Neural Computing and Applications
https://doi.org/10.1007/s00521-023-09040-8

ORIGINAL ARTICLE

DVC-Net: a new dual-view context-aware network for emotion recognition in the wild

Linbo Qing[1] · Hongqian Wen[1] · Honggang Chen[1] · Rulong Jin[1] · Yongqiang Cheng[2] · Yonghong Peng[3]

Received: 31 March 2022 / Accepted: 6 September 2023

Abstract
Emotion recognition in the wild (ERW) is a challenging task due to unknown and the unconstrained scenes in the wild environment. Different from previous approaches that use facial expression or posture for ERW, a growing number of researches are beginning to utilize contextual information to improve the performance of emotion recognition. In this paper, we propose a new dual-view context-aware network (DVC-Net) to fully explore the usage of contextual information from global and local views, and balance the individual features and context features by introducing the attention mechanism. The proposed DVC-Net consists of three parallel modules: (1) the body-aware stream to suppress the uncertainties of body gesture feature representation, (2) the global context-aware stream based on salient context to capture the global-level effective context, and (3) the local context-aware stream based on graph convolutional network to find the local discriminative features with emotional cues. Quantitative evaluations have been carried out on two in-the-wild emotion recognition datasets. The experimental results demonstrated that the proposed DVC-Net outperforms the state-of-the-art methods.

Keywords Emotion recognition · Attention mechanism · Graph convolutional network · Global–local scene feature

1 Introduction

Emotion recognition in the wild (ERW) aims to distinguish human emotional states in natural environment, such as happy, fear, surprise, and sad. ERW is a growing important research field of affective computing, which relies on the techniques developed in computer vision. The ERW has been broadly used in human–machine interaction [1], autonomous driving [2], health care [3], advertising recommendation [4], etc.

One of the key challenges in ERW is the unconstrained scenes in real-life environments, which is different from a controlled laboratory environment with predefined scenarios, e.g., indoor or outdoor environment. In real-life environment, the individuals are not restricted and their emotional states are more natural and real, which are more complicated to be analyzed. A number of works [5–7] have studied this challenge, and mainly fixate on facial expression, pose gesture, voice, and electroencephalogram (EEG) signal. With the establishment of increasingly natural scenes expression databases [8–10] and the advancement of deep learning, ERW has achieved great success. However,

✉ Honggang Chen
honggang_chen@scu.edu.cn

Linbo Qing
qing_lb@scu.edu.cn

Hongqian Wen
2020222050177@stu.scu.edu.cn

Rulong Jin
jinrulong@stu.scu.edu.cn

Yongqiang Cheng
yongqiang.cheng@sunderland.ac.uk

Yonghong Peng
y.peng@mmu.ac.uk

[1] College of Electronics and Information Engineering, Sichuan University, Chengdu, China

[2] Faculty of Technology, University of Sunderland, Sunderland, UK

[3] Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, UK

Published online: 04 October 2023

the performance suffers from the varying environment. On one hand, face or body features tend to be affected by illumination, occlusion and pose variations in natural scenes, which degrade the performance to some extent. On the other hand, the same behavior (e.g., facial expression and body gesture) may represent different emotional states in different scenes. For example, when we consider the postures, garments and surroundings for the common activities such as looking at the computer at home and in the office, we may encounter different emotional states.

To better tackle the above challenges, we propose a dual-view context-aware network (DVC-Net) to recognize human emotions. The main idea is to make full use of emotion relevant cues of individuals and the contextual information from static images. In this network, two different types of clues can be captured by convolutional neural network (CNN) in their respective branches. For individuals, we use an independent branch to extract features, and for scenes, a global–local structure is utilized to obtain emotion-related features. We design a dual attention with three modules in the DVC-Net to achieve better performance for ERW.

To suppress the uncertainties exhibited by individuals due to external disturbances (e.g., illumination, occlusion and pose variations), which affect the learnability of the facial and body features, a body-aware stream is first designed. Based on the facts that place or social situation affects one's emotional state reported in [11–13], a global context-aware stream is then included to capture the salient context and to reduce the influence of the redundant region. Specifically, the spatial attention mechanism is designed to guide the model to focus on the regions of interests for ERW. In order to further obtain the discriminative regions for the context, we introduce a local context-aware stream, where the image partition is utilized to "zoom-in" to local details. In other words, the image is partitioned into patches, for which a graph convolutional network (GCN) is designed to interlink each patch with local details. All patches are built as nodes of the graph, and the adjacent matrix is learned during the training process.

Our main contributions can be summarized as follows:

- We propose an end-to-end DVC-Net for ERW, which can effectively integrate global and local scene information to obtain more effective context feature representation. And based on the attention coefficient, the influence of individual features and context features is balanced.
- For local context information of the scene, we innovatively combine image partition and GCN to obtain local details and find the semantic correlation.
- Quantitative evaluations and extensive experiments have been carried out to demonstrate the effectiveness of the proposed method.

The rest of the paper is organized as follows: Sect. 2 introduces the methods related to emotion recognition in the wild. Section 3 describes the structure of DVC-Net and the corresponding implementation details. Section 4 conducts qualitative and quantitative experiments and analyzes the experimental results in detail. Section 5 summarizes the advantages of proposed method.

2 Related work

In human emotion recognition (HER), the main focus is on the physical aspects of individuals. Among them, facial emotion recognition (FER) is the mostly researched approach as it can provide the most obvious and intuitive emotional state. Conventionally, the feature and classifiers design heavily rely on expert knowledge. The most well-known feature extraction methods are histogram of oriented gradients (HoG), local binary pattern (LBP), and landmark related distance and angle features. These features are then fed into classifiers for training, such as Adaboost, SVM and random forest. With the breakthroughs in deep learning, CNN-based methods are now widely used for various computer vision tasks, which require a large amount of data for training. To solve this problem, datasets such as Extended Yale Face Database B (B+) [14] and FER2013 [15] have emerged to support related research. Jung et al. [16] employed two types of CNNs, one for extracting appearance features and the other for extracting geometric features, and finally, use the integrated model to recognize facial expression. Jain et al. [17] used an extended deep convolutional neural network to classify facial emotions into six categories. The model consists of convolutional layers and residual blocks, which are able to learn subtle features and distinguishing features. Along with the face, hands also carry rich source of information about body emotion language [18]. Body gestures [5] and body pose estimation [19] were also utilized for emotional body recognition [20]. Apart from these visual appearance based methods, Ferdinando et al. [21] used EEG signal for HER by exploiting multiple feature reprefile:///E:/xkh/jrlwhq/paper/bst/sn-mathphys.bstsentations. Qing et al. [22] further proposed an interpretable emotion recognition method based on EEG signals. Kwon et al. [23] learned salient features from the spectrogram of speech signals using CNNs which has improved speech emotion recognition (SER) performance. Although the methods by integrating speech and EEG are capable of complementing HER, they are limited to

laboratory conditions and their performance are significantly downgraded under unconstrained conditions. Therefore, the vision-based methods remain the primary method for HER in the wild environment. Wang et al. [24] applied a region attention network (RAN) to deal with pose variations and occlusion in the wild for FER. Suppressing uncertainties in unconstrained conditions by modifying uncertain samples with self-cure network was utilized in [25]. Huang et al. [26] designed a three-branch architecture to extract spatio-temporal features of body, skeleton and context to realize emotion recognition based on video in the wild. Video-based visual emotion recognition pays more attention to temporal features, while image-based visual emotion recognition pays more attention to scene information. Therefore, this paper focuses on visual emotion recognition based on static images.

In recent years, context-aware emotion recognition has gained intensive attentions, which is a way to recognize emotions in the presence of blurred faces or bodies by obtaining cues from the scene around an individual. For the utilization of context information, Machajdik et al. [27] employed psychology and art theory to extract image features specific to the domain of artworks with emotional expression. Alameda-Pineda et al. [28] introduced nonlinear matrix completion (NLMC) to recognize emotions from abstract paintings. CAER-S [12] and EMOTIC [13] datasets were proposed for the context-aware emotion recognition studies. Kosti et al. [13] proposed a two-branch network, one branch focuses on body-part and the other notices the context. In [29], the region proposal network (RPN) was used to extract relevant objects as the input of the graph convolution network (GCN) to construct an affective map. Several methods [12, 30–32] use an attention network to find the discriminative region of context. Overall, contextual information has been found to be very important for emotion recognition in the natural scene, but the contribution is yet to be further improved. A potential limitation is the contextual information has been directly used. Hence we have proposed a more sophisticated module to exploit both the global and local context and obtain a better representation of contextual features.

3 Proposed method

3.1 Overview

We present our proposed model, namely the DVC-Net, in this section. As shown in Fig. 1, the framework consists of three streams: body-aware stream, global context-aware stream and local context-aware stream. Specifically, a body-aware stream is used to extract individual features and learn a parameter λ from body attention module (BAM) to suppress the uncertainties. Global context-aware stream is implemented through attention branch network (ABN) to find emotion relevant features. Local context-aware stream is utilized to capture local scene information by image partition and GCN.

3.2 Body-aware stream

As mentioned earlier, individuals in natural scenes are susceptible to illumination, occlusion, etc., and directly extracting features for emotion recognition has low reliability. Therefore, the body attention module (BAM) is designed to avoid these uncertainties in the body-aware stream. Different from previous classical work [33] for attention mechanism, the BAM is directly applied to the body image rather than the feature map, which is more effective in determining the person's emotional confidence.

BAM endows with an emotional confidence weight for an individual. As shown in Fig. 2, given a body image $I_B \in \mathbb{R}^{3\times W\times H}$, where W and H are spatial dimensions. The module takes I_B as input and outputs the emotional confidence weight λ. Specifically, the BAM is constructed from a global mean pooling (GMP), two convolution layers with 1×1 kernels and a sigmoid function. The definition of emotional confidence weight can be formulated as,

$$\lambda = \sigma(W_B; I_B) \tag{1}$$

where λ is a learnable emotional confidence weight, σ is a nonlinear function, and W_B denotes the BAM parameters. After getting the attention weights λ, the feature vector v_B can be obtained by multiplying the weights with features of each sample, which is formulated as,

$$\vartheta_B = \lambda \cdot \mathcal{F}(W_B; I_B), \vartheta_B \in \mathbb{R}^{1\times K} \tag{2}$$

where $\mathcal{F}$ is forward propagation and W_B is the parameters of CNN feature extractor. K is the number of emotion categories. In our model, we use ResNet-50 [34] as a feature extractor for I_B, which is pre-trained on Image-Net [35].

3.3 Context stream

To exploit the emotional cues in the scene, we use two branches to extract global–local scene features. The global context-aware stream pays attention to the overall distribution of the scene, while the local context-aware stream focuses on the detailed information in the scene.

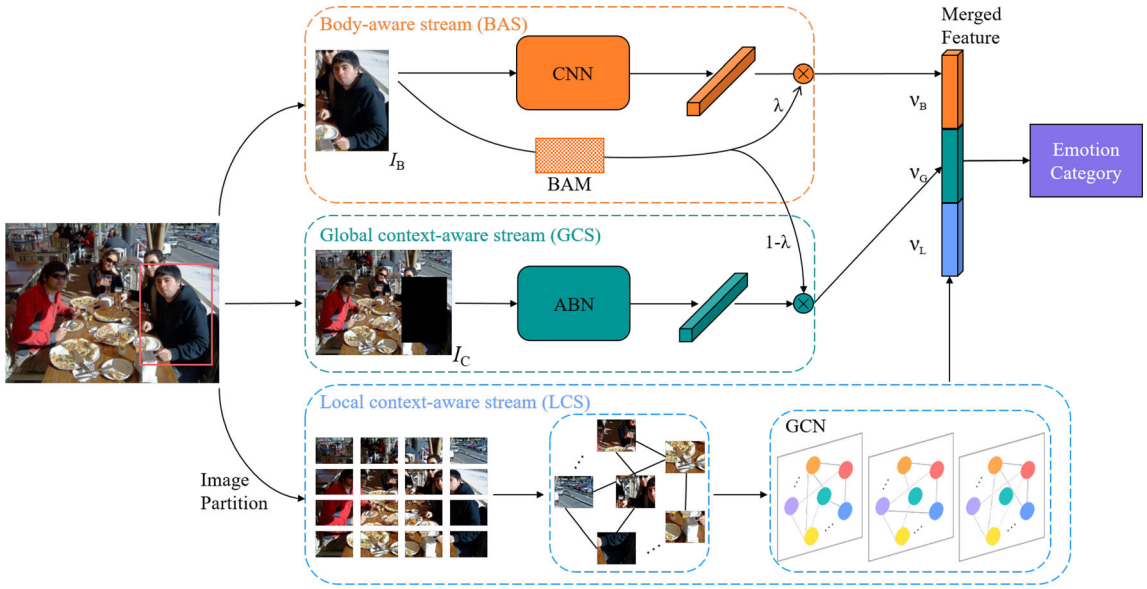

Fig. 1 The overview of dual-view context-aware network (DVC-Net), which consists of three modules. (1) Body-aware stream: is used to extract body features and learning a parameter λ from body attention module (BAM) to suppress the uncertainties. (2) global context-aware stream: attention branch network (ABN) is utilized to capture contextual information. And $1-\lambda$ is used to complement the features of the body-aware stream. (3) Local context-aware stream: the image is partitioned into patches to capture local scene information, and then this local information is connected using GCN. Finally, the features of the three modules are concatenated to predict the emotional state

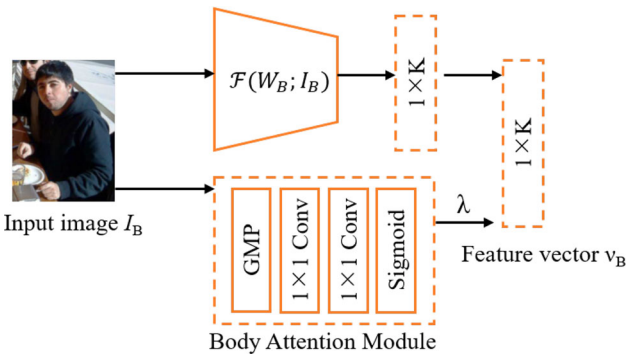

Fig. 2 Body-aware stream

3.3.1 Global context-aware stream

Since some images do not contain clear information about the emotions of the individual, it is worth considering the environment in which the individual is located.

In general, obtaining the scene features using the backbone network is not a good choice because the content of the scene is too complicated. Therefore, it is essential to selectively enhance the feature representation in major contents while suppressing the feature representation in minor contents, which is more in line with the pattern of human eyes to observe things.

To accomplish this issue, we integrate attention branch network (ABN) [36] model in a global context-aware stream. ABN aims to highlight the attention map for visual interpretation, which represents the important regions in image recognition. When emotion recognition is based on pictures with complex backgrounds, it is difficult to directly determine the emotion region in the background. ABN can learn the attention region through the feedback of CAM visual interpretation, and then enhance feature extraction by focusing on the attention region.

As shown in Fig. 3, ABN is composed of three modules: feature extractor, attention branch, and perception branch. The feature extractor is constructed by baseline model ResNet-18, which is utilized to obtain the feature map $g(I_C)$.

The attention branch outputs the attention map $M(I_C)$ through the attention mechanism of CAM [37]. Specifically, CAM has $K \times 3 \times 3$ convolution layers, 1×1 convolution layer, global average pooling (GAP), and softmax layer in the last three layers. We get the attention map $M(I_C)$ after $K \times 3 \times 3$ convolution layers, and feature vector $\vartheta_{G'} \in \mathbb{R}^{1\times K}$. Note that $\vartheta_{G'}$ is used in the training process for error back propagation.

The perception branch outputs the confidence of each category by feature map $g(I_C)$ and attention map $M(I_C)$. The structure for the perception branch is the same as the top layers of the feature extractor ResNet-18. This branch can be formulated as follows:

$$g^{'}(I_C) = (1 + M(I_C)) \cdot g(I_C) \tag{3}$$

$$\vartheta_{G''} = Softmax(\mathcal{F}^{'}(W_C^{'}; g^{'}(I_C))) \tag{4}$$

$$\vartheta_G = (1-\lambda) \cdot \vartheta_{G''} \tag{5}$$

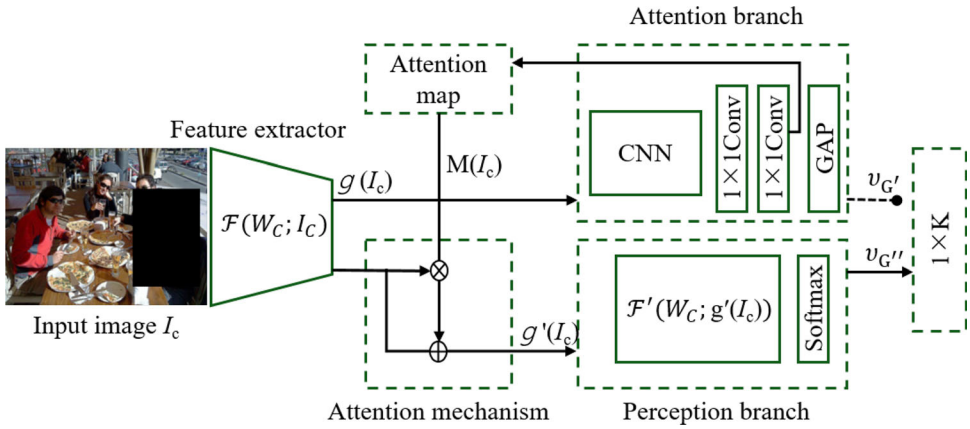

Fig. 3 Attention branch network (ABN)

Equation 3 indicates that the feature map $g^{'}(I_C)$ of the perception branch is generated by feature map $g(I_C)$ and attention map $M(I_C)$. In Eq. 4, $\mathcal{F}^{'}$ is forward propagation, $W_C^{'}$ denotes the perception branch parameters, and $\vartheta_{G''}$ describes the feature vector to the perception branch. Equation 5 gives the feature representation of the global context-aware stream. The reason why $\vartheta_{G''}$ multiplied by $1-\lambda$ because when the uncertainty factor λ of the characters in the scene increases, we should appropriately enlarge the effect of scene features for emotion recognition. Conversely, when the character features can assist well in emotion recognition, the role of scene features should be reduced appropriately.

3.3.2 Local context-aware stream

The utilization of global scene information is merely in a coarse-grained view. This will lead to the missing of some local emotion-related information in the scene and cannot fully exploit the scene information.

To address this issue, the local context-aware stream is designed to further exploit the contextual information in a fine-grained view. Previous studies [29] detects context elements to encode the local scenes, which requires detection algorithms to find emotion relevant elements with prior knowledge. To reduce the dependence on other algorithms and to further explore the relationships between the localities of the scenes, local context-aware stream innovatively uses image partition to find the local emotion clues. The research [38] has demonstrated that location information can enhance the feature presentation and improve performance on classification and detection tasks. However, image partition destructs the visual appearance of an image, which means noisy information also be introduced. To tackle this problem, we aggregate GCN to understand the relationship between each patch from the non-Euclidean data and focus on useful local emotional cues.

As shown in Fig. 4, given an input image $I \in \mathbb{R}^{3\times W\times H}$, we uniformly partition it into $N \times N$ patches. Each patch has $3 \times \frac{W}{N} \times \frac{H}{N}$ dimensions. Patches are put into feature extractor $\mathcal{F}(W_L; I)$ to obtain 1024-dimension feature vector $\vartheta_{L'}$ respectively. Note that N^2 feature extractors are sharing the same weights, which are comprised of two convolution layers with 3×3 kernels generating 64 and 64 feature channels, following a mean pooling.

For the classic method [38], $\vartheta_{L'}$ are processed by simply concatenating, then through the fully connected layer to obtain the final feature ϑ_L. However, this approach omits the relationship between each patch. To mitigate the problem, we utilize these $\vartheta_{L'}$ as nodes to construct the graph. Specifically, a GCN with nodes features $\mathcal{V} = [\vartheta_{L1'}, \vartheta_{L2'}, ..., \vartheta_{LN\times N'}] \in \mathbb{R}^{N\times N\times 1024}$ and an adjacency matrix $A \in \mathbb{R}^{N\times N}$ as inputs. For the *l-th* GCN layer H^l, it can be written as:

$$H^l = \sigma(AH^{l-1}W^{l-1}) \tag{6}$$

where $H^0 = \mathcal{V}$ and $W^{l-1} \in \mathbb{R}^{d^{'}\times d}$ is the weighted matrix in layer *l-1* with $d^{'}$ and d refers to the input and output feature dimension of *(l-1)-th* hidden layer. The adjacency matrix A is initialized by the all one matrix (ones (N)) and is learnable during the back-propagation process. Specifically, we use three GCN layers with 1024, 1024, and 512 output feature dimension, respectively. Each layer follows a Relu layer.

3.4 Feature fusion and training

Following the acquisition of all features from three parallel branches, early fusion is used to obtain the final emotional feature representation:

$$\vartheta = concatenate(\vartheta_B, \vartheta_G, \vartheta_L) \tag{7}$$

For the training of the output concatenated feature, we use two types of standard loss functions. Standard cross-entropy loss $\mathcal{L}_{CE}$ is utilized for CAER-S [12], and Euclidean

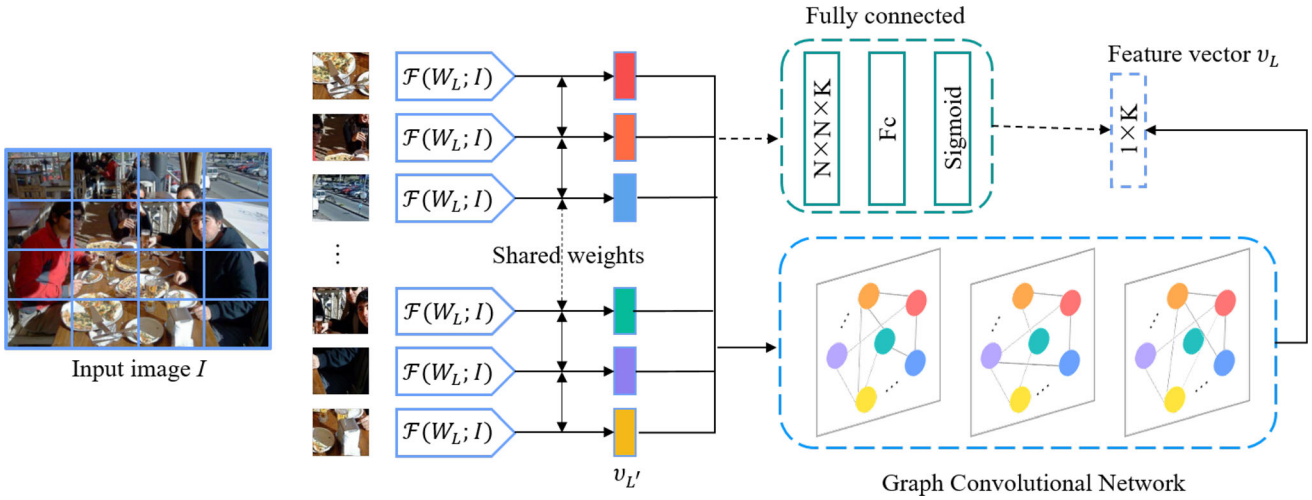

Fig. 4 Local context-aware stream. There are two methods to utilize $\vartheta_{L'}$, the classic method (green dashed box) concatenate all features, then use a fully connected layer to output the feature vector ϑ_L, our method (blue dashed box) use three GCN layers to link these features (color figure online)

loss $\mathcal{L}_{EL}$ is utilized in EMOTIC [13] due to it is a multi-label dataset. Two loss functions are defined as follows,

$$\mathcal{L}_{CE} = -ylog(\hat{y}) - (1-y)log(1-\hat{y}) \tag{8}$$

$$\mathcal{L}_{EL} = \frac{1}{K}\sum_{i=1}^{K} w_i(\hat{y_i} - y_i)^2 \tag{9}$$

where $\hat{y}$ is the predicted label, y is the ground truth and w_i is the weight of i-th category.

4 Experiments

In this section, we evaluate our method on two *in-the-wild* benchmark datasets, i.e., the CAER-S [12] and the EMOTIC [13]. First, the datasets and implementation details are introduced. Then, the classification results in comparisons with state-of-the-art methods are provided. Third, in order to illustrate the advantages of DVC-Net, ablation studies are also provided.

4.1 Datasets

The CAER-S [12] and EMOTIC [13] datasets used in the paper contain not only face or body emotions, but also the environment around the person. CAER-S contains 70K annotated static images, which are collected from 79 TV shows. Each image is annotated with seven emotion categories. EMOTIC contains 34,320 persons labelled in an unconstrained environment from 23,571 images. The annotation contains 26 discrete emotion categories and continuous VAD (Valence, Arousal, and Dominance) values. Each person is annotated with the corresponding bounding box and emotion labels, where each person is labelled with at least one emotion category. The datasets have been partitioned into training, validation, and testing subsets in the experiments. The details of the distribution is illustrated in Table 1, with sample images from the two datasets shown in Fig. 5.

4.2 Implementation details

During data preprocessing and augmentation, body image I_B and context image I_C are resized to 224 × 224. The color jitter is set to 0.4 and random horizontal flip is applied.

We use the Adam optimizer for hyper parameter settings. The learning rate is initialized as 0.0001, which then reduced by cosine annealing [39]. The batch size is set to 32, the maximum epoch is 100 and N is set to 4.

The proposed DVC-Net is implemented using Pytorch library [40] (version 1.5.1). The GPU model is GeForce RTX-2080Ti. The CPU model is Intel i7-9700 with 3.00 GHz. The operating system is Ubuntu 20.04 with 64 bits.

Table 1 Dataset distribution of CAER-S and EMOTIC datasets

Dataset	Partition	Number
CAER-S [12]	Train	49,007
	Val	–
	Test	20,992
EMOTIC [13]	Train	12,957
	Val	3,334
	Test	7,280

(a) CAER-S

(b) EMOTIC

Fig. 5 CAER-S and EMOTIC samples

4.3 Comparison to SOTA methods

We compare DVC-Net with some state-of-the-art (SOTA) approaches, which are described as follows:

- CAER-S-Net [12] built a two-stream encoding network, in which one stream focuses on extracting facial expression features and the other stream is designed to encode context information surrounding the person. Then an adaptive fusion network is utilized to fuse the two streams features.
- Kosti et al. [41] designed a dual branch network, one branch encodes body information and the other one focuses on the contextual information, two branches are fused via a fully connected layer.
- Zhang et al. [29] used Region Proposal Network (RPN) to extract contextual elements which then being fed into GCN to learn the affective relationship.
- EmotiCon [30] uses four branches to extract features for emotion recognition. Two modalities of faces and gaits were fused as multiple modalities stream, then concat with to background context stream and interactions stream.
- Bendjoudi et al. [42] built three modules for emotion classification and devised multi-label focal loss (MFL) to deal with imbalanced data.
- GRERN [32] combined GCN with gated recurrent unit (GRU) for context-aware emotion recognition. First, salient image regions via a bottom-up attention module are extracted. Then, GCN is utilized to construct the connections between those salient regions. Finally, redundant features of the graph are removed using GRU.
- Context-LGM [43] considers the object-context relation and models it in a hierarchical manner. First, the feature representation of the body and the context is obtained using the encoder, and then the emotion category is obtained using recognition net.
- CHAPNet [44] consists of two branches. The target branch get the features from body image. The situational context branch extracts event features, objects features and relation features. Then two branches' output is merged and gender, age and emotion are predicted.

Table 2 Comparison of the emotion recognition accuracy of SOTA methods on the CAER-S benchmark

Methods	Datasets		Years
	CAER-S Acc. (%)	EMOTIC mAP (%)	
CAER-S-Net [12]	73.51	20.84	2019
Kosti et al. [41]	–	27.38	2019
Zhang et al. [29]	76.73	28.42	2019
EmotiCon [30]	–	**35.48**	2020
Bendjoudi et al. [42]	–	28.33	2020
GRERN [32]	81.31	–	2021
Context-LGM. [43]	85.30	–	2021
CHAPNet. [44]	88.31	28.81	2021
DVC-Net (Ours)	**88.83**	29.96	–

The experimental results are shown in Table 2. In CAER-S dataset, an improvement of 15.32% over the baseline method (CAER-S-Net). Our method and GRERN both consider the global discriminative regions of context, so the performance is better than CAER-S-Net, and 7.52% over GRERN. Besides, the use of local information prompts our model to learn more helpful context information, which is a considerable additional module to increase the performance.

In EMOTIC dataset, we can notice that the task is very challenging due to the complexity of the environment plus the unbalanced dataset. DVC-Net has shown competitive results and achieved the highest mAP among all methods. Specifically, compared with [41] and [42], where contextual information is also included in the module, our DVC-Net achieved about 2.58% and 1.63% mAP improvement respectively. Clearly, these results illustrate the superiority of our proposed method which can better exploit the context information. As for the contextual elements-based method in [29], it is time-consuming and computationally expensive in the inference stage due to its two-stage approach that uses Faster-RCNN to detect objects in the scene and then feed them into GCN for emotional inference. Our DVC-Net still achieved an absolute 1.54% mAP improvement, which is a further evidence that combining global and local contextual information with attention mechanism can improve performance. It can be also aware that there is still some gap between our work and EmotiCon [30], the same as other more recent work (e.g. [42, 44]). The main reason is that EmotiCon makes full use of face, pose, context, and depth map information to learn emotional features. Compared with our single-stage method and other end to end methods, EmotiCon requires head detection, depth map generation and other steps in the data preprocessing phase, which is more complex and time-consuming. Figure 6 shows some sample results.

Moreover, Table 3 shows that DVC-Net obtains comparable performance in continuous VAD dimension values prediction. This is because the fact that our model focuses on the capture of global–local scene information, while the VAD metric focuses more on the status of the pose of the subjects themselves. In other words, our model is able to improve the performance of emotion recognition by effectively capturing scene features without compromising the individual features.

4.4 Ablation studies

To further demonstrate the performance of our DVC-Net quantitatively, we implement ablation studies to analyze the impacts from individual components in our framework.

Effects of different components To verify the respective contributions of the three modules of body-aware stream (BAS), global context-aware stream (GCS) and local context-aware stream (LCS), we remove modules one by one to conduct ablation experiments.

The results of experiments are summarised in Table 4. It can be seen that the combination of all three modules produces the best performance. The BAS achieves comparable performance due to the focus on the subject of emotional expression. In CAER-S dataset, by comparing BAS, BAS + LCS and BAS + GCS, it is clearly revealed that exploration of scene information from different views is more effective in improving emotion recognition than directly extracted features. Comparing BAS+LCS (w/o GCN) and BAS + LCS, BAS + GCS + LCS (w/o GCN) and BAS + GCS + LCS, it can be seen that the usage of LCS (w/o GCN) has limited gains. The main reason is

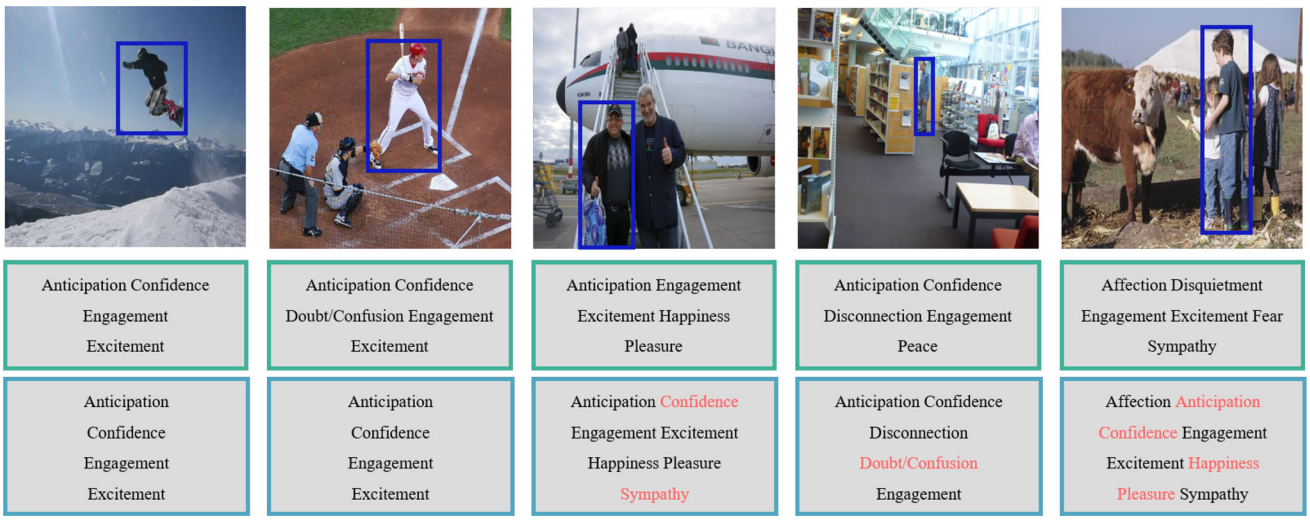

Fig. 6 Ground truth (green box) and prediction results (blue box) on images randomly selected on EMOTIC dataset (color figure online)

Table 3 The comparison on mean error rate on EMOTIC dataset

Dimension	In [41]	In [29]	In [42]	Ours
Valence	0.9	0.7	0.8	0.8
Arousal	1.2	1.0	1.0	1.0
Dominance	0.9	1.0	0.9	0.9
Mean	1.0	0.9	0.9	0.9

Table 4 Effects of different components on CAER-S (Acc.) and EMOTIC (mAP) datasets

Methods	Accuracy (%)	mAP (%)
BAS	81.79	28.99
BAS + LCS (w/o GCN)	81.96	28.82
BAS + LCS	82.16	28.37
BAS + GCS	85.03	29.04
BAS + GCS + LCS (w/o GCN)	86.53	29.39
DVC-Net (BAS + GCS + LCS)	88.83	29.96

learning the same features repeatedly is not helpful at the global and local levels. However, GCN is able to interlink relationships between different patches. The transmission of local information between the patches effectively facilitated the identification and got 0.2% and 2.3% improvement.

In EMOTIC dataset, the overall trend is similar to the CAER-S dataset. Note that BAS outperform BAS + LCS and BAS+LCS (w/o GCN). This may be due to the fact that patches of body is much higher than salient cues, and therefore, predicting the emotion of individuals from body may benefit from a simpler model. Instead, the relationship between context information is more semantic and may require a more refined model. It can also be noted that BAS+LCS (w/o GCN) achieves better recognition effect than BAS+LCS. The reason may be the complexity of context information in EMOTIC dataset. Local patches contain rich context features, and direct learning of local patches could get high mAP. When there is no global context view supplement, further inference of local graph does not obtain any contributing features, which indicates that the learning of single-view context features may be limited in different datasets. By comparing BAS, BAS + LCS and BAS + GCS, the results are indicating that scenario information may be limited from one perspective alone, but it can yield effective performance gains when used together in DVC-Net (BAS + GCS + LCS). It demonstrates the validity of our proposal to use dual-view for context information. BAS + GCS + LCS (w/o GCN) obtained a relatively good performance too. We consider this may be because the dataset contains much richer scene information. As previously mentioned, GCN can help propagate scene information among the patches while avoiding the interference of noisy patches. However, if each patch can help the model to recognize emotions, the role of GCN will be weakened.

Effects of attention module of DVC-Net To evaluate the impacts by the designed BAM and the ABN, a separate set of ablation experiments were performed.

As shown in Table 5, both BAM and ABN attention modules have improved the performance. When neither of the attention modules are used, the lack of emotion relevant cues in the body or scene causes a decrease of performance in the model. When only one of them is employed, the enhancement of the body or scene features also fails to coordinate the features of the other stream well. In CAER-S dataset, the performance of DVC-Net (w/o BAM) is lower than DVC-Net (w/o ABN), but the opposite is true in EMOTIC dataset. This indicates that when the two modules are not employed together, they may present different results due to the biases of different datasets.

Moreover, in order to further evaluate the collaborative performance of our model structure, we conduct a basic network, which is also composed of three channels, except that each channel uses only the backbone network to extract features. Specifically, the body-aware stream uses ResNet-50 without BAM, the global context-aware stream use ResNet-18 without ABN, and the local context-aware stream uses a two-layer convolutional network without GCN. The basic model using dual-view context information still obtains comparable performance, which provides an accuracy of 80.20%. Figure 7 illustrates the confusion matrix of backbone network only results and the fully featured DVC-Net ones. Our model has shown improvement in all categories, of which, *Angry* improved the most by 15%. In the basic model, *Angry*, *Happy*, and *Surprise* have a certain probability of being misjudged as *Neutral*, but this situation is significantly alleviated in the DVC-Net model.

Table 5 Effects of attention module of DVC-Net on CAER-S (Acc.) and EMOTIC (mAP) datasets

Methods	Accuracy (%)	mAP (%)
DVC-Net (w/o BAM w/o ABN)	87.91	29.09
DVC-Net (w/o BAM)	87.88	29.41
DVC-Net (w/o ABN)	88.21	28.66
DVC-Net	88.83	29.96

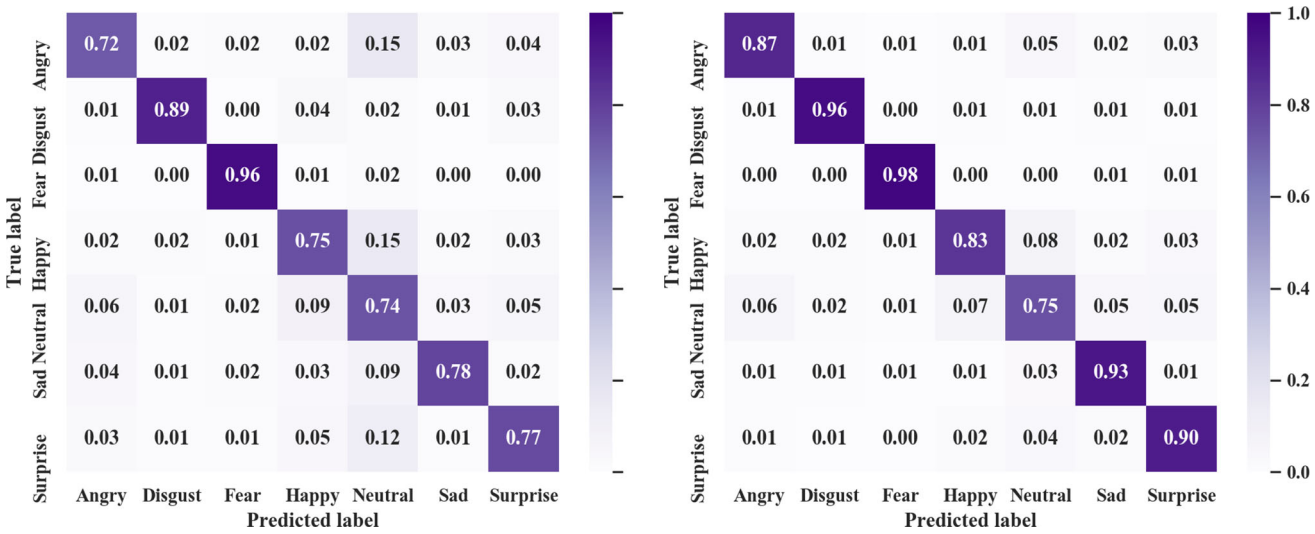

Fig. 7 Confusion matrix of without dual attention and graph network (left) and DVC-Net (right) on the CAER-S datasets

Effects of different values of N for image partition To search the appropriate value of N, we tried to change the values of N in a set of {2, 3,..., 9}, Fig. 8 shows the effects of different values of N. It should be noted that, there is no result for $N = 1$ due to it is a complete image, where no GCN is required. As shown in Fig. 8, The appropriate value of N is 4 for EMOTIC dataset due to it strikes a better balance between memory footprint and performance. The same experiment of different values of N also conduct on BAM + GSAM + LSAM (w/o GCN). We can see that DVC-Net with GCN performs better than the one without it. When $N = 9$, without GCN-based method drops significantly, but GCN-based method still tends to stable, which implied that GCN is beneficial for image partition in our model.

Effects of different positions of BAM in body-aware stream To search for the appropriate position of BAM, we tried to apply BAM to the body image and the feature map. Figure 9 shows the structure of body-aware stream by applying attention mechanism in different locations. Left shows

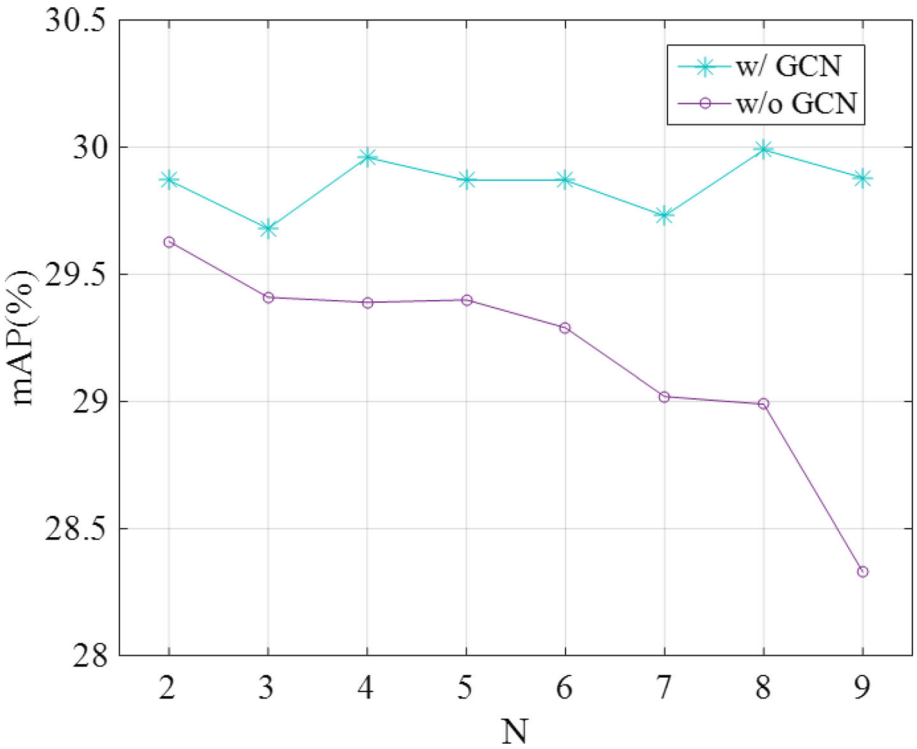

Fig. 8 The influence of partition number N and GCN on EMOTIC dataset

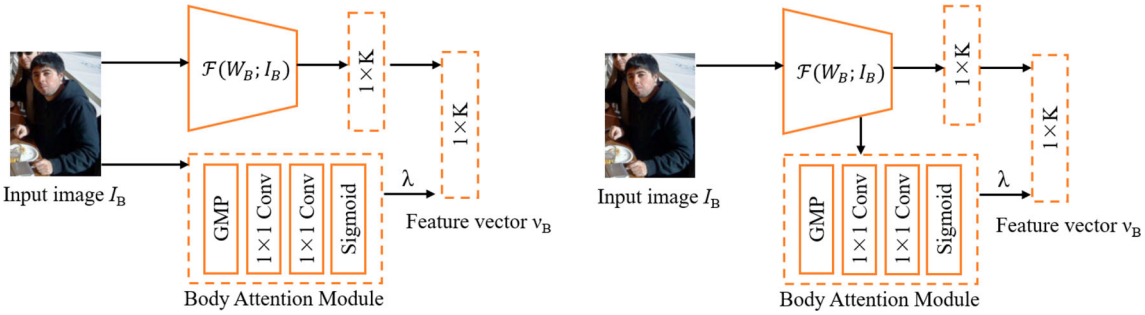

Fig. 9 Body-aware stream with BAM applied to body image (left) and BAM applied to feature map (right)

Table 6 Effects of different positions of BAM on CAER-S (Acc.) and EMOTIC (mAP) datasets

Methods	Accuracy (%)	mAP (%)
DVC-Net (BAM on body image)	**88.83**	**29.96**
DVC-Net (BAM on feature map)	88.56	29.65

The number with bold font is the best value, to better demonstrate the effectiveness of proposed work

the use of BAM directly on the body image, and right shows the use of BAM on the extracted feature map. The experimental results are shown in Table 6, the use of BAM on body image achieved better recognition accuracy on both CAER and EMOTIC datasets. Compared with the context information, the emotional clues expressed by body image are more obvious, applying BAM to low-dimensional information can obtain more effective confidence.

Effects of different layers of ResNet in body-aware stream We compare the accuracies of backbone networks for feature extraction in body-aware stream. We use ResNet 18, 34, 50, 101, 152 models on CAER-S and EMOTIC. Table 7 shows the accuracy and mAP of different feature extractors in BAS. In CAER-S dataset, it can be seen that with the increase of the number of network layers, the recognition accuracy does not monotonically increase or decrease, but has some fluctuations. In comparison, ResNet50 and ResNet152 achieved better results. In the EMOTIC dataset, the experimental results showed a peak at ResNet50. In general, ResNet152 is deeper and has a large number of model parameters, but the accuracy is not much improved. Therefore, ResNet50 is better used as the basic network.

Effects of different layers of ResNet in global context-aware stream We also compare the accuracies of baseline model for feature extraction in global context-aware stream. We use ResNet 18, 34, 50, 101, 152 models on CAER-S and EMOTIC to obtain the feature map. Table 8 shows the accuracy and mAP of different feature extractors in GCS. The experimental results show a similar distribution on both datasets. With the increase of the depth of

Table 7 Effects of different ResNet of body-aware on CAER-S (Acc.) and EMOTIC (mAP) datasets

Methods	Accuracy (%)	mAP (%)
DVC-Net (ResNet18 in BAS)	88.12	28.83
DVC-Net (ResNet34 in BAS)	88.13	29.06
DVC-Net (ResNet50 in BAS)	88.83	**29.96**
DVC-Net (ResNet101 in BAS)	88.64	29.31
DVC-Net (ResNet152 in BAS)	**88.90**	29.38

The number with bold font is the best value, to better demonstrate the effectiveness of proposed work

Table 8 Effects of different ResNet of global context-aware on CAER-S (Acc.) and EMOTIC (mAP) datasets

Methods	Accuracy (%)	mAP (%)
DVC-Net (ResNet18 in GCS)	**88.83**	**29.96**
DVC-Net (ResNet34 in GCS)	88.43	29.66
DVC-Net (ResNet50 in GCS)	88.00	29.39
DVC-Net (ResNet101 in GCS)	87.53	28.99
DVC-Net (ResNet152 in GCS)	87.16	28.83

The number with bold font is the best value, to better demonstrate the effectiveness of proposed work

ResNet, the recognition accuracy is declining. This may be because the attention map obtained based on the feature map is used to enhance training in the global context-aware stream, while the ability of the abstract feature map obtained in the too deep network to represent the important areas in the global context becomes poor.

Effects of different attention modules in global context-aware stream In our model, we use the ABN attention mechanism based on CAM to highlight the attention map based on feature map for visual interpretation and enhance the important areas in the learning context picture. In order to explore the recognition effects of different attention mechanisms, we used the other two most advanced attention mechanisms, DAT [45] and CrossFormer [46], for comparative experiments. DAT [45] proposed a novel deformable self-attention module to focus on relevant

Table 9 Effects of different attention mechanisms on CAER-S (Acc.) and EMOTIC (mAP) datasets

Methods	Accuracy (%)	mAP (%)
DVC-Net (CAM in GCS)	**88.83**	**29.96**
DVC-Net (DAT in GCS)	88.26	29.31
DVC-Net (CrossFormer in GCS)	88.49	28.77

The number with bold font is the best value, to better demonstrate the effectiveness of proposed work

regions and capture more informative features. CrossFormer [46] proposed cross-scale attention, learning cross-scale features and keeping both small-scale and large-scale features in the embeddings. The experimental results on EMOTIC and CAER-S datasets are shown in Table 9. Using CAM attention achieved best experimental results on both datasets. The main reason is that there is a certain difference between emotion recognition based on complex context pictures and general image classification in determining the effective area. For example, in the classification of subjects such as cats and dogs, the attention area is generally focused on the subject, while it is difficult to directly determine the emotion area in the context. ABN learns the attention region through the feedback of CAM visual interpretation, and then enhances feature extraction by focusing on the attention region, finally obtaining good performance.

4.5 Discussion

The success of CNN is attributed to its ability to extract features. However, feature learning becomes difficult when the task is too complex, such as scene recognition and segmentation. The proposed attention mechanism allows CNN to focus on the salient information in an image. The trick of image partition can be used to learn useful local information, and the relationship between patches can be connected by GCN to help the model learn the correlation between local information. Overall, exploring the scene information through dual views can effectively improve the ability of emotion recognition in the wild.

5 Conclusion

In this paper, we presented a new DVC-Net for emotion recognition in the wild, which consists of three parallel modules for better emotion recognition performance in a natural environment. The experiment has demonstrated the effectiveness of the modules being integrated in the proposed approach. The body-aware stream can suppress the uncertainties of body postures while obtaining body features in natural scenes. The global context-aware stream is able to focus on global contextual information and the local context-aware stream can further exploit the discriminative emotion features by combining image partition and GCN. The experimental results conducted on two in-the-wild benchmark datasets have shown that DVC-Net achieves competitive results, demonstrating the advantages of our proposed method. Note that our contribution is to extract context features from dual-view context-aware streams and balance individual features and context features based on attention coefficient, so the preexisting attention mechanism is directly applied. Better attention mechanism could improve the effect of feature extraction and additional visual cues can also be considered in future work.

Acknowledgements This work is supported by the National Natural Science Foundation of China [grant number 61871278].

Data availability The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Conflicts of interest The authors declare that there are no conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Hortensius R, Hekele F, Cross ES (2018) The perception of emotion in artificial agents. IEEE Trans Cogn Dev Syst 10 (4):852–864
2. Sini J, Marceddu AC, Violante M (2020) Automatic emotion recognition for the calibration of autonomous driving functions. Electronics 9 (3):518
3. Spekman ML, Konijn EA, Hoorn JF, The importance of coping appraisals and coping strategies (2018) Perceptions of healthcare robots as a function of emotion-based coping. Comput Hum Behav 85:308–318
4. Kao T-F, Du Y-Z (2020) A study on the influence of green advertising design and environmental emotion on advertising effect. J Clean Prod 242:118294
5. Luo Y, Ye J, Adams RB, Li J, Newman MG, Arbee Wang J.Z (2020) Towards automated recognition of bodily expression of emotion in the wild. Int J Comput Vis 128 (1):1–25
6. Dael N, Mortillaro M, Scherer KR (2012) Emotion expression in body action and posture. Emotion 12 (5):1085
7. Peng Y, Tang R, Kong W, Nie F A factorized extreme learning machine and its applications in EEG-based emotion recognition. In: International conference on neural information processing, pp 11–20 (2020). Springer
8. Kossaifi J, Tzimiropoulos G, Todorovic S, Pantic M (2017) AFEW-VA database for valence and arousal estimation in-the-wild. Image Vis Comput 65:23–36
9. Fabian Benitez-Quiroz C, Srinivasan R, Martinez AM Emotionet (2016) An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5562–5570
10. Li Y, Zeng J, Shan S, Chen X (2018) Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Trans Image Process 28 (5):2439–2450

11. Barrett LF, Mesquita B, Gendron M (2011) Context in emotion perception. Curr Dir Psychol Sci 20 (5):286–290
12. Lee J, Kim S, Kim S, Park J, Sohn K (2019) Context-aware emotion recognition networks. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10143–10152
13. Kosti R, Alvarez J.M, Recasens A, Lapedriza A (2017) Emotion recognition in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1667–1675
14. Georghiades AS, Belhumeur PN, Kriegman DJ (2001) From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell 23 (6):643–660
15. Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee D-H, et al (2013) Challenges in representation learning: a report on three machine learning contests. In: International conference on neural information processing, pp 117–124. Springer
16. Jung H, Lee S, Yim J, Park S, Kim J (2015) Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE international conference on computer vision, pp 2983–2991
17. Jain DK, Shamsolmoali P, Sehdev P (2019) Extended deep neural network for facial emotion recognition. Pattern Recogn Lett 120:69–74
18. Molchanov P, Gupta S, Kim K, Kautz J (2015) Hand gesture recognition with 3D convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1–7
19. Cao Z, Hidalgo G, Simon T, Wei S-E, Sheikh Y (2019) Openpose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell 43 (1):172–186
20. Noroozi F, Kaminska D, Corneanu C, Sapinski T, Escalera S, Anbarjafari G (2018) Survey on emotional body gesture recognition. IEEE Trans Affect Comput
21. Ferdinando H, Seppänen T, Alasaarela E (2017) Enhancing emotion recognition from ECG signals using supervised dimensionality reduction. In: ICPRAM, pp 112–118
22. Qing C, Qiao R, Xu X, Cheng Y (2019) Interpretable emotion recognition using EEG signals. IEEE Access 7:94160–94170
23. Kwon S et al (2020) A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors 20 (1):183
24. Wang K, Peng X, Yang J, Meng D, Qiao Y (2020) Region attention networks for pose and occlusion robust facial expression recognition. IEEE Trans Image Process 29:4057–4069
25. Wang K, Peng X, Yang J, Lu S, Qiao Y (2020) Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6897–6906
26. Huang Y, Wen H, Qing L, Jin R, Xiao L (2021)Emotion recognition based on body and context fusion in the wild. In: Proceedings of the IEEE/CVF conference on international conference on computer vision workshops (ICCVW), pp 3602–3610
27. Machajdik J, Hanbury A (2010) Affective image classification using features inspired by psychology and art theory. In: Proceedings of the 18th ACM international conference on multimedia, pp 83–92
28. Alameda-Pineda X, Ricci E, Yan Y, Sebe N (2016) Recognizing emotions from abstract paintings using non-linear matrix completion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5240–5248
29. Zhang M, Liang Y, Ma H (2019) Context-aware affective graph reasoning for emotion recognition. In: 2019 IEEE international conference on multimedia and expo (ICME), pp 151–156. IEEE
30. Mittal T, Guhan P, Bhattacharya U, Chandra R, Bera A, Manocha D (2020) Emoticon: Context-aware multimodal emotion recognition using Frege's principle. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14234–14243
31. Khan A.S, Li Z, Cai J, Tong Y (2021) Regional attention networks with context-aware fusion for group emotion recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1150–1159
32. Gao Q, Zeng H, Li G, Tong T (2021) Graph reasoning-based emotion recognition network. IEEE Access 9:6488–6497
33. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
34. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
35. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
36. Fukui H, Hirakawa T, Yamashita T, Fujiyoshi H (2019) Attention branch network: Learning of attention mechanism for visual explanation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10705–10714
37. Selvaraju R.R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
38. Noroozi M, Favaro P (2016) Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision, pp 69–84. Springer
39. Loshchilov I, Hutter F (2016) Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983
40. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32:8026–8037
41. Kosti R, Alvarez JM, Recasens A, Lapedriza A (2019) A context based emotion recognition using EMOTIC dataset. IEEE Trans Pattern Anal Mach Intell 42 (11):2755–2766
42. Bendjoudi I, Vanderhaegen F, Hamad D, Dornaika F (2021) Multi-label, multi-task CNN approach for context-based emotion recognition. Inf Fusion 76:422–428
43. Liu M, Sun X, Zhang F, Yu Y, Wang Y (2021) Context-LGM: leveraging object-context relation for context-aware object recognition. arXiv preprint arXiv:2110.04042
44. Singh A, Fan S, Kankanhalli M (2021) Human attributes prediction under privacy-preserving conditions. In: Proceedings of the 29th ACM international conference on multimedia, pp 4698–4706
45. Xia Z, Pan X, Song S, Li LE, Huang G (2022) Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4794–4803
46. Wang W, Yao L, Chen L, Cai D, He X, Liu W (2021) Crossformer: a versatile vision transformer based on cross-scale attention. arXiv e-prints, 2108