IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 26, 2021, accepted June 8, 2021, date of publication June 10, 2021, date of current version June 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3088340

D2C-Based Hybrid Network for Predicting Group Cohesion Scores

DANG XUAN TIEN[1], HYUNG-JEONG YANG[2], (Member, IEEE), GUEE-SANG LEE[2], (Member, IEEE), AND SOO-HYUNG KIM[2], (Member, IEEE)

[1]AISIA Research Laboratory, Ho Chi Minh City 700000, Vietnam
[2]Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, South Korea

Corresponding author: Soo-Hyung Kim (shkim@jnu.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) under Grant NRF-2020R1A4A1019191 and Grant NRF-2018R1D1A3A03000947.

ABSTRACT Group cohesiveness represents the bonding between members in a group. Indeed, a group with high cohesiveness may easily reach success in their task. Therefore, the most critical element that affects the success of a group is group cohesiveness, which is estimated by Group Cohesion Score (GCS). This study proposed an automatic GCS estimation system for the 7th Emotion Recognition in the Wild (EmotiW 2019) challenge in the task of the Group Cohesion Prediction. We proposed a multi-stream hybrid network based on scene-level, skeleton-level, UV coordinates-level, mid-fusion, and face-level, followed by late-fusion to combine these approaches. We also developed a joint training method called Discrete labels to Continuous scores (D2C), where discrete labels (categorical labels) directly participate in generating continuous scores. Our proposed method achieved 0.416 mean squared error on the testing set of the EmotiW 2019 dataset and became a state-of-the-art in this challenge. Furthermore, to confirm the ability of the proposed D2C method, we performed experiments on the AffectNet database and obtained relatively better results than state-of-the-art approaches.

INDEX TERMS Group cohesion, hybrid network, discrete labels, continuous labels, valence/arousal.

I. INTRODUCTION

Group cohesiveness is a fundamental theoretical concept in most studies of group behavior, social psychology, and sport psychology [1], [2]. Group cohesiveness represents the bonding of people in a group, where higher cohesiveness means stronger bonding among group members. Myers observed that while the successful teams showed improvement in cohesiveness, the opposite was true for the unsuccessful teams [3]. Based on this, group cohesiveness is the most critical factor and an essential measurement for the success of a group. According to psychological researches, various factors that affect group cohesion include members' similarity [4], group size [5], group success [6] and external competition and threats [7], [8]. Fig. 1 illustrates Group Cohesion Score (GCS) created to estimate cohesiveness among group members [9].

Nowadays, people share their photos on social networking services, such as Facebook, Instagram, and Twitter. Thousands of photos uploaded on these services are taken in formal and casual events like graduation ceremonies, birthday parties, family reunions, and social campaigns. These photos possess the valuable information that can be used to investigate group behavior, such as group cohesion. Besides that, AI technologies, especially Deep Learning methods, have flourished and continued showing their superior performances in various fields of computer vision, natural language processing, medical imaging, etc. The available resources motivate researchers to build an automatically estimate group cohesion system.

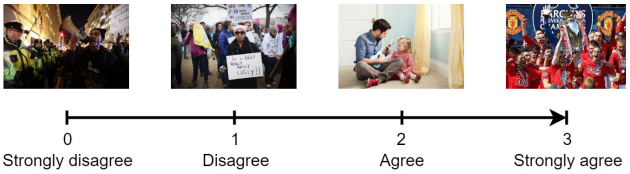


FIGURE 1. Group cohesion score.

Ghosh *et al.* [10] provided a database for group cohesion that was applied in the Group Cohesion sub-task of

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

the 7th Emotion Recognition in the Wild (EmotiW 2019) Challenge [11]. In Ghosh's paper, two approaches were proposed to build models for GCS prediction, namely image-level (scene-level) and face-level. Particularly, they showed that holistic scene and face-level information have valuable contributions to the perception of cohesion. Moreover, group emotion also affects group cohesion prediction while implementing the joint training method. This indicates that there is a correlation between group cohesion and group emotion.

The effect of scene-level and face-level play a vital role in the group cohesion prediction system, however, does not get high performances. This observation inspired us to exploit more approaches for system performance improvement. We discovered that the skeleton and UV coordinates (also known as texture coordinates which define a map of a 2D image onto a surface in 3D space [12]) also have positive contributions to the system. In this study therefore, we proposed a multi-stream hybrid network based on scene-level, skeleton-level, UV coordinates-level (UVC-level), mid-fusion, and face-level, followed by late-fusion being used to combine these approaches. We developed the Discrete labels to Continuous scores (D2C) method to predict labels in continuous dimensional from discrete categorical. The method's idea is an improvement of the joint training method in Ghosh's paper [10] where discrete labels directly participate in generating continuous labels instead. In this case, discrete labels and continuous labels were group emotion and group cohesion labels, correspondingly. Our method achieves superior performances among the Group Cohesion sub-challenge participants, a Mean Squared Error (MSE) of 0.517 and 0.416 on the validation and testing sets, respectively.

This study extends the work in conference paper [13] with detail explanation of D2C method and extensive experiments on AffectNet Database [14]. Moreover, we also perform the experiments to study the effect of parameters in the D2C's loss function. Our method can be applied not only in the group cohesion issues but also in any problem whose data contains two kinds of labels: discrete categorical and continuous dimensional. To demonstrate the ability of D2C method, we measured the performance of our proposed methods on the AffectNet database and achieved comparatively better results than other approaches.

The rest of the paper is organized as follows: We first revise current works related to the group-level cohesion issues in Section 2. Section 3 describes details of our proposed method. Experiments and results are discussed in Section 4. Section 5 gives conclusions and future research directions.

II. RELATED WORKS

Cohesiveness is a fascinating topic in social psychology because studies in this field major on group behavior, group dynamics [1], [2], [15]. Psychologists define cohesiveness as the number of members of a group liking each other or the amount of friendship between group members [16] or the ''glue'' that retains a group together [17]. Actually, a group with high cohesiveness may easily obtain success in their task [3]. Therefore, one of the most important factor that affects success in a group is group cohesiveness [18]. Many pieces of psychological research argued that group cohesion depends on several factors namely members' similarity [4], group size [5], group success [6] and external competition and threats [7], [8]. Furthermore, Treadwell *et al.* [9] created GCS (shown in Fig. 1) to evaluate cohesiveness among group members based on a 4-point scale of *strongly disagree (SD), disagree (D), agree (A) and strongly agree (SA)*, or can be represented by numbers from 0 to 3, correspondingly [10]. These studies have formed the fundamental knowledge for the evolution of social psychology especially, group cohesiveness.

Nowadays, the rapid growth of Machine Learning and Deep Learning inspired scientists to build an automatic system for estimating GCS. Hung *et al.* [19] applied Support Vector Machine (SVM) to estimate group cohesion on the audio-video of a group meeting data. This is one of the first studies that utilizes machine learning to predict group cohesion of a group of people in videos. Recently, the EmotiW 2019 presented the first study of group cohesion prediction in statics images [10], [11]. The EmotiW organizers provided a database for group cohesion, which is an upgrade version of Group Affect Database (GAF 3.0) [20] with group cohesion labels (GAF-Cohesion). GAF-Cohesion database contains not only the GC labels, but also the GE labels, namely *positive, neutral,* and *negative*. Besides that, they introduced two approaches to estimate GCS: (1) **Image-level**, trained on InceptionV3 [21] with ImageNet [22] weights; (2) **Face-level** applied transfer learning CapsNet [23], which are trained on RAF-DB [24], [25]. In our research, we added more approaches, such as skeleton-level and UVC-level, to enrich our hybrid network.

Using a hybrid network that is a combination of visual cues such as scene, skeleton and face has become a widespread in studies of predicting group cohesion [26], [27]. Guo *et al.* [26] and Zhu *et al.* [27] also used hybrid network, but their networks did not include blended features of visual cues which can be a critical factor to enhance the performance of hybrid networks. As demonstrated in our mid-fusion method, blended features have a remarkable contribution to the performance of our hybrid network. Ghosh *et al.* [10] demonstrated that joint training on both group emotion and group cohesion gives higher performances than individual training. Indeed, the most critical observation, which was presented in Ghosh's paper, is the correlation between group emotion and group cohesion. However, Ghosh's method and Guo's method [26] are hard to tune and improve performances because the group emotion do not directly join in predicting group cohesion, i.e., their performances cannot directly affect each other. In this paper therefore, we presented the D2C method to solve drawbacks in approaches of Ghosh and Guo.

To evaluate performance on GAF-Cohesion, the EmotiW 2019 organizers have suggested using Mean Squared Error (MSE). For AffectNet database [14], we use Root Mean Square Error (RMSE) and Concordance Correlation Coefficient (CCC) [28] to measure performance.

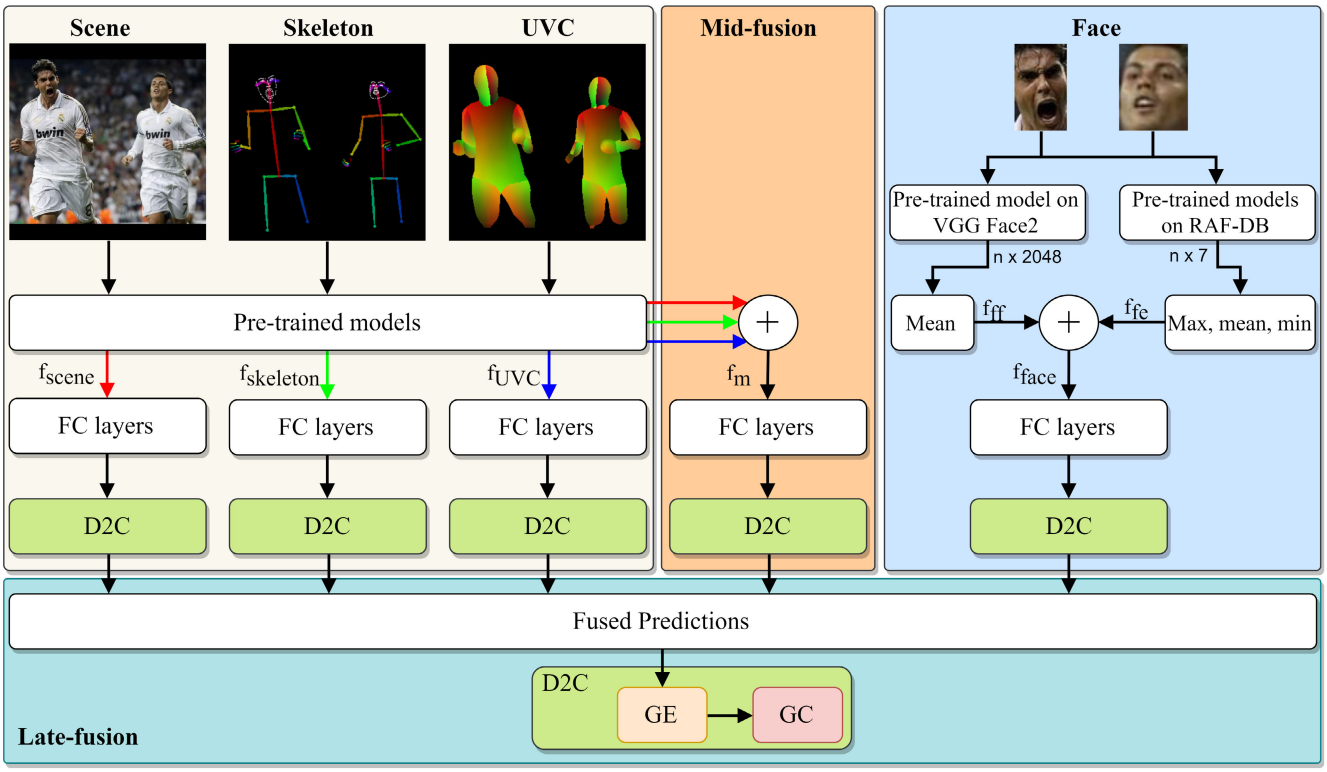


FIGURE 2. Pipeline of our multi-stream hybrid network. The network is a combination of multi-type of visual cues models, such as scene, skeleton, UVC, mid-fusion and faces. In face-level, *n* is the number of faces in an image.

III. PROPOSED METHOD

Fig. 2 illustrates the proposed hybrid network containing five independently trained streams; scene, skeleton, UV coordinates (UVC), mid-fusion, and face. The predicted values are fused by using the late-fusion method. We train a multi-layer perceptron (MLP) that comprises of fully connected layers and the D2C block, with deep features extracted from the state-of-the-art Convolutional Neural Network (CNN) models.

A. SCENE-LEVEL, SKELETON-LEVEL AND UVC-LEVEL ANALYSIS

Holistic scene and contextual information play an essential role in the group cohesion prediction system [10]. Therefore, the scene-level analysis is necessary to understand the scenario in images. In this approach, we used three pre-trained models, namely ResNet50 [29], InceptionV3 [21], and NasNet [30], to extract features of images. These networks have been state-of-the-art models of ImageNet database [22]. The dimensions of feature vectors extracted by ResNet50, InceptionV3, and NasNet are 2048, 2048, and 1056, respectively. The image features are then fed to an MLP to predict the GCS. The scene-level is illustrated in Fig. 2.

Skeleton analysis can have valuable contributions to the perception of cohesion. Skeleton features demonstrate the gestures and structures of a group of people. In this study, we used OpenPose [31]–[33] to obtain the skeleton of each image. OpenPose is an open-source library and widely used to extract the human body, facial, hand and foot keypoints in images.

Additionally, we realize that the 3D surface of a human body is also a crucial element that can improve the performance of our hybrid network. Thus, DensePose [34] library is applied to produce UV coordinates representing the 3D surface of the human body. Both skeleton-level and UVC-level networks are shown in Fig. 2. In these approaches, feature extractors, model modification, training parameters and training methods are the same as the scene-level.

B. FACE-LEVEL ANALYSIS

Ghosh *et al.* [10] showed that besides the scene-level, the facial expression of each face in an image is a useful cue for predicting group cohesion. Additionally, the facial features also have valuable contributions to face-level analysis. These observations motivated us to build a model for predicting the GCS based on a combination of facial features and facial expressions. Fig. 2 illustrates the structure of face-level approach.

We extract faces using recent state-of-the-art face detector, PyramidBox [35]. Nonetheless, in practice, PyramidBox cannot detect tiny faces, thus TinyFace [36] is used.

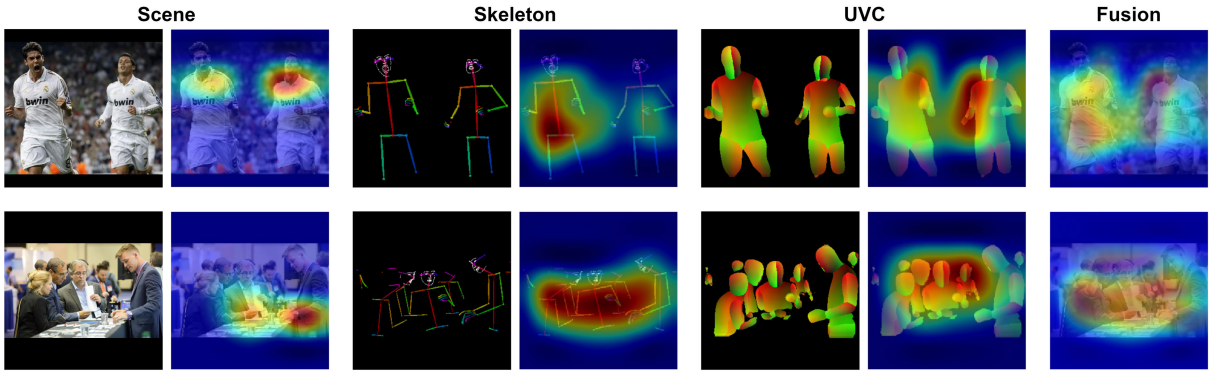


FIGURE 3. Feature maps of scene-level, skeleton-level, UVC-level and fusion are extracted by Grad-CAM [38]. The fusion feature becomes powerful because It gets the advantages of the other levels. Moreover, expanding the region of interest gives more knowledge about contextual and structure of a group in images, which can enhance the performance of our hybrid network. [Best viewed in color].

To get the facial features, we used the ResNet50 model (ResNet-VGGFace), which was trained on a large-scale face database introduced by Visual Geometry Group [37]. Each face has a feature vector with 2048 dimensions. Since the amount of faces in each image is variable, we pool the feature vectors by computing average to ensure the output vectors' sizes are the same on every image. $\mathbf{F}_{ff}$ denote the output vector of facial feature.

To extract facial expression, we first fine-tune models on large-scale facial expression databases, namely, AffectNet [14] and Real-world Affective Faces Database (RAF-DB) [24], [25]. We then utilized these models to get expression probability predictions for each of the faces present in an image. The facial expression vector created by calculating the average, maximum, and minimum on the facial expression vector of each face is denoted by $\mathbf{F}_{fe}$. The size of $\mathbf{F}_{fe}$ was 21.

After that, we concatenate all feature vectors, $\mathbf{F}_{face} = \mathbf{F}_{ff} \oplus \mathbf{F}_{fe}$. Where $\mathbf{F}_{face}$, whose dimension is 2069, is the combination vector of facial features and facial expressions. We created a MLP model for training our face-level network, as depicted in Fig. 2.

C. MID-FUSION METHOD

Scene features provide not only information about context (or environment) in images, but also the specific regions in images. Besides that, skeleton-level and UVC-level give information on the structure and the relation between members in a group. Accordingly, blending features can improve the performance of our hybrid network. As depicted in Fig. 3, the feature map of the fusion method gets the advantages of scene-level, skeleton-level, and UVC-level.

Fig. 2 illustrates the structure of the mid-fusion method where the combination feature, $\mathbf{F}_m$, is put to an MLP with three fully-connected layers linked with the D2C block. $\mathbf{F}_m$ is created by concatenating the features of scene ($\mathbf{F}_{scene}$), skeleton ($\mathbf{F}_{skeleton}$), and UVC ($\mathbf{F}_{UVC}$). $\mathbf{F}_m$ is defined by the following equation $\mathbf{F}_m = \mathbf{F}_{scene} \oplus \mathbf{F}_{skeleton} \oplus \mathbf{F}_{UVC}$.

D. LATE-FUSION METHOD

Since the results of the late-fusion method depended on the results of each model in the other approaches, the best models are needed. Firstly, we choose the best models based on differences between its validation and training loss values. The difference, $\mathcal{D}$, is defined as follows:

$$\mathcal{D} = \mathcal{V}_{\mathcal{L}} - \mathcal{T}_{\mathcal{L}} \tag{1}$$

Subject to constraints of $\mathcal{D} \leq \theta$ and $\mathcal{V}_{\mathcal{L}} \geq \mathcal{T}_{\mathcal{L}}$ Where θ is selection threshold. In this research, we chose θ in interval $[0.1, 0.3]$. $\mathcal{T}_{\mathcal{L}}$ and $\mathcal{V}_{\mathcal{L}}$ is training and validation loss values, correspondingly.

Secondly, we use these models to get GCS predictions. Finally, we concatenate these predictions and push them into an MLP model with D2C block.

E. D2C METHOD

D2C method utilizes the discrete labels (categorical labels) to estimate labels in continuous dimensional. It is based on the joint training method [10] where discrete labels immediately participate in generating continuous labels instead. Fig. 4 shows a typical MLP network with D2C block.

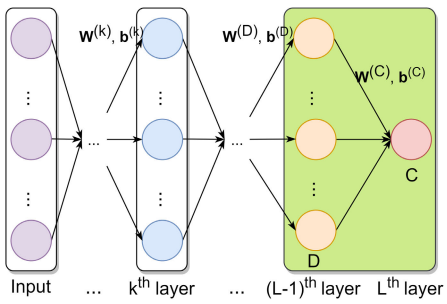


FIGURE 4. A typical MLP network with D2C block (in green box). k is the index of layers and $k = 1, 2, \ldots, (L-2)$.

We denote, L as the number of layers (excluded input layer), $d^{(l)}$ as the number of hidden units in layer l^{th}, $l = 1, 2, \ldots, L$, $\mathbf{z}^{(l)}$ as the input vector of hidden layer l^{th} (excluded bias), $\mathbf{a}^{(l)} \in \mathbb{R}^{d^{(l)}}$ as the output vector of layer l^{th}, $\mathbf{W}^{(l)} \in \mathbb{R}^{d^{(l-1)} \times d^{(l)}}$ and $\mathbf{b}^{(l)} \in \mathbb{R}^{d^{(l)}}$ are the weight matrix and bias matrix, respectively. D and C are the $(L-1)^{th}$ layer and the L^{th}, respectively. The number of hidden nodes in layer D is the number of classes in discrete categorical.

For example, a dataset $\left(\mathbf{x}, \mathbf{y}^{(D)}, \mathbf{y}^{(C)}\right)$ contains N samples. Where $\mathbf{x}$ is the input vectors, $\mathbf{y}^{(D)}$ is ground truth labels in discrete categorical, and $\mathbf{y}^{(C)}$ is ground truth labels in continuous labels.

Calculating feedforward for each sample $\left(\mathbf{x}_i, \mathbf{y}_i^{(D)}, \mathbf{y}_i^{(C)}\right)$ in the dataset, $i = 1, 2, \ldots, N$.

$$\mathbf{a}^{(0)} = \mathbf{x_i}$$
$$\mathbf{z}^{(k)} = \mathbf{W}^{(k)\mathsf{T}}\mathbf{a}^{(k-1)} + \mathbf{b}^{(k)}, k = 1, 2, \ldots, (L-2)$$
$$\mathbf{a}^{(k)} = f\left(\mathbf{z}^{(k)}\right), f(.) \text{ is an activation function}$$
$$\mathbf{z}^{(D)} = \mathbf{z}^{(L-1)} = \mathbf{W}^{(D)\mathsf{T}}\mathbf{a}^{(k)} + \mathbf{b}^{(D)}$$
$$\hat{\mathbf{y}}^{(D)} = \mathbf{a}^{(D)} = softmax\left(\mathbf{z}^{(D)}\right) \in \mathbb{R}^M, M \text{ is number of classes}$$
$$\mathbf{z}^{(C)} = \mathbf{z}^{(L)} = \mathbf{W}^{(C)\mathsf{T}}\mathbf{a}^{(D)} + \mathbf{b}^{(C)}$$
$$\hat{\mathbf{y}}^{(C)} = \mathbf{a}^{(C)} = tanh\left(\mathbf{z}^{(C)}\right) \in \mathbb{R}$$

While applying the D2C method, the final loss function is defined as:

$$\mathcal{L} = \alpha\mathcal{L}_D + (1-\alpha)\mathcal{L}_C \tag{2}$$

where α is weight or balance point, it determines the influence of $\mathcal{L}_D$ and $\mathcal{L}_C$ to the final loss function, $\mathcal{L}$, $\alpha \in [0, 1]$. $\mathcal{L}_D$ is the categorical cross-entropy loss function defined as follows:

$$\mathcal{L}_D = -\sum_{i=1}^{N}\sum_{j=1}^{M}\mathbf{y}_{ij}^{(D)}\log\left(\hat{\mathbf{y}}_{ij}^{(D)}\right) = -\sum_{i=1}^{N}\sum_{j=1}^{C}\mathbf{y}_{ij}^{(D)}\log\left(a_{ij}^{(D)}\right) \tag{3}$$

where M is the number of classes in categorical. $y_{ij}^{(D)}$ and $a_{ij}^{(D)}$ are the j^{th} in probability vector of $\mathbf{y}_i^{(D)}$ and $\mathbf{a}_i^{(D)}$, correspondingly.

$\mathcal{L}_C$ is loss functions that are usually used in regression problems such as MSE loss function or Euclidean loss function. In this case, we chose MSE loss function, therefore we have:

$$\mathcal{L}_D = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_i^{(C)} - \hat{\mathbf{y}}_i^{(C)}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_i^{(C)} - \mathbf{a}_i^{(C)}\right)^2 \tag{4}$$

In backward process, our goal is to update the parameters $\mathbf{W}$ and $\mathbf{b}$ as follows:

$$\mathbf{W}^{(l)} = \mathbf{W}^{(l)} - \eta\frac{\partial L}{\partial \mathbf{W}^{(l)}} \quad \mathbf{b}^{(l)} = \mathbf{b}^{(l)} - \eta\frac{\partial L}{\partial \mathbf{b}^{(l)}}$$

where η is learning rate. The best way to calculate the partial derivatives is to apply the backpropagation algorithm.

The partial derivatives at the output layer (layer C) will be the derivatives of $\mathcal{L}_C$ with respect to $\mathbf{W}^{(C)}$ and $\mathbf{b}^{(C)}$:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{W}^{(C)}} = (1-\alpha)\frac{\partial\mathcal{L}_C}{\partial\mathbf{W}^{(C)}} = (1-\alpha)\frac{\partial\mathcal{L}_C}{\partial\mathbf{z}^{(C)}}\cdot\frac{\partial\mathbf{z}^{(C)}}{\partial\mathbf{W}^{(C)}}$$
$$= (1-\alpha)\mathbf{a}^{(D)}\mathbf{p}^{(C)\mathsf{T}}$$

where $\mathbf{p}^{(C)} = \partial\mathcal{L}_C/\partial\mathbf{z}^{(C)} = \partial\mathcal{L}_C/\partial\mathbf{a}^{(C)}\cdot\partial\mathbf{a}^{(C)}/\partial\mathbf{z}^{(C)}$.

Similarly, we have:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{b}^{(C)}} = (1-\alpha)\frac{\partial\mathcal{L}_C}{\partial\mathbf{b}^{(C)}} = (1-\alpha)\mathbf{p}^{(C)}$$

Then, the derivatives at layer D are calculated as follows:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{W}^{(D)}} = \alpha\frac{\partial\mathcal{L}_D}{\partial\mathbf{W}^{(D)}} + (1-\alpha)\frac{\partial\mathcal{L}_C}{\partial\mathbf{W}^{(D)}}$$
$$\frac{\partial\mathcal{L}}{\partial\mathbf{b}^{(D)}} = \alpha\frac{\partial\mathcal{L}_D}{\partial\mathbf{b}^{(D)}} + (1-\alpha)\frac{\partial\mathcal{L}_C}{\partial\mathbf{b}^{(D)}}$$

We have

$$\frac{\partial\mathcal{L}_D}{\partial\mathbf{W}^{(D)}} = \frac{\partial\mathcal{L}_D}{\partial\mathbf{z}^{(D)}}\cdot\frac{\partial\mathbf{z}^{(D)}}{\partial\mathbf{W}^{(D)}} = \mathbf{a}^{(L-2)}\mathbf{q}^{(D)\mathsf{T}}$$
$$\frac{\partial\mathcal{L}_C}{\partial\mathbf{W}^{(D)}} = \frac{\partial\mathcal{L}_C}{\partial\mathbf{z}^{(D)}}\cdot\frac{\partial\mathbf{z}^{(D)}}{\partial\mathbf{W}^{(D)}} = \mathbf{a}^{(L-2)}\mathbf{p}^{(D)\mathsf{T}}$$

where $\mathbf{q}^{(D)} = \partial\mathcal{L}_D/\partial\mathbf{z}^{(D)} = \partial\mathcal{L}_D/\partial\mathbf{a}^{(D)}\cdot\partial\mathbf{a}^{(D)}/\partial\mathbf{z}^{(D)}$ and $\mathbf{p}^{(D)} = \partial\mathcal{L}_C/\partial\mathbf{z}^{(D)} = \partial\mathcal{L}_C/\partial\mathbf{a}^{(D)}\cdot\partial\mathbf{a}^{(D)}/\partial\mathbf{z}^{(D)}$.

Therefore

$$\frac{\partial\mathcal{L}}{\partial\mathbf{W}^{(D)}} = \alpha\mathbf{a}^{(L-2)}\mathbf{q}^{(D)\mathsf{T}} + (1-\alpha)\mathbf{a}^{(L-2)}\mathbf{p}^{(D)\mathsf{T}}$$

Similarly, we have:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{b}^{(D)}} = \alpha\mathbf{q}^{(D)} + (1-\alpha)\mathbf{p}^{(D)}$$

Furthermore, the derivatives at l^{th}, $l = (L-2), (L-3), \ldots, 1$ will be computed as follows:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{W}^{(l)}} = \alpha\mathbf{a}^{(l-1)}\mathbf{q}^{(l)\mathsf{T}} + (1-\alpha)\mathbf{a}^{(l-1)}\mathbf{p}^{(l)\mathsf{T}}$$
$$\frac{\partial\mathcal{L}}{\partial\mathbf{b}^{(l)}} = \alpha\mathbf{q}^{(l)} + (1-\alpha)\mathbf{p}^{(l)}$$

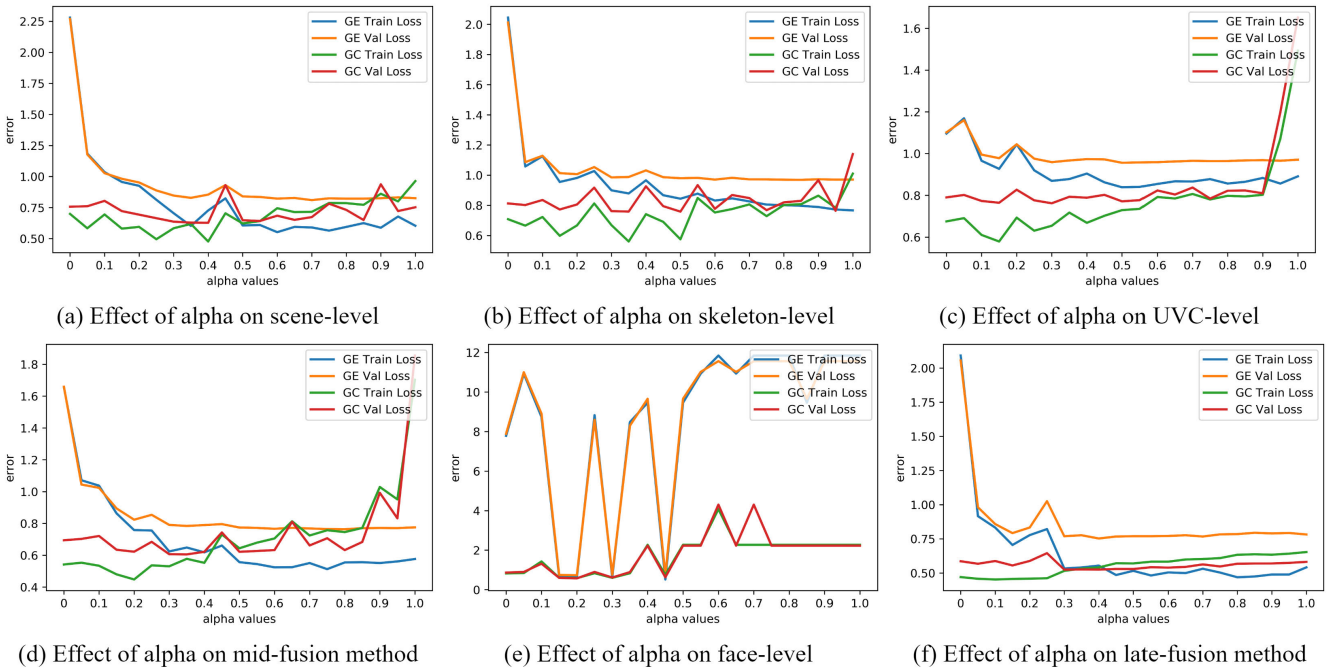
In summary, continuous labels are generated by discrete labels in feedforward. But, in backpropagation, the discrete labels are impacted by continuous labels. Moreover, α play an important role in our method, as it determines how much weight of $\mathcal{L}_D$ and $\mathcal{L}_C$ in $\mathcal{L}$.

IV. EXPERIMENTS AND RESULTS

In this paper, we used Keras [39] deep learning framework for all experiments. Class-weighted method was applied to solve the imbalance problem. Moreover, we carried out a comparison of Ghosh's joint training [10] and our D2C methods on two databases namely GAF-Cohesion [10], [11] and AffectNet [14].

A. DATASET

GAF-Cohesion, an extension of the GAF 3.0 database [20] with group cohesion labels, is the official database of the Group Cohesion sub-challenge in the EmotiW 2019 [11]. It contains group cohesion labels (discrete labels) and group cohesion labels (continuous labels). The values of group cohesion labels were marked in continuous domain [0, 3] corresponding to four group cohesion labels (0: SD, 1: D, 2: A, 3: SA). Besides that, three discrete categories were defined in the group emotion task as *negative, neutral,* and *positive*. The GAF-Cohesion database was divided into three parts:



(a) Effect of alpha on scene-level (b) Effect of alpha on skeleton-level (c) Effect of alpha on UVC-level

(d) Effect of alpha on mid-fusion method (e) Effect of alpha on face-level (f) Effect of alpha on late-fusion method

FIGURE 5. The alpha values influence the performances of each level in the group cohesion estimation system. The best value of α is around 0.3. When $\alpha < 0.3$ the graph of both GE and GC losses have the same downward trend. In contrast, $\alpha > 0.3$, the GE losses decreased while the opposite was true for the GC losses.

TABLE 1. Comparison between the performances of Ghosh's joint training [10] and our D2C methods in scene-level, skeleton-level, UVC-level and mid-fusion on the GAF-Cohesion validation set.

Approach	Val_{MSE}					
	ResNet50		InceptionV3		NasNet	
	Joint training	D2C	Joint training	D2C	Joint training	D2C
Scene	0.714	**0.633**	0.702	0.648	0.771	0.718
Skeleton	0.792	0.766	0.793	**0.765**	0.817	0.802
UVC	0.793	0.769	0.786	**0.762**	0.811	0.789
Mid-fusion	0.680	0.623	0.681	**0.607**	0.726	0.654

TABLE 2. Comparison between the performances of Ghosh's joint training and our D2C methods in the face-level approach on the GAF-Cohesion validation set. Where FF and FE models are the facial feature models and facial expression models, respectively.

		Val_{MSE}	
FF Models	FE Models	Joint training	D2C
ResNet-VGGFace	DenseNet201	0.654	0.689
ResNet-VGGFace	Inception-ResNet-v2	0.653	0.659
ResNet-VGGFace	ResNet-VGGFace	0.658	**0.616**

TABLE 3. The performance of late-fusion method on the GAF-Cohesion validation set. In the last case, we added two more face-level models therefore the number of models is 8.

θ	No. Models	Val_{MSE}
0.3	13	0.559
0.2	12	0.546
0.1	6	0.532
0.1	8	**0.525**

9300 images for training, 4244 images for validation, and 2899 for testing.

We further carried out experiments on the AffectNet database [14] to demonstrate the ability of the D2C method. The AffectNet is a large-scale facial expressions database containing both discrete categorical and continuous dimensional (valence and arousal) labels. Discrete categories include eleven classes, namely *Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, Contempt, None, Uncertain,* and *Non-face*. However, because the *uncertain* and *non-face* were not assigned valence/arousal, we did not use these classes in our experiments. In this study, we utilized 320739 and 4500 images for training and testing, correspondingly.

TABLE 4. The performance comparison on the GAF-Cohesion testing set.

Method	Test_{MSE}
Baseline [10]	0.5
Guo et al. [26]	0.438
Zhu et al. [27]	0.444
Ours	**0.416**

B. EXPERIMENTAL RESULTS ON GAF-COHESION DATABASE

In this study, the discrete labels and continuous labels were the group emotion (GE) labels and group cohesion (GC)

TABLE 5. The performance comparison on the AffectNet validation set.

	RMSE			CCC		
Method	Valence	Arousal	Mean value	Valence	Arousal	Mean value
SVR [14]	0.550	0.420	0.485	0.300	0.180	0.240
CNN [14]	**0.370**	0.410	0.390	**0.600**	0.340	0.470
Joint training + Euclidean loss	0.430	0.381	0.406	0.558	0.468	0.513
D2C + Euclidean loss (Ours)	0.424	0.385	0.405	0.554	0.477	0.516
Joint training + MSE loss	0.427	0.377	0.402	0.561	0.462	0.512
D2C + MSE loss (Ours)	0.405	**0.365**	**0.385**	0.582	**0.490**	**0.536**

labels, respectively. The α in Eqn. 2 is the weight or the balance point, it determines the influence of $\mathcal{L}_D$ and $\mathcal{L}_C$ to $\mathcal{L}$. Therefore, the best value of α is needed, which is the balance point where the differences between training losses and validation losses are minimized, and the validation losses also reach their minimal points in both GE and GC cases. Fig. 5 illustrates the effect of α values to each level in our system.

Moreover, we tuned activation functions (such as *sigmoid, linear and tanh*) at the output layer C, the *tanh* function generates the best results. The $\mathcal{L}_D$ and $\mathcal{L}_C$ are categorical cross-entropy loss function and MSE loss function, correspondingly. Table 1 displays the results of the scene-level, skeleton-level, UVC-level and mid-fusion.

The performances of the face-level and late-fusion are shown in Table 2 and Table 3, respectively.

Besides late-fusion models, we also used ensemble methods to fuse the predictions. In this paper, we used a weighted ensemble method defined as follow:

$$\hat{y}_w = \sum_{k=1}^{m} w_k \hat{y}_k \tag{5}$$

Subject to constraints of $\sum_{k=1}^{m} w_k = 1$, w_k is the weight of model k and $w_k \in \mathbb{R}$. Where, $\hat{y}_w$ is the output of the weighted ensemble method. m is the number of models that are joined in the progress. $\hat{y}_k$ denotes the outputs of model k.

Table 4 demonstrates a comparison between our method and the other methods on the GAF-Cohesion testing set. As shown in Table 4, our hybrid network with the D2C method achieved superior performance among the challenge participants.

C. EXPERIMENTAL RESULTS ON AffectNet DATABASE

In this research, the discrete labels and continuous labels were the facial emotion labels and valence/arousal labels, respectively. Facial features are extracted by using pre-trained ResNet-VGGFace model on the RAF-DB. We then applied an MLP with three fully-connected layers connected with the D2C block to estimate the valence/arousal values. In this case, we used $\alpha = 0.3$ and the activation function at output layer C as *tanh* function. Furthermore, we performed experiments with two cases of regression loss functions, namely MSE loss function and Euclidean loss function. Table 5 summarizes the performance of our method and other methods on the AffectNet validation set [14]. The results show the capability of the proposed D2C method and its valuable results in comparison with other methods, especially in the case of arousal.

V. CONCLUSION

In this paper, we presented a potential approach called D2C, which is to directly learn the interaction between continuous and discrete labels, for estimating group cohesion scores. We also proposed a multi-stream hybrid network combined with the D2C method to achieve state-of-the-art performance on the GAF-Cohesion database. Furthermore, the capability of the D2C was proved through the experimental results achieved on the AffectNet database. Therefore, our method can be applied not only in the group cohesion issues but also in any problem whose data contains two kinds of labels: discrete categorical and continuous dimensional. The focus of our future work will be the extension of this method into sequential data (e.g., video) whereas the cohesion scores are expressed in both spatial and temporal dimensions. In other words, we need to take into account the changing of human behaviors and context through time. In the future, a potential direction for our work is to discover and utilize the attention mechanism in the automatic GCS estimation system. The attention mechanism is able to point out the contribution of each area in the image, therefore, it can be a crucial factor to improve the performance of our system.

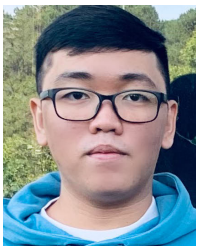
ACKNOWLEDGMENT

This research was done when D. X. Tien was with Pattern Recognition Laboratory, Chonnam National University.

REFERENCES

[1] N. J. Evans and P. A. Jarvis, "Group cohesion: A review and reevaluation," *Small Group Behav.*, vol. 11, no. 4, pp. 359–370, Nov. 1980.
[2] M. A. Hogg, "Group cohesiveness: A critical review and some new directions," *Eur. Rev. Social Psychol.*, vol. 4, no. 1, pp. 85–111, Jan. 1993.
[3] A. Myers, "Team competition, success, and the adjustment of group members," *J. Abnormal Social Psychol.*, vol. 65, pp. 325–332, Nov. 1962.
[4] H. Tajfel, *Social Identity and Intergroup Relations*, vol. 7. Cambridge, U.K.: Cambridge Univ. Press, 2010.
[5] A. V. Carron and K. S. Spink, "The group size-cohesion relationship in minimal groups," *Small Group Res.*, vol. 26, no. 1, pp. 86–105, Feb. 1995.
[6] S. J. Zaccaro and M. C. McCoy, "The effects of task and interpersonal cohesiveness on performance of a disjunctive group task1," *J. Appl. Social Psychol.*, vol. 18, no. 10, pp. 837–851, Aug. 1988.
[7] W. R. Thompson and D. P. Rapkin, "Collaboration, consensus, and détente: The external threat-bloc cohesion hypothesis," *J. Conflict Resolution*, vol. 25, no. 4, pp. 615–637, 1981.
[8] M. W. Rempel and R. J. Fisher, "Perceived threat, cohesion, and group problem solving in intergroup conflict," *Int. J. Conflict Manage.*, vol. 8, no. 3, pp. 216–234, 1997.

[9] T. Treadwell, N. Lavertue, V. Kumar, and V. Veeraraghavan, "The group cohesion scale-revised: Reliability and validity," *J. Group Psychotherapy, Psychodrama Sociometry*, vol. 54, no. 1, p. 3, 2001.
[10] S. Ghosh, A. Dhall, N. Sebe, and T. Gedeon, "Predicting group cohesiveness in images," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.
[11] A. Dhall, R. Goecke, S. Ghosh, and T. Gedeon, "EmotiW 2019: Automatic emotion, engagement and cohesion prediction tasks," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 546–550.
[12] J. F. Hughes, A. van Dam, M. McGuire, D. F. Sklar, J. D. Foley, S. K. Feiner, and K. Akeley, *Computer Graphics—Principles and Practice*, 3rd ed. Reading, MA, USA: Addison-Wesley, 2014.
[13] T. Xuan Dang, S.-H. Kim, H.-J. Yang, G.-S. Lee, and T.-H. Vo, "Group-level cohesion prediction using deep learning models with a multi-stream hybrid network," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 572–576.
[14] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.
[15] P. E. Mudrack, "Defining group cohesiveness: A legacy of confusion?" *Small Group Behav.*, vol. 20, no. 1, pp. 37–49, Feb. 1989.
[16] C. W. Langfred, "Is group cohesiveness a double-edged sword? An investigation of the effects of cohesiveness on performance," *Small Group Res.*, vol. 29, no. 1, pp. 124–143, Feb. 1998.
[17] D. R. Forsyth, *Group Dynamics*. Boston, MA, USA: Cengage Learning, 2018.
[18] D. J. Beal, R. R. Cohen, M. J. Burke, and C. L. McLendon, "Cohesion and performance in groups: A meta-analytic clarification of construct relations," *J. Appl. Psychol.*, vol. 88, no. 6, p. 989, 2003.
[19] H. Hung and D. Gatica-Perez, "Estimating cohesion in small groups using audio-visual nonverbal behavior," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 563–575, Oct. 2010.
[20] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: EmotiW 5.0," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 524–528.
[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[23] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3856–3866.
[24] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2584–2593.
[25] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.
[26] D. Guo, K. Wang, J. Yang, K. Zhang, X. Peng, and Y. Qiao, "Exploring regularizations with face, body and image cues for group cohesion prediction," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 557–561.
[27] B. Zhu, X. Guo, K. Barner, and C. Boncelet, "Automatic group cohesiveness detection with multi-modal features," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 577–581.
[28] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[30] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.
[31] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
[32] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1145–1153.
[33] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.
[34] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.
[35] X. Tang, D. K. Du, Z. He, and J. Liu, "PyramidBox: A context-assisted single shot face detector," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 797–813.
[36] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 951–959.
[37] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.
[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[39] F. Chollet. (2015). *Keras*. [Online]. Available: https://keras.io/



DANG XUAN TIEN received the B.S. degree from the Department of Mathematics and Computer Science, Ho Chi Minh City University of Science-VNUHCM, Vietnam, in 2017, and the M.S. degree from the School of Electronics and Computer Engineering, Chonnam National University, South Korea, in 2020. He is currently working as an AI Engineer in Vietnam. His research interests include pattern recognition and facial emotion analysis.

HYUNG-JEONG YANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Chonbuk National University, South Korea. She is currently a Professor with the Department of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea. Her main research interests include multimedia data mining, medical data analysis, social network service data mining, and video data understanding.

GUEE-SANG LEE (Member, IEEE) received the B.S. degree in electrical engineering and the M.S. degree in computer engineering from Seoul National University, South Korea, in 1980 and 1982, respectively, and the Ph.D. degree in computer science from Pennsylvania State University, in 1991. He is currently a Professor with the Department of Electronics and Computer Engineering, Chonnam National University, South Korea. His primary research interests include image processing, computer vision, and video technology.

SOO-HYUNG KIM (Member, IEEE) received the B.S. degree in computer engineering from Seoul National University, in 1986, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology, in 1988 and 1993, respectively. Since 1997, he has been a Professor with the School of Electronics and Computer Engineering, Chonnam National University, South Korea. His research interests include pattern recognition, document image processing, medical image processing, and deep learning.