Pattern Recognition 92 (2019) 177–191

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition

Siyue Xie, Haifeng Hu*, Yongbo Wu

School of Electronic and Information Technology, Sun Yat-sen University, Guangzhou, China

ARTICLE INFO

Article history:
Received 22 June 2018
Revised 19 February 2019
Accepted 21 March 2019
Available online 28 March 2019

Keywords:
Attention
Convolutional neural network
Facial expression recognition
Multi-Path variation-suppressing network
Salient expressional region descriptor

ABSTRACT

Facial Expression Recognition (FER) has long been a challenging task in the field of computer vision. In this paper, we present a novel model, named Deep Attentive Multi-path Convolutional Neural Network (DAM-CNN), for FER. Different from most existing models, DAM-CNN can automatically locate expression-related regions in an expressional image and yield a robust image representation for FER. The proposed model contains two novel modules: an attention-based Salient Expressional Region Descriptor (SERD) and the Multi-Path Variation-Suppressing Network (MPVS-Net). SERD can adaptively estimate the importance of different image regions for FER task, while MPVS-Net disentangles expressional information from irrelevant variations. By jointly combining SERD and MPVS-Net, DAM-CNN is able to highlight expression-relevant features and generate a variation-robust representation for expression classification. Extensive experimental results on both constrained datasets (CK+, JAFFE, TFEID) and unconstrained datasets (SFEW, FER2013, BAUM-2i) demonstrate the effectiveness of our DAM-CNN model.

1. Introduction

Facial Expression Recognition (FER) has long been an interesting and challenging topic in the field of computer vision. Researchers usually aim to construct a system that can automatically identify different expressions in images. Applications based on FER systems can be found in many cases, such as human-computer interaction (HCI) system [1], multimedia [2], surveillance [3] and driver safety [4].

There are various methods for solving the FER problem, which can be roughly grouped into two categories. The first category of methods classify expressions based on Action Units (AUs), which are tiny but discriminable muscle actions that are relevant to expressions. AU-based methods usually convert the FER problem to the task of AU detection [5]. However, local changes on faces are difficult to be detected, which makes it hard for computers to conduct accurate AU detections. Variations such as illuminations or pose changes can also degrade the performance of AU detection. The second kind of methods usually extract image features using hand-crafted patterns. Extracted features are used to represent an expressional image and train a classifier for FER. However, it is difficult for researchers to design a hand-crafted pattern that can be suitable for different conditions.

In recent years, Convolutional Neural Network (CNN) has been widely implemented in the research of computer vision and performs well in the FER task. CNN-based methods usually represent an image by feature maps in deep layers, which contain the global semantic information of the whole image. However, previous researches on FER have shown that expressional changes can be highly related to local facial regions [6–8], while a basic CNN is unable to learn meaningful patches for FER and always treats all regions equally in classification. This means that the basic CNN may not take full advantage of the extracted features for expression recognition. Intuitively, it will be more effective if we can give weights to different image patches or regions to indicate their importance for FER task.

In this paper, we propose a novel deep learning based model for solving the FER problem, called Deep Attentive Multi-path Convolutional Neural Network (DAM-CNN). As shown in Fig. 1, DAM-CNN consists of three parts, including the feature extraction module, the attention-based Salient Expressional Region Descriptor (SERD) and the Multi-Path Variation-Suppressing Network (MPVS-Net). In our method, we first extract features from expressional images using the VGG-Face network. Then we utilize SERD to adaptively produce a unique attentive mask for the extracted features, which discriminates the features that are related to expressions and helps locate some expression-sensitive regions. The MPVS-Net follows the SERD module to further generate a high-level representation by disentangling expressional information from different variations (e.g. genders, races, etc.). By jointly combining these modules,

* Corresponding author.
E-mail address: huhaif@mail.sysu.edu.cn (H. Hu).

https://doi.org/10.1016/j.patcog.2019.03.019
0031-3203/© 2019 Elsevier Ltd. All rights reserved.

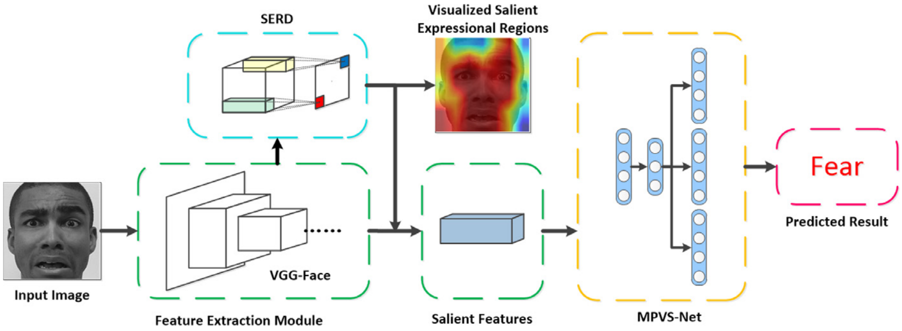


Fig. 1. The architecture of DAM-CNN.

DAM-CNN can effectively classify different expressions. The contributions of our work can be summarized as follows:

(1) We propose the DAM-CNN model for solving the FER problem. Different from traditional methods, DAM-CNN can adaptively locate some expression-sensitive regions for each expressional image and generate a high-level representation robust to different variations. Our proposed model is a generic framework, which can be generalized to similar classification tasks.
(2) The proposed SERD is able to adaptively quantify the importance of each image region for FER task. With SERD, we can filter out redundant features and retain the salient features that are efficient for the FER task.
(3) A new encoder-decoder framework, called MPVS-Net, is designed for suppressing multiple variations such as pose changes, gender differences etc. By jointly training a specific encoder and multiple decoders, MPVS-Net can disentangle expressional information from multiple variations. High-level representations learnt by the MPVS-Net can be more effective for expression classification.
(4) The proposed DAM-CNN model is evaluated on six different datasets including both constrained and unconstrained datasets. Experimental results on these datasets show that our model outperforms many existing works in FER task.

2. Related works

Expression recognition has long been a challenging task in the field of computer vision. In previous works, researchers usually attempt to solve the FER problem by detecting expression-related AUs or training a classifier with extracted expressional features. In recent years, deep learning algorithm is applied to expression analysis, which is regarded as a promising approach to recognize expressions.

AU-based methods are motivated by the work of Facial Action Coding System (FACS) [5], which aims to detect and analyze AUs on human faces. Tong et al. [9] model the mutual relationship among different AUs through a Dynamic Bayesian Network (DBN). Tian et al. [10] recognize seven upper face AUs using a multi-state facial component model. Sandbach et al. [11] focus on detecting AUs on 3D faces. Some works analyze facial behaviors and attempt to automatically detect AUs [12,13]. Pumarola et al. [14] even make use of AUs to generate vivid human expressions. However, tiny local changes on faces are difficult to be detected, which makes it hard to accurately identify different AUs. Some other researchers extract appearance features to train classifiers for expression recognition [15,16]. Some hand-crafted patterns, such as Local Binary Pattern (LBP) and Local Directional Pattern (LDP), can well represent some special properties of images and thus be introduced to solve the FER problem [17–19]. Some works fuse several kinds of features to incorporate different information, e.g. Majumder et al. [7] feed the extracted LBP and geometrical features to a 2-layer autoencoder, which generates fused features to represent an expressional image. Yan [20] utilizes different feature descriptors for feature extraction and learn multiple distance metrics for FER in videos. Moeini et al. [21] selected three types of handcrafted features to represent images. The extracted features are used to learn both identity and expression dictionaries simultaneously.

In recent years, CNN has become a popular tool for expression analysis. Zhao et al. [22] detect AUs through Deep Region and Multi-label Learning (DRML) algorithm. Mollahosseini et al. [23] use a CNN model with two convolutional layers and four inception layers FER. Their model performs well in many different expressional databases. Hamester et al. [24] construct a 2-channel network for feature extraction. Features extracted from these two channels are fused and then used for further classification. Yang et al. [25] propose the De-expression Residue Learning (DPL) to extract expression information by reconstructing a neutral face for each expressional face. Jung et al. [26] train a CNN as well as a deep neural network independently and fuse them through a fully connected layer by joint fine-tuning. Vielzeuf et al. [27] combine the VGG model and LSTM to conduct video emotion analysis. Zhu et al. [28] introduce Generative Adversarial Network (GAN) to make data augmentation for emotion classification. Their model can effectively classify some of the typical expressions. Zhang et al. [29] train their model with some generated expressions with different poses, which leads to a better performance on FER. Çuğu et al. [30] propose a lightweight model named MicroExpNet for expression recognition. Their model is trained following the concept of knowledge distillation and performs well in both run time and memory requirements.

Most aforementioned CNN-based methods represent an input image with the extracted feature maps. In general, all elements in feature maps are equally treated in classification. However, expressional analysis mainly focuses on local facial regions [6–8]. This fact reveals that only a part of extracted features are beneficial to FER. To highlight the most helpful information, some works introduce the attention mechanism to discriminate distinctive features. For example, Fu et al. [31] introduce a region-based attention algorithm into their image caption model. The attention algorithm induces their model to focus on different image regions in different time steps. Chen et al. [32] propose an attention model for image segmentation. Their model focuses on different objects as image scale varies. Sharma et al. [33] propose a visual attention mechanism to handle action recognition task. They weight some regions of the current frame based on their attention on some previous frames. Motivated by these works, we propose an attention-based module, named SERD, to focus our model on expression-related regions and discriminate features that are beneficial to FER task.

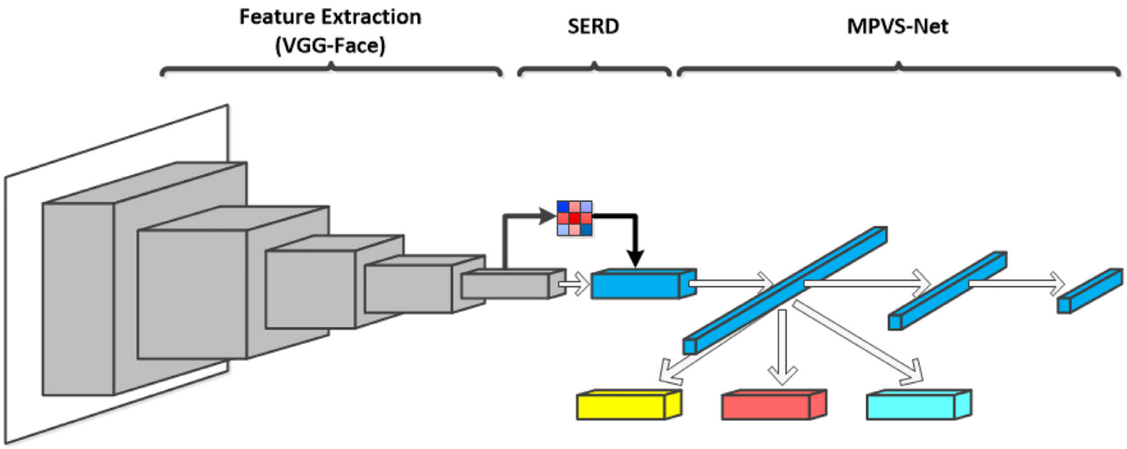


Fig. 2. Network architecture of DAM-CNN.

3. Proposed method

As shown in Fig. 2, the proposed DAM-CNN model consists of three modules, i.e., the VGG-Face network for extracting features, SERD for refining CNN features and highlighting salient expressional regions, and MPVS-Net for generating a high-level representation robust to multiple variations. The proposed model is built by two steps. In the first step, we construct a model by jointing VGG-Face with SERD, which is denoted as VGG-SERD in the following. In the second step, we build DAM-CNN by introducing MPVS-Net into the trained VGG-SERD model.

3.1. VGG-face For deep feature extraction

CNN has been successfully applied in image analysis. In our method, we apply the famous VGG-Face [34] model for feature extraction.

VGG-Face is a typical CNN model with 16 convolutional layers, 5 pooling layers and 3 fully-connected layers for face recognition. To extract expressional features, we fine-tune the pre-trained VGG-Face network on expressional datasets. Specifically, we reshape and reinitialize the last fully-connected layer to adapt it to FER task. For each facial image, feature maps from the last pooling layer with the size of $7 \times 7 \times 512$ are used to represent an image.

3.2. Salient expressional region descriptor for feature refinement

As FACS shows, most of the human expressions can be represented by a set of AUs, which reveals that expressional actions are relevant to a few local facial regions. However, VGG-Face merely extract features from all image regions without distinction. Extracted features inevitably contain some redundant information from expression-unrelated regions. Moreover, different facial regions contribute unequally for FER, which motivates us to treat them discriminatively in terms of their importance for the classification task.

In order to remedy the aforementioned defects of typical CNN on FER task, we propose the SERD to modify the VGG-Face model. SERD is an attention-based module, which is used to highlight discriminative features and detect expression-related regions in faces. The framework of SERD is shown in Fig. 3. In SERD, the extracted CNN features are firstly fed into the attention network, which outputs an attentive mask to quantify the importance of each position in feature maps. The extracted features will be weighted by the attentive mask and be activated before being delivered to the next module. Concretely, we denote the extracted CNN features as $\boldsymbol{h_c}$ and the attentive mask as $\boldsymbol{M}$. Similar to the work of [32,33], we adopt a one-layer convolutional model to obtain the attentive mask, which can be formulated as follows:

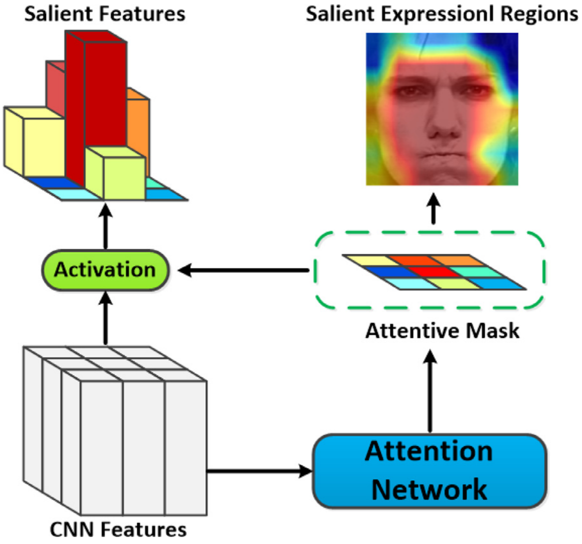


Fig. 3. Framework of the SERD.

$$\boldsymbol{M} = f_a(\boldsymbol{W_a} * \boldsymbol{h_c} + \boldsymbol{B_a}) \tag{1}$$

where $\boldsymbol{W_a}$ is the convolutional kernels of the attention network, $\boldsymbol{B_a}$ is the corresponding bias, " $*$ " indicates the operation of convolution and $f_a(x)$ is a nonlinear function. In our model, the size of $\boldsymbol{W_a}$ is set as $1 \times 1 \times 512 \times 1$. Therefore, each element (attentive weight) of the output mask is only related to the features that in the same position across channels. As features in the same position are corresponding to the same image region of the input, each attentive weight can reflect the degree of importance level of the corresponding region. In this way, the attentive mask is capable of preserving the spatial information of the input.

Intuitively, features extracted from expression-related regions can be more important in FER task, while features extracted from some marginal parts may be redundant. To quantify the importance of each feature and filter out redundant features simultaneously, we choose tanh as the nonlinear function in our attention module. By using tanh, the value of the attentive mask can be limited within the range of $(-1, 1)$. In our model, features with negative attentive weights will be regarded as redundant features, which should be filtered out. Features with positive weights will be regarded as expression-related features and therefore should be retained. Also, the region of the image with respect to a high attentive weight will be regarded as salient expressional region. In order to discriminate the features beneficial to FER, in the proposed SERD, extracted feature maps are firstly weighted by the attentive mask and then be activated by a Rectified Linear Unit (ReLU) function. The activated features, i.e., the salient features, can be expressed as:

$$\boldsymbol{h_s} = g(\boldsymbol{M} \odot \boldsymbol{h_c}), \tag{2}$$

where $g(x)$ is the ReLU function,"$\odot$" is the operation of element-wise multiplication and $\boldsymbol{h_s}$ indicates the salient features. By combining the function of tanh and ReLU, SERD is able to truncate redundant features and fine-tune expression-related features by weighting them with a positive attentive weight.

To avoid obtaining a trivial solution, we introduce a regularization term into the proposed attention mechanism. Concretely, the regularization term consists of two parts, which can be denoted as follows:

$$L_{reg} = \gamma_1 \|\boldsymbol{M}\|_1^2 + \gamma_2 \|\boldsymbol{1} - \boldsymbol{M} \odot \boldsymbol{M}\|_F^2, \tag{3}$$

where γ_1 and γ_2 are adjustable coefficients to balance the loss. The first part prevents the attentive mask from being saturated and keeps the mask sparse. On the contrary, the second part tends to assign each weight with a relatively large value. As these two parts have different effect on the attentive mask, SERD should learn to make a balance between them. Therefore, only some of the extracted features that are effective for FER will be assigned with large attentive weights. Features that have little effect on FER will be assigned with relatively small attentive weights. We evaluate the effect of each part of Eq. 3 in Section 4.2.2. Visualization results (Sections 4.2.1 and 4.2.2) show that SERD with the regularization term can really focus its attention on some expression-related regions.

By jointing VGG-Face with SERD, we can construct the model of VGG-SERD. In our training, we fine-tune VGG-SERD with expression images. The loss function of VGG-SERD can be expressed as:

$$L_{vs} = L_{ce-vs} + L_{reg} \tag{4}$$

$$L_{ce-vs} = -\frac{1}{n}\sum_{i=1}^{n} p_i \log \widetilde{p}_i \tag{5}$$

where L_{ce-vs} is the cross-entropy corresponding to the predicted results of VGG-SERD, n is the number of samples, p_i is the ground-truth label of the ith sample and $\widetilde{p}_i$ is the predicted result with respect to the ground-truth class. By introducing the regularization term into the loss function, SERD is forced to discriminate features that are not only related to expressions but also effective to the final classification task.

3.3. Multi-Path variation-suppressing network

By introducing the SERD into VGG-Face, we can filter out some redundant features of images. However, as samples are different with each other, variations such as genders, appearance or skin color can be different in expression-related regions among different images. Examples are shown in Fig. 4. Therefore, features retained by SERD still involve some information about variations, which is unnecessary to FER and may degrade the performance of the model. Inspired by Ghifary et al. [35], we propose the MPVS-Net to handle different variations. MPVS-Net is constructed by a three-layer network, where the first layer is based on a Multi-Path AutoEncoder (MPAE) and the following are two fully-connected layers. In our method, we replace all fully-connected layers of VGG-SERD with the MPVS-Net. The model that combines VGG-SERD and MPVS-Net is denoted as DAM-CNN. With the help of MPVS-Net, we can disentangle expressional information from different variations and generate a high-level representation robust to multiple variations for each image.

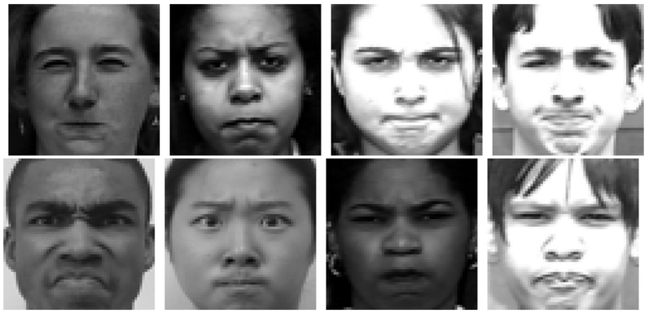


Fig. 4. An example about variations (sampled from the CK+ dataset): all subjects perform the same expression (anger), but faces differ in appearance, genders, complexion, etc.

3.3.1. Multi-path autoencoder

The most important part of MPVS-Net is the MPAE. As shown in Fig. 5, the MPAE consists of three parts: the input layer, the hidden layer (an encoder) and the decoding layer (multiple decoders).It can be regarded as an extended autoencoder with multiple decoding targets. The purpose of MPAE is to jointly achieve good reconstruction of different samples among all decoders when given the same input sample. The decoding targets and the input sample should have the same expression label but different variations. By reconstructing different samples, we expect the encoder to learn an intermediary representation that is robust to different variations while decoders can learn the knowledge of different variations. In this way, the learned intermediary representation can serve as the high-level representation for further classification.

The MPAE can be formulated as follows. The input sample of MPAE is denoted as $\boldsymbol{s_i}$, where $i (i = 1, 2, \ldots, n)$ indicates the ith sample. In the training stage, we construct a sample pair $\{\boldsymbol{s}_i, \boldsymbol{s}_i^j\}$ for each input sample, where $\boldsymbol{s}_i^j$ is the expected reconstruction targets of the jth $(j = 1, 2, \ldots, N)$ decoder with respect to $\boldsymbol{s}_i$. $\boldsymbol{s}_i$ and $\boldsymbol{s}_i^j$ should have the same class label. In our model, all samples used in MPAE are represented by salient features. To obtain the intermediary representation, the operation of the encoder f_E can be formulated as follows:

$$\boldsymbol{h}_i = f_E(\boldsymbol{s}_i) = sig(\boldsymbol{W_E} \cdot \boldsymbol{s}_i + \boldsymbol{B_E}), \tag{6}$$

where $\boldsymbol{h}_i$ is the intermediary representation with respect to $\boldsymbol{s}_i$, $\boldsymbol{W_E}$ is the weight between the input layer and the hidden layer, $\boldsymbol{B_E}$ is the corresponding bias. $sig(x)$ is the sigmoid function, which serves as a activation function in the model. To reconstruct the decoding target, the operation of the jth decoder f_D^j can be formulated as:

$$\hat{\boldsymbol{s}}_i^j = f_D^j(\boldsymbol{h}_i) = sig(\boldsymbol{V}^j \cdot \boldsymbol{h}_i + \boldsymbol{B}_{\boldsymbol{D}}^j) \tag{7}$$

where $\hat{\boldsymbol{s}}_i^j$ is the decoded result of the jth decoder, $\boldsymbol{V}^j$ and $\boldsymbol{B_D}^j$ are the corresponding weight and bias. Each decoder is expected to reconstruct a different sample using the same encoded representation $\boldsymbol{h}_i$.

In our method, we use L_2-norm to measure the reconstruction performance. Therefore, the reconstruction loss of the jth decoder

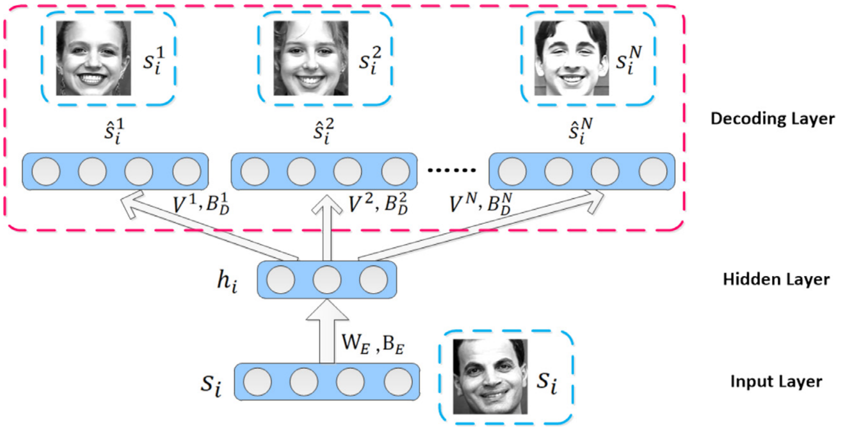


Fig. 5. Architecture of the MPAE.

can be defined as:

$$\begin{aligned} J_j(\boldsymbol{W_E}, \boldsymbol{B_E}, \boldsymbol{V}^j, \boldsymbol{B_D^j}) &= \frac{1}{n}\sum_{i=1}^{n} \|\hat{\boldsymbol{s}}_i^j - \boldsymbol{s}_i^j\|_2^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \|f_D^j(\boldsymbol{h}_i) - \boldsymbol{s}_i^j\|_2^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \|f_D^j(f_E(\boldsymbol{s}_i)) - \boldsymbol{s}_i^j\|_2^2 \end{aligned} \tag{8}$$

where n is the number of samples. As labels of specific variations are unavailable in most expression datasets, it is impracticable to assign each decoder with a specific variation. In our experiment, the decoding target $\boldsymbol{s}_i^j$ is randomly selected from the training set. As we have mentioned above, the only same information shared between $\boldsymbol{s}_i$ and $\boldsymbol{s}_i^j$ is their expression. In order to achieve good reconstruction of the target, the encoder should learn to generate a variation-robust representation while each decoder should learn the knowledge of different variations. Therefore, the weights of the encoder and decoders can be optimized by minimizing the overall reconstruction loss J_{all} as follows:

$$\begin{aligned} [\boldsymbol{W_E}, \boldsymbol{B_E}, \boldsymbol{V}^j, \boldsymbol{B_D^j}] &= \underset{\boldsymbol{W_E}, \boldsymbol{B_E}, \boldsymbol{V}^j, \boldsymbol{B_D^j}}{\arg\min}\ J_{all} \\ &= \underset{\boldsymbol{W_E}, \boldsymbol{B_E}, \boldsymbol{V}^j, \boldsymbol{B_D^j}}{\arg\min} \sum_{j=1}^{N} J_j(\boldsymbol{W_E}, \boldsymbol{B_E}, \boldsymbol{V}^j, \boldsymbol{B_D^j}) \end{aligned} \tag{9}$$

Following this training rule, expressional information can be effectively disentangled from variations. The encoder is induced to learn a representation that is only related to expressions, which can be regarded as a high-level representation for further classification.

To better explain the effect of MPAE, in the rightmost column of Fig. 6, we visualize the distribution of high-level representations of samples from two constrained datasets (CK+ and JAFFE) and two unconstrained datasets (BAUM-2i and SFEW). Visualizations are conducted using t-SNE [36], which is a widely used tool for visualizing high-dimensional data. The goal of this visualization study is to show the performance of MPAE in reducing the influence of variations. To make a comparison, the distributions of samples represented by raw pixels (original image) and salient features (extracted through VGG-SERD) are also shown in Fig. 6. These three types of image representations, i.e., raw pixels, salient features and high-level representation, are denoted as $\boldsymbol{R_1}$, $\boldsymbol{R_2}$ and $\boldsymbol{R_3}$ respectively. In the left column of Fig. 6, we can clearly see that samples represented by $\boldsymbol{R_1}$ distribute irregularly in the space. This is reasonable as variations have not been suppressed in original images. It is hard to separate different expressional images in the raw pixel space. By employing VGG-SERD, the distribution of $\boldsymbol{R_2}$ (the middle column of the Fig. 6) tends to form several clusters in the space. Most samples in the same cluster are from the same expression class. Since variations are not well handled by VGG-SERD, overlap can still be observed among different clusters. In contrast to other two types of representations, the distribution of $\boldsymbol{R_3}$ (the rightmost column of Fig. 6) are clearly separated into several clusters except for that of SFEW. MPAE performs better especially when there are limited variations (e.g., in CK+ and JAFFE). This fact reveals that variations have been well suppressed, which verifies the effectiveness of MPAE. As variations in SFEW are more challenging than that of other three datasets, samples from different classes still overlap with each other when represented by $\boldsymbol{R_3}$. However, as Fig. 6 illustrates, most samples of anger, sadness and happiness represented by $\boldsymbol{R_3}$ have already been distinguished from samples of other classes. In other words, MPAE still works even though there are some challenging variations in images.

3.3.2. Classification using multi-path variation-suppressing network

Having constructed the MPAE, we can effectively disentangle expressions from different variations. As we can obtain the disentangled representation through the hidden layer of MPAE, we then joint two fully-connected layers with the hidden layer to build up the MPVS-Net, as shown in Fig. 2. In the proposed model, the first fully-connected layer will be activated by the ReLU function while the second one is followed by a softmax classifier to conduct expression classification.

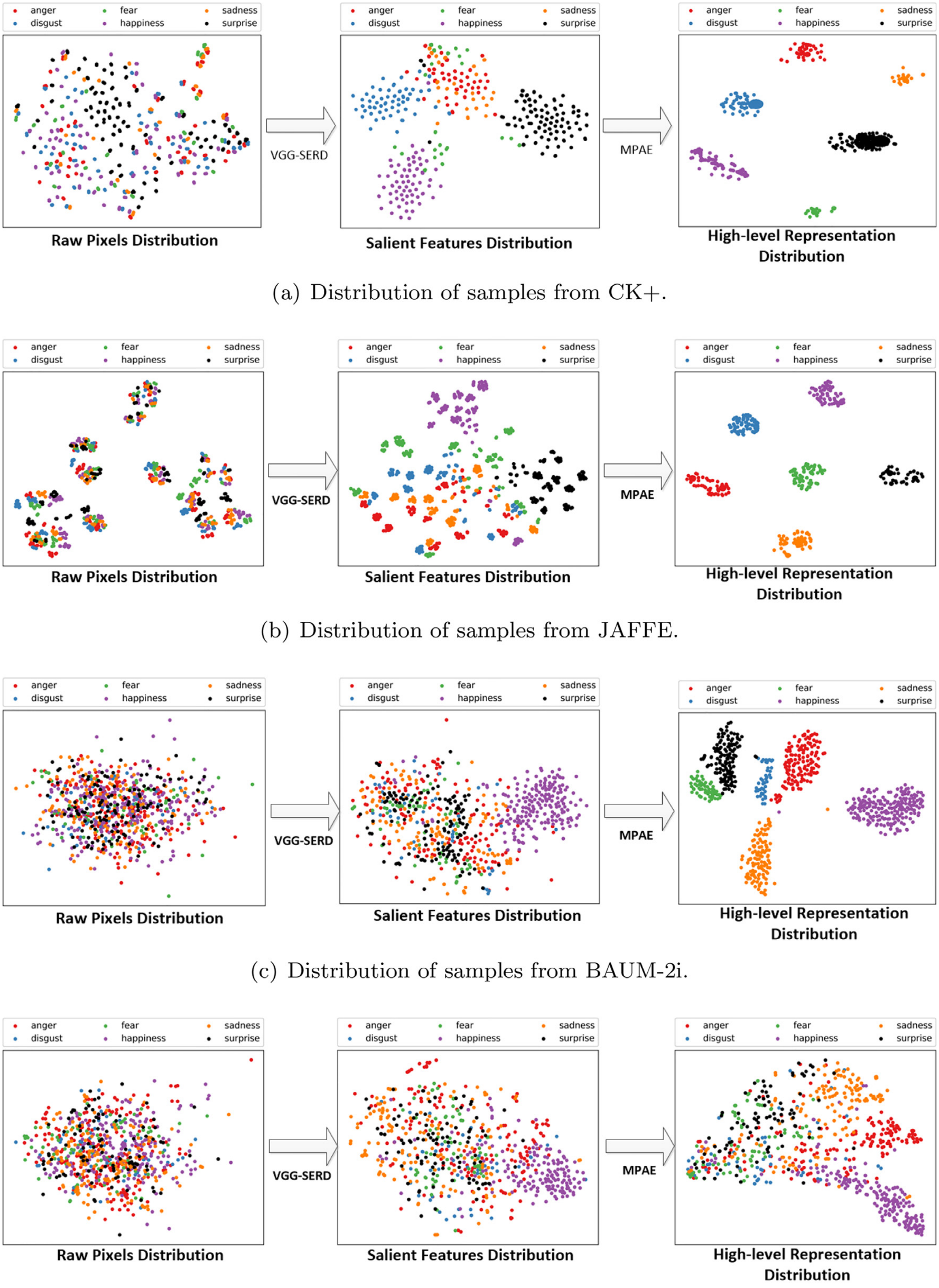
The MPAE and fully-connected layers are updated following different rules. As for the two fully-connected layers, parameters are updated as follows:

$$[\boldsymbol{W}, \boldsymbol{B}] = \underset{\boldsymbol{W}, \boldsymbol{B}}{\arg\min} L_{ce-D}(\boldsymbol{W}, \boldsymbol{B}), \tag{10}$$

where L_{ce-D} indicates the cross-entropy of the classification results of DAM-CNN, $\boldsymbol{W}$ and $\boldsymbol{B}$ are the weight and bias with respect to these two fully-connected layers.When training the MPAE, we modify its loss function to make it adaptive to our classification task. Other than the reconstruction loss (Eq. 8), we additionally introduce the cross-entropy of the classification result to regularize MPAE. Thus the parameters of MPAE can be updated as follows:

$$\begin{aligned} [\boldsymbol{W_E}, \boldsymbol{B_E}, \boldsymbol{V^j}, \boldsymbol{B_D^j}] &= \underset{\boldsymbol{W_E}, \boldsymbol{B_E}, \boldsymbol{V^j}, \boldsymbol{B_D^j}}{\arg\min}\ J_{all} + \lambda L_{ce-D} \\ &= \underset{\boldsymbol{W_E}, \boldsymbol{B_E}, \boldsymbol{V^j}, \boldsymbol{B_D^j}}{\arg\min} \sum_{j=1}^{N} J_j + \lambda L_{ce-D} \end{aligned}, \tag{11}$$

where λ is a coefficient for tuning the impact between the task of classification and reconstruction. Regularized by the cross-entropy



(a) Distribution of samples from CK+.

(b) Distribution of samples from JAFFE.

(c) Distribution of samples from BAUM-2i.

(d) Distribution of samples from SFEW.

Fig. 6. A visualization of the sample distribution under three types of representations. Left to right in each panel: samples represented by the raw pixels (R_1), salient features (R_2) and high-level representation (R_3). Visualizations are conducted on two constrained datasets (with limited variations): (a) CK+ and (b) JAFFE, and two unconstrained datasets (with many variations): (c) BAUM-2i and (d) SFEW. (Best viewed in color.).

of the classification result, MPAE should learn to generate a high-level representation that is not only robust to multiple variations, but also suitable for the classification task. Such processing makes our model more effective to the FER task.

Concretely, we train DAM-CNN by two steps. The training algorithm is summarized in Algorithm 1. In the first step, we fine-tune the VGG-SERD with expressional images. The kernels and biases of the SERD are randomly initialized. The parameters of the last fully-connected layer of VGG-SERD will be reinitialized to match the FER task. We fine-tune VGG-SERD until its loss function (Eq. 4) converges. In the second step, we introduce the MPVS-Net into VGG-SERD to construct DAM-CNN. The parameters of feature extraction module and SERD will be fixed when we train MPVS-Net (i.e., only MPVS-Net will be updated). As we have trained VGG-SERD in the first step, we can obtain all $\boldsymbol{s_i}$ and $\boldsymbol{s_i^j}$ before starting the second step. In our experiments, we store all sample pairs, i.e., $(\boldsymbol{s}_i, \boldsymbol{s}_i^j)$, in preparation for the subsequent training. We train DAM-CNN until MPVS-Net's loss function (Eq. 11) converges. Having considered the risk of overfitting, we introduce dropout [37] to train our model. In these two training stages, dropout is applied to the fully-connected layers as well as the MPAE to enhance the generalization ability of our model.

Algorithm 1 Training of DAM-CNN.

```
Require: Training data $\{(x_i, y_i), i = 1, 2, \dots, n\}$, $x_i$ is the $i$-th expressional image, $y_i$ is the corresponding class label. Learning rate $\alpha_S$ for VGG-SERD, $\alpha_M$ for MPAE and $\alpha_{fc}$ for fully-connected layers of MPVS-Net. Coefficients $\gamma_1$, $\gamma_2$ and $\lambda$.
Ensure: Parameters of DAM-CNN
while loss of VGG-SERD does not converge do
    Fine-tune $\boldsymbol{W_a}$ and $\boldsymbol{B_a}$ with $\alpha_S$ by minimizing Eq. (4);
end while
Fix parameters of feature extraction module and the SERD;
while loss of MPVS-Net dose not converge do
    for $j = 1$ to $N$ do
        Update $\boldsymbol{W_E}$ and $\boldsymbol{B_E}$ with $\alpha_{\boldsymbol{M}}$ by Eq. (11);
        Update $\boldsymbol{V^j}$ and $\boldsymbol{B_D^j}$ with $\alpha_{\boldsymbol{M}}$ by Eq. (11);
        Update $\boldsymbol{W}$ and $\boldsymbol{B}$ with $\alpha_{fc}$ by Eq. (10);
    end for
end while
```

4. Experimental evaluation

In this section, extensive experiments are carried out on six public expression datasets to evaluate the proposed method. The evaluation works have the following objectives:

(1) Investigate the various properties of the two novel modules in DAM-CNN, i.e., the SERD and the MPVS-Net.
(2) Evaluate the performance of our DAM-CNN model on expression recognition. The recognition accuracy will be compared with some other competitive conventional methods as well as deep-based approaches.
(3) Evaluate the generalization ability of DAM-CNN across different datasets.

Before presenting the experimental results, we first give a brief introduction of all datasets and the detailed configuration of our model. Classification results and the corresponding analysis are presented in the following sections.

4.1. Experimental setup

4.1.1. Datasets

We evaluate our model on six public expressional datasets, which are captured in both constrained and unconstrained scenarios. The constrained datasets include the Extended Cohn-Kanade (CK+) dataset [38], the Japanese Female Facial Expression (JAFFE) database [39], and the Taiwanese Facial Expression Image Database (TFEID) [40]. The unconstrained datasets include the BAhcesehir University Multilingual Affective Face Database (BAUM-2i) [41], the Static Facial Expressions in the Wild (SFEW) [42] and the Facial Expression Recognition 2013 (FER2013) [43]. Some examples of these datasets are shown in Fig. 7. We briefly introduce these six datasets in the following.

CK+. The CK+ dataset contains 593 image sequences across 123 subjects. All these sequences are from neutral face to the peak expression. In this dataset, only 309 sequences are labeled with one of the six prototypical expressions and thus are selected out for experiments. For each selected sequence, we pick out the last three image frames to construct the training and testing sets. Additionally, the first frame of each selected sequence are extracted as neutral face. Therefore, 927 images (6 expressions) or 1236 images (7 expressions) are involved in our experiments.

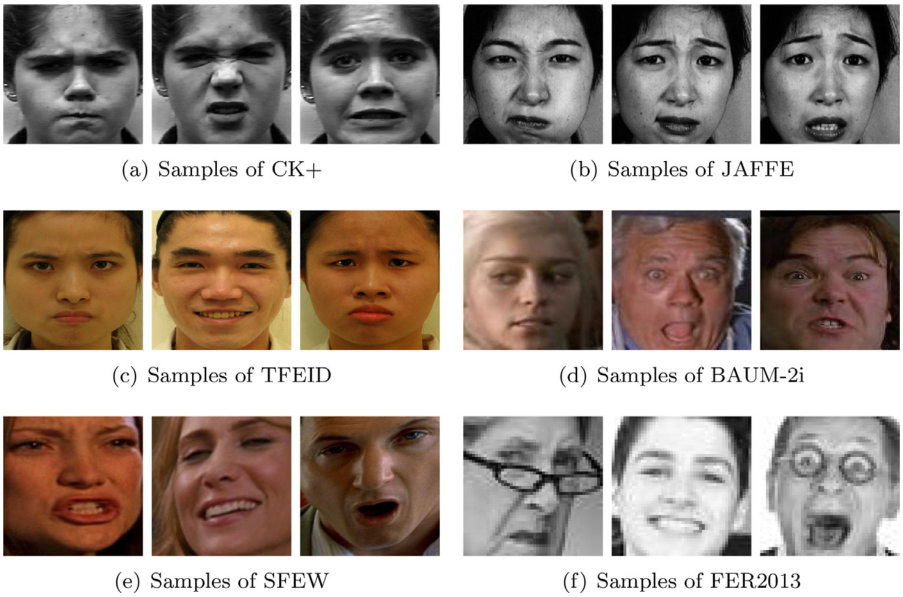
JAFFE. The JAFFE database consists of 213 images sampled from 10 Japanese female models. Only 183 images are labeled with one of the basic expressions and others are neutral images. As there are too few images available for training, we augment the data before experiment. We flip all images to obtain their corresponding mirror images. We rotate the original images by the angle of 5° in clockwise and counterclockwise respectively. Additionally, Gaussian noise with zero mean and 0.01 variance is added to the original image. Therefore, we finally obtain 1065 images for our experiments.

TFEID. The images of TFEID are captured from 40 models under two kinds of intensities (high and slight). For each subject, eight classes of expressions are collected, which includes neutral, anger, contempt, disgust, fear, happiness, sadness and surprise. In our experiments, we only pick out the images that are labeled with one of the six basic expression or the neutral. Therefore, 580 images of TFEID are used in our experiments.

BAUM-2i. The BAUM-2i is a static expression dataset. All images of BAUM-2i are extracted from BAUM-2, which is a dataset of audio-visual affective facial clips collected from movies and TV series. Each image of BAUM-2i is labeled with one of the eight expressions (neutral, anger, contempt, disgust, fear, happiness, sadness and surprise) as well as the genders (male or female). Since all images are in the close-to-real-life conditions (i.e., with pose, age and illumination variations, etc.), the BAUM2-i is more challenging than those constrained databases. In our experiments, 998 images labeled with one of the seven expressions (six basic expression and the neutral) are used to evaluate our model.

SFEW. SFEW is developed by selecting frames from Acted Facial Expressions in the Wild (AFEW), which is a dynamic facial expression dataset extracted from movies. SFEW dataset covers unconstrained facial expressions, varied head poses, large age range, occlusions, varied focus, different resolution of face and close to real world illumination. All images are labeled with one of the seven expressions (six basic expression and the neutral). The training set contains 847 images and the validation set contains 409 images.

FER2013. The FER2013 consists of 35,887 images, each of which is labeled with one of the seven expressions (six basic expression and the neutral). Samples of this dataset are all grayscale image with the size of 48 × 48. The training set consists of 28,709 images while the validation and testing set both contains 3589 images.


(a) Samples of CK+

(b) Samples of JAFFE

(c) Samples of TFEID

(d) Samples of BAUM-2i

(e) Samples of SFEW

(f) Samples of FER2013

Fig. 7. Some examples of the six datasets (i.e., CK+, JAFFE, TFEID, BAUM2-i, SFEW and FER2013).

Table 1
The network configuration of DAM-CNN (N: decoders number, K: class number).

Input: 224 × 224 × 3	
VGG-Face(ConvNet)	Conv(3 × 3 × 3 × 64)
	Conv(3 × 3 × 64 × 64), ReLU
	Maxpool(2 × 2)
	Conv(3 × 3 × 64 × 128), ReLU
	Conv(3 × 3 × 128 × 128), ReLU
	Maxpool(2 × 2)
	Conv(3 × 3 × 128 × 256), ReLU
	Conv(3 × 3 × 256 × 256), ReLU × 2
	Maxpool(2 × 2)
	Conv(3 × 3 × 256 × 512), ReLU
	Conv(3 × 3 × 512 × 512), ReLU × 2
	Maxpool(2 × 2)
	Conv(3 × 3 × 512 × 512), ReLU × 3
	Maxpool(2 × 2)
SERD	Conv(1 × 1 × 512 × 1), tanh
MPVS-Net	MPAE Encoder(25088 × 4096), Sigmoid
	Decoders(4096 × 25088), Sigmoid × N
	FC-Layer(4096 × 4096), ReLU
	FC-Layer(4096 × K)
	Softmax

In the preprocessing stage, facial regions are detected by the Viola-Jones faces detector [44]. All images used in our experiments are resized to 224 × 224 to match the input size of our model.

4.1.2. Validation setup

To evaluate the performance of our model, cross-validation is applied to all of our experiments. To compare with other competitive methods, we implement 10-fold cross-validation on the experiments on CK+ and JAFFE. 5-fold cross-validation is implemented on TFEID, which is the same as the setting of [45]. 7-fold cross-validation is implemented on BAUM-2i following the PPI protocol of [41]. As for SFEW and FER2013, our model are trained using the predefined training set and evaluated on the public validation set (SFEW, FER2013) and testing set (FER2013).

In the experiments, we set $\alpha_a = 1 \times 10^{-5}$, $\alpha_M = \alpha_{fc} = 1 \times 10^{-4}$, $\gamma_1 = 1 \times 10^{-3}$, $\gamma_2 = 1 \times 10^{-4}$ and $\lambda = 20$. The dropout rate is 0.5. The architecture of our model is shown in Table 1.

Table 2
Effect of SERD by averaged recognition accuracy and standard deviation (%).

Datasets	VGG-Face	VGG-SERD
CK+	93.59 ± 4.72	93.98 ± 2.51
JAFFE	95.05 ± 4.84	96.37 ± 4.12
TFEID	91.47 ± 2.96	92.01 ± 2.59
BAUM-2i	58.01 ± 2.93	57.89 ± 3.79
FER2013(testing set)	62.99 ± 0.37	62.22 ± 0.42
SFEW	39.04 ± 0.30	40.75 ± 1.27

4.2. Analysis of SERD

To verify the effectiveness of SERD on FER task, we make an analysis on the SERD module in this section. We first make an comparison on the performance of VGG-SERD and VGG-Face and visualize the regions where SERD focuses on. Then, we conduct a series of ablation experiments to evaluate the effect of each part in the regularization term (refer to Eq. 3).

4.2.1. Performance of SERD

In our method, we propose SERD to discriminate salient features and locate expression-related regions. To evaluate this attention-based module, we compare the performance of the following two models: the VGG-Face and the VGG-SERD. Experiments are conducted on all six aforementioned datasets.

The comparison results are displayed in Table 2. We can observe that the VGG-SERD performs slightly better than VGG-Face in all three constrained datasets (i.e., CK+, JAFFE and TFEID), which demonstrates the effectiveness of the proposed attention mechanism. Furthermore, the standard deviation of the accuracy of VGG-SERD is lower than that of the VGG-Face. This means that VGG-SERD performs more stably than the VGG-Face. This result is reasonable as SERD filters out features from expression-unrelated regions of an image. Information of these regions contributes little to FER and sometimes may burden the classification system. Removal of these redundant features makes the model more discriminative to expressions. In unconstrained conditions, VGG-SERD performs better than VGG-Face only in the SFEW dataset. This is because many extreme variations are involved in the unconstrained datasets, which make it much more challenging than that

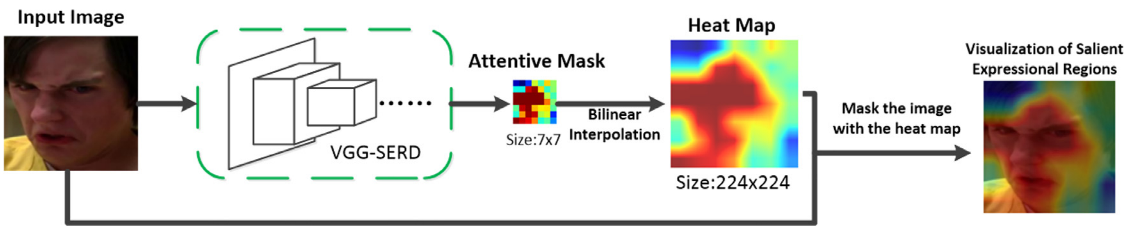
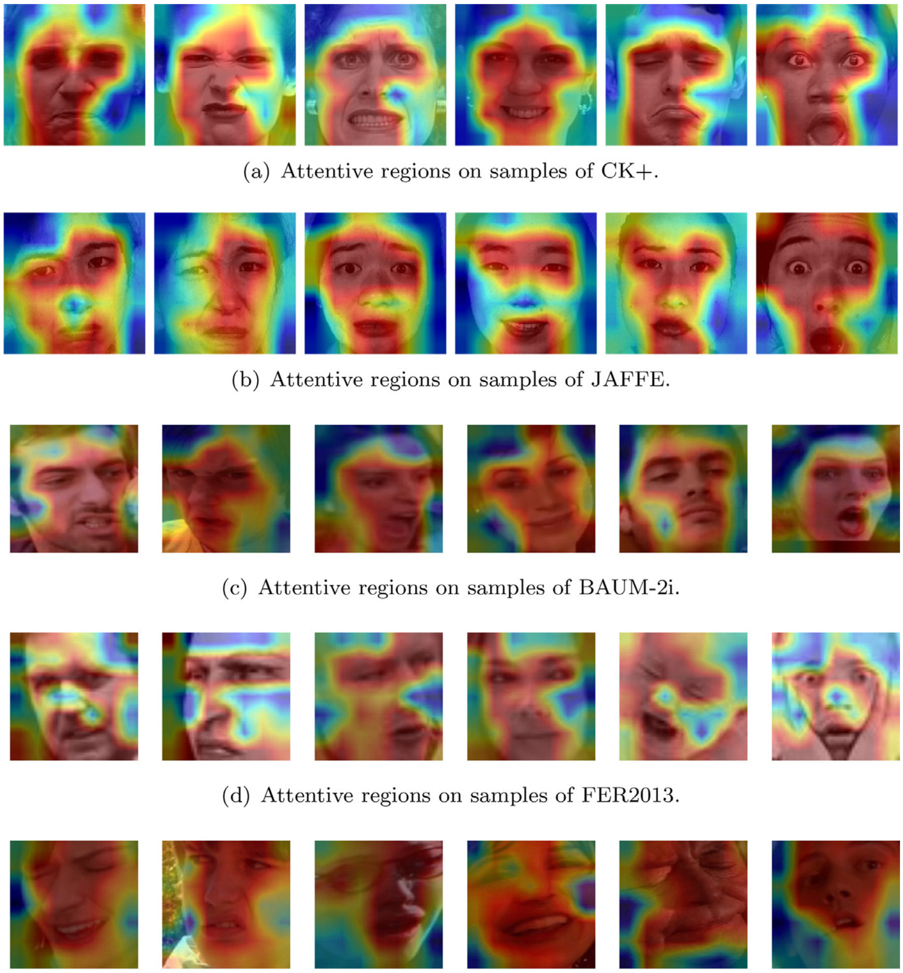


Fig. 8. The pipeline of visualizing salient expressional regions.



(a) Attentive regions on samples of CK+.

(b) Attentive regions on samples of JAFFE.

(c) Attentive regions on samples of BAUM-2i.

(d) Attentive regions on samples of FER2013.

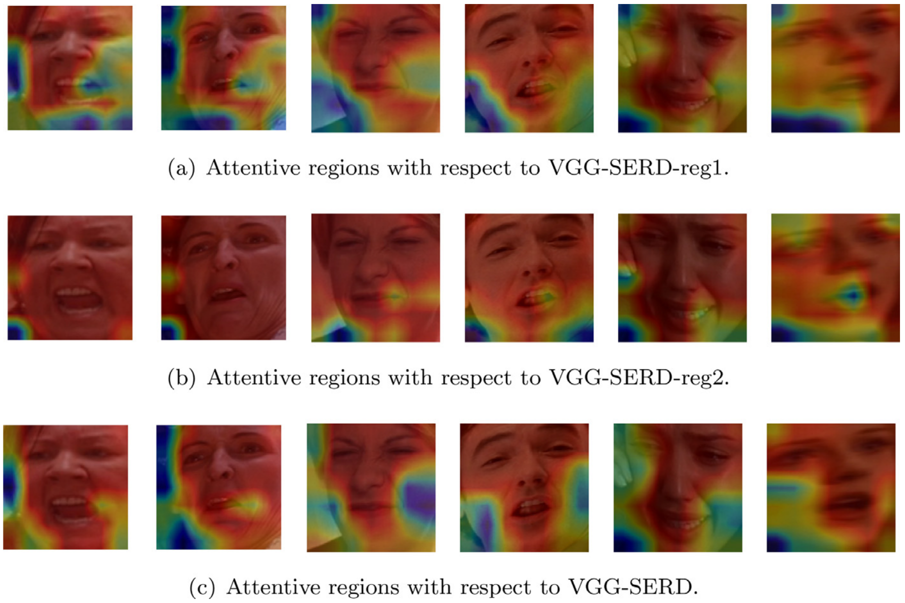
(e) Attentive regions on samples of SFEW.

Fig. 9. Visualization of the salient expressional regions in constrained dataset (CK+ and JAFFE) and unconstrained dataset (BAUM-2i, FER2013 and SFEW). Left to right in each panel: anger, disgust, fear, happiness, sadness, surprise.

of constrained dataset. In addition, the SERD only consists of a convolutional layer. It is hard for such a shallow network to handle all complex variations, which may lead to the degradation of VGG-SERD in BAUM-2i and FER2013.

To interpret the effect of SERD in a more clear perspective, we visualize the salient expressional regions that SERD focuses on. The pipeline of our visualization work is schematized in Fig. 8. Concretely, VGG-SERD will first yield an attentive mask (size: 7×7) for each image. Each element of the mask indicates the importance of a certain region in the input image. In order to match with the input image, we directly resize the mask by bilinear interpolation, which yields a heat map with the size of 224×224. In this way, we can simply mask the heat map on the original image to roughly show the salient expressional regions on the face. As shown in Fig. 9, warm-toned parts of an image correspond to regions with large attentive weights while cold-toned parts correspond to regions with relatively low attentive weights. Some properties of SERD can be observed from the distribution of the visualization works:

(1) The SERD can adaptively learn unique salient expressional regions for each image, which indicates that the SERD can well adapt to different subjects or expressional images.
(2) The SERD usually focuses its attention on regions near mouse, eyes, brows and forehead, where wrinkles often appear. Flat regions (e.g. cheek or forehead in anger or disgust expression) are usually assigned with lower attention. Some marginal or expression-unrelated regions, such as ear and the apex of nose, are assigned with relatively lower



(a) Attentive regions with respect to VGG-SERD-reg1.

(b) Attentive regions with respect to VGG-SERD-reg2.

(c) Attentive regions with respect to VGG-SERD.

Fig. 10. Some examples of the attentive regions on SFEW dataset under different model settings. Left to right in each panel: anger, disgust, fear, happiness, sadness surprise.

attentive weights. This fact implies that the SERD can automatically locate regions that are relevant to FER task.

(3) The SERD is able to filter out some redundant features or repetitive information for FER task. We can observe that some of the salient expressional regions distribute asymmetrically in faces. This is because human faces are symmetric in nature, which means that some extracted features from symmetric parts can be redundant in FER. SERD has learnt of the symmetry of faces and automatically ignore some redundant features extracted from symmetric regions. The visualization result demonstrates that SERD is able to discriminate important features for FER task.

(4) The attentive regions of those three unconstrained datasets (i.e., BAUM-2i, FER2013 and SFEW) mostly concentrate on some expression-related regions (e.g., neighborhood of eyes or mouth) even though faces are not in frontal view. This fact reveals that SERD is able to deal with facial images with some pose variations.

4.2.2. Ablation study of SERD

In Section 3.2, we regularize the SERD module with Eq. 3, which helps generate a significant attentive mask for the model. To investigate the effect of each component in this regularization term (i.e., Eq. 3), we conduct two ablation experiments to evaluate SERD. In the first experiment, we train VGG-SERD with only using the first part of the regularization term, denoted as VGG-SERD-reg1. In the second experiment, the VGG-SERD is trained with only using the second part, denoted as VGG-SERD-reg2. Experiments are conducted on the SFEW dataset. All experiments follow the aforementioned validation settings.

Table 3 shows the performance of these two models as well as the VGG-SERD. From Table 3 we can find that the VGG-SERD performs slightly better than the VGG-SERD-reg1 and VGG-SERD-reg2. This means that combining both parts of the regularization term can improve the performance of our model in some extent.

Fig. 10 visualizes some examples of the attentive regions of these three models. From Fig. 10(a), we can find that VGG-SERD-reg1 attempts to discriminate some important regions. However, the attentive regions distribute dispersedly and even randomly in some examples. This means that VGG-SERD-reg1 is unable to well focus on some specific and continuous regions. Different from VGG-SERD-reg1, VGG-SERD-reg2 can focus its attention on continuous regions, as Fig. 10(b) shows. However, the SERD of VGG-SERD-reg2 tends to assign all regions with a large attentive weights, which implies that the model fails to estimate the importance of each extracted feature for FER task. Compared with VGG-SERD-reg1 and VGG-SERD-reg2, VGG-SERD assigns larger attentive weights to the regions where expressions may occur (Fig. 10(c)). It means that SERD can discriminate some important features and adaptively focus its attention on expression-related regions. The visualization results demonstrate the fact that jointly combining these two parts of regularization term can really help our model find salient expressional regions.

Table 3
Performance of SERD on SFEW dataset under different settings (%).

Model	Accuracy
VGG-SERD-reg1	39.61
VGG-SERD-reg2	39.20
VGG-SERD	40.75

4.3. Analysis of MPVS-Net

To verify the effectiveness of MPVS-Net, we have conducted some experiments to evaluate the performance of DAM-CNN when assigned with different number of decoding paths. Additionally, we conduct another experiment on BAUM-2i to investigate MPVS-Net when dealing with specific variation (gender).

4.3.1. MPVS-net With different number of paths

In this section, we intend to investigate how path numbers in MPVS-Net will affect the performance of DAM-CNN. In experiments, we assign MPVS-Net with different number of decoding paths. Experiments are conducted in all six aforementioned datasets. All experiments follow the validation settings we specified in Section 4.1.2. Limited by computational resource, the decoding path number only varies from two to five. In addition, we compare the performance with VGG-SERD, which can be regarded as the model without MPVS-Net.

Table 4
The Recognition Accuracy (%) of DAM-CNN versus different number of paths in MPVS-Net (the best performance on each dataset is marked in bold).

Dataset	VGG-SERD	DAM-CNN			
		2-path	3-path	4-path	5-path
CK+(6-class)	93.98	94.15	94.96	95.81	**95.88**
JAFFE(6-class)	96.37	97.55	98.23	98.14	**99.22**
TFEID(7-class)	92.01	91.78	92.82	**93.36**	93.20
BAUM-2i	57.89	57.54	59.91	58.98	**61.52**
FER2013(testing set)	62.22	65.32	**66.20**	65.82	65.96
FER2013(validation set)	60.65	64.33	64.49	63.86	**65.31**
SFEW	40.75	40.59	41.89	42.14	**42.30**

The experimental results are shown in Table 4. From the result, we can observe some characters of MPVS-Net, which are summarized as follows:

(1) The model that is jointed with more than two decoding paths performs much better than VGG-SERD, which demonstrates the effectiveness of our proposed MPVS-Net. Compared with VGG-SERD, DAM-CNN further refines extracted features by generating a high-level representation. Such processing makes DAM-CNN more discriminative to different expressions, which improves the performance.

(2) The recognition accuracy of DAM-CNN tends to rise when the number of decoding paths increases. As decoder number increases, the DAM-CNN is capable to learn more information about different variations, which helps the model disentangle expressions from more variations and therefore lead to a better performance.

(3) Although the performance of DAM-CNN improves as the number of decoders increases, the gains on improvement gradually diminish. This may result from the sample assignment strategy of MPVS-Net. As decoding targets of each decoders are randomly selected, the same sample or samples of the same subject are more likely assigned to different decoders when decoding paths increase. Therefore, MPVS-Net may learn overlapped variations among different decoders. One plausible solution is to correlate each decoder with a specific variation.

(4) The recognition accuracy of DAM-CNN is expected to meet a saturation point if decoding paths continue increasing. However, as computational resource limited, we are unable to probe this point at the present stage.

4.3.2. MPVS-net with gender variation

We conduct another experiment to evaluate DAM-CNN when decoders are assigned with specific variations. This experiment is implemented on BAUM-2i. Other than expression label, gender label (male or female) is also available in BAUM-2i for each image. Therefore, in this experiment, we can construct a two-path MPVS-Net with respect to gender variation for DAM-CNN. Each decoder is only required to reconstruct a sample with a specific gender. The DAM-CNN model under this setting is denoted as DAM-CNN-g. The corresponding model with randomly selected decoding targets is denoted as DAM-CNN-r. Validation setting follows the protocol we aforementioned in Section 4.1.2.

In our experiment, an averaged recognition accuracy of 58.25% is obtained by DAM-CNN-g. As for DAM-CNN-r, the averaged recognition accuracy is 57.54%. We can observe that DAM-CNN-g performs better than the DAM-CNN-r. This result reveals that learning the knowledge of a specific variation, such as gender, can be beneficial for our model. Compared with DAM-CNN-r, the reconstruction object of MPAE is more definite in DAM-CNN-g because each decoder only need to reconstruct the input sample by a specific gender. Therefore, DAM-CNN-g is less likely learn similar or overlapped variations among different decoders, which leads to a better performance. The experimental result also demonstrates that MPVS-Net is able to disentangle variations from extracted features.

Table 5
Confusion Matrix of DAM-CNN on CK+ (%) (An: anger, Di: disgust, Fe: fear, Ha: happiness, Sa: sadness, Su: surprise).

	An	Di	Fe	Ha	Sa	Su
An	98.89	0	0	0	1.11	0
Di	1.11	97.96	0	0.37	0.56	0
Fe	2.22	0	74.44	11.11	3.33	8.89
Ha	0	0	0	100	0	0
Sa	11.48	0	2.96	0	84.07	1.48
Su	0.14	0	0.42	0.69	0.69	98.01

Table 6
Performance Comparison on CK+ (%) (loso: leave-one-subject-out, 5(10)-fold: 5(10)-fold-cross validation).

Method	Accuracy	Validation Settings
SPSD [46]	88.52	loso
DSNMF [47]	90.92	loso
DTAN [26]	91.44	10-fold
MCSPL [48]	91.53	10-fold
DTGN [26]	92.35	10-fold
Ali-Net [23]	93.20	5-fold
FMF [49]	94.48	loso
PHOG+LBP [50]	94.63	5-fold
SFPL [6]	94.69	10-fold
DSNGE [47]	94.82	loso
LPQ-SLPM-NN [51]	95.90	loso
CDMML [20]	96.60	–
DTAGN [26]	97.25	10-fold
DF-KSOM [7]	98.95	10-fold
DAM-CNN(proposed)	95.88	10-fold

4.4. Quantitative evaluation results

We have evaluated DAM-CNN on both constrained datasets (CK+, JAFFE, TFEID) and unconstrained datasets (BAUM-2i, SFEW, FER2013). In this section, we specify the experimental results on each aforementioned dataset.

4.4.1. Results on constrained datasets

Results on CK+. DAM-CNN is implemented to classify six prototypical expressions in the experiments on CK+. The detailed recognition accuracies (confusion matrix) of each expression are listed in Table 5. An averaged recognition accuracy of 95.88% is obtained by DAM-CNN with the highest recognition of 100% in the expression of happiness. DAM-CNN performs well in classifying most expressions other than that in the class of fear. This may be due to lack of training samples in the expression of fear. In our experiments, only 75 samples of fear are used to train our model. This sample size is the least among all prototypical expressions, which may bias the training and result in learning more knowledge of other expressions than that of the fear.

We compare the performance of our model with some existing methods. Details of the comparison results are listed in Table 6. The models of DTAN [26], DTGN [26], DTAGN [26] and Ali-Net [23] are deep learning based methods while others are conventional models. From Table 6, we can clearly observe that the proposed DAM-CNN outperforms many other competitive methods. Compared with conventional methods, DAM-CNN extracts features through a deep learning based network without much prior knowledge, which makes it adaptive to more complex tasks. Compared with those competitive deep learning based methods, DAM-CNN can discriminate some expression-related regions and generate a

Table 7
Confusion Matrix of DAM-CNN on JAFFE (%) (An: anger, Di: disgust, Fe: fear, Ha: happiness, Sa: sadness, Su: surprise).

	An	Di	Fe	Ha	Sa	Su
An	99.67	0.03	0	0	0	0
Di	2.00	96.50	0	0	1.50	0
Fe	0	0	99.00	0	0	1.00
Ha	0	0	0	100	0	0
Sa	0.03	0	0	0	99.33	0
Su	0	0	0	0	0	100

variation-robust representation for each input, which makes DAM-CNN more effective in FER task. Among all compared methods, three models (CDMML, DTGAN and DF-KSOM) perform better than DAM-CNN in recognition accuracy. This can be explained as follows. First, all these three models extracted more than one type of features to represent an expressional image, while DAM-CNN represents an image only with the extracted CNN features. Empirically, representations fused by multiple different features can be more useful than that of using single type of features. However, compared with the work of [26], our model still performs better when images are only represented by single type of feature, e.g., the model of DTAN (represented by appearance features) and DTGN (represented by geometrical features). This fact demonstrates the superiority of the DAM-CNN. In the work of CDMML [20] and DTAGN [26], the input are all image sequences. This enables their model to learn some extra motional or temporal information about expression changes, which can be beneficial to the FER task. As for the work of DF-KSOM [7], the number of samples used in their experiments (3718 images with the size of 300 × 300) are much more than that of ours (927 images with the size of 224 × 224), which may be another reason to explain for their great performance.

Results on JAFFE. Similar to the experiments on CK+, in JAFFE, we use DAM-CNN to classify six prototypical expressions. The confusion matrix is shown in Table 7. DAM-CNN obtains an averaged recognition accuracy of 99.32%, with the highest recognition of 100% in both case of happiness and surprise. The result demonstrates the effectiveness of DAM-CNN model on the JAFFE database. In addition, the numbers of samples in each expression class are more evenly distributed (compared with the CK+ dataset). Therefore, our model can equally learn knowledge for all expressions without much biases or preference in the experiment on JAFFE.

The proposed method is compared with many competitive methods. The comparison results are shown in Table 8. The models of Sobel-CNN [24], SAE [52], CAE-CNN [24] and DCNN [53] are deep learning based methods, while others are conventional models. From Table 8 we can observe that DAM-CNN outperforms nearly all other competitive methods, which verifies the superiority of our model.

Results on TFEID. Experiments of six-class and seven-class (six prototypical expression as well as the neutral) classification are both conducted on the TFEID dataset. Comparison results are shown in Table 9, where DAM-CNN can obtain an averaged recognition accuracy of 93.65% and 93.20% in the task of six-class and seven-class classification respectively. From the result, we can see that the performance of DAM-CNN is comparable to some state-of-the-art methods, which demonstrates the effectiveness of our model. The recognition accuracy on seven-class classification only mildly lower than that on six-class classification task, which means that our model can effectively discriminate different prototypical expressions.

Table 8
Performance comparison on JAFFE (%) (loo: leave-one-out, 5(10)-fold: 5(10)-fold cross-validation).

Method	Accuracy	Validation Settings
SenTion [54]	88.83	10-fold
GSP [55]	92.10	loo
FERME [56]	92.20	10-fold
Sobel-CNN [24]	92.60	10-fold
SFPL [6]	92.63	10-fold
FMF [49]	92.93	10-fold
SAE [52]	94.01	loo
CAE-CNN [24]	94.10	10-fold
EPFEC [57]	94.37	loo
DCNN [53]	96.10	–
DDL [21]	98.00	10-fold
MMSC [58]	98.57	loo
PBMFEA [59]	98.73	10-fold
PCA+LDA+LS-SVM(Polynomial) [60]	98.86	5-fold
PCA+LDA+LS-SVM(Linear) [60]	99.33	5-fold
PCA+LDA+LS-SVM(RBF) [60]	99.46	5-fold
DAM-CNN(proposed)	99.32	10-fold

Table 9
Performance comparison on TFEID (%) (lopo: leave-one-person-out, 5(10)-fold: 5(10)-fold cross-validation).

Classification	Method	Accuracy	Validation Settings
Six-Class	MFA [61]	91.70	–
	SDM [61]	92.58	–
	DSNGE [62]	93.89	lopo
	SLPM [61]	94.32	–
	DAM-CNN(proposed)	93.65	5-fold
Seven-Class	SVM [45]	71.58	5-fold
	McFIS [45]	75.00	5-fold
	LGBPHS [51]	93.66	10-fold
	LTeP [51]	95.15	10-fold
	DAM-CNN(proposed)	93.36	5-fold

4.4.2. Results on unconstrained datasets

In order to evaluate the performance of our model when dealing with more complex variations, we conduct several experiments on three unconstrained expression datasets, i.e., BAUM-2i, SFEW and FER2013. In the experiments on SFEW and FER2013, seven-class classification is conducted to evaluate the performance of our model. As for the experiments on BAUM-2i, both six-class and seven-class classification are conducted to evaluate DAM-CNN. The performance of our model as well as the comparison results are listed in Table 10.

The analyses about the experimental result are summarized as follow:

(1) In the experiments on SFEW and BAUM-2i, the recognition accuracies of our model are much better than that of some competitive works such as AUDN [69], CNN-VA [70] and SLPM [61]. The performance on FER2013 is also comparable to some existing works, which verifies the effectiveness of DAM-CNN.

(2) The proposed DAM-CNN always achieves good performance on these datasets even though there are many complex variations in images. This may owing to the proposed MPVS-Net. By adopting MPVS-Net, DAM-CNN can generate a variation-robust representation for each expressional image, which well improves the performance of our model under complex situations. The result demonstrates that DAM-CNN is capable to handle different variations.

(3) In the experiments on FER2013, the recognition accuracy of DAM-CNN is lower than those of the state-of-the-art method. This may be due to the image resolution of FER2013. The size of images in FER2013 is 48 × 48, which only contains limited local details. However, the feature

Table 10
Performance Comparison on Unconstrained Datasets (%) (val: validation set of FER2013, test: testing set of FER2013).

Dataset	Method	Accuracy
FER2013	MTCNN [63]	60.70
	CNN-Ensemble [64]	65.03
	RTCNN [65]	66.00
	DFEITF [66]	69.87(val)
	DFEITF [66]	71.10(test)
	DAM-CNN(proposed)	65.31(val)
	DAM-CNN(proposed)	66.20(test)
SFEW	AURF [67]	24.98
	JPEM [29]	26.58
	LAIBP [68]	29.70
	AUDN [69]	30.14
	CNN-GAP [70]	32.80
	HDLBP [66]	36.93
	CNN-base [70]	38.50
	DAG [28]	39.07
	CNN-VA [70]	40.00
	DAM-CNN(proposed)	42.30
BAUM-2i(6 class)	LAP [51]	56.38
	LBP [51]	59.46
	LMP [51]	60.64
	LGBPHS [51]	62.41
	SLPM [61]	63.62
	DAM-CNN(proposed)	67.92
BAUM-2i(7 class)	LAP [51]	54.97
	LMP [51]	57.54
	Baseline [41]	57.77
	LGBPHS [51]	57.99
	LBP [51]	58.99
	DAM-CNN(proposed)	61.52

extraction module of our model (i.e., VGG-Face) is pre-trained on images with the size of 224 × 224. It is hard for DAM-CNN to learn enough knowledge about expressions when lacking of image details. Additionally, the proposed SERD attempts to focus on expression-related regions, which is more sensitive to local details. When DAM-CNN is evaluated on other two datasets (i.e., SFEW and BAUM-2i) that have high resolution images, it performs much better.

4.4.3. Results on cross-datasets evaluation

To evaluate the generalization ability of our model, we conduct several experiments across different datasets. Four datasets are involved in the experiments, which include CK+, JAFFE, TFEID and SFEW. To be specific, we train DAM-CNN with the samples of one dataset and then test it on the other dataset. Seven class classification is conducted on each experiments.

The experimental results are shown in Table 11. From the result, we can observe that DAM-CNN always outperforms most of the listed models. It is reasonable as our model can discriminate the expression-related features and generate a variation-robust representation for FER, which effectively improves its performance across different datasets. Compared with other methods, DAM-CNN has a better generalization ability. On one hand, the feature extraction module of our model, i.e., VGG-Face, has been pre-trained on a large face dataset. We only fine-tune it in the expression dataset instead of training it starting from scratch. By transferring the learned knowledge about faces to FER task, our model can be more adaptive to different situations. On the other hand, the proposed MPVS-Net can effectively disentangle expression information from variations, which helps generate variation-robust representations for the FER task. These advantages effectively prevent our model from severe overfitting. Furthermore, dropout is introduced to train our model, which partly improves the generalization ability of our model.

Table 11
Performance on Cross-datasets Evaluation (%).

Train	Test	Method	Accuracy
CK+	JAFFE	LGIP [51]	41.30
		LBP-SVM [71]	41.30
		CCEC [72]	42.30
		DAM-CNN(proposed)	43.38
	SFEW	AUDN [73]	29.43
		DAM-CNN(proposed)	24.44
JAFFE	CK+	LMP [51]	37.32
		CCEC [72]	48.20
		DAM-CNN(proposed)	49.10
	TFEID	SVM [45]	20.33
		McFIS [45]	37.73
		LGBPHS [51]	60.09
		DAM-CNN(proposed)	52.75
TFEID	JAFFE	SVM [45]	13.92
		McFIS [45]	31.45
		MBC_P [51]	47.28
		DAM-CNN(proposed)	43.38

4.5. Discussions

From the experimental results in Sections 4.4.1 and 4.4.2, we can see that DAM-CNN performs well in both constrained and unconstrained datasets. In the visualization work, SERD has shown its ability on locating salient expressional regions. In Table 4, results also demonstrate the effectiveness of MPVS-Net.

Although DAM-CNN performs well in FER task, there are still some room for improvement, which can be summarized as follows:

(1) In unconstrained dataset, SERD may fail to focus on salient expressional regions when the image is with some complex and extreme variations (e.g., the input face is with an extreme pose, or only half or a part of a face is captured in the image). Since the architecture of SERD only consists of one convolutional layer, it is too shallow to handle complex variations. One possible solution is to construct a deeper network for SERD, which can partly improve the robustness on different variations.

(2) In MPVS-Net, the architecture of encoder and decoders are all fully-connected networks. Therefore, one limitation of DAM-CNN is that it will contain a large number of trainable parameters if MPVS-Net is assigned with many decoding paths. One approach to address this problem is to replace the fully-connected architecture with deconvolutional networks. As parameters are shared among different neurons in deconvolutional layers, the number of trainable parameters can be effectively decreased, which can also reduce the risk of overfitting.

5. Conclusions

In this paper, we propose a novel model named DAM-CNN for FER. The proposed model consists of three modules, i.e., a feature extraction module (VGG-Face), the SERD and the MPVS-Net. In our method, features are extracted by the VGG-Face and fed into the SERD module, which adaptively highlights the features that are highly-relevant to FER task and helps locate expression-sensitive regions. Visualizations of the located regions show that SERD is able to automatically focus on some expression-related regions such as the neighborhood of eyes and mouth. The other module, MPVS-Net, is proposed to handle different variations. Based on the encoder-decoder architecture, the high-level representations yielded by MPVS-Net are robust to multiple variations, which is effective for FER task. By jointly combining the SERD and MPVS-Net, DAM-CNN can be more discriminative to different expressions.

Experimental results on both constrained and unconstrained datasets demonstrate the effectiveness of our model.

The proposed DAM-CNN can still be improved in some aspects. In our future work, we intend to modify our training strategy: converting from the current two-stage training to an end-to-end manner. By end-to-end training, we expect to better develop the advantages of SERD and MPVS-Net. Also, we plan to optimize the network structure following the approaches we mentioned in Section 4.5 in order to improve the robustness and efficiency of our model. In addition, DAM-CNN is a generic model for classification. It can be extended to other recognition task such as face recognition, which can be one of our future work to further investigate the effectiveness of DAM-CNN.

Acknowledge

This work was supported in part by the National Natural Science Foundation of China under Grants 61673402, 61273270, and 60802069, in part by the Natural Science Foundation of Guangdong under Grants 2017A030311029, 2016B010109002, in part by the Science and Technology Program of Guangzhou under Grants 201704020180, and in part by the Fundamental Research Funds for the Central Universities of China under Grant 17lgzd08.

References

[1] A. Ryan, J. F. Cohn, S. Lucey, J. Saragih, P. Lucey, F. D. la Torre, A. Rossi, Automated facial expression recognition system, in: Proceedings of the 43rd Annual International Carnahan Conference on Security Technology, Zurich, Switzerland 2009, pp. 172–177. doi:10.1109/CCST.2009.5335546.

[2] A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: survey of an emerging domain, Image Vis. Comput. 27 (12) (2009) 1743–1759, doi:10.1016/j.imavis.2008.11.007.

[3] Q. Wang, K. Jia, P. Liu, Design and implementation of remote facial expression recognition surveillance system based on PCA and KNN algorithms, in: Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), IEEE, 2015, pp. 314–317. doi:10.1109/IIH-MSP.2015.54.

[4] E. Vural, M. Çetin, A. Erçil, G. Littlewort, M. Bartlett, J. Movellan, Automated Drowsiness Detection for Improved Driving Safety, Ford Otosan, 2008.

[5] P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Palo Alto: Consulting Psychologists, 1978.

[6] S. Happy, A. Routray, Automatic facial expression recognition using features of salient facial patches, IEEE Trans. Affect Comput. 6 (1) (2015) 1–12, doi:10.1109/TAFFC.2014.2386334.

[7] A. Majumder, L. Behera, V.K. Subramanian, Automatic facial expression recognition system using deep network-based data fusion, IEEE Trans. Cybern. 48 (1) (2018) 103–114, doi:10.1109/TCYB.2016.2625419.

[8] W.J. Baddar, Y.M. Ro, Bilateral hemiface feature representation learning for pose robust facial expression recognition, in: Proceedings of the Signal and Information Processing Association Annual Summit and Conference (APSIPA) Asia-Pacific, IEEE, 2016, pp. 1–4.

[9] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1683–1699, doi:10.1109/TPAMI.2007.1094.

[10] Y.-l. Tian, T. Kanade, J.F. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 97–115, doi:10.1109/34.908962.

[11] G. Sandbach, S. Zafeiriou, M. Pantic, D. Rueckert, Recognition of 3D facial expression dynamics, Image Vis. Comput. 30 (10) (2012) 762–773, doi:10.1016/j.imavis.2012.01.006.

[12] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Fully automatic facial action recognition in spontaneous behavior, in: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition FGR, IEEE, 2006, pp. 223–230. doi:10.1109/FGR.2006.55.

[13] J.F. Cohn, L.I. Reed, Z. Ambadar, J. Xiao, T. Moriyama, Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior, in: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, vol. 1, IEEE, 2004, pp. 610–616. doi:10.1109/ICSMC.2004.1398367.

[14] A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfeliu, F. Moreno-Noguer, Ganimation: anatomically-aware facial animation from a single image, in: Proceedings of the 15th European Conference Computer Vision - ECCV, Munich, Germany Part X, 2018, pp. 835–851. 10.1007/978-3-030-01249-6_50.

[15] A. Majumder, L. Behera, V.K. Subramanian, Emotion recognition from geometric facial features using self-organizing map, Pattern Recognit. 47 (3) (2014) 1282–1293, doi:10.1016/j.patcog.2013.10.010.

[16] Q. Mao, Q. Rao, Y. Yu, M. Dong, Hierarchical Bayesian theme models for multipose facial expression recognition, IEEE Trans. Multimed. 19 (4) (2017) 861–873, doi:10.1109/TMM.2016.2629282.

[17] M.-W. Huang, Z.-w. Wang, Z.-L. Ying, A new method for facial expression recognition based on sparse representation plus LBP, in: Proceedings of the 3rd International Congress on Image and Signal Processing (CISP), vol. 4, IEEE, 2010, pp. 1750–1754. 10.1109/CISP.2010.5647898.

[18] T. Jabid, M.H. Kabir, O. Chae, Robust facial expression recognition based on local directional pattern, ETRI J. 32 (5) (2010) 784–794, doi:10.4218/etrij.10.1510.0132.

[19] A. Tawari, M.M. Trivedi, Face expression recognition by cross modal data association, IEEE Trans. Multimed. 15 (7) (2013) 1543–1552, doi:10.1109/TMM.2013.2266635.

[20] H. Yan, Collaborative discriminative multi-metric learning for facial expression recognition in video, Pattern Recognit. 75 (2018) 33–40.

[21] A. Moeini, K. Faez, H. Moeini, A.M. Safai, Facial expression recognition using dual dictionary learning, J. Vis. Commun. Image Represent. 45 (2017) 20–33.

[22] K. Zhao, W.-S. Chu, H. Zhang, Deep region and multi-label learning for facial action unit detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3391–3399. doi:10.1109/CVPR.2016.369.

[23] A. Mollahosseini, D. Chan, M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2016, pp. 1–10. 10.1109/WACV.2016.7477450.

[24] D. Hamester, P. Barros, S. Wermter, Face expression recognition with a 2-channel convolutional neural network, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE, 2015, pp. 1–8. doi:10.1109/IJCNN.2015.7280539.

[25] H. Yang, U.A. Ciftci, L. Yin, Facial expression recognition by de-expression residue learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPRSalt Lake City, UT, USA, 2018, pp. 2168–2177. doi:10.1109/CVPR.2018.00231.

[26] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2983–2991. doi:10.1109/ICCV.2015.341.

[27] V. Vielzeuf, S. Pateux, F. Jurie, Temporal multimodal fusion for video emotion classification in the wild, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ACM, 2017, pp. 569–576.

[28] X. Zhu, Y. Liu, J. Li, T. Wan, Z. Qin, Emotion classification with data augmentation using generative adversarial networks, In Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Cham, 2018, pp. 349–360.

[29] F. Zhang, T. Zhang, Q. Mao, C. Xu, Joint pose and expression modeling for facial expression recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPRSalt Lake City, UT, USA, 2018, pp. 3359–3368. doi:10.1109/CVPR.2018.00354.

[30] İ. Çuğu, E. Şener, E. Akbaş, Microexpnet: An extremely small and fast model for expression recognition from frontal face images, 2017, pp. 1–9. arXiv preprint 1711.07011.

[31] K. Fu, J. Jin, R. Cui, F. Sha, C. Zhang, Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2321–2334, doi:10.1109/TPAMI.2016.2642953.

[32] L.-C. Chen, Y. Yang, J. Wang, W. Xu, A.L. Yuille, Attention to scale: scale-aware semantic image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3640–3649. doi:10.1109/CVPR.2016.396.

[33] S. Sharma, R. Kiros, R. Salakhutdinov, Action recognition using visual attention, In NIPS workshop on Time Series, 2015, pp. 1–11.

[34] O.M. Parkhi, A. Vedaldi, A. Zisserman, et al., Deep face recognition, in: Proceedings of the BMVC, vol. 1, 2015, p. 6. doi:10.5244/C.29.41.

[35] M. Ghifary, W. Bastiaan Kleijn, M. Zhang, D. Balduzzi, Domain generalization for object recognition with multi-task autoencoders, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2551–2559. doi:10.1109/ICCV.2015.293.

[36] L.v.d. Maaten, G. Hinton, Visualizing data using T-sne, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.

[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[38] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended Cohn-Kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2010, pp. 94–101.

[39] M. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, Coding facial expressions with Gabor wavelets, in: Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, 1998, pp. 200–205.

[40] The taiwanese facial expression image database (tfeid), (http://bml.ym.edu.tw/tfeid/).

[41] C.E. Erdem, C. Turan, Z. Aydin, Baum-2: a multilingual audio-visual affective face database, Multimed. Tools Appl. 74 (18) (2015) 7429–7459.

[42] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Static facial expression analysis in tough conditionsv: Data, evaluation protocol and benchmark, in: Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2011, pp. 2106–2112.

[43] Challenges in representation learning: Facial expression recognition challenge (fer2013), (http://www.kaggle.com/c/challengesin-representation-learning-facial-expression-recognitionchallenge).

[44] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154, doi:10.1023/B:VISI.0000013087.49260.fb.

[45] K. Subramanian, V.B. Radhakrishnan, S. Ramasamy, Database independent human emotion recognition with meta-cognitive neuro-fuzzy inference system, in: Proceedings of the IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), IEEE, 2014, pp. 1–6.

[46] S. Taheri, Q. Qiu, R. Chellappa, Structure-preserving sparse decomposition for facial expression analysis, IEEE Trans. Image Process. 23 (8) (2014) 3590–3603, doi:10.1109/TIP.2014.2331141.

[47] Y.-H. Tu, C.-T. Hsu, Dual subspace nonnegative matrix factorization for person-invariant facial expression recognition, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR), IEEE, 2012, pp. 2391–2394.

[48] L. Zhong, Q. Liu, P. Yang, J. Huang, D.N. Metaxas, Learning multiscale active facial patches for expression analysis, IEEE Trans. Cybern. 45 (8) (2015) 1499–1510, doi:10.1109/TCYB.2014.2354351.

[49] L. Zhang, D. Tjondronegoro, Facial expression recognition using facial movement features, IEEE Trans. Affect. Comput. 2 (4) (2011) 219–229, doi:10.1109/T-AFFC.2011.13.

[50] S. Happy, A. Routray, Robust facial expression classification using shape and appearance features, in: Proceedings of the Eighth International Conference on Advances in Pattern Recognition (ICAPR), IEEE, 2015, pp. 1–5. 10.1109/ICAPR.2015.7050661.

[51] C. Turan, K. Lam, Histogram-based local descriptors for facial expression recognition (FER): a comprehensive study, J. Visual Commun. Image Represent. 55 (2018) 331–341, doi:10.1016/j.jvcir.2018.05.024.

[52] B. Huang, Z. Ying, Sparse autoencoder for facial expression recognition, in: Proceedings of the IEEE 12th International Conference on Ubiquitous Intelligence and Computing and IEEE 12th International Conference on Autonomic and Trusted Computing and IEEE 15th International Conference on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), IEEE, 2015, pp. 1529–1532. 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.274.

[53] A. Uçar, Deep convolutional neural networks for facial expression recognition, in: Proceedings of the IEEE International Conference on Innovations in Intelligent Systems and Applications (INISTA), IEEE, 2017, pp. 371–375. doi:10.1109/INISTA.2017.8001188.

[54] R. Islam, K. Ahuja, S. Karmakar, F. Barbhuiya, Sention: A framework for sensing facial expressions, 2016, pp. 1–6. arXiv preprint 1608.04489.

[55] H. Meena, K. Sharma, S. Joshi, Improved facial expression recognition using graph signal processing, Electron Lett. 53 (11) (2017) 718–720, doi:10.1049/el.2017.0420.

[56] H. da Cunha Santiago, T.I. Ren, G.D. Cavalcanti, Facial expression recognition based on motion estimation, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE, 2016, pp. 1617–1624. 10.1109/IJCNN.2016.7727391.

[57] Y. Rahulamathavan, M. Rajarajan, Efficient privacy-preserving facial expression classification, IEEE Trans. Dependable Secure Comput. 14 (3) (2017) 326–338, doi:10.1109/TDSC.2015.2453963.

[58] O. Krestinskaya, A.P. James, Facial emotion recognition using min-max similarity classifier, In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 752–758.

[59] M. Dahmane, J. Meunier, Prototype-based modeling for facial expression analysis, IEEE Trans. Multimed. 16 (6) (2014) 1574–1584, doi:10.1109/TMM.2014.2321113.

[60] N.B. Kar, K.S. Babu, A.K. Sangaiah, S. Bakshi, Face expression recognition system based on ripplet transform type ii and least square SVM, Multimed. Tools Appl. (2017) 1–24.

[61] C. Turan, K.-M. Lam, X. He, Soft Locality Preserving Map (SLPM) for Facial Expression Recognition, 2018, pp. 1–21. arXiv preprint 1801.03754.

[62] H.-W. Kung, Y.-H. Tu, C.-T. Hsu, Dual subspace nonnegative graph embedding for identity-independent expression recognition, IEEE Trans. Inf. Forensics Secur. 10 (3) (2015) 626–639.

[63] J. Xiang, G. Zhu, Joint face detection and facial expression recognition with MTCNN, in: Proceedings of the 4th International Conference on Information Science and Control Engineering (ICISCE), IEEE, 2017, pp. 424–427.

[64] K. Liu, M. Zhang, Z. Pan, Facial expression recognition with CNN ensemble, in: Proceeding of the International Conference on Cyberworlds (CW), IEEE, 2016, pp. 163–166.

[65] O. Arriaga, M. Valdenegro-Toro, P. Plöger, Real-time convolutional neural networks for emotion and gender classification, 2017, pp. 1–5. arXiv preprint 1710.07557.

[66] S. Munasinghe, C. Fookes, S. Sridharan, Deep features-based expression-invariant tied factor analysis for emotion recognition, in: Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2017, pp. 546–554.

[67] M. Liu, S. Li, S. Shan, X. Chen, Au-aware deep networks for facial expression recognition, in: Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 2013, pp. 1–6.

[68] B. Santra, D.P. Mukherjee, Local saliency-inspired binary patterns for automatic recognition of multi-view facial expression, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 624–628.

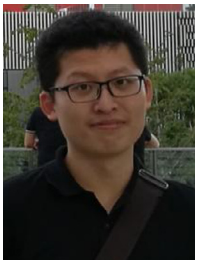
[69] K. Radlak, B. Smolka, High dimensional local binary patterns for facial expression recognition in the wild, in: Proceedings of the 18th Mediterranean Electrotechnical Conference (MELECON), IEEE, 2016, pp. 1–5.

[70] W. Sun, H. Zhao, Z. Jin, A visual attention based ROI detection method for facial expression recognition, Neurocomputing 296 (2018) 12–22.

[71] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.

[72] F.A.M. da Silva, H. Pedrini, Effects of cultural characteristics on building an emotion classifier through facial expression analysis, J. Electron Imaging 24 (2) (2015) 023015.

[73] M. Liu, S. Li, S. Shan, X. Chen, Au-inspired deep networks for facial expression feature learning, Neurocomputing 159 (2015) 126–136.



Siyue Xie is currently a graduate student in the School of Electronics and Information Engineering, Sun Yat-sen University, China. His major research interests include computer vision and pattern recognition. One particular interest is facial expression recognition.

Haifeng Hu received the Ph.D. degree from Sun Yat-sen University in 2004, and now he is a professor of School of Electronics and Information Engineering at Sun Yat-senUniversity. His research interests are in computer vision, pattern recognition, image processing and neural computation. He has published about 120 papers since 2000.

Yongbo Wu is currently a graduate student in the School of Electronics andInformation Engineering, Sun Yat-sen University, China. His major research interests include computer vision and pattern recognition. His interestsare robust face recognition and facial expression recognition.