huang, Zhoufeng Liu, and Liang Liao. Fg-agr: Fine-grained associative graph representation for facial expression recognition in the wild. *IEEE TCSVT*, 34(2):882–896, 2023.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[Liu *et al.*, 2023] Juan Liu, Min Hu, Ying Wang, Zhong Huang, and Julang Jiang. Symmetric multi-scale residual network ensemble with weighted evidence fusion strategy for facial expression recognition. *Symmetry*, 15:1228, 2023.

[Liu *et al.*, 2024] Shuai Liu, Shichen Huang, Weina Fu, and Jerry Chun-Wei Lin. A descriptive human visual cognitive strategy using graph neural network for facial expression recognition. *INT J MACH LEARN CYB*, 15:19–35, 2024.

[Lucey *et al.*, 2010] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPR 2010*, pages 94–101. IEEE, 2010.

[Mao *et al.*, 2024] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, Aibin Huang, and Yigang Wang. Poster++: A simpler and stronger facial expression recognition network. *PATTERN RECOGN*, 2024.

[Mollahosseini *et al.*, 2017] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE T AFFECT COMPUT*, 10(1):18–31, 2017.

[Nguyen *et al.*, 2023] Xuan-Bac Nguyen, Chi Nhan Duong, Xin Li, Susan Gauch, Han-Seok Seo, and Khoa Luu.

Micron-bert: Bert-based facial micro-expression recognition. In *CVPR*, pages 1482–1492, 2023.

[Pham *et al.*, 2021] Luan Pham, The Huynh Vu, and Tuan Anh Tran. Facial expression recognition using residual masking network. In *ICPR 2020*, pages 4513–4519. IEEE, 2021.

[Psaroudakis and Kollias, 2022] Andreas Psaroudakis and Dimitrios Kollias. Mixaugment & mixup: Augmentation methods for facial expression recognition. In *CVPR*, pages 2367–2375, 2022.

[Quadir and Tanveer, 2024] A Quadir and M Tanveer. Granular ball twin support vector machine with pinball loss function. *IEEE TCSS*, 2024.

[Roy and Etemad, 2023] Shuvendu Roy and Ali Etemad. Active learning with contrastive pre-training for facial expression recognition. In *ACII*, pages 1–8. IEEE, 2023.

[Ruan *et al.*, 2021] Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Shen, and Hanzi Wang. Feature decomposition and reconstruction learning for effective facial expression recognition. In *CVPR*, pages 7660–7669, 2021.

[Singh *et al.*, 2023] Rajesh Singh, Sumeet Saurav, Tarun Kumar, Ravi Saini, Anil Vohra, and Sanjay Singh. Facial expression recognition in videos using hybrid cnn & convlstm. *Int. J. Inf. Technol.*, 15(4):1819–1830, 2023.

[Sun *et al.*, 2023a] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *ACM MM 2023*, pages 6110–6121, 2023.

[Sun *et al.*, 2023b] Zhe Sun, Hehao Zhang, Jiatong Bai, Mingyang Liu, and Zhengping Hu. A discriminatively deep fusion approach with improved conditional gan (im-cgan) for facial expression recognition. *PATTERN RECOGN*, 135:109157, 2023.

[Taheri *et al.*, 2013] Sima Taheri, Vishal M Patel, and Rama Chellappa. Component-based recognition of facesand facial expressions. *IEEE T AFFECT COMPUT*, 4(4):360–371, 2013.

[Tao and Duan, 2024] Huanjie Tao and Qianyue Duan. Hierarchical attention network with progressive feature fusion for facial expression recognition. *NEURAL NETWORKS*, 170:337–348, 2024.

[Wang *et al.*, 2024] Zhifeng Wang, Kaihao Zhang, Wenhan Luo, and Ramesh Sankaranarayana. Htnet for micro-expression recognition. *NEUROCOMPUTING*, 602:128196, 2024.

[Wu and Cui, 2023] Zhiyu Wu and Jinshi Cui. La-net: Landmark-aware learning for reliable facial expression recognition under label noise. In *ICCV*, pages 20698–20707, 2023.

[Wu *et al.*, 2023a] Zhenqian Wu, Xiaoyuan Li, Yazhou Ren, Xiaorong Pu, Xiaofeng Zhu, and Lifang He. Self-paced neutral expression-disentangled learning for facial expression recognition. In *ACAIT*. IEEE, 2023.

[Wu *et al.*, 2023b] Zhenqian Wu, Yazhou Ren, Xiaorong Pu, Zhifeng Hao, and Lifang He. Generative neutral features-disentangled learning for facial expression recognition. In *ACM MM*, pages 4300–4308, 2023.

[Xia and Wang, 2021] Bin Xia and Shangfei Wang. Micro-expression recognition enhanced by macro-expression from spatial-temporal domain. In *IJCAI*, 2021.

[Xia *et al.*, 2023] Shuyin Xia, Dawei Dai, Long Yang, Zhany Li, Danf Lan, Hao Zhu, and Guoy Wang. Graph-based representation for image based on granular-ball. *arXiv preprint arXiv:2303.02388*, 2023.

[Xie *et al.*, 2024] Qin Xie, Qinghua Zhang, Shuyin Xia, Fan Zhao, Chengying Wu, Guoyin Wang, and Weiping Ding. Gbg++: A fast and stable granular ball generation method for classification. *IEEE TETCI*, 2024.

[Zeng *et al.*, 2022] Dan Zeng, Zhiyuan Lin, Xiao Yan, Yuting Liu, Fei Wang, and Bo Tang. Face2exp: Combating data biases for facial expression recognition. In *CVPR*, pages 20291–20300, 2022.

[Zhang and Yu, 2022] Jing Zhang and Huimin Yu. Improving the facial expression recognition and its interpretability via generating expression pattern-map. *PATTERN RECOGN*, 129:108737, 2022.

[Zhang *et al.*, 2021] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *NeurIPS*, 34:17616–17627, 2021.

[Zhang *et al.*, 2023] Qinghua Zhang, Chengying Wu, Shuyin Xia, Fan Zhao, Man Gao, Yunlong Cheng, and Guoyin Wang. Incremental learning based on granular ball rough sets for classification in dynamic mixed-type decision system. *IEEE Trans Knowl Data Eng*, 35:9319–9332, 2023.

[Zhao *et al.*, 2011] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *IMAGE VISION COMPUT*, 29(9):607–619, 2011.

[Zhao *et al.*, 2021] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE T IMAGE PROCESS*, 30:6544–6556, 2021.

[Zhao *et al.*, 2024] Daqi Zhao, Jingwen Wang, Haoming Li, and Deqiang Wang. Landmark-based adaptive graph convolutional network for facial expression recognition. *IEEE Access*, 2024.

[Zheng *et al.*, 2023] C Zheng, M Mendieta, and C Chen. Poster: a pyramid cross-fusion transformer network for facial expression recognition. in 2023 ieee. In *ICCVW*, 2023.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.

[Zhu *et al.*, 2022] Jie Zhu, Yuan Zong, Hongli Chang, Yushun Xiao, and Li Zhao. A sparse-based transformer network with associated spatiotemporal feature for micro-expression recognition. *IEEE SIGNAL PROC LET*, 29:2073–2077, 2022.