

Manuscript version: Author's Accepted Manuscript

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:

<http://wrap.warwick.ac.uk/152750>

How to cite:

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

Discriminative Attention-augmented Feature Learning for Facial Expression Recognition in the Wild

Linyi Zhou^a, Xijian Fan^{a*}, Tardi Tjahjadi^b, Sruti Das Choudhury^c

^aCollege of Information Science and Technology, Nanjing Forestry University, Nanjing, Jiangsu, 210037, China

^bSchool of Engineering, University of Warwick Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom.

^cSchool of Natural Resources, Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68583-0987, USA.

*Corresponding author, Email address: xijian.fan@njfu.edu.cn

Abstract

Facial expression recognition (FER) in-the-wild is challenging due to unconstrained 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-Centre (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.

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], Histogram of Oriented Gradients (HOG) [3] and 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, 8, 9, 10]. It has also been shown that various CNN architectures can achieve promising results in FER [11, 12, 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 neighbourhood. 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 modelling 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 modelling 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 centre 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 centre for each class to obtain compact features. Both of them aim at learning discriminative feature representations. Inspired by these two losses, Triplet-Centre loss (TC loss) is proposed for 3D object retrieval in [20]. By combining the merits of both triplet loss and centre 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, 22, 23, 24, 25, 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

2.1. Deep Learning for FER

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 Boltzman 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, 31, 32].

2.2. Attention Mechanisms in Neural networks

Attention has been widely adopted for modelling 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 modelling 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.

2.3. Deep metric learning loss

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., centre loss and triplet loss. Centre loss [19]

was proposed as an auxiliary for softmax loss to learn more discriminative features. In the training process, centre loss learns a centre for the features of each class and pulls features of the same class to its corresponding centre. Through the joint supervision of softmax loss, centre loss is able to learn compact features that are close to their centres. However, centre loss does not consider the inter-class separability 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 optimises 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.

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 form 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 emphasising 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.

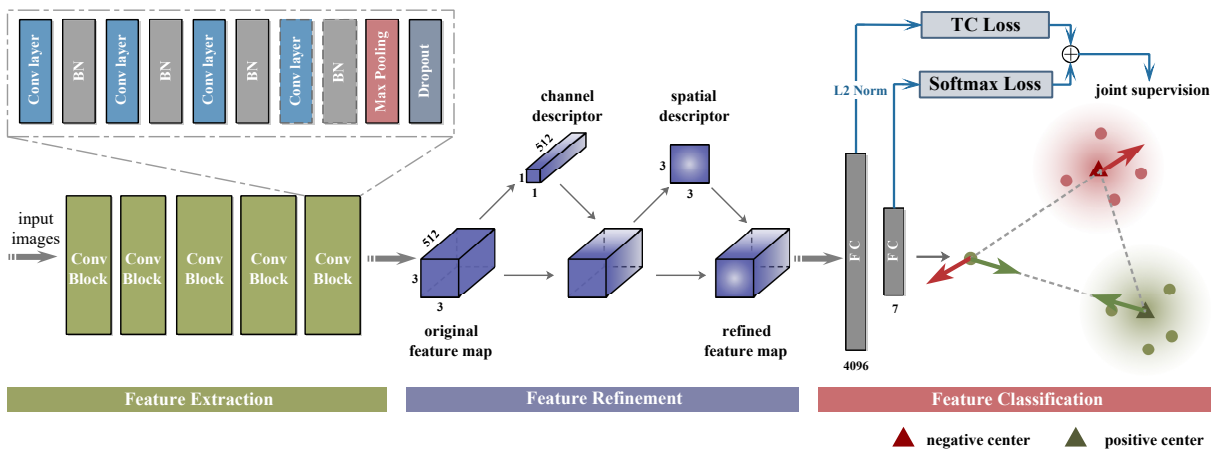


Figure 1: An overview of the proposed DAF-CNN framework. The generated feature representations are learned through three stages, i.e., feature extraction, feature refinement, and feature classification.

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 neighbourhood, 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 parameterised 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 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 $X \in \mathbb{R}^{W \times H \times C}$, 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 $V_{avg} \in \mathbb{R}^{1 \times 1 \times C}$ and $V_{max} \in \mathbb{R}^{1 \times 1 \times C}$. They are then fed to a network consisting of two hidden layers FC_1 and FC_2 . FC_1 reduces the feature dimension to $1 * 1 * C/r$, 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

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

and

$$\text{Net}[V(x)] = FC_2 [FC_1(V(x))], \quad (2)$$

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

$$\mathcal{A}_s = Sigmoid\{\otimes [Concat(M_{avg}(x); M_{max}(x))]\}, \quad (3)$$

where \otimes 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

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

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

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 centre 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 centre loss, the learned clusters are

likely to be overlapped since centre 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 centre 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 $p \in \{1, 2, \dots, |\mathcal{K}|\}$, 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 centre, thereby obtaining $\mathcal{C} = \{c_1, c_2, \dots, c_{|\mathcal{K}|}\}$, where $c_y \in \mathbb{R}^d$ denotes the centre vector for samples with label y , and $|\mathcal{K}|$ is the number of centres. For simplicity, we adopt e_i to represent $E(x_i)$ in this paper. In terms of triplet loss, the input triplet (x_i^a, x_i^+, x_i^-) is constituted of samples. While in TC loss, we select the i -th sample x_i , its corresponding positive centre c^p , and its nearest negative centre c_{min}^q to reconstruct the triplet input as (x_i, c^p, c_{min}^q) . 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 centres as its similarity metric, TC loss avoids direct interaction with samples of poor quality such as mislabelled 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 centre c^p to represent the degree of deviation among expressions of the same class, which is formulated as

$$D(e_i, c^p) = \frac{1}{2} \|e_i - c^p\|_2^2. \quad (6)$$

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

$$D(e_i, c_{min}^q) = \frac{1}{2} \|e_i - c_{min}^q\|_2^2. \quad (7)$$

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

$$\frac{1}{2} \|e_i - c^p\|_2^2 + m < \frac{1}{2} \|e_i - c_{min}^q\|_2^2. \quad (8)$$

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

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

3.3.2. Backward propagation

To compute the back-propagation gradients of the input feature embeddings and the corresponding centres, we assume the following notations for demonstration: $\delta[\text{condition}]$ 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.,

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

These cluster centres are updated based on mini-batches similar to the practice in centre 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 centre $\frac{\partial L_{tc}}{\partial c_j}$ are determined as follows:

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

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

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 centres efficiently. At the same time, deep-metric losses target at learning compact and separate representations by explicitly modelling 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 embeddings [23, 20, 41]. Therefore, an effective approach for improvement is to combine the classification and similarity constraints to form a joint optimisation strategy. The final loss function is defined as

$$L_{total} = \lambda L_{tc} + L_{softmax}, \quad (13)$$

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 labelled 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 labelled 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 labelling 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 Section 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 grey 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 λ are respectively set to 11 and 0.007 for FER2013, and respectively set to 13 and 0.013 for SFEW. We initialized the centres 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 $\lfloor N/128 \rfloor$ 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. Fig. 2 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].

Angry	0.65	0.01	0.10	0.02	0.11	0.02	0.09
Disgust	0.18	0.73	0.02	0.02	0.02	0.02	0.02
Fear	0.10	0.00	0.57	0.03	0.14	0.07	0.09
Happy	0.01	0.00	0.02	0.89	0.02	0.02	0.04
Sad	0.10	0.00	0.09	0.04	0.61	0.01	0.15
Surprise	0.01	0.00	0.08	0.03	0.01	0.85	0.01
Neutral	0.06	0.00	0.04	0.04	0.12	0.02	0.71
	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral

Figure 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

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. 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: Performance comparison with the state-of-the-art methods on FER2013 testing set.

Method	Accuracy(%)	
	Single	Ensemble
Devries et al. [47]	67.21	-
Tang [11]	-	71.2
Guo et al. [48]	71.33	-
Mollahosseini et al. [13]	66.4	-
Hua et al. [49]	68.18	71.91
Yu et al. [12]	70.3	72.1
Kim et al. [50]	70.58	72.72
Connie et al. [51]	72.1	73.58
Shao et al. [51]	71.14	-
Proposed method (DAF-CNN)	72.39	-

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 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

Angry	0.53	0.06	0.04	0.09	0.05	0.05	0.17
Disgust	0.09	0.04	0.09	0.13	0.22	0.13	0.30
Fear	0.26	0.00	0.13	0.13	0.15	0.11	0.22
Happy	0.11	0.01	0.00	0.75	0.03	0.03	0.07
Sad	0.12	0.00	0.05	0.10	0.58	0.04	0.11
Surprise	0.11	0.05	0.05	0.09	0.12	0.41	0.16
Neutral	0.06	0.01	0.04	0.05	0.08	0.06	0.70
	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral

Figure 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.)

diagonal values shows that happiness and neutral have the highest recognition rates, while the recognition accuracy for disgust and fear are 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 (a) DAF-CNN_NOatt&tcl, (b) DAF-CNN_NOtcl, (c) DAF-CNN_NOatt, and (d) DAF-CNN. As illustrated in Fig. 4 and Fig. 5, the learned features are clustered according to the seven expressions, each cluster denoted by a different colour 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 and Fig. 5 are the most significantly scattered and mixed in the feature space. Since softmax does not regulate the 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 Fig. 4 and Fig. 5 does not explicitly perform

Table 3: Performance comparison with the state-of-the-art methods on SFEW validation set.

Method	Accuracy(%)	
	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	-

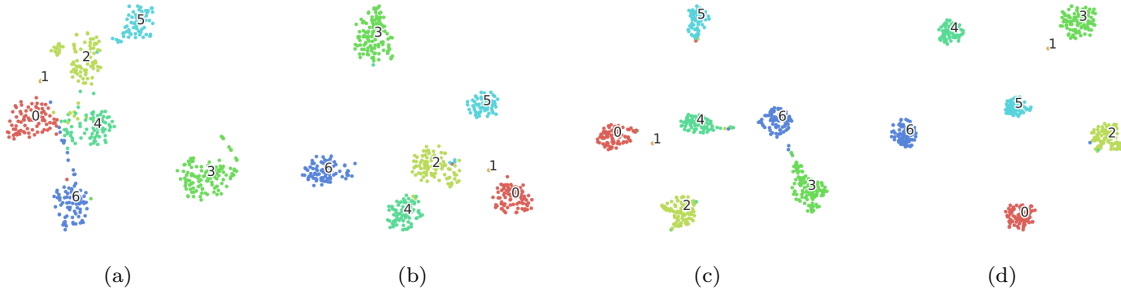


Figure 4: A visualization of deeply-learned features on FER2013 training set learned by (a) DAF-CNN_NOatt&tcl (b) DAF-CNN_NOtcl (c) DAF-CNN_NOatt, and (d) DAF-CNN, including 512 samples from the training data set of FER2013. Note that the features learned by (d) are the most compact and separate. Best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

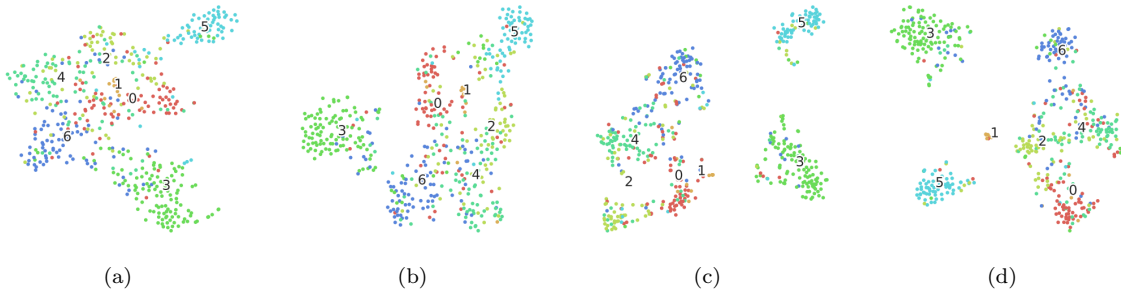


Figure 5: The distribution of deeply-learned features on the FER2013 testing set. (Best viewed in colour.)

similarity constraint to regulate the feature distribution, it still learns better representations that are more compact and separate than (a) in Fig. 4 and Fig. 5. As shown in Fig. 4(b), learned clusters are less likely to overlap than in Fig. 4(a). This improvement of discriminability can be ascribed to the effective feature refinement performed by attention. Third, in Fig. 4 and Fig. 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. 5(c), features are prone to aggregate to a denser point, since TC loss learns a corresponding centre for each expression class. While in Fig. 5(b), 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 Fig. 4 and Fig. 5 achieve the best intra-class compactness and inter-class discrepancy in both training and testing datasets. Compare (c) with (d) in Fig. 4 and Fig. 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.

Author Contributions: Xijian Fan is the corresponding author who designed the algorithm and revised the paper. Linyi Zhou conceived of, designed and performed the experiments, analyzed the data and wrote this paper. Tardi Tjahjadi provided some comments and suggestions, and also revised the paper. Srui Das Choudhury provided some suggestions and comments for the performance improvement of the recognition algorithm.

Acknowledgments: The work is supported by National Science Foundation of CHINA (Grant No. 61902187), Jiangsu Province Innovative and Entrepreneurial Talent Project, Nanjing Forestry University High-level Start-up Foundation for Research, and National College Students Innovation and Entrepreneurship Training Program (Grant No. 2019NFUSPITP0548).

Conflicts of interest: The authors declare no conflict of interest.

References

- [1] S. Li, W. Deng, Deep facial expression recognition: A survey, arXiv preprint arXiv:1804.08348.
- [2] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, *IEEE transactions on pattern analysis and machine intelligence* 29 (6) (2007) 915–928.
- [3] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), Vol. 1, IEEE, 2005, pp. 886–893.
- [4] Z. Zhang, M. Lyons, M. Schuster, S. Akamatsu, Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron, in: *Proceedings Third IEEE International Conference on Automatic face and gesture recognition*, IEEE, 2015, pp. 454–459.
- [5] 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: 2010 IEEE computer society conference on computer vision and pattern recognition-workshops, IEEE, 2010, pp. 94–101.
- [6] M. Pantic, M. Valstar, R. Rademaker, L. Maat, Web-based database for facial expression analysis, in: 2005 IEEE international conference on multimedia and Expo, IEEE, 2005.
- [7] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [9] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.
- [10] Q. Ye, J. Yang, F. Liu, C. Zhao, N. Ye, T. Yin, L1-norm distance linear discriminant analysis based on an effective iterative algorithm, *IEEE Transactions on Circuits and Systems for Video Technology* 28 (1) (2016) 114–129.
- [11] Y. Tang, Deep learning using linear support vector machines, arXiv preprint arXiv:1306.0239.
- [12] Z. Yu, C. Zhang, Image based static facial expression recognition with multiple deep network learning, in: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ACM, 2015, pp. 435–442.
- [13] A. Mollahosseini, D. Chan, M. H. Mahoor, Going deeper in facial expression recognition using deep neural networks, in: 2016 IEEE Winter conference on applications of computer vision (WACV), IEEE, 2016, pp. 1–10.
- [14] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: *European conference on computer vision*, Springer, 2016, pp. 483–499.
- [15] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [16] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, Cbam: Convolutional block attention module, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19.
- [17] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.-S. Chua, Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, 2017, pp. 6298–6306. doi:10.1109/CVPR.2017.667.
- [18] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.
- [19] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: *European conference on computer vision*, Springer, 2016, pp. 499–515.
- [20] X. He, Y. Zhou, Z. Zhou, S. Bai, X. Bai, Triplet-center loss for multi-view 3d object retrieval, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1945–1954.

- [21] T. Connie, M. Al-Shabi, W. P. Cheah, M. Goh, Facial expression recognition using a hybrid cnnsift aggregator, in: *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, Springer, 2017, pp. 139–149.
- [22] G. Levi, T. Hassner, Emotion recognition in the wild via convolutional neural networks and mapped binary patterns, in: *Proceedings of the 2015 ACM on international conference on multimodal interaction*, ACM, 2015, pp. 503–510.
- [23] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.
- [24] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O’Reilly, Y. Tong, Island loss for learning discriminative features in facial expression recognition, in: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, IEEE, 2018, pp. 302–309.
- [25] Q. Ye, Z. Li, L. Fu, Z. Zhang, W. Yang, G. Yang, Nonpeaked discriminant analysis for data representation, *IEEE transactions on neural networks and learning systems* 30 (12) (2019) 3818–3832.
- [26] L. Fu, Z. Li, Q. Ye, H. Yin, Q. Liu, X. Chen, X. Fan, W. Yang, G. Yang, Learning robust discriminant subspace based on joint l_2 , p -and l_2 , s -norm distance metrics, *IEEE Transactions on Neural Networks and Learning Systems*.
- [27] S. Rifai, Y. Bengio, A. Courville, P. Vincent, M. Mirza, Disentangling factors of variation for facial expression recognition, in: *European Conference on Computer Vision*, Springer, 2012, pp. 808–822.
- [28] S. Reed, K. Sohn, Y. Zhang, H. Lee, Learning to disentangle factors of variation with manifold interaction, in: *International Conference on Machine Learning*, 2014, pp. 1431–1439.
- [29] S. Ge, C. Li, S. Zhao, D. Zeng, Occluded face recognition in the wild by identity-diversity inpainting, *IEEE Transactions on Circuits and Systems for Video Technology* 30 (10) (2020) 3387–3397. doi:10.1109/TCSVT.2020.2967754.
- [30] K. Zhu, Z. Du, W. Li, D. Huang, Y. Wang, L. Chen, Discriminative attention-based convolutional neural network for 3d facial expression recognition, in: *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, 2019, pp. 1–8. doi:10.1109/FG.2019.8756524.
- [31] X. Zhou, K. Jin, Y. Shang, G. Guo, Visually interpretable representation learning for depression recognition from facial images, *IEEE Transactions on Affective Computing* 11 (3) (2020) 542–552. doi:10.1109/TAFFC.2018.2828819.
- [32] X. Zhou, Z. Wei, M. Xu, S. Qu, G. Guo, Facial depression recognition by deep joint label distribution and metric learning, *IEEE Transactions on Affective Computing* (2020) 1–1doi:10.1109/TAFFC.2020.3022732.
- [33] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473.
- [34] M.-T. Luong, H. Pham, C. D. Manning, Effective approaches to attention-based neural machine translation, arXiv preprint arXiv:1508.04025.
- [35] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: *International conference on machine learning*, 2015, pp. 2048–2057.
- [36] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.
- [37] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.
- [38] S. Jetley, N. A. Lord, N. Lee, P. H. Torr, Learn to pay attention, arXiv preprint arXiv:1804.02391.
- [39] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: *bmvc*, Vol. 1, 2015, p. 6.
- [40] S. Horiguchi, D. Ikami, K. Aizawa, Significance of softmax-based features in comparison to distance metric learning-based features, arXiv preprint arXiv:1712.10151 2.
- [41] Z. Meng, P. Liu, J. Cai, S. Han, Y. Tong, Identity-aware convolutional neural network for facial expression recognition, in: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, IEEE, 2017, pp. 558–565.
- [42] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H.

- Lee, Challenges in representation learning: A report on three machine learning contests, in: International Conference on Neural Information Processing, Springer, 2013, pp. 117–124.
- [43] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Collecting large, richly annotated facial-expression databases from movies, *IEEE multimedia* 19 (3) (2012) 34–41.
- [44] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, T. Gedeon, Video and image based emotion recognition challenges in the wild: Emotiw 2015, in: Proceedings of the 2015 ACM on international conference on multimodal interaction, ACM, 2015, pp. 423–426.
- [45] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.
- [46] H. Hoffmann, A. Scheck, T. Schuster, S. Walter, K. Limbrecht, H. C. Traue, H. Kessler, Mapping discrete emotions into the dimensional space: An empirical approach, in: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2012, pp. 3316–3320.
- [47] T. Devries, K. Biswaranjan, G. W. Taylor, Multi-task learning of facial landmarks and expression, in: 2014 Canadian Conference on Computer and Robot Vision, IEEE, 2014, pp. 98–103.
- [48] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, D. Tao, Deep neural networks with relativity learning for facial expression recognition, in: 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2016, pp. 1–6.
- [49] W. Hua, F. Dai, L. Huang, J. Xiong, G. Gui, Hero: Human emotions recognition for realizing intelligent internet of things, *IEEE Access* 7 (2019) 24321–24332.
- [50] B.-K. Kim, J. Roh, S.-Y. Dong, S.-Y. Lee, Hierarchical committee of deep convolutional neural networks for robust facial expression recognition, *Journal on Multimodal User Interfaces* 10 (2) (2016) 173–189. doi:10.1007/s12193-015-0209-0.
- [51] S. Jie, Q. Yongsheng, Three convolutional neural network models for facial expression recognition in the wild, *Neurocomputing* 355. doi:10.1016/j.neucom.2019.05.005.
- [52] H.-W. Ng, V. D. Nguyen, V. Vonikakis, S. Winkler, Deep learning for emotion recognition on small datasets using transfer learning, in: Proceedings of the 2015 ACM on international conference on multimodal interaction, ACM, 2015, pp. 443–449.
- [53] Y. Ji, Y. Hu, Y. Yang, F. Shen, H. T. Shen, Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network, *Neurocomputing* 333 (2019) 231 – 239.
- [54] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, *Journal of machine learning research* 9 (Nov) (2008) 2579–2605.