Neural Computing and Applications (2022) 34:925–936
https://doi.org/10.1007/s00521-021-06045-z

SPECIAL ISSUE ON ADVANCES OF NEURAL COMPUTING PHASING CHALLENGES IN THE ERA OF 4TH INDUSTRIAL REVOLUTION

Discriminative attention-augmented feature learning for facial expression recognition in the wild

Linyi Zhou[1] · Xijian Fan[1] · Tardi Tjahjadi[2] · Sruti Das Choudhury[3]

Received: 26 February 2021 / Accepted: 13 April 2021 / Published online: 29 April 2021

Abstract
Facial expression recognition (FER) in-the-wild is challenging due to unconstraint settings such as varying head poses, illumination, and occlusions. In addition, the performance of a FER system significantly degrades due to large intra-class variation and inter-class similarity of facial expressions in real-world scenarios. To mitigate these problems, we propose a novel approach, Discriminative Attention-augmented Feature Learning Convolution Neural Network (DAF-CNN), which learns discriminative expression-related representations for FER. Firstly, we develop a 3D attention mechanism for feature refinement which selectively focuses on attentive channel entries and salient spatial regions of a convolution neural network feature map. Moreover, a deep metric loss termed Triplet-Center (TC) loss is incorporated to further enhance the discriminative power of the deeply-learned features with an expression-similarity constraint. It simultaneously minimizes intra-class distance and maximizes inter-class distance to learn both compact and separate features. Extensive experiments have been conducted on two representative facial expression datasets (FER-2013 and SFEW 2.0) to demonstrate that DAF-CNN effectively captures discriminative feature representations and achieves competitive or even superior FER performance compared to state-of-the-art FER methods.

Keywords Facial expression recognition · Salient features · Metric learning · Convolution neural network · Attention mechanism

1 Introduction

Facial expression recognition (FER) plays a crucial role in non-verbal communication among humans by providing profuse information related to their emotions. Many studies have been conducted on FER due to its extensive applications, e.g., human–computer interaction, medical treatment, driver fatigue surveillance [1], etc. Although many advances have been made, achieving accurate FER is still very challenging due to the subtlety, complexity, and variability of facial expressions.

Much progress has been made on extracting discriminative features to represent facial patterns in order to boost the performance of a FER system. In general, feature extraction methods can be categorized into extracting handcrafted features and deeply-learned features. Handcrafted methods obtain facial features with prescribed descriptors, such as Local Binary Patterns (LBP) [2],

✉ Xijian Fan
xijian.fan@njfu.edu.cn

Linyi Zhou
zhoulinyi@njfu.edu.cn

Tardi Tjahjadi
t.tjahjadi@warwick.ac.uk

Sruti Das Choudhury
s.d.choudhury@unl.edu

[1] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, Jiangsu, China

[2] School of Engineering, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK

[3] Department of Computer Science and Engineering School of Natural Resources, University of Nebraska, Lincoln, NE 68583, USA

Histogram of Oriented Gradients (HOG) [3], Gabor wavelet [4], etc. These methods have achieved impressive performance on several benchmarks collected under controlled laboratory settings including CK + [5] and MMI [6].

However, the handcrafted descriptors require manual selection of facial features and depend much on prior knowledge. In addition, such handcrafted methods are not robust and thus lack generalization ability when faced with unconstrained settings and real-world scenarios. Recently, deep learning techniques, especially the success of Convolution Neural Network (CNN), have yielded excellent performance on a wide range of image classification tasks [7–10]. It has also been shown that various CNN architectures can achieve promising results in FER [11–13]. However, employing CNN in FER is not always satisfactory due to the design of its receptive fields. Since these receptive fields are local, the information of input images is processed within a restricted neighborhood. Thus, the network fails to capture long-range contextual correlations that are of crucial importance for better recognition performance. In addition, the performance of a FER system, especially in real-world scenarios, often suffers from unconstrained challenges (e.g., varying illumination and head poses), preventing the CNN from extracting useful (e.g., expression-related) features.

Therefore, directly exploiting convolutional features to perform recognition of expression can lead to sub-optimal results because of the local operators and various challenges. Recently, self-attention emerges as a potential solution and has achieved promising results in sequence modeling and semantic segmentation [14]. By exploiting global operators (such as global max pooling), this attention mechanism has been used [15] to extract useful information which is contained in local descriptors to enhance global representations. These attention-based methods shed much light on the spatial aspect of enhancement.

Nevertheless, there exist other aspects of attention that are worth investigating, which can be analyzed via a CNN architecture. The pipeline of a CNN starts from a convolution layer which scans the input images with a collection of filters and outputs a series of response maps that are further processed sequentially by the subsequent convolutional layers. During this process, the channel axis is introduced to extend the CNN feature from two-dimensional (2D) into three-dimensional (3D) domain. Since the convolution filter performs as a pattern detector which captures both the low-level visual cues (e.g., edges and corners) and high-level semantic pattern, each 2D channel slice of 3D feature maps spatially encodes the information related to a certain pattern. Hence, the CNN features are inherently 3D representations.

However, most existing attention-based methods merely focus on the spatial dimension while limited work has paid attention to both aspects [16, 17]. Therefore, to fully exploit the information within a feature map, in this paper, we introduce a dual-facet attention mechanism for FER which performs both spatial and channel-wise feature recalibration.

With respect to the channel-wise attention, it helps to highlight the usefulness of expression-related patterns encoded in specific channel maps. Given a CNN feature map, not all the feature channels are of equal importance for expression recognition. Some channels of high numerical values correspond to their related expressions, enabling the capture of expression-related feature representations and facilitating the expression recognition. While other entries are neither informative nor expression-related, some of which may even cause interference, degrading the discriminability of the extracted features. Thus, to obtain salient feature representations, it is necessary to perform channel-wise attention, where attentive feature channels are emphasized, and non-informative channels are suppressed. Also, by explicitly modeling the interdependencies among channels, the proposed method is able to gather long-range contextual correlations. The refined feature maps generated by channel attention unit can be further exploited for better recalibration, noting that not all partial regions along the spatial dimension are informative. Some facial sub-regions are critical to emotion recognition due to their high response to certain expressions.

For example, raised cheeks are expressive and can thereby be easily identified in happy faces, referred as expression-related features. While other spatial areas such as irrelevant facial parts and non-informative background only generate low responses that are not expression-related. Motivated by the above observations, we further incorporate spatial attention to selectively focus on expression-related localities out of an emphasized feature channel. With the enhancement of these salient features, richer contextual abstractions within the spatial dimension could further be captured. Moreover, since both the background and irrelevant facial regions are suppressed, the proposed method is able to disentangle non-informative factors and generate more discriminative feature representations.

With respect to feature classification, some effort has been made on designing effective classifiers for FER, which is also crucial to achieve good FER results. Conventionally, most deep learning methods minimize cross-entropy loss and employ the softmax activation function for prediction. Despite its popularity, the softmax loss is not capable of dealing with the problems existing exclusively in FER, i.e., images for FER tend to have both high intra-class variation and high inter-class similarity.

For example, surprise expression can be either positive with a wide-open smile or negative with a tensed mouth, revealing the high intra-class variation, while fearful and disgusted faces are often confused due to the similar displayed patterns, e.g., curved mouth and tensed eyes, indicating the high inter-class similarity. Since the softmax loss only focuses on seeking a decision boundary to keep different classes apart, it merely encourages the separateness of learned features. As a result, in the embedding feature space, clusters of different classes are likely to be overlapped while features of the same class are scattered within one individual cluster.

Thus, features learned by softmax loss are not discriminative and robust in nature. As the key task of FER requires dealing with high intra-class variation and inter-class similarity, softmax-based features are not sufficient for accurate predictions, necessitating the CNN network to learn more effective and discriminative representations.

More recently, the emerging deep metric learning methods have been investigated for image retrieval and person re-identification with large intra-class variations. This suggests that deep metric learning may offer more pertinent representations for FER. The triplet loss [18] and center loss [19] are two representative losses used in deep metric learning methods, the former develops a triplet constraint to reduce the intra-class variation and inter-class similarity, while the latter learns a center for each class to obtain compact features. Both of them aim at learning discriminative feature representations.

Inspired by these two losses, Triplet-Center loss (TC loss) is proposed for 3D object retrieval in [20]. By combining the merits of both triplet loss and center loss, TC loss targets directly on addressing intra-class variation and inter-class similarity by minimizing the intra-class distance and maximizing inter-class distance at the same time. Motivated by these approaches, to further enhance the discriminative power of the expression representations and address the similar intra- and inter-class problems in FER, we employ TC loss to enhance the discriminability and robustness of feature representation. Unlike existing work [21–26] that only consider feature extraction or feature classification stage separately, we propose a novel approach with discriminative feature learning in both stages which combines the attention mechanism and deep metric learning into an end-to-end fashion.

In the feature extraction stage, 3D attention mechanism is augmented to exploit global information within the feature map to emphasize salient and meaningful expression-related features that are more discriminative. In the feature classification stage, TC loss is integrated to explicitly target on intra-class and inter-class distances to learn compact and separate features. Thus, the discriminative power of the refined features is further enhanced. Extensive experiments have been conducted to evaluate our method on two well-known in-the-wild datasets, i.e., FER2013 and SFEW. Promising accuracy results have been achieved that surpass most of the existing methods, demonstrating the effectiveness of the proposed method.

In summary, the major contributions of this paper are as follows:

(1) We propose a novel framework augmented with 3D attention mechanism, which highlights the usefulness of both expression-related features and emotional salient regions to generate more discriminative representations.
(2) We introduce TC loss for FER, to learn discriminative features that are both compact and separate in the feature space, explicitly addressing the problem of high inter-class similarity and intra-class variation.
(3) We develop a Discriminative Attention-augmented Feature Learning Convolution Neural Network (DAF-CNN) integrated with the proposed 3D attention and TC loss for discriminative feature learning, unifying the expression-related feature learning, and deep metric learning to jointly boost the performance of FER.

2 Related work

During the transition of FER from laboratory-controlled to unconstrained in-the-wild conditions, deep learning techniques have, in recent years, been increasingly applied to FER that have achieved promising results. The winning system of the FER-2013 Challenge [11] uses SVM classifier as an alternative to the cross-entropy loss, which shows that switching from traditional softmax layer to a linear SVM top layer is beneficial for some deep architectures. To disentangle interfering factors in face images such as head pose, illumination, and facial morphology, the following methods were proposed. Rifai et al. [27] proposed a multi-scale contractive CNN to obtain local-translation-invariant representations, and designed auto-encoders to separate discriminative expression information from subject identity and pose, while Reed et al. [28] constructed a Bolzman machine to model high-order interactions of expression and put forward training strategies for disentangling. Ge et al. [29] addresses the occlusion problem for face recognition in the wild as a related task. Besides, representation learning and metric learning for FER have gained much interest from researchers recently [30–32].

Attention has been widely adopted for modeling sequences due to its ability to capture long-range interactions. Bahdanau et al. [33] first combined attention with

Recurrent Neural Network for alignment in neural machine translation (NMT). To further improve the effectiveness of NMT, Luong et al. [34] proposed an effective attention-based method which introduced two different classes of mechanisms, i.e., global attention and local attention. In addition, various attention mechanisms have been proposed for visual tasks such as image captioning, visual question answering, and image classification.

Visual attention was first proposed by Xu et al. [35] in image captioning, where both soft and hard attention mechanisms are exploited. As for visual question answering, Yang et al. [36] introduced Question-guided image attention to solve the task. Considered as an effective solution, attention model has also been applied to classification task. Wang et al. [37] proposed Residual Attention Network employing a hourglass network to generate 3D attention maps for intermediate features, which demonstrated its robustness to noisy labels. Hu et al. [15] proposed Squeeze-and-Excitation network to perform channel-wise attention via modeling the inter-channel relationship, while Jetley et al. [38] measured spatial attention by considering the feature maps at various layers in the CNN and produce a 2D matrix of scores for each map. These attention mechanisms [15, 38] aimed to specifically address the weakness of convolutions.

In order to learn more robust and discriminative features, deep metric learning has been widely adopted. Much attention has been paid to two representative losses, i.e., center loss and triplet loss. Center loss [19] was proposed as an auxiliary for softmax loss to learn more discriminative features. In the training process, center loss learns a center for the features of each class and pulls features of the same class to its corresponding center. Through the joint supervision of softmax loss, center loss is able to learn compact features that are close to their centers.

However, center loss does not consider the inter-class reparability explicitly, which may lead to inter-class overlap. Alternatively, triplet loss [18] was proposed for face recognition, using triplets as input, each of which consists of an anchor, and a positive and a negative examples. Specifically, the triplet loss optimizes a constraint function which forces the distance between positive and negative pairs to be larger than a fixed margin. With deep embedding, it is capable of learning both compact and separate clusters in the feature space. The effectiveness of triplet loss has been demonstrated in [18, 39]. However, due to the complexity of triplet construction and inefficiency of hard-sample mining, the training process can be unstable and slow in convergence.

In order to face sophisticated problems related to facial expressions in real-world scenarios, this paper proposes a novel DAF-CNN architecture, which learns discriminative expression-related representations for FER. This approach is based on a 3D attention mechanism for feature refinement and on a deep metric loss (TC loss) which further enhances the discriminative power of the deeply-learned features, using an expression-similarity constraint. The introduced novel approach, simultaneously minimizes the intra-class distance and maximizes the inter-class distance in order to learn both compact and separate features. It is an efficient model that combines integration of an attention mechanism with deep metric learning, in order to capture more discriminative expression-related features and to lead to the significant improvement of FER accuracy.

3 Methodology

3.1 Overview

As is illustrated in Fig. 1, the proposed DAF-CNN framework consists of three components. In the feature extraction stage, VGG style convolutional blocks from the CNN backbone. The first two blocks comprise three convolution layers while each of the next three blocks comprises four convolution layers, each followed by a batch normalization (BN) layer. The generated feature map is then fed to the attention module, which performs feature refinement by emphasizing attentive channels and salient regions sequentially. In the feature classification stage, a classifier (i.e., consisting of two fully connected (FC) layers) with similarity constraint learning is employed, where a joint objective function including the TC loss and softmax loss is imposed to learn more discriminative expression representations during the learning process.

3.2 3D Attention mechanism

3.2.1 Channel attention

In order to exploit inter-channel discriminability, the spatial information of each slice in a 3D feature map is aggregated. In general, the channel importance can be measured based on two criteria. The first is global average pooling, which is adopted extensively due to its effectiveness in computing spatial statistics [15].

The second is max pooling. Since max pooling units are very sensitive to the maximum value in the neighborhood, they are good at preserving the strongest features. We exploit both criteria by utilizing a neural network with two hidden layers to balance their decision power.

The network functions as a parameterized combination of the two pooling methods, serving as a more effective criterion for weighting the discriminability of all feature entries. It is worth noting that the spatial information is encoded in a learnable way, which adaptively redistribute

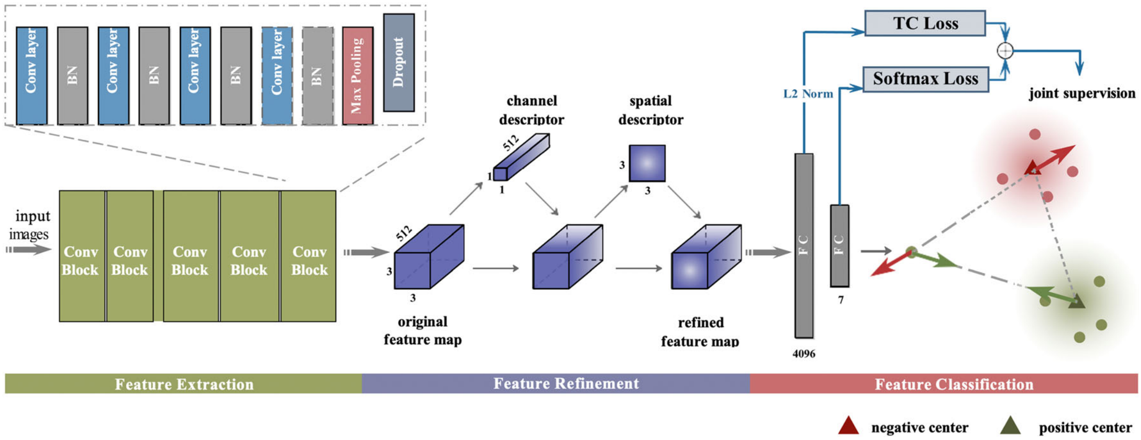

Fig. 1 An overview of the proposed DAF-CNN framework. The generated feature representations are learned through three stages, i.e., feature extraction, feature refinement, and feature classification

the weights to get richer expression-related features and task-oriented clues. Therefore, the representative power of the network is enhanced.

Denote the original input features as $X \in R^{W \times H \times C}$, where W, H, and C denote width, height and the number of channels, respectively. First they are forked to be max and mean pooled in parallel and then reshaped into two channel feature vectors $V_{avg} \in R^{1 \times 1 \times C}$ and $V_{max} \in R^{1 \times 1 \times C}$. They are then fed to a network consisting of two hidden layers FC_1 and FC_2. FC_1 reduces the feature dimension to $1 \times 1 \times C/r$, where r is the reduction ratio. FC_2 with C units reshapes the dimension to fit the original size. Generated by the last sigmoid layer, the channel attention $\mathcal{A}_c$ is

$$\mathcal{A}_c = \text{Sigmoid}\left[\text{Net}\left(V_{avg}(x)\right) \oplus \text{Net}\left(V_{max}(x)\right)\right] \quad (1)$$

and

$$\text{Net}[V(x)] = \text{FC}_2[\text{FC}_1(V(x))] \quad (2)$$

where $\oplus$ denotes the element-wise summation, and *Net* is the sigmoid activation function.

3.2.2 Spatial attention

Given an emphasized feature entry, not all the entire scope of the 2D map is informative. Generally, an expression only corresponds to part of the facial localities of an image. Some regions are not expression-related or useless for recognition and should be suppressed. Thus, we incorporate spatial attention mechanism for further recalibration. Instead of considering all local spatial regions equally, the spatial attention unit assigns more weight to attentive regions and less to non-attentive ones.

By re-weighting the feature map where expression-related regions are assigned higher scores, the spatial attention unit helps to generate emotional salient features that are more discriminative. Similarly, we perform both the mean pooling and max pooling along the width to measure the spatial importance. In general, the mean pooling operation averages the relevant degree of spatial locations while the max pooling selects the most attentive attributes to enhance the sub-region importance.

We then obtain two weighed matrices M_{avg} and $M_{max} \in \mathbb{R}^{W \times H}$. They are concatenated along the width dimension and fed to a convolution layer. Since the encoded weights are spatial, it is natural to perform a convolutional operation to fuse the information. Following the strategy in the channel attention, emotional salient regions can be detected in a learnable way as the weights of receptive field are updated throughout the training process. The spatial attention $\mathcal{A}_s$ is defined as

$$\mathcal{A}_s = \text{sigmoid}\left\{\circledast\left[\text{Concat}\left(M_{avg}(x); M_{max}(x)\right)\right]\right\} \quad (3)$$

where $\circledast$ denotes the convolution operator, and *Concat* denotes concatenation of its input.

3.2.3 Spatial-channel attention

Both the channel and spatial attention modules can take the same feature map x as input and fork into two parallel processes. However, we argue that the feature map is enhanced if the two attention modules are cascaded. Empirically, we perform the channel-weighted attention and spatial-weighted attention sequentially, instructing the network what and where to focus in order. Thus, the 3D attention module is an effective unification of two separate attention modules, generating the channel attention $\mathcal{A}_c$ and spatial attention $\mathcal{A}_s$, respectively. With the effective integration of two separate modules, the cascaded attention mechanisms not only emphasize expression-related attentive feature channels, but also highlight the usefulness of expression-sensitive facial regions. Overall, the final attention function is defined as

$$M' = \mathcal{A}_c(x) \otimes x \tag{4}$$

$$M = \mathcal{A}_s(M') \otimes M' \tag{5}$$

where $\otimes$ denotes element-wise product, and M' and M respectively represent the intermediate and ultimate refined feature maps.

3.3 TC loss for FER

Due to high inter-subject variations introduced by person-specific attributes such as gender, age, and various appearances, different facial expressions are prone to share similar personal characteristics while the same expression may have diverse representations. Thus, intra-class distances are likely to be larger than the inter-class distances, making it challenging to distinguish facial expressions. In addition, the extracted features may contain subject-dependent information that is not expression-related, leading to insufficient clues to generate accurate predictions of expressions.

Therefore, to enhance the discriminative power of the embedded features, we further exploit deep metric losses for expression-similarity mining in the embedding feature space. Two representative deep metric losses (i.e., triplet loss and center loss) have shown their superiority over the traditional softmax loss in reducing the intra-class variation and inter-class similarity.

However, these two losses still have a few limitations. With respect to center loss, the learned clusters are likely to be overlapped since center loss does not consider the inter-class separability explicitly. Regarding the triplet loss, it is subjected to the complexity of triplet construction and inefficiency of hard-sample mining (i.e., active samples that contribute to improve the model by violating the triplet constraint). To address the above-mentioned limitations, we introduce the Triplet-Center loss [20] for FER to mitigate the influence of both intra-class variation and inter-class similarity efficiently.

3.3.1 Forward propagation

The fundamental philosophy behind TC loss is to combine the advantages of triplet loss and center loss, i.e., to efficiently achieve intra-class compactness and inter-class dispersion of the learned features simultaneously. Given the training dataset $\{(x_i, y_i)\}_{i=1}^{N}$ which consists of N samples $x_i \in \mathcal{X}$ with the corresponding labels $p \in \{1, 2, \ldots, |\mathcal{K}|\}$, these samples are mapped to d-dimensional vectors in the embedding feature space with a neural network embedder denoted by $E(\cdot)$.

In TC loss, it is assumed that the features of 3D shapes from the same class share one corresponding center, thereby obtaining $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{K}|}\}$, where $c_y \in \mathbb{R}^d$ denotes the center vector for samples with label y, and $|\mathcal{K}|$ is the number of centers. For simplicity, we adopt e_i to represent $E(x_i)$ in this paper. In terms of triplet loss, the input triplet $\left(x_i^a, x_i^+, x_i^-\right)$ is constituted of samples. While in TC loss, we select the i sample x_i, its corresponding positive center c^p, and its nearest negative center $c_{\min}^q$ to reconstruct the triplet input as $\left(x_i, c^p, c_{\min}^q\right)$. Compared to triplet loss in which the number of triplets is $O(N^3)$, only N triplets will be formed for TC loss. Consequently, TC loss avoids the complexity of triplet construction and the necessity for hard-sample mining. Moreover, by utilizing centers as its similarity metric, TC loss avoids direct interaction with samples of poor quality such as mislabeled faces and noises that are prone to perturb or dominate the hard positives and negatives. Therefore, the stability of the training process and the robustness of the model are enhanced.

To measure the expression-similarity among facial expressions, we adopt the Euclidean distance of the mapped i-th embedded sample e_i and its positive center c^p to represent the degree of deviation among expressions of the same class, which is formulated as

$$D(e_i, c^p) = \frac{1}{2} e_i - c^{p}_{2^2} \tag{6}$$

The degree of resemblance among different expression categories is similarly defined as

$$D\left(e_i, c_{\min}^q\right) = \frac{1}{2} e_i - c^{q2}_{\min 2} \tag{7}$$

Accordingly, we develop the expression-similarity constraint to ensure that the distance from e_i to its positive center c^p is larger than that to its nearest negative center $c_{\min}^q$ by a fixed margin m, which is defined as

$$\frac{1}{2} e_i - c_2^{p2} + m < \frac{1}{2} e_i - c^{q2}_{\min 2} \tag{8}$$

Finally, given a batch of training data with M samples, the TC loss function is given as

$$L_{tc} = \sum_{i=1}^{M} \max\left(D(e_i, c^p) + m - D\left(e_i, c_{\min}^q\right), 0\right) \tag{9}$$

3.3.2 Backward propagation

To compute the back-propagation gradients of the input feature embedding and the corresponding centers, we assume the following notations for demonstration: δ[condition] is an indicator function which outputs 1 if the condition is satisfied and outputs 0 otherwise, and $\tilde{L_i}$ represents the TC loss of i-th sample, i.e.,

$$\tilde{L_i} = \max\left(D(e_i, c^p) + m - D\left(e_i, c_{\min}^q\right), 0\right) \tag{10}$$

These cluster centers are updated based on mini-batches similar to the practice in center loss.

The partial derivatives of our TC loss of Eq. 9 with respect to the feature embedding of i-th sample $\frac{\partial L_{tc}}{\partial e_i}$ and j-th center $\frac{\partial L_{tc}}{\partial c_j}$ are determined as follows:

$$\begin{aligned}\frac{\partial L_{tc}}{\partial e_i} &= \left(\frac{\partial D(e_i, c^p)}{\partial e_i} - \frac{\partial D\left(e_i, c^q_{\min}\right)}{\partial e_i}\right) \cdot \delta\left[\tilde{L_i} > 0\right] \\ &= \left(c^q_{\min} - c^p\right) \cdot \delta\left[\tilde{L_i} > 0\right]\end{aligned} \tag{11}$$

$$\begin{aligned}\frac{\partial L_{tc}}{\partial c_j} &= \frac{\sum_{i=1}^{M}\left(e_i - c_j\right) \cdot \delta\left[\tilde{L_i} > 0\right] \cdot \delta[p = j]}{1 + \sum_{i=1}^{M} \delta\left[\tilde{L_i} > 0\right] \cdot \delta[p = j]} \\ &\quad - \frac{\sum_{i=1}^{M}\left(e_i - c_j\right) \cdot \delta\left[\tilde{L_i} > 0\right] \cdot \delta[q = j]}{1 + \sum_{i=1}^{M} \delta\left[\tilde{L_i} > 0\right] \cdot \delta[q = j]}\end{aligned} \tag{12}$$

3.3.3 Joint supervision with softmax loss

The softmax loss directly encourages the separability of different classes and often converges faster than deep metric-based losses, thus providing a guidance for seeking better centers efficiently. At the same time, deep metric losses target at learning compact and separate representations by explicitly modeling the cross-expression relationship. According to the recent work presented in [40], softmax loss and deep metric-based loss could be complementary to each other, and the combination of these two losses could achieves more discriminative and robust embedding. Empirically, these two losses can be combined to achieve more discriminative and robust feature embedding [20, 23, 41]. Therefore, an effective approach for improvement is to combine the classification and similarity constraints to form a joint optimization strategy. The final loss function is defined as

$$L_{total} = \lambda L_{tc} + L_{softmax} \tag{13}$$

where λ is a trade-off hyper-parameter to balance the two terms.

4 Experiments

4.1 Experimental datasets

To evaluate the performance of the proposed method, extensive experiments are conducted on two well-known facial expression databases: FER2013 [42] and SFEW [43]. The FER2013 database is a large, publicly available database collected automatically by the Google image search API. It contains 28,709 training images, 3,589 validation images, and 3,589 test images with seven expression labels (i.e., anger, disgust, fear, happiness, sadness, surprise, and neutral). Every image is registered and resized to 48*48 pixels after rejecting incorrectly labeled frames and adjusting the cropped region.

The dataset is challenging since the depicted faces vary significantly with respect to the subject's age, face pose, and other factors, reflecting realistic conditions. The accuracy of expression classification by humans on this dataset is about 65.5% [42]. The SFEW 2.0 database was created from Acted Facial Expressions in the Wild (AFEW) [43] using the key-frame extraction method. It contains 891 training samples, 431 validation samples, and 372 test samples. These images are extracted from film clips, and labeled with six basic expressions of angry, disgust, fear, happy, sad, surprise, and the neutral class.

It targets for unconstrained facial expressions with large variations, reflecting real-world conditions such as different head poses, occlusions, and backgrounds. Since SFEW 2.0 is a dataset for the 2015 competition Challenges in the Wild (EmotiW) [44], test sample labeling is private and held back by the challenge organizer. Since we do not have access to the testing data, the evaluation results are reported in terms of the validation data.

4.2 Implementation details

Our experiments were conducted on a server with a Tesla P100 GPU provided by Google Colab. As introduced in Sect. 3.1, the structure of DAF-CNN has three parts, i.e., feature extraction block, 3D attention module and TC loss classifier. The detailed network structure is illustrated in Fig. 1. The input images are preprocessed by MTCNN, scaled to 96*96 pixels and normalized to [0, 1] by dividing each pixel gray level by 255. This is insufficient for training a deep CNN with limited training data.

To avoid overfitting, a data augmentation strategy is employed to train the CNN models both for FER2013 and SFEW. A dropout rate of 0.5 is employed for the last two FC layers. For the attention part, the reduction ratio r is set to 8, and the kernel size of the convolution in Eq. 3 is set to 7*7. As to the TC loss classifier, the margin m and the trade-off parameter A are respectively set to 11 and 0.007 for FER2013, and respectively set to 13 and 0.013 for SFEW. We initialized the centers with a Gaussian distribution, and the mean and standard deviation are (0, 0.01). The network is trained with Adam optimizer [45] with a min-batch of 128 for FER2013 and 32 for SFEW.

The initial learning rate was set to 0.001, while the minimum learning rate was set to 1e-5. Each training epoch had [N/128] batches, with the training samples randomly selected from the training set. The trained network parameters and accuracy at each epoch were recorded. If the validation accuracy did not increase by at least 0.0005 for 13 epochs, the learning rate was reduced by a factor of

0.2, and the previous model with the best accuracy was reloaded.

4.3 Results

4.3.1 Results on FER2013

The confusion matrix of the proposed DAF-CNN model on FER2013 dataset is shown in Fig. 2, where the leading diagonal entries represent the recognition accuracy for each expression. Table 1 shows that surprise, happiness, and disgust are the emotions with the three highest recognition rates. However, confusion frequently occurs among anger, fear, and sadness because these emotions are often presented by similar facial expressions [46].

Ablation study. To better evaluate the effectiveness of the proposed method (i.e., DAF-CNN), we conducted an ablation study to verify the contribution of each of its component to the performance of its whole network. In addition to the proposed DAF-CNN, three different methods are developed, i.e.,

1. DAF-CNN_NOatt&tcl, which denotes the proposed network without incorporating the 3D attention module and TC loss function (i.e., still using the softmax loss);
2. DAF-CNN_NOatt, which denotes the proposed network without the attendance of the 3D attention module; and
3. DAF-CNN_NOtcl, which denotes the proposed network without the supervision of TC loss.

The results in Table 1 show that either augmenting the proposed 3D attention or incorporating the TC loss function significantly boosts the recognition accuracy. This demonstrates the effectiveness of the two components of the proposed method, which can be employed individually to improve the performance of FER.

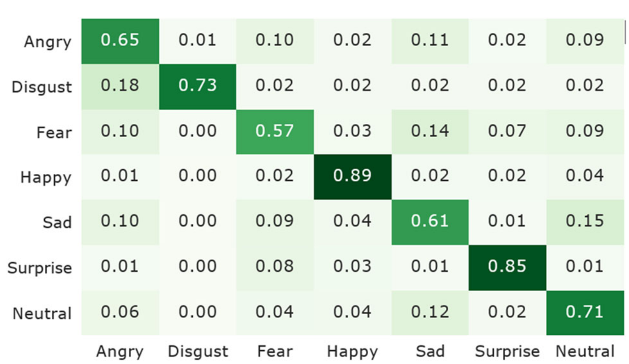

Fig. 2 Confusion matrix of the proposed DAF-CNN method evaluated on FER2013 test set. (The ground truth and the predicted expression labels are given by the first column and the first row, respectively)

Table 1 Ablation study result on the FER2013 testing set

Method	Accuracy (%)
DAF–CNN_NOatt&tcl	70.41
DAF–CNN_NOatt	71.38
DAF–CNN_NOtcl	71.80
DAF–CNN	72.39

Moreover, with the combination of these two promising components, the proposed DAF-CNN achieves the highest accuracy performance with a notable margin over the DAF-CNN_NOatt&tcl. This is because each component plays a complementary role in providing useful clues for FER from different perspectives.

The former exploits the CNN feature map to generate salient features while the latter attends in the embedding feature space to learn better representations Table 2.

4.3.2 Results on SFEW

We also validated the proposed method on the SFEW 2.0 dataset. Considering that deep CNNs are prone to overfit when they are trained with a small amount of data (891 images in SFEW training set), our strategy is to pre-train the model on the FER2013 training set and then fine-tune on the SFEW training set. Since there exist biases in two different datasets, the customized hyperparameters of TC loss for FER2013 dataset are not necessarily optimal for SFEW dataset.

Empirically, the pre-trained model equipped with attention module and supervised by softmax loss has a superior generalization ability. While in the fine-tuning

Table 2 Performance comparison on FER2013 testing set

	Accuracy (%)	
Method	single	Ensemble
Devries et al. [47]	67.21	–
Tang [11]	–	71.20
Guo et al. [48]	71.33	–
Mollahosseini et al. [13]	66.40	–
Hua et al. [49]	68.18	71.91
Yu et al. [12]	70.30	72.10
Kim et al. [50]	70.58	72.72
Connie et al. [51]	72.10	73.58
Shao et al. [51]	71.14	–
Proposed method (DAF–CNN)	72.39	–

stage, the softmax loss is replaced by TC loss to better suit the characteristics of the SFEW dataset.

The confusion matrix of the proposed method on the SFEW validation set is shown in Fig. 3.

The leading diagonal values shows that happiness and neutral have the highest recognition rates, while the recognition accuracy for disgust and fear is much lower than the others. These results are also observed in other published works.

Comparison with the state-of-the-art. The performance comparisons between the proposed method and the state-of-the-art FER methods are shown in Table 3. The table shows the proposed DAF-CNN model outperforms the baseline method of SFEW (35.93% on the validation set) by a large margin.

With respect to the comparison of single network performances, DAF-CNN ranks the first with an accuracy of 52.98%. Even compared with voting-based methods, the performance of our method is still competitive or comparable, demonstrating the effectiveness and robustness of the proposed method under real-world conditions.

4.4 Visualization analysis

To further demonstrate the effectiveness of our proposed method, we used t-SNE [54], a widely employed method for visualizing high dimensional data, to compare feature representations learned by DAF-CNN_NOatt&tcl, DAF-CNN_NOtcl, DAF-CNN_NOatt, and DAF-CNN.

As illustrated in Figs. 4 and 5, the learned features are clustered according to the seven expressions, each cluster denoted by a different color and a numeral. Some characteristics of the results can be observed from the comparison which is worth further detailed analysis. First, among all the evaluated models, features learned by softmax, i.e., (a) in Fig. 4

In Fig. 5 are the most significantly scattered and mixed in the feature space. Since softmax does not regulate the

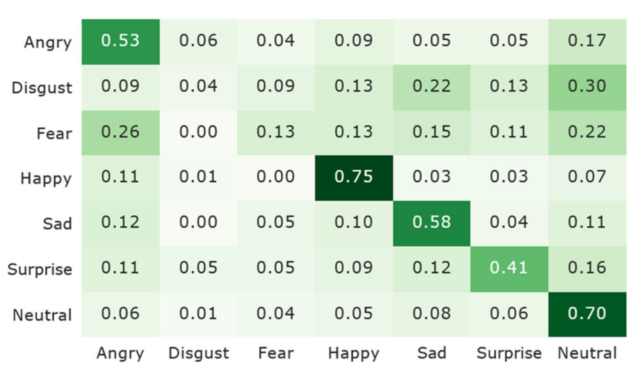

Fig. 3 Confusion matrix of the proposed DAF-CNN method evaluated on SFEW validation set. (The ground truth and the predicted labels are given by the first column and the first row, respectively)

Table 3 Performance comparison on SFEW validation set

	Accuracy (%)	
Method	single	Ensemble
Dhall et al. (baseline of SFEW) [44]	35.93	–
Yu et al. [12]	52.29	55.96
Ng et al. [52]	48.5	–
Mollahosseini et al. [13]	47.7	–
Levi et al. [22]	44.73	51.75
DLP–CNN [23]	51.05	–
IACNN [41]	50.98	–
IL–CNN [24]	51.83	52.52
Ji et al. [53]	51.2	–
DAF–CNN	52.98	–

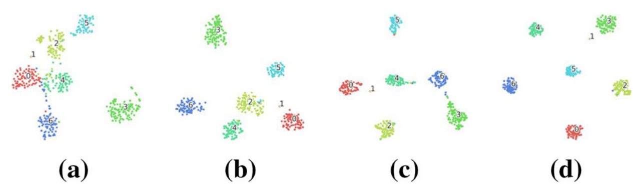

Fig. 4 A visualization of deeply-learned features on FER2013 training set learned by **a** DAF-CNN_NOatt&tcl **b** DAF-CNN_NOtcl **c** DAF-CNN_NOatt, and **d** DAF-CNN, including 512 samples from the training data set of FER2013. Note that the features learned by **d** are the most compact and separate. Best viewed in color)

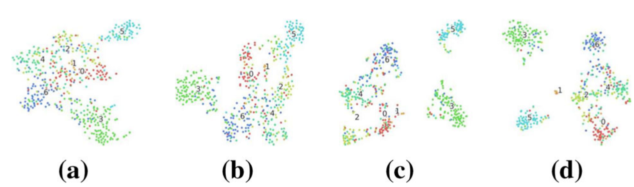

Fig. 5 The distribution of deeply-learned features on the FER2013 testing set. (Best viewed in color)

distances between embedded features, clusters of different classes are easily overlapped while features within the same cluster are scattered.

Second, although the method for (b) in Figs. 4 and 5 does not explicitly perform similarity constraint to regulate the feature distribution, it still learns better representations that are more compact and separate than (a) in Figs. 4 and 5. As shown in Fig. 4b, learned clusters are less likely to overlap than in Fig. 4a. This improvement of discriminability can be ascribed to the effective feature refinement performed by attention. Third, in Figs. 4 and 5, both (b) and (c) can learn discriminative representations but they differ from each other. TC loss can keep the learned features compact and isolated simultaneously. As can be observed in Fig. 5c, features are prone to aggregate to a denser point, since TC loss learns a corresponding center for each expression class. While in Fig. 5b, features are likely to be scattered without a clear centroid, rendering its lack of compactness.

This comparison demonstrates the effectiveness of integrated TC loss, which explicitly forces the combined similarity and deviation constraints to minimize the intra-class distance and maximize inter-class distance simultaneously. Fourth, notably, with the integration of both attention module and TC loss, features learned by (d) in Figs. 4 and 5 achieve the best intra-class compactness and inter-class discrepancy in both training and testing datasets. Compare (c) with (c) with (d) in Figs. 4 and 5, it can be concluded that jointly performing discriminative feature learning by exploiting both the feature map and the embedding feature space is effective and complementary, thus benefits can be accumulated to greatly enhance the representative power of the network.

5 Conclusion

This paper presents a novel deep learning approach for FER. The approach captures more comprehensive expression-related representations through a DAF-CNN. The proposed 3D attention mechanism not only emphasizes expression-related attentive feature channels, but also highlights the usefulness of expression-sensitive facial regions. In addition, the TC loss forces a similarity constraint on the learned features to simultaneously minimize the intra-class distance and maximize inter-class distance. Overall, it can be concluded that joint integration of attention mechanism and deep metric learning effectively captures more discriminative expression-related features and leading to the significant improvement of FER accuracy.

The introduced method simulates and realistically models a complex environment, using a small volume of labeled data. It performs novel adjustment of its hyper parameters based on the target data, and it achieves high-precision classification compared to other sophisticated methods [55, 56]. An important innovation is the employment of attention-augmented [57, 58] to large intra-class variation and inter-class similarity of facial expressions [59, 60] in real-world scenarios. The performance of the proposed system has been tested on a multi-dimensional complex dataset. The obtained high-precision results, greatly enhance the introduced methodology.

Future improvements of the system, should focus on further optimizing the hyper parameters of the proposed method. This will result in an even more efficient, accurate, and faster classification process. Also, it will be very important to study the extension of this method for the analysis and classification of real-time facial expressions. Finally, the proposed algorithm will be extended to operate in a fully self-determined manner by self-attention network [61, 62].

Acknowledgements The work is supported by National Science Foundation of CHINA (Grant No. 61902187 and 61801241), Natural science foundation of Jiangsu Province (BK20180746), Jiangsu Province Innovative and Entrepreneurial Talent Project, and National College Students Innovation and Entrepreneurship Training Program (Grant No. 2019NFUSPITP0548).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Li S, Deng W (2018) Deep facial expression recognition: a survey. arXiv preprint arXiv:1804.08348
2. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans Pattern Anal Mach Intell 29:915–928
3. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, 1: 886–893
4. Zhang Z, Lyons M, Schuster M, Akamatsu S (2015) Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. Proceedings third IEEE In: International conference on automatic face and gesture recognition. IEEE, 454–459
5. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. 2010 ieee computer society conference on computer vision and pattern recognition-workshops. IEEE, 94–101
6. Pantic M, Valstar M, Rademaker R, Maat L (2005) Web-based database for facial expression analysis. 2005 IEEE In: International conference on multimedia and Expo. IEEE
7. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
8. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, 1–9
9. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
10. Ye Q, Yang J, Liu F, Zhao C, Ye N, Yin T (2016) L1-norm distance linear discriminant analysis based on an effective iterative algorithm. IEEE Trans Circuits Syst Video Technol 28:114–129
11. Tang Y (2013) Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239
12. Yu Z, Zhang C (2015) Image based static facial expression recognition with multiple deep network learning. Proceedings of the 2015 ACM on In: International conference on multimodal interaction. ACM, 435–442

13. Mollahosseini A, Chan D, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. 2016 IEEE winter conference on applications of computer vision (WACV). IEEE, 1–10
14. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. European conference on computer vision. Springer, 483–499
15. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. proceedings of the IEEE conference on computer vision and pattern recognition, 7132–7141
16. Woo S, Park J, Lee JY, So Kweon I Cbam: convolutional block attention module. Proceedings of the European conference on computer vision (ECCV), 3–19
17. Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua TS (2017) SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. https://doi.org/10.1109/CVPR.2017.667
18. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. Proceedings of the IEEE conference on computer vision and pattern recognition, 815–823
19. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. European conference on computer vision. Springer, 499–515
20. He X, Zhou Y, Zhou Z, Bai S, Bai X (2018) Triplet-center loss for multi-view 3d object retrieval. Proceedings of the IEEE conference on computer vision and pattern recognition, 1945–1954
21. Connie T, Al-Shabi M, Cheah WP, Goh M (2017) Facial expression recognition using a hybrid CNN–SIFT aggregator. In: International workshop on multi-disciplinary trends in artificial intelligence. Springer, 139–149
22. Levi G, Hassner T (2015) Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. Proceedings of the 2015 ACM on In: International conference on multimodal interaction. ACM, 503–510
23. Li S, Deng W, Du J (2017) Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. Proceedings of the IEEE conference on computer vision and pattern recognition, 2852–2861
24. Cai J, Meng Z, Khan AS, Li Z, O'Reilly J, Tong Y (2018) Island loss for learning discriminative features in facial expression recognition. 2018 13th IEEE In: International conference on automatic face & gesture recognition (FG 2018). IEEE, 302–309
25. Ye Q, Li Z, Fu L, Zhang Z, Yang W, Yang G (2019) Nonpeaked discriminant analysis for data representation. IEEE trans neural netw learn syst 30:3818–3832
26. Fu L, Li Z, Ye Q, Yin H, Liu Q, Chen X, Fan X, Yang W, Yang G (2020) Learning robust discriminant subspace based on joint L2, p-and L2, s-norm distance metrics. IEEE transactions on neural networks and learning systems
27. Rifai S, Bengio Y, Courville A, Vincent P, Mirza M (2012) Disentangling factors of variation for facial expression recognition. European conference on computer vision. Springer, 808–822
28. Reed S, Sohn K, Zhang Y, Lee H (2014) Learning to disentangle factors of variation with manifold interaction. In: International conference on machine learning, 1431–1439
29. Ge S, Li C, Zhao S, Zeng D (2020) Occluded face recognition in the wild by identity-diversity inpainting. IEEE Trans Circuits Syst Video Technol 30:3387–3397. https://doi.org/10.1109/TCSVT.2020.2967754
30. Zhu K, Du Z, Li W, Huang D, Wang Y, Chen L (2019) Discriminative attention-based convolutional neural network for 3D facial expression recognition. 2019 14th IEEE In: International conference on automatic face gesture recognition (FG 2019), 1–8 https://doi.org/10.1109/FG.2019.8756524
31. Zhou X, Jin K, Shang Y, Guo G (2020) Visually interpretable representation learning for depression recognition from facial images. IEEE Trans Affect Comput 11:542–552. https://doi.org/10.1109/TAFFC.2018.2828819
32. Zhou X, Wei Z, Xu M, Qu S, Guo G (2020) Facial depression recognition by deep joint label distribution and metric learning. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2020.3022732
33. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473
34. Luong MT, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025
35. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. In: International conference on machine learning, 2048–2057
36. Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked attention networks for image question answering. Proceedings of the IEEE conference on computer vision and pattern recognition, 21–29
37. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017) Residual attention network for image classification. Proceedings of the IEEE conference on computer vision and pattern recognition, 3156–3164
38. Jetley S, Lord NA, Lee N, Torr PH (2018) Learn to pay attention. arXiv preprint arXiv:1804.02391
39. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. BMVC 1:6
40. Horiguchi S, Ikami D, Aizawa K (2017) Significance of softmax-based features in comparison to distance metric learning-based features. arXiv preprint arXiv:1712.10151, 2
41. Meng Z, Liu P, Cai J, Han S, Tong Y (2017) Identity-aware convolutional neural network for facial expression recognition. 2017 12th IEEE In: International conference on automatic face & gesture recognition (FG 2017). IEEE, 558–565
42. Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee DH (2013) Challenges in representation learning: a report on three machine learning contests. In: International conference on neural information processing. Springer, 117–124
43. Dhall A, Goecke R, Lucey S, Gedeon T (2012) Collecting large, richly annotated facial-expression databases from movies. IEEE Multimed 19:34–41
44. Dhall A, Ramana Murthy O, Goecke R, Joshi J, Gedeon T (2015) Video and image based emotion recognition challenges in the wild: emotiw 2015. Proceedings of the 2015 ACM on In: International conference on multimodal interaction. ACM, 423–426
45. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
46. Hoffmann H, Scheck A, Schuster T, Walter S, Limbrecht K, Traue HC, Kessler H (2012) Mapping discrete emotions into the dimensional space: an empirical approach. 2012 IEEE In: International conference on systems, man, and cybernetics (SMC). IEEE, 3316–3320
47. Devries T, Biswaranjan K, Taylor GW (2014) Multi-task learning of facial landmarks and expression. 2014 Canadian conference on computer and robot vision. IEEE, 98–103
48. Guo Y, Tao D, Yu J, Xiong H, Li Y, Tao D (2016) Deep neural networks with relativity learning for facial expression recognition. 2016 IEEE In: International conference on multimedia & expo workshops (ICMEW). IEEE, 1–6
49. Hua W, Dai F, Huang L, Xiong J, Gui G (2019) HERO: Human emotions recognition for realizing intelligent Internet of Things. IEEE Access 7:24321–24332

50. Kim BK, Roh J, Dong SY, Lee SY (2016) Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. J Multimodal User Interfaces 10:173–189. https://doi.org/10.1007/s12193-015-0209-0
51. Jie S, Yongsheng Q (2019) Three convolutional neural network models for facial expression recognition in the wild. Neurocomputing. https://doi.org/10.1016/j.neucom.2019.05.005
52. Ng HW, Nguyen VD, Vonikakis V, Winkler S (2015) Deep learning for emotion recognition on small datasets using transfer learning. Proceedings of the 2015 ACM on In: International conference on multimodal interaction. ACM, 443–449
53. Ji Y, Hu Y, Yang Y, Shen F, Shen HT (2019) Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network. Neurocomputing 333:231–239
54. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J mach learn res 9:2579–2605
55. Demertzis K, Iliadis L (2020) GeoAI: A model-agnostic meta-ensemble zero-shot learning method for hyperspectral image analysis and classification. Algorithms 13:61. https://doi.org/10.3390/a13030061
56. Demertzis K, Iliadis L, Pimenidis E (2020) Large-scale geospatial data analysis: geographic object-based scene classification in remote sensing images by GIS and deep residual learning. In: Iliadis L, Angelov P, Jayne C, Pimenidis E (eds) proceedings of the 21st EANN (engineering applications of neural networks) 2020 conference. EANN 2020. Proceedings of the In: International neural networks society, vol 2. Springer, Cham https://doi.org/10.1007/978-3-030-48791-1_21
57. Ly NT, Nguyen CT, and Nakagawa M (2020) "Attention augmented convolutional recurrent network for handwritten japanese text recognition," 2020 17th In: International conference on frontiers in handwriting recognition (ICFHR), Dortmund, Germany, 163–168 https://doi.org/10.1109/ICFHR2020.2020.00039
58. Bello I, Zoph B, Le Q, Vaswani A, and Shlens J (2019) "Attention augmented convolutional networks," 2019 IEEE/CVF In: International conference on computer vision (ICCV), Seoul, Korea (South), 3285–3294 https://doi.org/10.1109/ICCV.2019.00338
59. Chen Y, Wang J, Chen S, Shi Z, and Cai J (2019) "Facial motion prior networks for facial expression recognition," 2019 IEEE visual communications and image processing (VCIP), Sydney, NSW, Australia, 1–4 https://doi.org/10.1109/VCIP47243.2019.8965826
60. Yi J, Sima Y, Zhou M, and Yang J (2019) "Facial expression sequence interception based on feature point movement," 2019 IEEE 11th In: International conference on advanced infocomm technology (ICAIT), Jinan, China, 58–62 https://doi.org/10.1109/ICAIT.2019.8935902
61. Kim M, Kim T, and Kim D (2020) "Spatio-temporal slowfast self-attention network for action recognition," 2020 IEEE In: International conference on image processing (ICIP), Abu Dhabi, United Arab Emirates, 2206–2210 https://doi.org/10.1109/ICIP40778.2020.9191290
62. He N, Fang L, Li Y, and Plaza A (2019) "High-order self-attention network for remote sensing scene classification," IGARSS 2019–2019 IEEE In: International geoscience and remote sensing symposium, Yokohama, Japan, 3013–3016, https://doi.org/10.1109/IGARSS.2019.8898320