IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 8, 2022, accepted February 27, 2022, date of publication March 3, 2022, date of current version March 14, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3156598

Ad-Corre: Adaptive Correlation-Based Loss for Facial Expression Recognition in the Wild

ALI POURRAMEZAN FARD AND MOHAMMAD H. MAHOOR, (Senior Member, IEEE)
Ritchie School of Engineering and Computer Science, University of Denver, Denver, CO 80208, USA

Corresponding authors: Mohammad H. Mahoor (mmahoor@du.edu) and Ali Pourramezan Fard (ali.pourramezanfard@du.edu)

This work was partially supported by an internal grant from the University of Denver.

ABSTRACT Automated Facial Expression Recognition (FER) in the wild using deep neural networks is still challenging due to intra-class variations and inter-class similarities in facial images. Deep Metric Learning (DML) is among the widely used methods to deal with these issues by improving the discriminative power of the learned embedded features. This paper proposes an Adaptive Correlation (Ad-Corre) Loss to guide the network towards generating embedded feature vectors with high correlation for within-class samples and less correlation for between-class samples. Ad-Corre consists of 3 components called Feature Discriminator, Mean Discriminator, and Embedding Discriminator. We design the Feature Discriminator component to guide the network to create the embedded feature vectors to be highly correlated if they belong to a similar class, and less correlated if they belong to different classes. In addition, the Mean Discriminator component leads the network to make the mean embedded feature vectors of different classes to be less similar to each other. We use Xception network as the backbone of our model, and contrary to previous work, we propose an embedding feature space that contains k feature vectors. Then, the Embedding Discriminator component penalizes the network to generate the embedded feature vectors, which are dissimilar. We trained our model using the combination of our proposed loss functions called Ad-Corre Loss jointly with the cross-entropy loss. We achieved a very promising recognition accuracy on AffectNet, RAF-DB, and FER-2013. Our extensive experiments and ablation study indicate the power of our method to cope well with challenging FER tasks in the wild. The code is available on Github.

INDEX TERMS Facial expression recognition, facial emotion recognition, Ad-Corre loss, loss function, convolutional neural network.

I. INTRODUCTION

Automated Facial Expression Recognition (FER) is one of the most important visual recognition technologies to detect human emotions, a universal signal that is used by humans for non-verbal communication [1], [2]. Six expressions -angry, disgust, fear, happy, sad, and surprise- are defined by Ekman *et al*. [3] as the basic universal emotional expressions. Although automated FER has been a topic of study for decades, its widely-used applications in Human-Computer Interaction (HCI), driver monitoring for autonomous driving, education, healthcare, and psychological treatments has brought more attention to it more recently. Although the Convolutional Neural Networks (CNNs) based methods have achieved a promising accuracy in a wide range of applications [4]–[6], automated FER is still considered a challenging task specifically when it comes to practical applications.

Deep metric learning (DML) methods are proposed to improve the discrimination between the embedded feature vectors with respect to the class categories. Another benefit of DML, where the model is trained discriminatively, is that the network can learn the semantically meaningful embedded feature vectors which tend to be robust against intra-class variations [7]. To achieve this goal, The widely-used triplet loss [4] was proposed to increase the similarities between the embedded feature vectors of similar expressions while increasing the differences between different expressions. In other words, each time an anchor image, a positive sample image (having the same expression as the anchor image), and a negative sample image (having a different expression as the anchor image) are chosen. The triplet loss tends to penalize the network to minimize the distance (mostly the euclidean) between the anchor and the positive samples while

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

increasing the distances between the anchor image and the negative images.

Although triplet loss can enhance the generation of the embedded feature vectors and accordingly the performance of the classification task which is done mostly using cross-entropy (CE) loss, there are some issues with it. The most challenging issue of triplet loss [4] is its sensitivity to the process of choosing the anchor, positive and negative samples. In other words, unless the so-called *hard-positive* and *hard-negative* samples are chosen correctly, the network converges slowly or may never converge [4]. Moreover, a misclassified or a poor-quality image can likely be chosen as a hard-positive or a hard-negative candidate. While online mini-batch selecting strategy could alleviate the mentioned issues, the problem can still be devastating specifically for the task of FER where the intra-class variations and inter-class similarities of the facial expressions are dramatically high.

In this paper, we introduce Ad-Corre Loss to improve the discriminative power of the deep embedded feature vectors. Contrary to the triplet loss [4], the correlation loss is agnostic to the process of selecting the triplets. More specifically, inspired by the definition of the Correlation Matrix, we define Ad-Corre Loss to make the embedded feature vectors belonging to the same expression class to be highly correlated, while those belonging to the different expression classes be less correlated (see Fig. 1). Ad-Corre Loss consists of 3 different components called Feature Discriminator (FD), Mean Discriminator (MD), and Embedding Discriminator (ED) components.

We propose the FD component (see Fig. 2) to lead the network to generate embedded feature vectors with high correlation if they belong to the same classes and low correlation if they belong to different classes. The FD component uses all the samples within a mini-batch and calculates the correlation for all pairs of the embedded feature vectors. Such characteristic enables the Ad-Corre Loss to be tolerant to mislabeled images. Moreover, we propose *Adaptive Attention Map*, which is designed to monitor the performance of the model during the training phase and generate an attention map to penalize the model more for the expressions that are frequently misclassified compared to the ones with less frequent misclassification rate. The FD component utilizes the Adaptive Attention Map to direct the network toward generating more discriminative embedded feature vectors by monitoring the performance of the classification task.

In addition, we express that the high intra-class variations and inter-class similarities for FER in the wild might not be well discriminated by using only one embedded feature vector per input image. Accordingly, for an input image we introduce one *embedding space* containing k different embedded feature vectors. We propose the ED component (see Fig. 2) to penalize the model to generate the embedded feature vectors within the embedding space to be less correlated to each other, which leads the network towards extracting a wider range of features from the input image. In other words, since for each input image, the model generates k different

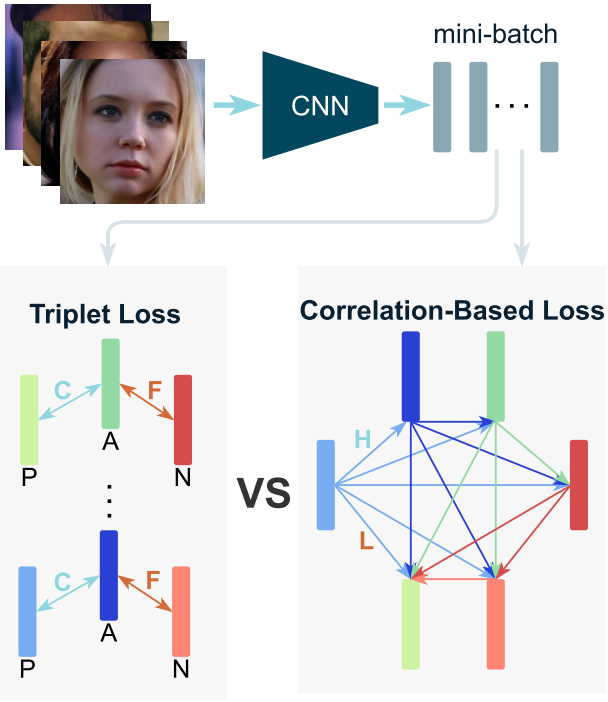

FIGURE 1. Correlation-based Loss versus Triplet Loss. Contrary to Triplet-based losses, Correlation Loss considers all the samples in a mini-batch and leads the network to create embedded feature vectors such that for similar classes they are highly correlated, and for different classes, they are less correlated. A:anchor, P:positive, N:negative, C:close, F:far, H: highly correlated, L: less correlated.

embedded feature vectors, the more dissimilar the embedded feature vectors the more discriminative they can be. As a consequence, for each input image, it is more likely that different embedded feature vectors tend to represent different features from the images. We define k equal to the number of facial expressions (which is seven in the context of this paper).

Considering FER as the task of clustering different embedded feature vectors, we want the model to create clusters that are compact and far from each other. In this manner, we propose the MD component (see Fig. 3), which calculates the mean of the embedded feature vectors for each class of expressions and penalizes the model to make the means of each class to be less correlated to the other means.

The contributions of our approach are summarized as follows.

- We propose Ad-Corre Loss, a correlation-based loss function with 3 main components designed to increase the discriminative power of the model.
- FD component designed to guide the model to generate the embedded feature vectors that are highly correlated if they belong to a similar class, and less correlated if belong to different classes. *Adaptive Attention Map* is proposed to guide the FD component to penalize the model more, for more frequent misclassified expressions compared to the less frequent misclassified classes.
- MD component designed to make the network generate the embedded feature vectors such that the means of

each class are less correlated, resulting in compactness of the clusters.
- ED component penalizes the network to make the embedded feature vectors to be as less correlated as possible to extract the wider range of features from the input image.

The remainder of this paper is organized as follows. Sec. II reviews the related work in FER. Sec. III describes our proposed Deep Metric-based model, the proposed loss function, and describes how it improves the accuracy of FER in the wild. Sec. IV provides the experimental results, and finally, Sec. VI concludes the paper with some discussions on the proposed method and future research directions.

II. RELATED WORK

In this section, we first review the previous work in FER and then discuss the use of DML for general image classification tasks. Afterward, we review the work used DML in FER.

FER in General: FER has been studied widely and a broad variety of approaches are proposed. In the following, we review some of the important work such as [8]–[19] in this context.

In order to extract the spatial relations within facial images and the temporal relations between different frames in the video, Hasani *et al*. [8] proposed a CNN including 3D Inception-ResNet layers. In another work, Georgescu *et al*. [9] proposed to fuse the automatically extracted features by utilizing a CNN and handcrafted features extracted by the bag-of-visual-words to design a model for FER. In addition, Hoang *et al*. [10] expressed that the general background data can also be considered as complementary cues for emotion recognition, and proposed a method using the visual relationship between the main target and the adjacent objects in the background to facial emotion recognition.

Recognition of human emotion recognition using audio and visual features is also studied in some previously proposed work. Recently, Schoneveld *et al*. [11] used deep feature representations of the audio and visual modalities to improve the accuracy of the FER task. In addition, Zhou *et al*. [12] explored audio features using speech-spectrogram and Log Mel-spectrogram and evaluated facial features with different CNNs and different emotion pretrained strategies.

Proposing new CNNs is another widely used method for FER which is studied in the following researches: Hasani *et al*. [13] proposed a CNN architecture using a function with bounded derivative instead of a simple shortcut path in the residual units for automatic recognition of facial expressions. Yu *et al*. [14] proposed a multi-task framework for the global-local representation of facial expressions, where a shallow module is responsible for extracting information from local regions and the global image, and then a part-based module process the critical local regions. Shi *et al*. [15] proposed a multiple branch cross-connected convolutional neural network (MBCC-CNN) for facial expression recognition, constructed based on the Network-in-Network, and tree structure approaches to extract features from the facial image more effectively. Liu *et al*. [17] proposed a hybrid CNN including Spatial Attention CNN, designed to extract expressional features from an input face image, as well as a series of Long Short-term Memory Networks with Attention mechanism, designed for the potential use of facial landmark points for FER. Dharanya *et al*. [16] proposed Auxiliary Classifier Generative Adversarial Network (AC-GAN) based model which regenerates the basic facial emotions from an input face image and then classifies them. Zhang *et al*. [18] proposed a weakly supervised local-global attention network which is designed to extract and combine the local and the global features from input facial images. Also, their proposed architecture is designed to use the attention mechanism to deal with part location and feature fusion problems. Liu *et al*. [19] proposed a framework including a face alignment method to reduce the intra-class difference, a feature extraction module to obtain the semantic information, and a backbone model for FER.

DML for Classification in General: DML has a wide variety of application in computer vision including image classification [20], [21], [22], face recognition [23], [4], and re-identification [24], [25], visual search [26], vehicle re-identification [27], [28], [29] and so on.

For image classification task, Hoffer *et al*. [20] proposed a triplet network for image classification which is trained on triplets of data with anchor points, a positive that belongs to the same class, and a negative that belongs to a different class. In another work, Deng *et al*. [21] proposed a method to first create a feature vector for the labeled samples and then use them to classify the unlabeled samples. Zhe *et al*. [22] proposed an algorithm for learning a robust discriminative hyper-spherical feature space.

For the face recognition task, Chopra *et al*. [23] proposed the first methods for training a similarity metric from data. They proposed a discriminative loss function such that the similarity metric is small for pairs of faces from the same person and large for pairs from different persons. In another work, FaceNet [4] proposed the triplet loss which uses the triplet embedded feature vector for face verification and recognition. In addition, Additive Angular Margin Loss proposed by Deng *et al*. [30] designed to obtain highly discriminative features from the input image to improve the accuracy of the face recognition task.

For face re-identification, Ding *et al*. [24] proposed a distance-driven feature learning which tends to maximize the relative distance between the matched pair and the mismatched pair for each triplet unit. More recently, Circle loss [25] introduced a unified perspective for learning with class-level labels and pair-wise labels for face recognition, person re-identification.

For vehicle re-identification application, Deep Relative Distance Learning is proposed by Liu *et al*. [27] in which a two-branch CNN is used to project raw vehicle images

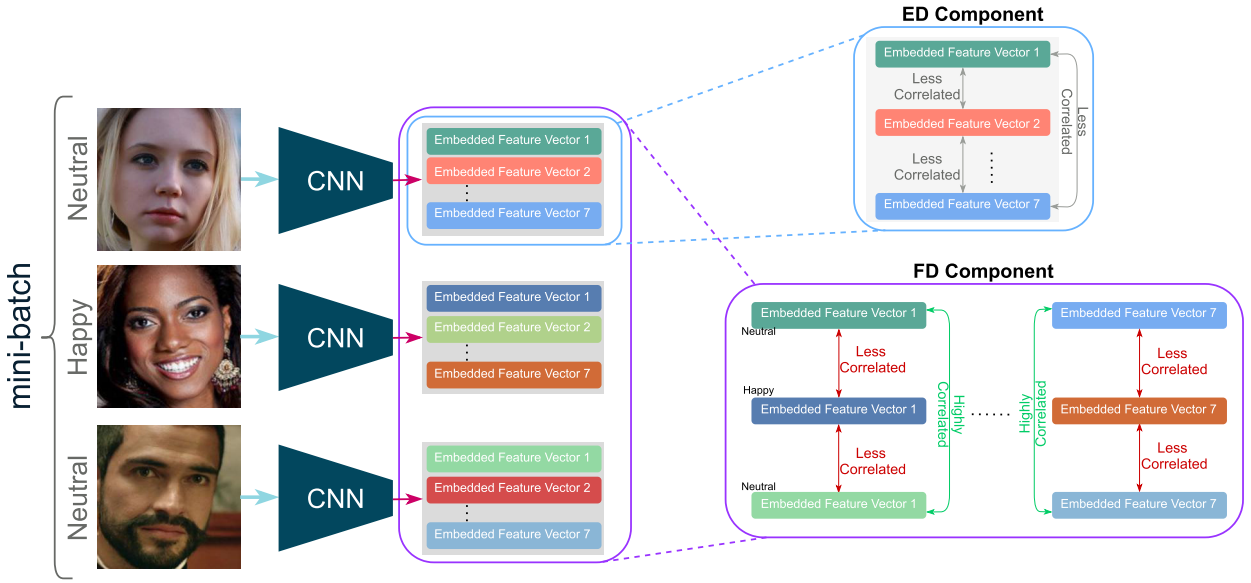

FIGURE 2. **FD component:** For each pair of images in a mini-batch, the FD component of the Ad-Corre Loss penalize the network to generate the corresponding embedded feature vectors such that they are highly correlated if belonging to a similar emotion class, and less correlated if belonging to different emotion classes. **ED component:** ED component of Ad-Corre Loss is proposed to lead the network such that for an input image, the embedded feature vectors in the embedding space be less correlated. Thus, each embedded feature vector will decode different features of the facial input image.

into a Euclidean space measure the similarity of arbitrary two vehicles. Angular loss proposed by Wang *et al*. [29] for learning better similarity metrics by constraining the angle at the negative point of triplet triangles. VP-ReID proposed by Wei *et al*. [28] extracts robust visual descriptors by learning and fusing complementary regional and global features with multi-branch CNN.

In another work, Bell *et al*. [26] proposed a multi-domain embedding for visual search in interior design. Besides, multi-similarity loss [31] proposed the general pair weighting framework which tends to convert the DML sampling problem into a unified view of pair weighting.

DML for FER: Many of the previous DML-based work in FER have either proposed a custom loss function or an algorithm that result in better discrimination of the embedded feature vectors.

The following researches have proposed custom loss functions: Wen *et al*. [32] proposed Center loss which simultaneously learns a center for embedded feature vectors of each class and penalizes the distances between the embedded feature vectors and their corresponding class centers. The Island loss [33] is also proposed to reduce the intra-class variations while enlarging the inter-class difference by maximizing the cosine distance between the class centers in the embedding space. Separate loss *et al*. [34] which is a cosine version of both center and Island loss functions. While the intra-loss maximizes the cosine similarity between the features belonging to a class, the inter-loss minimize the cosine similarity between the class centers in the embedding space. Meng *et al*. [35] proposed an identity-aware CNN as well as an identity-sensitive contrastive loss which learns identity-related information to alleviate variations that are introduced by personal attributes and achieve better FER performance.

Besides, the following researches tends to provide algorithms to improve the discriminative power of the embedded feature vectors: Liu *et al*. [36] combined the deep metric loss and softmax loss using a unified two fully connected layer with joint optimization to improve the performance of FER. A multi-scale CNN with an attention mechanism is proposed by Li *et al*. [37] to learn the importance of different convolutional receptive fields using both softmax loss and a regularized version of the center loss [32] to discriminate features in the embedding space. Li *et al*. [38] proposed Deep Locality-Preserving CNN being trained using Locality-Preserving loss to enforce the intra-class compactness by locally clustering deep features using the k-nearest neighbor algorithm. Discriminant distribution-agnostic loss [39] enforces the inter-class dissimilarity which can be useful while dealing with extremely imbalanced datasets. Hayale *et al*. [40] proposed an algorithm for automated FER to preserve the local structure of images in the embedding similarity space.

Despite achieving good accuracy, the majority of these work need the selection of sample pairs either online or offline which requires extra work and process. Moreover, they are sensitive to the pair selection process as well, while our proposed method is capable of coping well with this issue.

III. METHODOLOGY

In this section, we first explain the network architecture. Then, we describe our proposed Ad-Corre Loss and its main FD, MD, and ED components. Moreover, we explain how

each component of AD-Corre loss contributes to improving the accuracy of FER.

A. NETWORK ARCHITECTURE

We use Xception [41], an efficient CNN, as the backbone of our network architecture. After the Global Average Pooling layer, we introduce an embedding feature space containing k independent embedded feature vectors. Then, we concatenate all the embedded feature vectors followed by a dropout layer to prevent the model from over-fitting. Finally, we use a fully Connected layer with softmax activation to generate the class probabilities.

Contrary to the majority of the previously proposed CNNs, we use more than one embedded feature vector to improve the capacity of the CNN as well as the accuracy of FER (in Sec IV-D we study the effect of using multiple embedded feature vectors). Although using more embedded feature vectors results in improvement of the accuracy, it increases the number of parameters of the model too. We empirically define k, the number of the embedded feature vector in the embedding space, the same as the number of the expression classes to put a trade-off between the accuracy and the efficiency of the model.

Similar to FaceNet [4], we normalize each embedded feature vector using the L2 normalization method with a size of 256. The goal of using more than one embedded feature vector is to force each vector to extract different features from the input image. To achieve this goal, we propose the ED component (see Sec. III-B3) which penalizes the model to generate the embedded feature vectors with fewer similarities for an input image.

B. AD-CORRE LOSS

The correlation of two d-dimensional random variables (defined in $[-1, +1]$ range) is a measure that indicates the joint variability between them. Specifically, positive covariance tends to show similar behavior between two random variables, while a negative covariance indicates the opposite behavior. According to Eq. 1, the correlation between the $X_{d\times 1}$ and $Y_{d\times 1}$ is 1 when X and Y are identical and is -1 when they are uncorrelated. The $\bar{x}$ and $\bar{y}$ are the mean of the X and Y vectors, respectively.

$$\mathrm{COR}(X, Y) = \frac{\sum_{k=1}^{d}(X_k - \bar{x})(Y_k - \bar{y})}{\sqrt{(\sum_{k=1}^{d} X_k - \bar{x})(\sum_{k=1}^{d} Y_k - \bar{y})}} \tag{1}$$

Thus, the correlation matrix between n number of d-dimensional random variables represents the joint variability between each possible $n \times n$ pairs. In other words, we define $\mathrm{CORM}_{n\times n}$ as the correlation matrix between n numbers of d dimensional random variables such that $\mathrm{CORM}[i, j]_{n\times n}$ for $i, j \in \{0, 1, \ldots, n\}$ indicates the joint variability between the i^{th} and the j^{th} variables if $i \neq j$, and the variance of the i^{th} variable if $i = j$. More specifically, if we define a set of n numbers of k dimensional variables as $\mathrm{V}_{n\times d} = \{\mathrm{V}_1, \mathrm{V}_2, \ldots, \mathrm{V}_n\}$, we can define the correlation matrix according to Eq. 2:

$$\mathrm{CORM}_{n\times n} = \begin{bmatrix} \mathrm{COR}(\mathrm{V}_1, \mathrm{V}_1) & \ldots & \mathrm{COR}(\mathrm{V}_1, \mathrm{V}_n) \\ \vdots & \ddots & \vdots \\ \mathrm{COR}(\mathrm{V}_n, \mathrm{V}_1) & \ldots & \mathrm{COR}(\mathrm{V}_n, \mathrm{V}_n) \end{bmatrix} \tag{2}$$

Using the correlation matrix, it is possible to compare the similarity between each pair within a mini-batch. We use this characteristic to introduce each component of our proposed Ad-Corre Loss. For the FD component, we use the correlation matrix to compare the correlation between the embedded feature vectors within a mini-batch. Likewise, for the MD component, we use the correlation matrix to measure the correlation between the mean vectors of each expression class. For ED, for each input image, we use the correlation matrix to measure the correlation between the embedded feature vectors in the embedding space.

1) FD COMPONENT OF AD-CORRE LOSS

Inspired by the definition of correlation matrix (see Equations 1, 2), we propose FD component of Ad-Corre Loss. Assume we define our mini-batch to have size n. It is presented in Fig. 2, for a mini-batch of n input images, there will be k different embedded feature vectors in the embedding space. We define the set of class labels within a mini-batch as $\mathrm{Labels} = \{l_1, l_2, \ldots, l_n\}$, where $l_i \in \{0, 1, \ldots, 6\}$ (we only consider 7 human expressions in the context of this paper) represent the facial expression for the i^{th} image. Then, we define the $\mathrm{npSign}(l_i, l_j)$ function in Eq. 3:

$$\mathrm{npSign}(l_i, l_j) = \begin{cases} +1 & \text{If } l_i = l_j \\ -1 & \text{otherwise} \end{cases} \tag{3}$$

Then, we define the $\Phi_{n\times n}$ in Eq. 4:

$$\Phi_{n\times n} = \begin{bmatrix} \mathrm{npSign}(l_1, l_1) & \ldots & \mathrm{npSign}(l_1, l_n) \\ \vdots & \ddots & \vdots \\ \mathrm{npSign}(l_n, l_1) & \ldots & \mathrm{npSign}(l_n, l_n) \end{bmatrix} \tag{4}$$

to illustrate, for any possible pair of the embedded feature vectors within a mini-batch, if the selected pair belongs to the same human expression classes, the corresponding item in $\Phi_{n\times n}$ matrix will be $+1$, and -1 if the pair belongs to different expression classes. Now, based on the definition of the correlation matrix in Eq.2, we define the *Naive* version of the FD (NFD) component in Equations 5, 6:

$$\mathrm{Loss}_{\mathrm{NFD}} = \frac{1}{kn^2}\sum_{l=0}^{k}\sum_{i=0}^{n}\sum_{j=0}^{n} \beta[i, j]\,|\Phi[i, j] - \mathrm{CORM}_l[i, j]| \tag{5}$$

$$\beta_{n\times n} = 1_{n\times n} - I_{n\times n} \tag{6}$$

where $1_{n\times n}$ is a matrix where all the elements are 1, $I_{n\times n}$ is the *identity* matrix, and CORM_l represents the correlation matrix of the l^{th} embedded feature vector (as mentioned in Sec. III-A, we define k as 7). Besides, $\beta_{n\times n}$ is defined to set the diagonal of the correlation matrix to zero since such

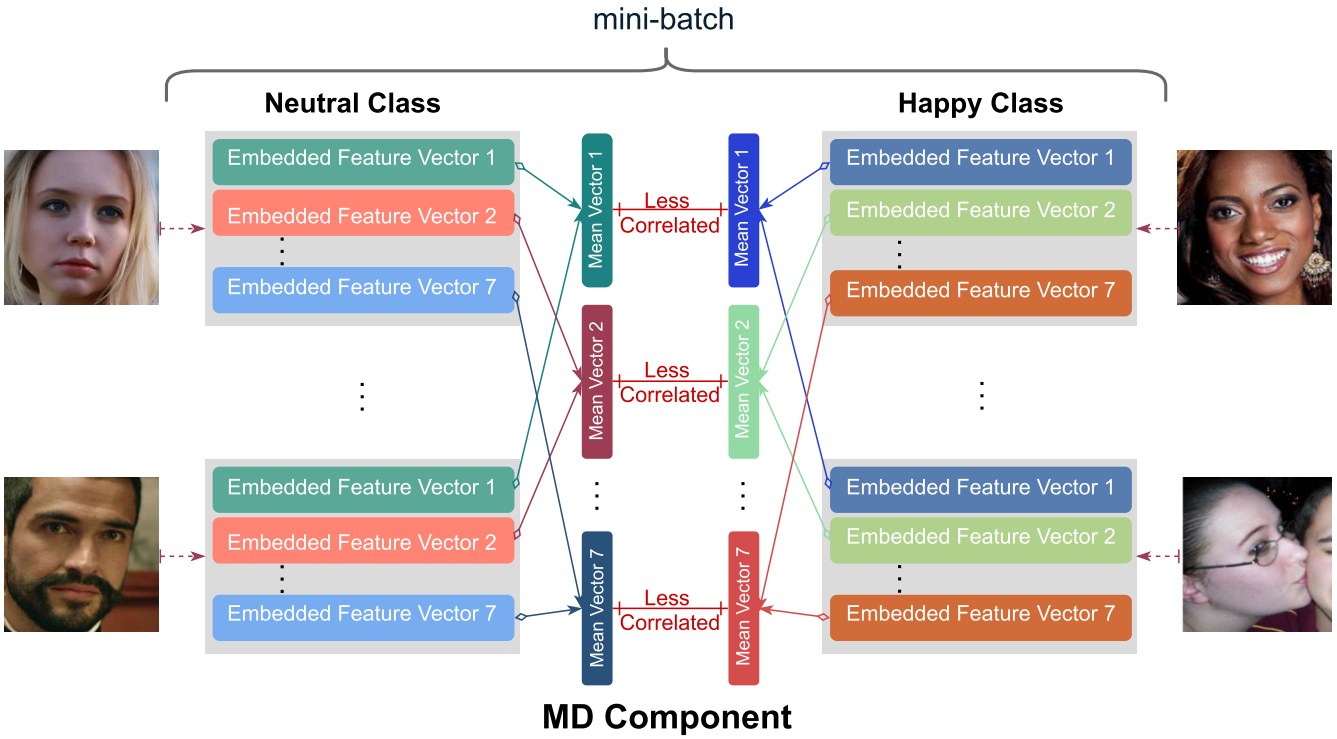

FIGURE 3. We propose MD component of Ad-Corre Loss to calculate the mean of the embedded feature vectors with respect to their expression class ($C_1, \ldots, C_k$), and lead the network to make this mean vectors to be dissimilar.

values represent the variances. We calculate the correlation loss for the k embedded feature vectors. NFD component penalizes the model to generate embedded feature vectors that are highly correlated if they belong to a similar class, and uncorrelated if they belong to different classes.

While the *triplet*-based loss function [4] families are sensitive to the process of selecting *anchor*, *positive*, and *negative* triplets, the proposed NFD component takes the advantages of all the samples within a mini-batch. Such a characteristic makes the NFD component to be tolerant to mislabeled, and low-quality images, which can be selected mistakenly as the *Hard-Negative* pairs.

Adaptive Attention Map: To further improve the accuracy of the model, we define Adaptive Attention Map, which penalizes the model more, for more frequent misclassified expressions compared to less frequent ones. In other words, if the model tends to misclassify *disgust* for 30% of the input images, while such rate for *happy* is 10%, our proposed Adaptive Attention Map will penalize the model much more for the former expression class. Consequently, we define a hyper-parameter called δ as a factor of the mini-batch size. Then, for δ most recent iterations, we save both the ground-truth and the predicted labels during the training process. Using the predicted and the ground-truth labels, we calculate the training confusion matrix -where rows are ground truth labels and accordingly it is normalized by rows- and call it $\text{CONFM}_{7\times7}$. The value of δ affects the confusion matrix update speed. Likewise, for a small δ, the confusion matrix becomes so sensitive to the mini-batch, while for the bigger δ, it keeps many historical samples and might not represent the current accuracy of the model. We empirically choose δ as $5\times n$ to make the confusion matrix represent an updated status of the model while being robust to the randomly selected mini-batch. We define the $w(l_i)$ function in Eq.7:

$$w(l_i) = 1 - \text{CONFM}[l_i, l_i] + \epsilon \tag{7}$$

where $l_i \in \{0, 1, \ldots, 6\}$ is the facial expression label for the i^{th} image in the current mini-batch, and $\epsilon = 10e^{-7}$. To illustrate, the main diagonal of the $\text{CONFM}_{7\times7}$ shows the accuracy (in percentage) of the corresponding classes. Accordingly, $w(l_i) \in [\epsilon, 1+\epsilon]$ will assign a greater weight to the expressions for which the model performs less accurately, while the weight is smaller for the expressions that model performs more accurately. We use ϵ to make sure that the output of Eq.7 is greater than zero. Afterward, we define the Adaptive Attention Map and call it $\Omega_{n\times n}$ in Eq. 8:

$$\Omega_{n\times n} = \begin{bmatrix} w(l_1)+w(l_1) & \ldots & w(l_1)+w(l_n) \\ \vdots & \ddots & \vdots \\ w(l_n)+w(l_1) & \ldots & w(l_n)+w(l_n) \end{bmatrix} \tag{8}$$

$\Omega_{n\times n}$ defines the attention map for NFD component. To be more detailed, for any possible i^{th} and j^{th} pair in the mini-batch, $\Omega[i, j]$ considers the accuracy of the model for the corresponding class of both i, and j and accordingly calculates the weight based on Eq. 7. We define FD component of

Ad-Corre Loss in Eq. 9:

$$\text{Loss}_{\text{FD}} = \frac{1}{kn^2}\sum_{l=0}^{k}\sum_{i=0}^{n}\sum_{j=0}^{n} \beta[i,j]\ \Omega[i,j]\ |\Phi[i,j] - \text{CORM}_l[i,j]| \quad (9)$$

FD component is proposed to improve the discriminative power of the embedded feature vectors that are defined in the embedding feature space. Moreover, monitoring the performance of the network during the training process and updating the proposed Adaptive Attention Map forces the model to perform accurately for dramatically imbalanced datasets. We further (see Sec. IV-D) show the effect of the Adaptive Attention Map in improving the accuracy.

2) MD COMPONENT OF AD-CORRE LOSS

We propose MD component in the Ad-Corre Loss to make the network generate the embedded feature vectors such that the means of different classes are uncorrelated. As Fig. 3 shows, MD component is calculated for each of the k embedded feature vectors in the embedding space. First, we calculate the mean of the embedded feature vectors having the same class within a mini-batch and call it $\text{M\_Set}_{k\times m} = \{\text{m\_efv}_1, \text{m\_efv}_2, \ldots, \text{m\_efv}_k\}$, where k is the number of the human facial expressions (which is 7), and m is the size of the embedded feature vectors (which is 256). Then, using the definition of the correlation matrix in Eq. 2, we define the $\text{MeanM}_{k\times k}$ as the correlation matrix of the mean set, M_Set. We define Loss_{MD} in Eq. 10:

$$\text{Loss}_{\text{MD}} = \frac{1}{k^3}\sum_{l=0}^{k}\sum_{i=0}^{k}\sum_{j=0}^{k} (1_{k\times k}[i,j] - I_{k\times k}[i,j])\ \ |1 + \text{MeanM}_k[i,j]| \quad (10)$$

where MeanM_l is the correlation matrix of the mean set of the l^{th} embedded feature vector. Besides, the $(1_{k\times k}[i,j]$ term set diagonal to be zero as they are the variances, and the $|1 + \text{MeanM}_k[i,j]|$ term guides the network to generate the embedding since the correlation of the mean feature vectors with different labels be as close to 1, which indicates they are less correlated. We further show in Sec.IV-D that the MD component of Ad-Corre Loss results in better discrimination of the generated embedded feature vectors in the embedding space.

3) ED COMPONENT OF AD-CORRE LOSS

As Fig. 2 shows, our proposed architecture contains k different embedded feature vectors being defined in a so-called embedding space. Particularly, we define k the same as the numbers of the facial expressions (which is 7 in this paper), and thus for an input image, the model creates 7 independent embedded feature vectors. Since there is no relation between the embedded feature vectors generated for and input image I, there is no guarantee that the different embedded feature vectors represent different features of I. Consequently, we proposed the ED component of Ad-Corre Loss to force the model to generate the embedded feature vectors which are as less correlated to each other as possible. In other words, the ED component is designed to guide the model to generate the embedded feature vectors representing different features from an input image.

We define the embedding space as $ES = \{EFV_1, \ldots, EFV_k\}$, where EFV_i is i^{th} embedded feature vector. Followed by the definition of the correlation matrix in Eq. 2, we define the *embedding correlation matrix* called $\text{EmbM}_{k\times k}$ such that $\text{EmbM}[i,j]$ shows the correlation between the EFV_i and EFV_j. We define ED component, Loss_{ED} in Eq. 11:

$$\text{Loss}_{\text{ED}} = \frac{1}{nk^2}\sum_{l=0}^{n}\sum_{i=0}^{k}\sum_{j=0}^{k} (1_{k\times k}[i,j] - I_{k\times k}[i,j])\ \ |1 + \text{EmbM}_l[i,j]| \quad (11)$$

where EmbM_l represents the embedding correlation matrix of the l^{th} sample in the mini-batch. In addition, $1_{k\times k}$ is a matrix with all elements as 1, and $I_{k\times k}$ is the *identity* matrix.

In the best scenario, if all the embedded feature vectors in the embedding space are uncorrelated, all the elements of EmbM are -1. Accordingly, the term $|1 + \text{EMB}_{k\times k}[i,j]|$ in Eq. 11 will be zero (consider that we use the first term, $1_{k\times k}[i,j] - I_{k\times k}[i,j]$), to set the diagonal to zero since they represent the variance). In contrast, the highest amount of the ED component is when all the embedded feature vectors are identical, such that all elements of EmbM are 1.

4) AD-CORRE LOSS

We use the CE loss function for classification. In addition, to improve the model accuracy specifically to cope with the intra-class variation as well as inter-class similarities of the facial expressions, we train our proposed model using CE jointly with our proposed Ad-Corre Loss, ($\text{Loss}_{\text{Ad-Corre}}$) in Eq.12:

$$\text{Loss}_{\text{Ad-Corre}} = \text{CE} + \lambda\ (\text{Loss}_{\text{FD}} + \text{Loss}_{\text{MD}} + \text{Loss}_{\text{ED}}) \quad (12)$$

where Loss_{FD}, Loss_{MD}, and Loss_{ED} are the three components of the Ad-Corre Loss proposed to improve the discriminative power of the model by forcing it to generate the embedded feature vectors which are similar for the images having the same expression, and dissimilar for the images with different expressions. Besides, the value of the hyper-parameter λ can dramatically affect the accuracy of the model. We further investigate the effect of λ in Sec IV-D and accordingly set it as 0.5. Likewise, in Sec IV-D, we show that the Ad-Corre Loss performs much better than the baseline CE loss function.

IV. EXPERIMENTAL RESULTS

In this section, we first explain the datasets used in our experiments. We then present the implementation details of the model. Afterward, we compare the accuracy of our proposed model trained using Ad-Corre Loss (see Eq. 12) with

the state-of-the-art methods. Finally, we present a detailed ablation study.

A. DATASET

We conducted our experiments using FER-2013 [42], RAF-DB [43] and AffectNet [44] that are the widely-used wild FER datasets. Since the images in such datasets contain broad diversity across age, gender, pose, image quality, and illumination, the models require to be much more robust compared to lab-controlled datasets.

TABLE 1. Number of annotated images for each expression on the studied databases.

Expressions	FER-2013	RAF-DB	AffectNet
Neutral	6,198	3204	80,276
Happy	8,989	5957	146,198
Sad	6,077	2469	29,487
Surprise	4,002	1619	16,288
Fear	5,121	355	8,191
Disgust	547	877	5,264
Anger	4,953	867	28,130

AffectNet [44] is by far the largest publicly available wild FER dataset that provides both categorical and Valence-Arousal annotations. It contains 450,000 facial images that are manually annotated with eight basic expression labels and gathered from the Internet by querying expression-related keywords in three search engines. Following many state-of-the-art FER methods, we exclude the contempt expression in our experiments. AffectNet [44] has an imbalanced test set, a balanced validation set, and an imbalanced training set. As the test set is not released by the authors, we report accuracy on the validation set where each category contains 500 samples. Plus, following many state-of-the-art FER methods, we exclude the contempt expression in our experiments.

FER-2013 [42] contains 28,709 training images, 3589 validation images, and 3589 test images, annotated with six basic human facial expressions as well as neutral. Following the other researches, we report our accuracy on the combination of validation and test set. FER-2013 has an imbalanced test set, validation set, and training set.

RAF-DB [43] contains 29,672 facial images that are annotated with basic or compound expressions by 40 trained human annotators.. The dataset has two parts: the single-label subset (basic emotions) and the two-tab subset (compound emotions). We only use images with 6 basic emotions as well as neutral, including 12,271 images as training set and 3,068 images as testing set. Both training and test sets are imbalanced in RAF-DB [43].

B. IMPLEMENTATION DETAILS

For the training set in each dataset, we cropped all the images and extracted the face region according to the provided bounding boxes. Then the facial images are scaled to 224×224 pixels. We augmented the images in terms of rotation (from -45 to 45 degrees), crop, contrast, and brightness to add robustness to the network. We used the Adam optimizer [57] for training the networks with a learning rate of 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $decay = 10^{-5}$. We then trained the networks for about 20 epochs for AffectNet [44] and 60 epochs for FER-2013 [42], and RAF-DB [43] with a batch size of 60. We implemented our networks using the TensorFlow library and ran them on an NVidia 1080Ti GPU. The code and all the pretrained models are available on Github.

Since FD and MD components of Ad-Corre Loss are calculated within a mini-batch, we need to investigate if the batch size can affect the final performance of the model. Moreover, the FD and the MD component are not able to function as expected if we define the batch size very small number such as 1. To deal with this issue, we propose to use the virtual-batch technique. In this technique, we define a virtual-batch size as b multiplied by the size of the real batch. Then, we save the embedded feature vectors b times, and then calculate the Ad-Corre Loss and perform the backpropagation. In our experiments, we define b to be 1, 5, and 10 which result in the virtual-batch size being 60, 300, and 600, and the final changes in the performance of the network were negligible. However, this technique can be utilized when the batch size is very small due to the available GPU memory.

C. CLASSIFICATION RESULTS AND COMPARISON

Tables 2, 3, 4 present our results and compare them with several state-of-the-art methods on AffectNet [44], RAF-DB [43] and FER-2013 [42], respectively. On AffectNet [44], our proposed model achieves a classification accuracy of 63.36% which is the state-of-the-art among the recently proposed methods.

Since the test set of RAF-DB [43] is imbalanced, we follow some of the previous work and report the average accuracy, which is the mean of diagonal values in the confusion matrix. The classification accuracy of our proposed model on this dataset is 86.96% which is comparable to the accuracy (87.03%) reported in [49]. We achieve the average accuracy of 79.01%, which is the highest among the methods that have reported this metric. On FER-2013 [42] dataset, our model achieves the classification accuracy of 72.03%, which outperforms the highest previous reported accuracy of 71.53%, using BReG-NeXt [13] by a margin of 0.5%.

Fig. 4 shows the confusion matrix obtained by our proposed model trained with Ad-Corre Loss on AffectNet [44], RAF-DB [43] and FER-2013 [42] respectively. On AffectNet [44], the most confusion occurred between *Fear* and *Surprise*, and *Disgust* and *Anger* which are 18% and 17%, respectively since both pairs are very similar. The same pattern of confusion occurs on RAF-DB [43], where the model is mostly confused between *Fear* and *Surprise* (about 18%). The second highest confusion (about 16%) happens between *Disgust* and *Neutral* which can be caused by extremely low number of samples in training set for *Disgust*. On FER-2013 [42], we see the pattern of confusion again between *Disgust* and *Anger* (about 13%), while most confusion, about 18%, occurred between *Sad* and *Neutral*.

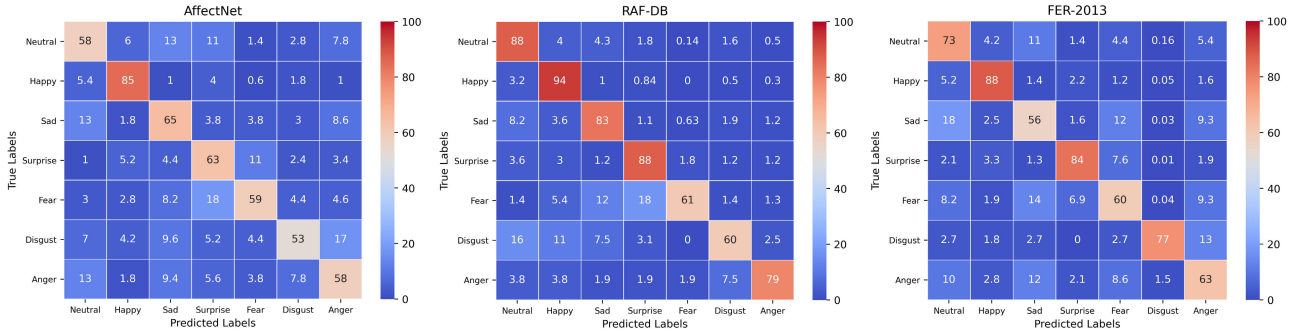

FIGURE 4. Confusion matrices of the proposed model being trained by Ad-Corre Loss on AffectNet [44], RAF-DB [43], and FER-2013 [42] datasets.

TABLE 2. Comparison to the state-of-the-art results on AffectNet [44] dataset.

Method	Upsample [44]	IR50 [12]	paCNN [45]	IPA2LT [46]	Weighted loss [44]	gaCNN [47]	separate loss [34]	RAN [48]	SCN [49]	LHF [9]	[11]	PSR [50]	Ad-Corre
Accuracy (%)	47.00	53.92	55.33	55.71	58.00	58.78	58.89	59.5	60.23	61.6	63.31	**63.77**	63.36

TABLE 3. Comparison to the state-of-the-art results on RAF-DB [43] dataset.

Method	Accuracy(%)	Average Accuracy(%)
DLP-CNN [43]	84.22	74.20
FSN [51]	81.10	72.46
paCNN [45]	83.27	-
gaCNN [47]	85.07	-
IPA2LT [46]	86.77	-
ALT [52]	84.50	76.50
separate loss [34]	86.38	77.25
WS-LGAN [53]	85.07	-
RAN [48]	86.90	-
DLN [18]	86.4	-
SCN [49]	**87.03**	-
Ad-Corre	86.96	**79.01**

TABLE 4. Comparison to the state-of-the-art results on FER-2013 [42] dataset.

Method	shao *et al.* [54]	vielzeuf *et al.* [55]	DNNRL [56]	BReG-NeXt [13]	MBCC [15]	Ad-Corre
Accuracy (%)	71.14	71.2	71.33	71.53	71.52	**72.03**

TABLE 5. Precision, Recall, and F1-score of our proposed Ad-Corre Loss on AffectNet [44] dataset.

Emotion	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger
Precision	53.10	79.77	58.78	57.14	69.57	70.32	57.93
Recall	58.20	86.00	65.60	63.20	59.00	52.60	58.40
F1-score	55.53	82.77	62.00	60.01	63.85	60.18	58.16

In addition, we report the precision, recall, and f1-score of our proposed Ad-Corre Loss on AffectNet [44], RAF-DB [43] and FER-2013 [42] datsets in Tables 5, 6, 7 respectively.

In order to show the accuracy of the proposed model trained by Ad-Corre Loss, we provide some correctly-classified as well as some misclassified samples from AffectNet (see Figures 7 and 7), RAF-DB (see Figures 7 and 7), and FER-2013(see Figures 7 and 7). By reviewing the depicted samples, it can be figured out that the model performs accurately in many challenging cases with extreme head pose and image illumination, and brightness.

To compare the discriminative power of Ad-Corre and CE loss, We use t-SNE [58] to visualize the embedded feature vectors. As Fig. 5 shows, while the feature vectors generated using CE loss are not easily distinguishable for different facial expressions, the feature vectors generated using Ad-Corre Loss are more disentangled from each other. Training model with Ad-Corre Loss resulted in a more compact cluster of the embedded feature vectors compared to the clusters generated training the model using CE loss. Moreover, the distance between the clusters using Ad-Corre Loss is much further from each other compared to the clusters created using CE loss. Thus, we can observe that the Ad-Corre Loss can increase the compactness of the model, by successfully decreasing the intra-class difference. Moreover, Ad-Corre Loss enhances inter-class separability which leads to the discriminative power of the model.

TABLE 6. Precision, Recall, and F1-score of our proposed Ad-Corre Loss on RAF-DB [43] dataset.

Emotion	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger
Precision	83.14	93.15	84.93	85.49	77.19	70.45	84.45
Recall	87.74	93.95	83.65	88.16	60.27	59.23	79.61
F1-score	85.38	93.55	84.29	86.80	67.69	64.35	81.96

TABLE 7. Precision, Recall, and F1-score of our proposed Ad-Corre Loss on FER-2013 [42] dataset.

Emotion	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger
Precision	63.50	90.75	61.96	80.42	61.91	75.67	63.77
Recall	73.22	88.09	56.09	83.65	59.46	77.06	62.82
F1-score	68.01	89.40	58.88	82.00	60.66	76.36	63.29

D. ABLATION STUDY

To investigate the effect of each proposed loss function and the proposed embedded feature vectors, we conducted five different experiments on the AffectNet [44] dataset. In the first experiment, we trained the model using CE as the loss function and call it CE. We conducted another experiment called CE + Loss_{MD}, where we trained the network using both CE loss and just MD component of Ad-Corre Loss. In the next experiment called CE + Loss_{ED}, we trained our proposed model using CE loss and just the ED component of Ad-Corre Loss. Likewise, CE + Loss_{FD} indicates using CE loss jointly with just FD component of Ad-Corre Loss.

As Table 8 shows, the accuracy of the model which is trained using only the CE loss function is 56.46%. Using CE + Loss_{MD} slightly improves the model accuracy by

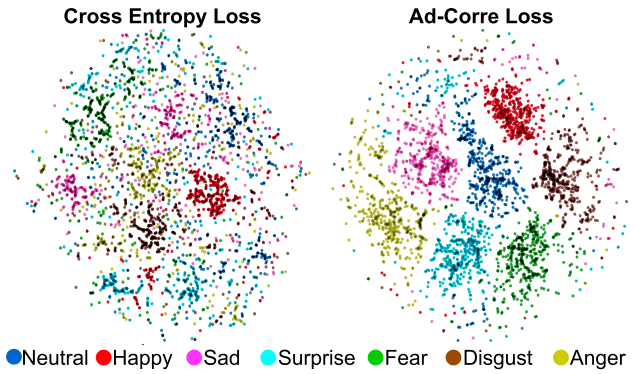

FIGURE 5. Visualization of the concatenation of the 7 embedded feature vectors using t-SNE [58] on AffectNet [44] dataset. Training the model using Ad-Corre Loss results in a more distinguishable embedded feature vectors compared to the model trained with the cross-entropy loss.

TABLE 8. Studying the Affect of the proposed loss functions on the accuracy of the model.

Model	CE	CE + Loss_{MD}	CE + Loss_{ED}	CE + Loss_{FD}	Ad-Corre
Accuracy (%)	56.46 (baseline)	57.75 (↑ 1.29%)	59.51 (↑ 3.05%)	61.86 (↑ 5.4%)	**63.36** (↑ 6.9%)

1.29% (increasing from 56.46% to 57.75%). Training the model using CE + Loss_{ED} results in more accuracy improvement from 56.46% to 59.51% (3.05%). Using CE + Loss_{FD} improves the accuracy of the model much more compared to the previous experiments by 5.4% (increasing from 56.46% to 61.86%). Finally, the accuracy of the model which is trained using Ad-Corre Loss (see Eq. 12) has improved by about 6.9% compared to the baseline model trained with just CE loss, indicating that our proposed Ad-Corre Loss is capable of improving the discriminative power of our defined embedded feature vectors which results in better classification of facial expressions.

To investigate the effect of the number of the embedded feature vectors in the embedding space, we define 2 different models called Emb_3, and Emb_10 with 3, and 10 embedded feature vectors respectively. We train Emb_3, and Emb_10 using Ad-Corre Loss. According to Table 9, the accuracy of FER using Emb_3 model is 59.88% which is around 3.48% lower than the accuracy of the baseline model, which has 7 embedded feature vectors. Moreover, the accuracy of Emb_10 is 63.56% which indicates only around 0.2% improvements compared to the baseline model. Although increasing the number of the embedded feature vectors can increase the accuracy of the model, our experiment shows that the amount of increase is very small. Moreover, the computational overhead should be taken into the account too (see Sec. IV-E and Table 11).

We defined the hyper-parameter λ in our defined Ad-Corre Loss in Eq. 12 to put a balance between the CE loss and our proposed Ad-Corre. As the final goal of the model is to classify the facial expressions of the human faces, setting the value of λ can play a crucial role in the accuracy of the model. In other words, we define λ to put a balance between the task of clustering the embedded feature vectors

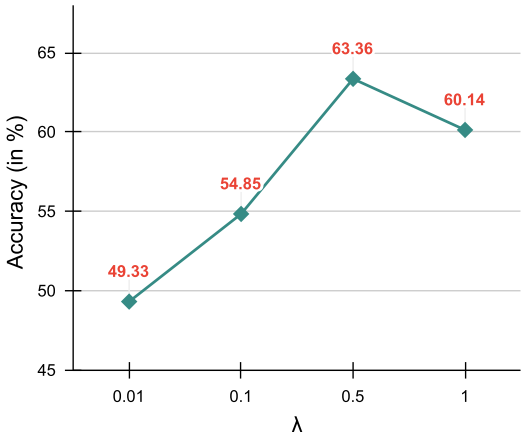

FIGURE 6. The affect of different values of the hyper-parameter λ on the accuracy of the model.

TABLE 9. Studying the Affect of the number of embedded feature vectors defined in the embedding space on the accuracy of the model.

Model	Ad-Corre (7 embedded feature vectors)	Emb_3	Emb_10
Accuracy (%)	63.36 (baseline)	59.88 (↓ 3.48%)	63.56 (↑ 0.2%)

and the classification of the facial expressions. Although better clustering of the embedded feature vectors can result in a better classification, a *poorly-chosen* λ can lead the network towards *only* learning either classification or the clustering task. We conducted 4 different experiments and set the value of λ to be 0.01, 0.1, 0.5, and 0.1 and trained the network using Ad-Corre Loss for 20 epochs on AffectNet [44] dataset.

As Fig. 6 shows, setting λ as 0.01 results in 49.33% accuracy which indicates that the model has focused more on the clustering of the embedded feature vectors. In other words, the classification task has been neglected by the model which resulted in poor accuracy. In the second experiment, we set λ to 0.1 and we see a better classification accuracy (54.85%). Following the trend, we increased the value of λ to 0.5, and 1 which result in 63.36% and 60.14% accuracy receptively. We can conclude that while setting λ as 0.5 might put a balance between the clustering task and the classification task, increasing λ to 1 leads the network towards paying much attention to the classification task which can cause poor clustering of the embedded feature vectors. Hence, we choose λ as 0.5 in our proposed Ad-Corre Loss in Eq.12.

In another experiment, we investigated if the proposed Ad-Corre Loss can be applied to other CNNs and improve the accuracy of the model. We used Resnet50 [59] as our backbone model and trained two model instances, one using the CE loss and another one using the Ad-Corre loss. We trained the model following the configuration reported in Sec.IV-B and used 7 embedded feature vectors.

As Table 10 shows, the accuracy of the model trained using the CE loss is 68.25%, 82.13%, and 55.57% on FER-2013 [42], RAF-DB [43], and AffectNet [44] dataset, respectively. Training the model using the proposed Ad-Corre Loss results in about 3.23% (from 68.25% to 71.48%),

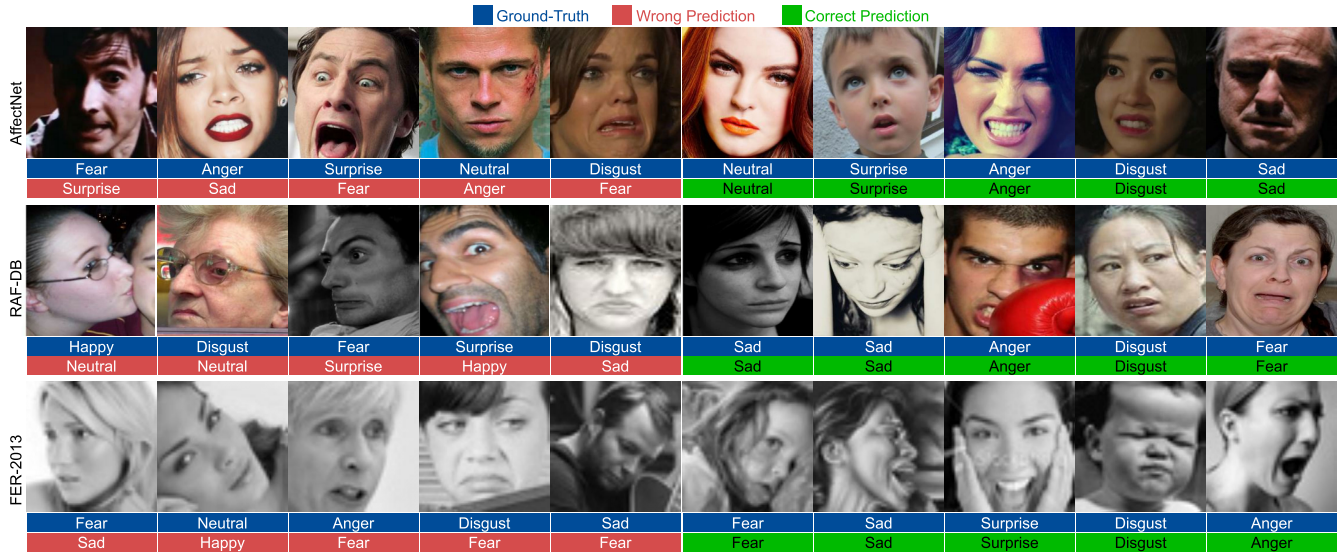

FIGURE 7. Some correctly-classified, and misclassified samples fom AffectNet [44], RAF-DB [43], and FER-2013 [42] dataset.

TABLE 10. We used Resnet50 as the backbone CNN and trained using CE loss, while we trained $\text{Resnet50}_{\text{Ad-Corre}}$ using Ad-Corre Loss. The accuracy of the latter is much higher compared to the former on FER-2013 [42], RAF-DB [43], and AffectNet [44] dataset.

Model	FER-2013	RAF-DB	AffectNet
Resnet50	68.25	82.13	55.57
$\text{Resnet50}_{\text{Ad-Corre}}$	71.48	85.93	62.78

3.8% (from 82.13% to 85.93%), and 7.21% (from 55.57% to 62.78%) accuracy improvement on FER-2013 [42], RAF-DB [43], and AffectNet [44] dataset, respectively. This experiment shows that Ad-Corre Loss is applicable to different CNN models and compared to widely used CE loss, it can improves the discriminating power of the model.

E. MODEL EFFICIENCY

In this section, we study the numbers of the parameters as well as the number of the floating-point operations (FLOPs) of our proposed model with respect to the numbers of embedded feature vectors. In Table 11, we show the number of the parameters of the official Xception [41] model which is about 20.86 million (M). Adding 1 embedded feature vector to the model (EMB_1) results in around 0.52M increase in the numbers of the model parameters (from 20.86M to 21.38M). The corresponding increase in the FLOPs is only about 0.58M (from around 4, 554, 344k to 4, 554, 919k). Adding 3 embedded feature vectors (EMB_3) results in around 1.57M increase in the model parameters (from 20.86M to 22.43M), and around 1.62M in the FLOPs compared to the those of the official Xception [41]. Using 7 embedded feature vectors (EMB_7), the model parameters increases from 20.86M to 24.54M by around 3.68M, and the FLOPs by around 3.73M compared to the Xception [41]. Despite a small increase in model size and its FLOPs, as we show in Table 8, using Ad-Corre Loss with 7 embedded feature vectors results in around 6% increase in classification accuracy.

TABLE 11. Number of the model parameters and the FLOPs of the official Xception [41] model, compared to our proposed model. We report these parameters considering different number of the embedded feature vectors.

Model	Xception [41]	Emb_1	Emb_3	Emb_7	Emb_10
#Parameters	20,861,480	21,386,281	22,437,419	24,545,839	26,132,530
#FLOPs	4,554,344,528	4,554,919,764	4,555,971,420	4,558,080,876	4,559,668,344

V. DISCUSSION

In this section, we discuss the similarities and the differences between our proposed loss functions and the two widely-used loss functions for classification, the Center Loss [32] and the Island Loss [33].

The Center Loss [32] predicts the center of each cluster and penalizes the model to generate the embedded feature vectors such that the distance between each embedded feature vector and the corresponding predicted center vector is as small as possible. In contrast, our proposed Correlation Loss calculates the distance between each pair of the embedded feature vectors within a mini-batch and penalize the model to generate the embedded feature vectors belonging to similar emotion class to be highly correlated, and those belonging to different classes to be less correlated. Hence, the model generates the embedded feature vectors within a similar class to be compact implicitly. Moreover, in Center Loss [32], there is no guarantee that the generated center vectors be far from each other, while the Correlation Loss penalizes the network to generate the embedded feature vector from different emotion classes to be less correlated from each other, and hence, the center of different emotion classes will be less correlated.

The Island Loss [33] was proposed to improve the performance of the Center Loss [32] by penalizing the network to increase the distance between the predicted center of the clusters from each other. Contrary to the Island Loss [33], our proposed Mean Loss calculates the center of each cluster and penalizes them to be as less correlated as possible. We believe calculating the center of each cluster (the Mean Loss) and

penalizing the network according to their correlation is much easier than predicting the center of each cluster and then penalizing the network to increase the distance between them.

In addition, our proposed Adaptive Attention Map puts weights on the emotion classes which are misclassified by the model more frequently, compared to the emotions that are easier to be recognized. There is no similar mechanism in either the Center Loss [32] or the Island Loss [33].

VI. CONCLUSION

In this paper, we proposed a method for facial expression recognition in the wild. We used the widely used Xception [41] and Resnet50 [59] CNN architectures as our backbone models and proposed an embedding feature space containing k different embedded feature vectors. We introduced Ad-Corre Loss which consists of FD, ED, and MD components and the CE loss. Our experiment shows that regardless of the backbone model choice, the proposed Ad-Corre Loss improves the discriminative power of the generated embedded feature vectors. Our proposed model trained using the Ad-Corre Loss achieved a very promising recognition accuracy on AffectNet [44] and RAF-DB [43], and FER-2013 [42]. Ad-Corre Loss can be easily used in other classification tasks in future work.

REFERENCES

[1] C. Darwin, *The Expression of the Emotions in Man and Animals*. Chicago, IL, USA: Univ. Chicago press, 2015.

[2] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[3] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Personality Social Psychol.*, vol. 17, no. 2, p. 124, 1971.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[5] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, "Img2pose: Face alignment and detection via 6DoF, face pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7617–7627.

[6] A. P. Fard, H. Abdollahi, and M. Mahoor, "ASMNet: A lightweight deep neural network for face alignment and pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1521–1530.

[7] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4004–4012.

[8] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 30–40.

[9] M. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019.

[10] M.-H. Hoang, S.-H. Kim, H.-J. Yang, and G.-S. Lee, "Context-aware emotion recognition based on visual relationship detection," *IEEE Access*, vol. 9, pp. 90465–90474, 2021.

[11] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognit. Lett.*, vol. 146, pp. 1–7, Jun. 2021.

[12] H. Zhou, D. Meng, Y. Zhang, X. Peng, J. Du, K. Wang, and Y. Qiao, "Exploring emotion features and fusion strategies for audio-video emotion recognition," in *Proc. Int. Conf. Multimodal Interact.*, 2019, pp. 562–566.

[13] B. Hasani, P. S. Negi, and M. Mahoor, "BReG-NeXt: Facial affect computing using adaptive residual networks with bounded gradient," *IEEE Trans. Affect. Comput.*, early access, Apr. 13, 2020, doi: 10.1109/TAFFC.2020.2986440.

[14] M. Yu, H. Zheng, Z. Peng, J. Dong, and H. Du, "Facial expression recognition based on a multi-task global-local network," *Pattern Recognit. Lett.*, vol. 131, pp. 166–171, Mar. 2020.

[15] C. Shi, C. Tan, and L. Wang, "A facial expression recognition method based on a multibranch cross-connection convolutional neural network," *IEEE Access*, vol. 9, pp. 39255–39274, 2021.

[16] D. V., A. N. Joseph Raj, and V. P. Gopi, "Facial expression recognition through person-wise regeneration of expressions using auxiliary classifier generative adversarial network (AC-GAN) based model," *J. Vis. Commun. Image Represent.*, vol. 77, May 2021, Art. no. 103110.

[17] C. Liu, K. Hirota, J. Ma, Z. Jia, and Y. Dai, "Facial expression recognition using hybrid features of pixel and geometry," *IEEE Access*, vol. 9, pp. 18876–18889, 2021.

[18] W. Zhang, X. Ji, K. Chen, Y. Ding, and C. Fan, "Learning a facial expression embedding disentangled from identity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6759–6768.

[19] J. Liu, Y. Feng, and H. Wang, "Facial expression recognition using pose-guided face alignment and discriminative features based on deep learning," *IEEE Access*, vol. 9, pp. 69267–69277, 2021.

[20] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition*, A. Feragen, M. Pelillo, and M. Loog, Eds. Cham, Switzerland: Springer, 2015, pp. 84–92.

[21] B. Deng, S. Jia, and D. Shi, "Deep metric learning-based feature embedding for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1422–1435, Feb. 2020.

[22] X. Zhe, S. Chen, and H. Yan, "Directional statistics-based deep metric learning for image classification and retrieval," *Pattern Recognit.*, vol. 93, pp. 113–123, Sep. 2019.

[23] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, vol. 1, Jun. 2005, pp. 539–546.

[24] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, Oct. 2015.

[25] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6398–6407.

[26] S. Bell and K. Bala, "Learning visual similarity for product design with convolutional neural networks," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 1–10, Jul. 2015.

[27] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2167–2175.

[28] L. Wei, X. Liu, J. Li, and S. Zhang, "VP-ReID: Vehicle and person re-identification system," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2018, pp. 501–504.

[29] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2593–2601.

[30] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4690–4699.

[31] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5022–5030.

[32] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 499–515.

[33] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 302–309.

[34] Y. Li, Y. Lu, J. Li, and G. Lu, "Separate loss for basic and compound facial expression recognition in the wild," in *Proc. Asian Conf. Mach. Learn.*, 2019, pp. 897–911.

[35] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Proc. 12th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, May 2017, pp. 558–565.

[36] X. Liu, B. V. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. workshops*, Jul. 2017, pp. 20–29.

[37] Z. Li, S. Wu, and G. Xiao, "Facial expression recognition by multi-scale CNN with regularized center loss," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3384–3389.
[38] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.
[39] A. H. Farzaneh and X. Qi, "Discriminant distribution-agnostic loss for facial expression recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 406–407.
[40] W. Hayale, P. S. Negi, and M. Mahoor, "Deep Siamese neural networks for facial expression recognition in the wild," *IEEE Trans. Affect. Comput.*, early access, May 4, 2021, doi: 10.1109/TAFFC.2021.3077248.
[41] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
[42] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing*. Berlin, Germany: Springer, 2013, pp. 117–124.
[43] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2852–2861.
[44] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan./Mar. 2017.
[45] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-gated CNN for occlusion-aware facial expression recognition," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2209–2214.
[46] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 222–237.
[47] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.
[48] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.
[49] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 6897–6906.
[50] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131988–132001, 2020.
[51] S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen, "Feature selection mechanism in CNNs for facial expression recognition," in *Proc. BMVC*, 2018, p. 317.
[52] C. Florea, L. Florea, M.-S. Badea, C. Vertan, and A. Racoviteanu, "Annealed label transfer for face expression recognition," in *Proc. BMVC*, 2019, p. 104.
[53] H. Zhang, W. Su, and Z. Wang, "Weakly supervised local-global attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 37976–37987, 2020.
[54] J. Shao and Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild," *Neurocompting*, vol. 355, pp. 82–92, Aug. 2019.
[55] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 569–576.
[56] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, and D. Tao, "Deep neural networks with relativity learning for facial expression recognition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2016, pp. 1–6.
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[58] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

ALI POURRAMEZAN FARD received the M.Sc. degree in computer engineering from the Iran University of Science and Technology, Tehran, Iran, in 2015. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the Department of Electrical and Computer Engineering, University of Denver. He is also a Graduate Teaching Assistant with the Department of Electrical and Computer Engineering, University of Denver. His research interests include computer vision, machine learning, and deep neural networks, especially on face alignment, and facial expression analysis.

MOHAMMAD H. MAHOOR (Senior Member, IEEE) received the M.S. degree in biomedical engineering from the Sharif University of Technology, Iran, in 1998, and the Ph.D. degree in electrical and computer engineering from the University of Miami, Coral Gables, FL, USA, in 2007. He is currently a Professor of electrical and computer engineering at the University of Denver. He does research in the area of computer vision and machine learning, including visual object recognition, object tracking, affective computing, and human–robot interaction (HRI), such as humanoid social robots for interaction and intervention of children with autism and older adults with depression and dementia. He has received over $7M in research funding from state and federal agencies, including the National Science Foundation and the National Institute of Health. He has published over 158 conferences and journal papers.