Triplet loss with multistage outlier suppression and class-pair margins for facial expression recognition

Weicheng Xie, Haoqian Wu, Yi Tian, Mengchao Bai, Linlin Shen*

***Abstract*—Deep metric based triplet loss has been widely used to enhance inter-class separability and intra-class compactness of network features. However, the margin parameters in the triplet loss for current approaches are usually fixed and not adaptive to the variations among different expression pairs. Meanwhile, outlier samples like faces with confusing expressions, occlusion and large head poses may be introduced during the selection of the hard triplets, which may deteriorate the generalization performance of the learned features for normal testing samples. In this work, a new triplet loss based on class-pair margins and multistage outlier suppression is proposed for facial expression recognition (FER). In this approach, each expression pair is assigned with an order-insensitive or two order-aware adaptive margin parameters. While expression samples with large head poses or occlusion are firstly detected and excluded, abnormal hard triplets are discarded if their feature distances do not fit the model of normal feature distance distribution. Extensive experiments on seven public benchmark expression databases show that the network using the proposed loss achieves much better accuracy than that using the original triplet loss and the network without using the proposed strategies, and the most balanced performances among state-of-the-art algorithms in the literature.**

***Index Terms*—Facial expression recognition, deep metric learning, hard triplet selection, multistage outlier suppression, adaptive class-pair margin.**

I. Introduction

THE softmax loss is widely employed in convolutional neural networks (CNNs) to measure the difference between the network output and supervision signal [1], while the triplet loss [2] and a number of variants have been proposed to further enforce intra-class compactness and inter-class separability of the learned features. The quadruplet loss [3] motivated from the triplet loss, is proposed to further enlarge inter-class distance. To boost the performance of the triplet loss, the triplet selection and the margin parameter adjustment get a lot of attention.

In order to minimize the intra-class distance and maximize the inter-class distance during the network optimization, the hardest sample triplet is considered. Actually, these triplet losses help the network learn more sufficiently from the difficult samples in the hardest triplets. Song et al. [4] selected the hardest sample pair, i.e. the sample pairs with maximum intra-class distance or minimum inter-class distance, in the training samples. It was stated that the mining of hard triplets [5] and outliers [6] is beneficial for improving the performance of original triplet loss. The hardest sample pair is selected based on the identity information [7], while the distance from the anchor sample is used as the metric to select the hardest triplet. Wu et al. [8] further used the distance distribution to select the hard triplet. The triplet with the most uniform sample-pair distance is deemed as the hardest one. Yu et al. [9] converted the selection of hard samples into a problem of sample weighting with a hard-aware loss to assign bigger weights to harder samples. Zhou et al. [10] proposed an improved triplet loss with auxiliary class centers of hard samples to consistently minimize the intra-class distance in the training process. To learn more discriminative features from visually similar classes, Ge et al. [11] introduced a new violate margin based on the hierarchical tree to automatically select meaningful hard samples with the guide of global context. Instead of using explicit distance as the metric, the implicit feature embedding is learned adaptively to obtain the distribution shift for triplet selection [12], [13]. While the negative samples are further divided into three groups, i.e. easy, semi-hard and hard samples [14], Sohn [15] suggested to select the negative examples that interact with each other to improve the convergence by a multi-class N-pair loss.

In order to dynamically adapt the triplet loss according to the running condition, the setting of the loss parameter, i.e. the margin parameter, is often considered. The margin parameter was introduced by Hadsell et al. [16] to filter out relatively hard samples for training with the contrastive loss, whose self-adaptive model has drawn lots of attentions. Wang et al. [17] introduced an adaptive margin parameter in listwise loss to assign larger margins to harder negative samples. To dynamically update the margin parameter, Li et al. [18] used the correlation between the inter-class distances of the projected image features and the semantic representations, Wang et al. [19] and Chen et al. [3] used the inter-class and intra-class distances. Chen and Deng [20] introduced an adaptive large margin constraint to convert a fixed margin into a local-adaptive angular margin.

A. Related Works

For facial expression recognition (FER) with deep metrics, Zhang et al. [21] proposed an identity loss in an auxiliary task to supervise the training process, as well as enhance robustness of the main task without stopping in the whole training stage.

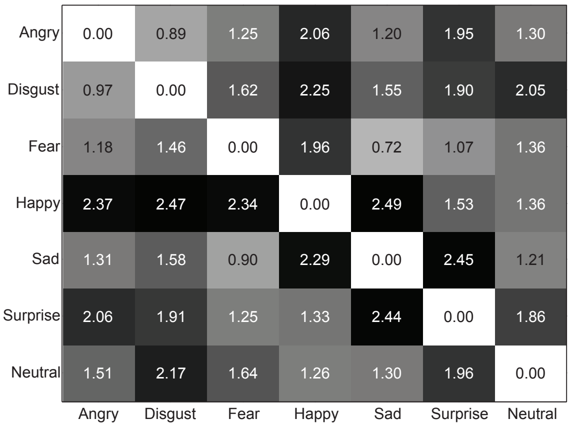
The authors are with Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen, 518060, China. W. Xie was also a visiting research fellow in School of Computer Science, University of Nottingham, Nottingham, UK. E-mail: wcxie@szu.edu.cn; 1910272008@email.szu.edu.cn; 2172272045@email.szu.edu.cn; baimengchao2017@email.szu.edu.cn; llshen@szu.edu.cn; Corresponding author: Prof. Linlin Shen, Tel: 86-0755-86935089, Fax: 86-0755-26534078, llshen@szu.edu.cn

Kim et al. [22] designed the contrastive representation in the feature space of an encoder network based on contrastive metric learning and a supervised reconstruction for FER task. Triplet loss [23] and the $(N+M)$-tuplet loss [24] (N negative and M positive examples in a mini-batch) were proposed to take into account the expression variations of different classes. Li and Deng [25] surveyed the state-of-the-art works related with the deep metric losses and the corresponding variants.

Integrating the hard samples in the network training can boost the performance of the triplet loss for testing dataset by learning more discriminative features. However, in current algorithms for hard sample selection, the hard samples are assumed to be normal, i.e. they are centralized around the center of the same class and do not entangle with samples from other classes. Wu et al. [26] proposed a novel sampling to mine intra-class, i.e. positive samples with local sample distribution, to improve the deep embedding in the context of large intra-class variations. In this work, we initialize the value of the margin parameter based on the distribution of feature distance and adapt it dynamically for pairwise expressions. Meanwhile, inter-class distance is used to update the margin parameter, which is estimated with the distances between the class centers. In this way, the runtime cost of the inter-class distance computation is reduced. The robustness of the margin parameter update is improved, since the centers are updated according to all the visited training samples using global information.

To adapt the triplet loss to different running conditions, self-adaptive margin parameter is employed. However, a fixed margin is not applicable when inter-class differences appear with significantly different scale intensities [27]. For FER task, the difference between 'fear' and 'happy' expressions is more apparent than that between 'fear' and 'sad'. Dynamic margin parameter for each expression pair during network training is rarely studied. Thus, pairwise FER is introduced to reduce the influence of inconsistency of expression pair variations [28]. Motivated from the pairwise FER [28], adaptive margin parameter for each expression pair is introduced to take into account the scale inconsistency of different expression pairs.

Meanwhile, for tasks like FER, confusing samples are popular and behave very diversely among different person identities, which can easily introduce confusing samples. While hard samples are widely believed to increase the robustness of network during the softmax loss based training, it may introduce noises during triplet sample selection and thus result in poor generalization performance for normal testing expression samples. Tian et al. [29] use primary feature distance distribution to exclude outliers from the hard triplet selection. However, the largely posed or occluded faces can also be recognized as the hard samples. Full learning of these samples, i.e. putting more emphasis on these samples, may provide misguidance for frontal and non-occluded expressions. Meanwhile, the class order is not sufficiently explored for the expression-pair difference representation. In this work, multistage outlier suppression is proposed to take into account face occlusion, head pose and feature distance distribution, while class order-aware margins are designed to more accurately depict the expression-pair differences. Furthermore, the experimental evaluation of the proposed algorithm is greatly enhanced.



	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	0.00	0.89	1.25	2.06	1.20	1.95	1.30
Disgust	0.97	0.00	1.62	2.25	1.55	1.90	2.05
Fear	1.18	1.46	0.00	1.96	0.72	1.07	1.36
Happy	2.37	2.47	2.34	0.00	2.49	1.53	1.36
Sad	1.31	1.58	0.90	2.29	0.00	2.45	1.21
Surprise	2.06	1.91	1.25	1.33	2.44	0.00	1.86
Neutral	1.51	2.17	1.64	1.26	1.30	1.96	0.00

Fig. 1. The bias $\{S_{i,j}\}$ defined in equation (1) for all the expression pairs $\{(i,j)|1 \leq i,j \leq 7\}$, the network used and the experimental setting are same with Section III-A and the experiment in Table II.

B. Motivation

To study the diversity among the differences of various expression pairs, Fig. 1 shows the biases of the average inter-class and intra-class distances for 42 pairs of the seven expressions of the FER2013 database [30]. The average distance bias is defined as follows

$$S_{i,j} = \frac{1}{N_{i,j}} \sum_{x_a^{(i)}, x_p^{(i)}, x_n^{(j)}} d(f(x_n^{(j)}), f(x_a^{(i)}))^2 - d(f(x_p^{(i)}), f(x_a^{(i)}))^2, \tag{1}$$

where $x_p^{(i)}$ and $x_n^{(j)}$, i.e. the positive and negative samples, having the same (the i-th class) and different (the j-th class) class labels with the anchor sample $x_a^{(i)}$ $(x_p^{(i)} \neq x_a^{(i)})$, compose a triplet, and $N_{i,j}$ is the number of triplets in terms of the expression pair of (i,j); $d(f(x_n^{(j)}), f(x_a^{(i)})) = ||f(x_n^{(j)}) - f(x_a^{(i)})||_2$ is the L_2-norm distance; $f(x_a^{(i)})$ is the embedded feature vector, i.e. output of the fully connected (FC) layer, for the input of the anchor $(x_a^{(i)})$. Please note that the triplet $(x_a^{(i)}, x_p^{(i)}, x_n^{(j)})$ is not random and traverses all the possible sample combination, thus, $S_{i,j} \neq S_{j,i}$ reflects the bias of the inter-class and the intra-class distances.

As shown in Fig. 1, there is large variation among the biases between inter-class and intra-class distances, i.e. the largest bias (2.49) is three times the smallest bias (0.72). Since different expressions have different deform intensities and scales, a different margin parameter for each expression pair is beneficial for the recognition.

To illustrate the motivation of the outlier-suppressed triplet selection, Fig. 2 shows three abnormal 'happy' expressions (c), (d) and (e). As shown in Fig. 2, a person may present an expression significantly different from others due to the identity diversity (Fig. 2(c)), large face occlusion (Fig. 2(d)) or head pose (Fig. 2(e)), while network trained with these samples in the proposed triplet loss may result in model over-fitting and poor generalization performance on other normal

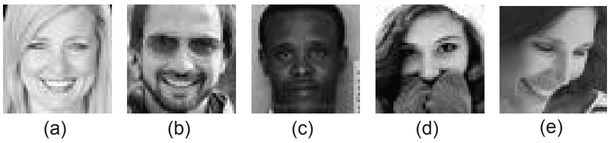


Fig. 2. The abnormal 'happy' expressions with confusing appearance (c), occlusions (d) and poses (e) in the FER2013 expression database [30].

samples. Thus, abnormal hard positive and negative samples shall be excluded to reduce the influence of outlier triplets.

C. Contribution

In this work, a new triplet loss with class-pair margins and multistage outlier suppression is proposed for FER. A different margin parameter is assigned for each expression pair to address the deform intensity variations among different expression classes, and adjusted dynamically according to the distances between class centers. Furthermore, hard outlier samples, i.e. largely posed, occluded faces and samples behaving significantly different from the same-class samples, are excluded during hard triplet selection to improve the generalization performance on the normal samples. This work makes the following contributions:

- Class-pair margins, either class order-insensitive or order-aware, are introduced to address deform intensity inconsistency among different classes, which are adaptively adjusted according to the inter-class distances;
- Outlier hard samples, i.e. largely posed, occluded faces and abnormally-offset samples, are detected in the feature distance space, and excluded from the triplet loss based network training;
- The proposed algorithm achieves competitive performance on seven public expression databases, when compared with the various triplet loss variants and the state-of-the-art approaches.

This paper is structured as follows. Section II gives a description about the proposed algorithm step by step. The experimental results of the proposed algorithm on public databases are presented in Section III. Finally, discussions and conclusions are addressed in Section IV.

II. The Proposed Algorithm

In this section, the proposed self-adaptive class-pair margins, the selection of hard triplets based on occlusion and pose detection and feature distance distribution, as well as the network training and optimization are introduced.

A. Self-adaptive Class-pair Margins

In order to boost the network discrimination ability for difficult samples, the original triplet loss is presented as follows

$$\mathcal{L}_t^{ori} = \frac{1}{2}\sum_{x_a}[d(f(x_a), f(x_p))^2 - d(f(x_a), f(x_n))^2 + \alpha]_+, \quad (2)$$

where α is the margin parameter determining the hardness of the selected samples, x_a and $f(x_a)$ are the abbreviations of

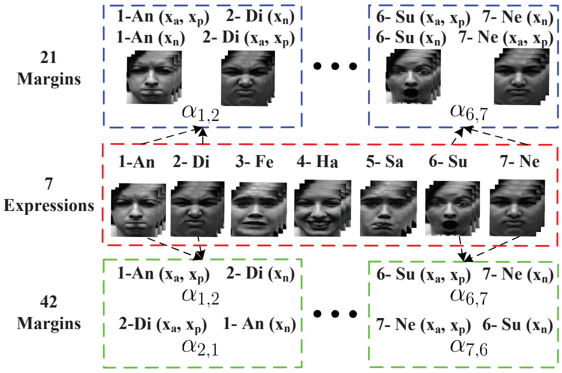

21
Margins
1-An (x_a, x_p) 2- Di (x_n)
1-An (x_n) 2- Di (x_a, x_p)
$\alpha_{1,2}$
6- Su (x_a, x_p) 7- Ne (x_n)
6- Su (x_n) 7- Ne (x_a, x_p)
$\alpha_{6,7}$
7
Expressions
1-An
2- Di
3- Fe
4- Ha
5- Sa
6- Su
7- Ne
42
Margins
1-An (x_a, x_p) 2- Di (x_n)
$\alpha_{1,2}$
2-Di (x_a, x_p) 1- An (x_n)
$\alpha_{2,1}$
6- Su (x_a, x_p) 7- Ne (x_n)
$\alpha_{6,7}$
7- Ne (x_a, x_p) 6- Su (x_n)
$\alpha_{7,6}$

Fig. 3. The order-insensitive class-pair margins with respect to (w.r.t.) $21 = \#class(\#class-1)/2$ (the 1st row) and the order-aware class-pair margins w.r.t. 42 (the 3rd row) expression pairs ($\#class = 7$). For example, $\alpha_{1,2}$ corresponds to the triplets that x_a and x_p belong to the 1st class, while x_n belongs to the 2nd class. 'An', 'Di', 'Fe', 'Ha', 'Sa', 'Su' and 'Ne', are the abbreviations of 'Angry', 'Disgust', 'Fear', 'Happy', 'Sad', 'Surprise' and 'Neutral', respectively.

$x_a^{(i)}$ and $f(x_a^{(i)})$ in equation (1), which determine the gradient direction for back propagation; x_p and x_n are randomly generated pair of positive and negative samples for each anchor example x_a; $[\cdot]_+ \equiv max(\cdot, 0)$ is the hinge function.

Suppose the face pair x_p and x_n are presenting expressions i and j ($1 \leq i, j \leq \#class$), respectively, an adaptive margin $\alpha_{i,j}$ of class order-insensitive or order-aware can be used to consider the scale characteristics of each expression pair.

- As shown in the first row of Fig. 3, when the order of i, j is not considered, i.e. $\alpha_{i,j} = \alpha_{j,i}$, a total number of $(\#class(\#class-1)/2)$ margins are available.
- As shown in the third row of Fig. 3, when the order of i, j is considered, i.e. $\alpha_{i,j} \neq \alpha_{j,i}$, a total number of $(\#class(\#class-1))$ margins need to be defined.

The class-pair triplet loss is now formulated as follows

$$\mathcal{L}_t = \frac{1}{2}\sum_{x_a}[d(f(x_a), f(x_p))^2 - d(f(x_a), f(x_n))^2 + \alpha_{i,j}]_+, \quad (3)$$

where $\alpha_{i,j}$ is a margin associated with the expressions of the triplet (x_a, x_p, x_n). Thus, $\alpha_{i,j}$ reflects the feature variation intensity of the samples belonging to the group of triplets, i.e. $\{(x_a, x_p, x_n) | Class(x_a) = Class(x_p) = i, Class(x_n) = j\}$.

Larger margin parameter $\alpha_{i,j}$ encourages harder triplets, i.e. the loss in equation (3) is an increasing function of $\alpha_{i,j}$, which makes the gradient descent-based optimization w.r.t. $\alpha_{i,j}$ unreasonable. A heuristic margin updating method is proposed to use the updated centers of the expression classes:

$$\begin{cases} nr = \beta_{max} min(\frac{\beta_{thresd}}{NumIter}, 1), \\ \alpha_{i,j}^{new} = [||c_i - c_j||_2 - \frac{1}{N_a}\sum_{x_a, x_p} d(f(x_a), f(x_p))^2]_+, \\ \alpha_{i,j}^{final} = (1 - nr) \cdot \alpha_{i,j}^{old} \frac{\gamma^{new}}{\gamma^{old}} + nr \cdot \alpha_{i,j}^{new}, \\ \alpha_{i,j}^{old} = \alpha_{i,j}^{new}. \end{cases} \quad (4)$$

where $\alpha_{i,j} \leftarrow \alpha_{i,j}^{final}$ is the class-pair margin used in the current iteration; γ^{new}, γ^{old} are the L_2-norms of the embedded features of current and the last iterations before feature normalization. To restrict sample's feature representation onto a hypersphere, the embedded feature $f(x)$ is normalized to $\frac{\gamma \cdot f(x)}{||f(x)||_2}$, where γ

is a network optimization variable, which is updated with the feature $f(x)$, during network back-propagation. The initialization of γ and the optimization details are presented in the study [31]; N_a is the number of the intra-class sample pairs; The weight nr is introduced to use the preceding information for margin update; Since the margin updated in the preceding β_{thresd} iterations is not stable, $\alpha_{i,j}^{new}$ contributes to $\alpha_{i,j}^{final}$ with the maximum weight of $nr = \beta_{max}$. While $\alpha_{i,j}^{old}$ becomes stable, the contribution of $\alpha_{i,j}^{new}$, i.e. nr, is gradually decreased to make the margin parameter stagnate to fixed value. β_{thresd} and β_{max} are set to 1,000 and 0.5, respectively.

During the adaptive renewal of $\alpha_{i,j}$, the class centers $\{c_i\}$ are also updated based on the center loss $\mathcal{L}_C$ [32] as follows

$$\mathcal{L}_C = \frac{1}{2}\sum_k d(f(x_{[k]}), c_{y_k})^2, \tag{5}$$

where y_k is the expression label of the k-th sample $x_{[k]}$, c_{y_k} is the center feature vector of the y_k-th class, i.e. $\{f(x_{[k]})\}$.

Regarding to the margin update strategy (4), when a sample is far away from its class center, the variance of $||c_i - c_j||_2 - d(f(x_a), f(x_p))^2$ is relatively large. The offsets smaller than zero are discarded with the operator $[\cdot]_+$, while the values larger than zero increase the loss $\mathcal{L}_t$. Consequently, the feature movement toward the corresponding center is encouraged with equation (4), which is beneficial to decrease the intra-class distance.

The self adaptive update process presented in equation (4) not only removes the direct computation of the inter-class distances, but also improves the robustness of inter-class distance approximation, since current centers are updated based on the information of all the visited training samples in the preceding iterations.

B. Outlier-suppressed Hard Triplet Selection

During the training of network using all the samples and softmax loss, outlier (abnormal) samples may misguide the network to correlate the outlier features with expressions, and yielding wrong prediction for normal samples. Thus, the outlier samples are first detected, and further suppressed to reduce the misguidance information for normal samples.

Definition 1. *Outlier Expression Samples are samples whose face consists of large proportion of non-face regions or whose facial deforms are significantly different with common deforms of the labeled expression.*

As largely occluded or posed faces usually consist of large proportion of non-face regions, they are decided as outlier samples. For outlier expressions, distance distribution of deep features is used to detect significantly different deforms.

1) Occlusion and Pose Outlier: To reduce the misguidance of the abnormal samples during triplet loss training, largely posed and occluded faces are treated as outliers and excluded from the candidate hard samples based on linear regression and multi-task CNN, respectively.

A face database with various head poses [33] is used to train the linear regressor. The head pose database is a benchmark consisting of 2,790 monocular face images from 15 person identities, the pan and tilt angles vary from $-\pi/2$ to $\pi/2$. Example faces from the database are presented in Fig. 4.

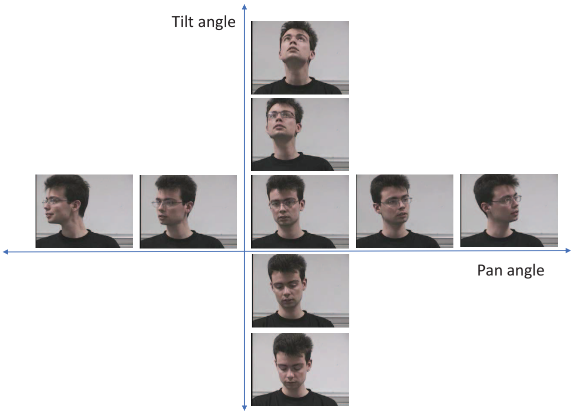


Fig. 4. Example faces with different pan and tilt angles in the posed face database [33].

The 2D Face Alignment Network (FAN) [34] is employed to locate the five landmark points, while the failure cases are categorized as largely occluded and posed faces and excluded from training. We denote the landmark points of the training and testing samples as matrix X and vector z, respectively. By minimizing the objective function $||z - Xw||_2^2$, the weights w are obtained with the least square estimation as $w = \left(X^T X\right)^{-1} X^T z$. The index of the posed face with the most similar head pose as the testing sample is predicted as $i_0 = arg\max w$. The testing sample is deemed as an outlier pose face, i.e. largely posed face, if the following condition holds

$$max(TILT_{i_0}, PAN_{i_0}) \geq \kappa \tag{6}$$

where $TILT_{i_0}$, PAN_{i_0} denotes the 'tilt' and 'pan' angles of the i_0-th sample, κ is the predefined threshold.

To detect and exclude largely occluded faces, the occlusions of various key facial parts are jointly detected with a multi-task CNN [35]. The general procedure of the detector can be summarized as three steps [35], i.e. the pre-training, fine tuning and multi-task occlusion identification with multi-layer perception (MLP). After the training of the multi-task CNN for occlusion detection, four tasks for respective facial parts are followed to judge whether left eye, right eye, nose or mouth is occluded or not. The testing sample is deemed as an outlier occlusion face, i.e. largely occluded face, if the following condition holds

$$\sum_i L_{i,1} \geq \beta \tag{7}$$

where L_i denotes the FC layer output in the task for the i-th facial part, while $L_{i,1}$ is the predicted probability of the i-th part occlusion after the Softmax activation; β is the predefined threshold.

2) Feature-Distance-Distribution Outlier: After the elimination of largely occluded and posed faces, easily-confusing samples are also treated as outliers, and excluded from hard triplets based on feature distance distribution.

Based on the maximal intra-class and minimal inter-class

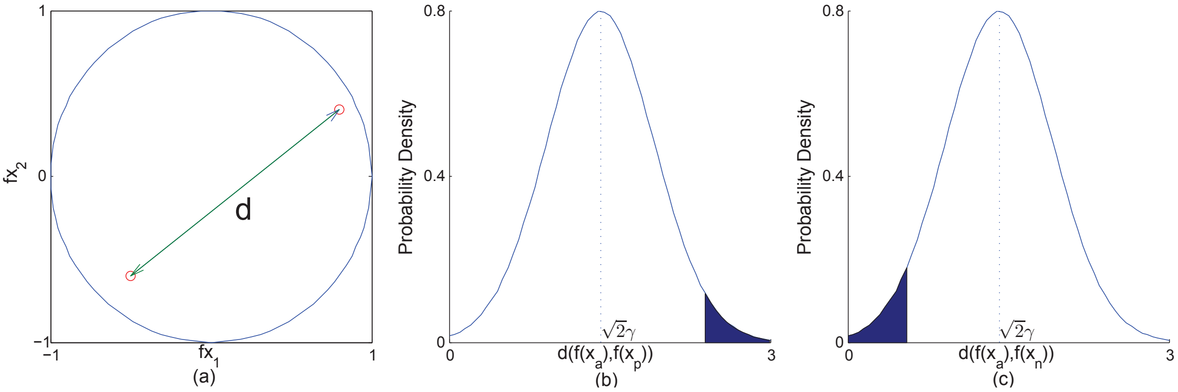


Fig. 5. (a) Random distance d with $n=2$, where $f(x)=(fx_1, fx_2)$. (b),(c) The feature distance distribution and the rejection regions (blue solid regions) for hard positive and negative sample selection with significance levels of $\tau_p=0.025$ and $\tau_n=0.05$, respectively.

feature distances, the hardest positive and negative samples in the study [7] are selected as follows

$$\begin{cases} x_p^* = arg\max_{x_p} d(f(x_a),f(x_p))^2, \\ x_n^* = arg\min_{x_n} d(f(x_a),f(x_n))^2. \end{cases} \tag{8}$$

where x_p and x_n are selected from a training batch. However, learning from the hardest triplet may misguide the network training due to abnormally-offset samples, such as the example expressions in Fig. 2. Kumar B G et al. [36] proposed to suppress the outlier sample in a triplet based on a margin hyper-parameter. In this work, we detect abnormal hard triplets according to the accurate distribution of feature distance and discard them in advance.

When the feature dimension, i.e. n, is large enough, it is induced in the studies [37], [8] (see Lemma 1 of Section V) that the random variable of distance of two embedding feature vectors, i.e. d, approximately obeys the following normal distribution as follows

$$d \sim \mathcal{N}(\sqrt{2}\gamma, \frac{\gamma}{\sqrt{2n}}), \tag{9}$$

where γ is the L_2-norm of the embedded feature; $\sqrt{2}\gamma$ and $\frac{\gamma}{\sqrt{2n}}$ are the mean and standard variance. An example 2D feature distance variable, i.e. d, is presented in Fig. 5(a).

For the detection of outlier samples, the null (H_0) and alternative (H_1) hypothesises of normal samples are first constructed as follows

$$\begin{cases} H_0 : \{\mu_{d(f(x_a),f(x_p))} \leq \sqrt{2}\gamma\}, \\ H_1 : \{\mu_{d(f(x_a),f(x_p))} > \sqrt{2}\gamma\}. \end{cases} \tag{10}$$

where $\mu_{d(f(x_a),f(x_p))}$ denotes the mean of the random variable $d(f(x_a),f(x_p))$. A selected positive sample x_p is deemed to be normal if the distance $d(f(x_a),f(x_p))$ falls in the acceptance region of the null hypothesis under a significance level τ_p, while abnormal if this distance lies in the corresponding rejection region, i.e. the corresponding alternative hypothesis is accepted. Similarly, a selected negative sample is an outlier if the corresponding null hypothesis $H_0 : \{\mu_{d(f(x_a),f(x_n))} \geq \sqrt{2}\gamma\}$ is rejected under a given significance level τ_n. Thus, the triplet (x_a, x_p, x_n) is discarded when $d(f(x_a),f(x_p))$ or $d(f(x_a),f(x_n))$ lies in the corresponding rejection region, i.e. one of the following rejection conditions is satisfied

$$\begin{cases} d(f(x_a),f(x_p)) \geq \sqrt{2}\gamma + \frac{\gamma}{\sqrt{2n}}F^{-1}(1-\tau_p), \\ d(f(x_a),f(x_n)) \leq \sqrt{2}\gamma - \frac{\gamma}{\sqrt{2n}}F^{-1}(\tau_n), \end{cases} \tag{11}$$

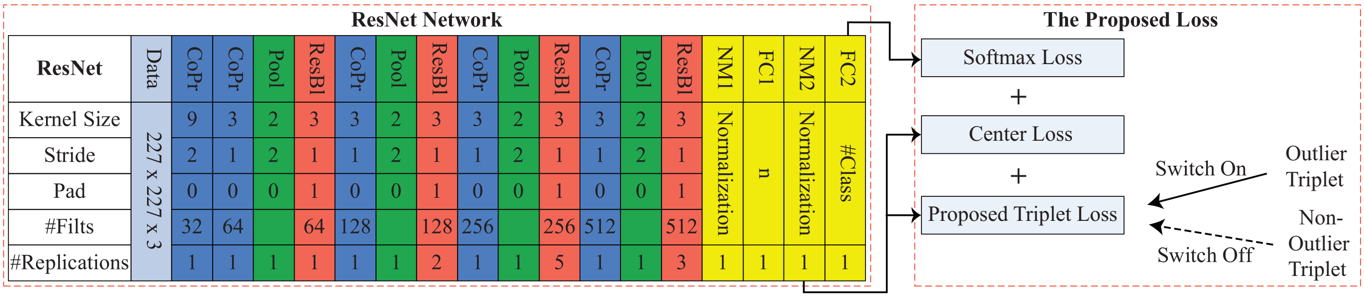
where $F^{-1}(1-\tau_p)$ is the inverse of the cumulative probability distribution of the normal distribution in equation (9) with cumulative probability being $1-\tau_p$, i.e. $P_F\{d \leq F^{-1}(1-\tau_p)\} = 1-\tau_p$; τ_p, τ_n are the significance levels of the positive and negative samples, respectively. The rejection regions described in equation (11) are shown in Figs. 5(b) and 5(c). Since the categories of negative samples are more diverse and there are larger variations among them, larger proportion of negative samples are assumed to be outliers, i.e. the significance level of negative pairs is larger than that of positive pairs. More precisely, during excluding of outliers from hard triplets, the requirement for negative samples to be normal is stronger than that of positive samples.

By providing additional upper and lower bounds for the selections of positive and negative samples according to the feature distance distribution, the proposed outlier-suppressed method in equation (11) can further decrease the influence of outlier samples, compared with the bias constraint in equation (3).

C. Network Training

The employed network structure is presented in Fig. 6, where the residual network (ResNet18 [1]) with slight modification [38], i.e. the dimension of the last FC layer output is set to #class, is employed. The same image preprocessing as the study [38] is employed. The self-adaptive normalization layer [31] is added after the last but one FC layer, i.e. the L_2-norm of the FC output vector $f(x)$ is normalized to a value γ with a self-adaptive mode. To fully make use of the already trained models, the fine tuning of an available face recognition model is employed.

Since hard samples are excluded from the proposed triplet loss only, they are still involved in the softmax loss-based training process. While the proposed triplet loss is used for highlighting the information of normal samples during training, the softmax loss $\mathcal{L}_S$ and the center loss are used to



ResNet Network

ResNet	Data	CoPr	CoPr	Pool	ResBl	CoPr	Pool	ResBl	CoPr	Pool	ResBl	CoPr	Pool	ResBl	NM1	FC1	NM2	FC2
Kernel Size	227 x 227 x 3	9	3	2	3	3	2	3	3	2	3	3	2	3	Normalization	n	Normalization	#Class
Stride		2	1	2	1	1	2	1	1	2	1	1	2	1				
Pad		0	0	0	1	0	0	1	0	0	1	0	0	1				
#Filts		32	64		64	128		128	256		256	512		512				
#Replications		1	1	1	1	1	1	2	1	1	5	1	1	3	1	1	1	1

Fig. 6. The network structure of ResNet. *CoPr* denotes the convolution layer followed by the PReLU activation function. *Pool* is the MaxPooling layer. *ResBl* is a residual block with output $ResOutput = PoolOutput + CoPr(CoPr(PoolOutput))$. $NM1$ denotes the 1st normalization layer [31]. *#Replications* denotes the times the same block is replicated. *#Filts* denotes the number of feature maps. $n = 512$ and *#class* denote the dimension of embedded feature and the number of expression classes.

boost the discriminative ability of the learned features, aiming at all the samples. The final loss is then formulated as follows

$$\mathcal{L} = \mathcal{L}_S + \lambda_C \mathcal{L}_C + \lambda_t \mathcal{L}_t. \tag{12}$$

where $\mathcal{L}_C$ and $\mathcal{L}_t$ are the center and triplet loss presented in equations (5) and (3); λ_C and λ_t are the regularization coefficients.

To avoid network non-convergence introduced by the instability of the class center, i.e. c_{y_k} in equation (5), in the preliminary iterations, a scale factor ρ is introduced to reduce the influence of such instability for the margin update (equation (4)). The scale factor is gradually increased from 0 to its maximum, which is used to scale the loss weight of λ_t in equation (3) as follows

$$\lambda_t \leftarrow \rho(\#iter) \cdot \lambda_t \equiv \frac{\lambda_t}{1 + 10e^{-\frac{\#iter}{3000}}}, \tag{13}$$

where $\#iter$ is the number of algorithm iterations.

For the network optimization, the gradient of $\mathcal{L}$ w.r.t. each variable is calculated, where the gradients associated with the proposed triplet loss (3), i.e. $\mathcal{L}_t$, are presented as follows

$$\begin{cases} sgn = 1_{d(f(x_a),f(x_p))^2 + \alpha_{i,j} > d(f(x_a),f(x_n))^2}, \\ \frac{\partial \mathcal{L}_t}{\partial f(x_a)} = (f(x_n) - f(x_p)) \cdot sgn, \\ \frac{\partial \mathcal{L}_t}{\partial f(x_p)} = (f(x_p) - f(x_a)) \cdot sgn, \\ \frac{\partial \mathcal{L}_t}{\partial f(x_n)} = (f(x_a) - f(x_n)) \cdot sgn. \end{cases} \tag{14}$$

where $1_{\{\cdot\}}$ is the 0-1 sign function.

For clarity, the entire optimization framework for the proposed triplet loss is illustrated in Algorithm 1.

III. Experimental Results

A. Experimental Setting and Databases

We perform the experiments using four-kernel Nvidia TITAN GPU Card and CAFFE package. The parameter settings of the proposed algorithm are presented in Table I. The network in Fig. 6 is chosen as the backbone network, which is trained for 120 epochs on each dataset via stochastic gradient descent (SGD) with initial learning rate of 0.01 (decayed by a factor of 0.5 for each 30 epochs), while the momentum and the weight decay value are set to 0.9 and 0.0001, respectively.

Algorithm 1 Network training with the proposed triplet loss.

```
Set the hyper-parameters $\kappa$, $\beta$, $\beta_{max}$, $\beta_{thresd}$, $\lambda_C$, $\lambda_t$, $MaxIter$.
Initialize the class-pair margins [3], class centers and network parameters.
for $s = 0, \cdots, MaxIter$ do
    Update the margin parameters according to the updated centers with equation (4);
    Generate the candidate samples excluding the largely posed and occluded faces detected in equations (6) and (7);
    Obtain the hard triplets with the strategy (8) from the candidate samples based on the distance-distribution constraint (11);
    Perform network forward to obtain the triplet loss with equation (3);
    Perform network backward to compute the gradients of the proposed triplet loss w.r.t. the embedded features with equation (14);
    Perform stochastic gradient descent (SGD) to update the embedded features and network parameters;
end for
Output the trained model for testing.
```

The proposed algorithm is tested on the expression databases of the FER2013 [30], AFEW [39], Extended Cohn-Kanade (CK+) [40], AffectNet [41], Oulu-CASIA [42], MMI [30] and BU-3DFE [43], whose examples are presented in Fig. 7.

The FER2013 database [30] is an expression database collected from the internet and used for a challenge. The database consists of 35,887 grayscale face images, while the training set consists of 28,709 examples, both the validation (the public test) and testing (the private test) datasets contain 3,589 expression images. Each face was labeled with neutral (Ne) or one of the six typical expressions, i.e. angry (An), disgust (Di), fear (Fe), happy (Ha), sad (Sa) and surprise (Su).

Acted Facial Expressions in the Wild (AFEW-6.0) [39] is a data corpus of dynamic temporal facial expressions labeled with neutral and the six typical expressions. The images are extracted from movies, where 757 and 365 of 1,122 sequences are used for training and testing, respectively. The three peak frames from each sequence of the original validation dataset

TABLE I
THE PARAMETER SETTING OF THE PROPOSED ALGORITHM.

Parameter Name	Value	Parameter Name	Value
λ_C in Eq. (12)	0.1	λ_t in Eq. (12)	1e-5
Learning rate	1e-2	Batch size	64
τ_p in Eq. (11)	0.025	τ_n in Eq. (11)	0.05
κ in Eq. (6)	$\pi/3$	β in Eq. (7)	1.5

are selected as the testing dataset.

The CK+ database [40] consists of 593 expression sequences from 123 subjects, which are labeled with the six typical expressions and contempt (Co). For our testing, the ‘contempt’ expression is not considered, the three peak and neutral frames from each of the remaining 415 expression sequences, from 106 person identities were selected as the testing dataset.

The AffectNet [41] database contains about 420,300 manually annotated facial expression images, i.e. 28,7651 training samples and 4,000 validation samples with neutral, contempt and the six typical expressions. For our testing, the ‘contempt’ expression is not employed. Since the original testing dataset is not made public, the original validation set is used for testing, while the similar division as the study [41] is performed on the original training dataset to generate our training and validation datasets, with 283,901 and 3,500 samples, respectively.

The Oulu-CASIA NIR&VIS expression database [42] contains videos of 80 subjects, which are captured with two imaging systems, *NIR* (Near Infrared) and *VIS* (Visible light), under three different illumination conditions, i.e. normal (*Strong*) indoor illumination, weak illumination and dark illumination. Each face sequence presents one of the six typical expressions, where the three peak expressions of the database of *VIS* or *Strong* are used. A simple augmentation with 16 different crops for each face is conducted to generate 23,040 images.

The MMI database [30] includes 31 person identities with ages vary from 19 to 62, which is either a European, Asian, or South American. The faces present the six typical expressions. The peak frames with the top three deform intensities in each of 205 expression sequences are employed for testing, and the selected faces are further augmented to generate 15,675 images.

The BU-3DFE database [43] consists of 2,500 pairs of 3D face models and texture images from 100 subjects, i.e. 56 female and 44 male. Each subject displayed one of the six typical expressions with four intensity levels. Following the test protocol in [44], only the texture images with the top two deform intensities were selected for the testing.

For the recognition of each testing sample, majority voting based on the probabilities of augmented face regions is employed. For the following experiments, the same strategy as the state of the arts, i.e. the person-independent strategy with multiple-fold cross validation, is employed for testing and comparison.

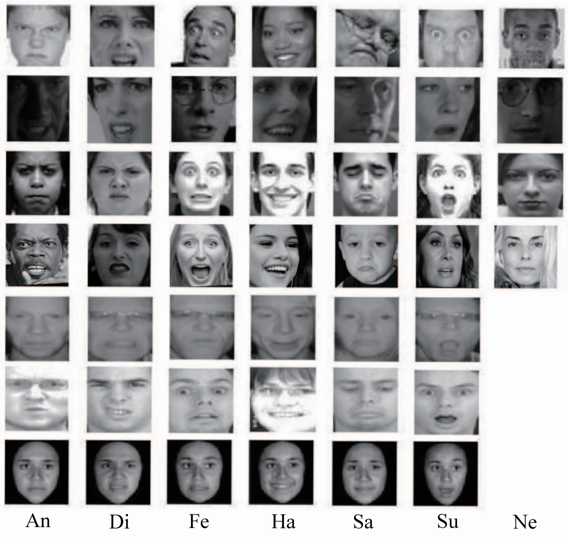


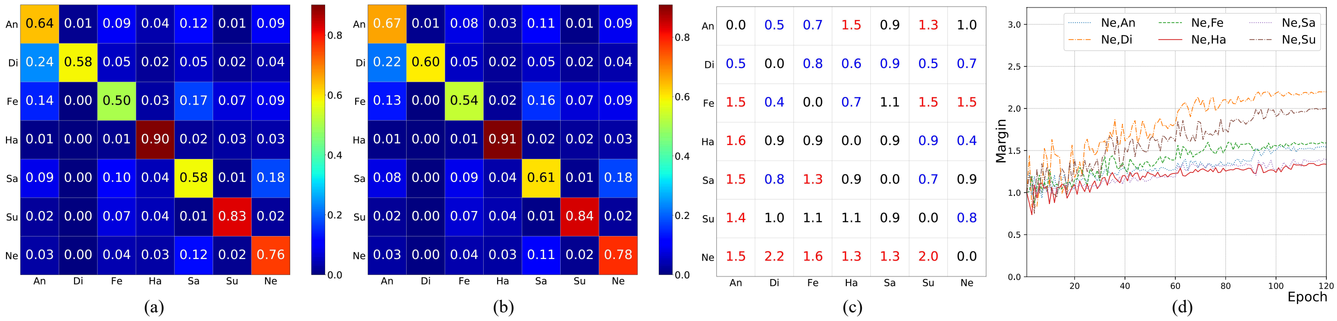
Fig. 7. Example images of the seven public datasets, i.e. FER2013, AFEW, CK+, AffectNet, Oulu-CASIA, MMI and BU-3DFE.

B. Algorithm Analysis

To evaluate the overall performance, the confusion matrix of the proposed algorithm for the FER2013 database is presented in Fig. 8(b), while the confusion matrices of the other six databases are presented in Fig. 9. The confusion recognition rates in Figs. 8(b) and 9 show that the expressions ‘angry’, ‘disgust’, ‘fear’ and ‘sad’ are relatively more difficult than the other three expressions, and are easier to be confused with each other due to the smaller differences. Meanwhile, the confusion accuracies among different expression pairs present strong inconsistency.

To study the usefulness of the proposed class-pair margins for inconsistent performances among different expression pairs, Fig. 8 also presents the confusion matrices with and without the proposed class-pair margins. Fig. 8 shows that the improvement with the class-pair margins for difficult classes is more significant than that for relatively easy classes, i.e. the improvement for the ‘Fear’ expression is about 4%, which is almost 4 times the improvement for the ‘Happy’ expression. Thus, the class-pair margins are more beneficial for the expression classes with finer deform intensity.

To study the correlation between the confusion probabilities and the adjusted margins, Figs. 8(b) and 8(c) further present the confusion probabilities and the updated pairwise margins.



(a)

	An	Di	Fe	Ha	Sa	Su	Ne
An	0.64	0.01	0.09	0.04	0.12	0.01	0.09
Di	0.24	0.58	0.05	0.02	0.05	0.02	0.04
Fe	0.14	0.00	0.50	0.03	0.17	0.07	0.09
Ha	0.01	0.00	0.01	0.90	0.02	0.03	0.03
Sa	0.09	0.00	0.10	0.04	0.58	0.01	0.18
Su	0.02	0.00	0.07	0.04	0.01	0.83	0.02
Ne	0.03	0.00	0.04	0.03	0.12	0.02	0.76

(b)

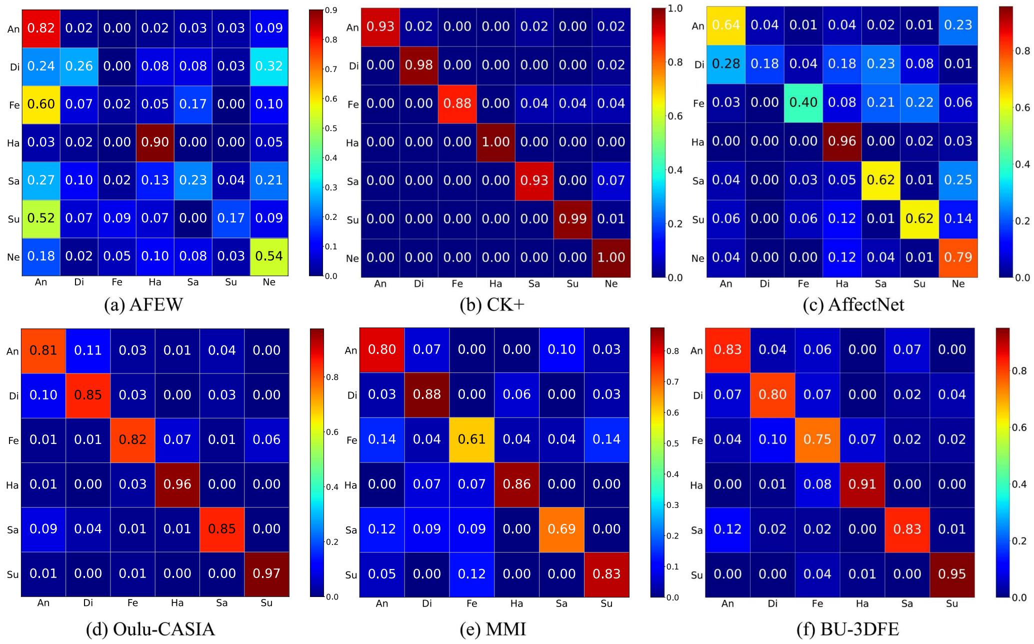
	An	Di	Fe	Ha	Sa	Su	Ne
An	0.67	0.01	0.08	0.03	0.11	0.01	0.09
Di	0.22	0.60	0.05	0.02	0.05	0.02	0.04
Fe	0.13	0.00	0.54	0.02	0.16	0.07	0.09
Ha	0.01	0.00	0.01	0.91	0.02	0.02	0.03
Sa	0.08	0.00	0.09	0.04	0.61	0.01	0.18
Su	0.02	0.00	0.07	0.04	0.01	0.84	0.02
Ne	0.03	0.00	0.04	0.03	0.11	0.02	0.78

(c)

	An	Di	Fe	Ha	Sa	Su	Ne
An	0.0	0.5	0.7	1.5	0.9	1.3	1.0
Di	0.5	0.0	0.8	0.6	0.9	0.5	0.7
Fe	1.5	0.4	0.0	0.7	1.1	1.5	1.5
Ha	1.6	0.9	0.9	0.0	0.9	0.9	0.4
Sa	1.5	0.8	1.3	0.9	0.0	0.7	0.9
Su	1.4	1.0	1.1	1.1	0.9	0.0	0.8
Ne	1.5	2.2	1.6	1.3	1.3	2.0	0.0

(d)

Fig. 8. The confusion matrices without (a) and with (b) the proposed class-pair margins, the values of the adjusted pairwise margins (c) for the FER2013 database and their evolution curves (d). The top 15 smallest and largest nonzero margins are marked with blue and red colors, respectively. The evolution curves of six pairwise margins, i.e. $\alpha_{Ne,An}$, $\alpha_{Ne,Di}$, $\alpha_{Ne,Fe}$, $\alpha_{Ne,Ha}$, $\alpha_{Ne,Sa}$ and $\alpha_{Ne,Su}$, are visualized.



(a) AFEW

	An	Di	Fe	Ha	Sa	Su	Ne
An	0.82	0.02	0.00	0.02	0.03	0.03	0.09
Di	0.24	0.26	0.00	0.08	0.08	0.03	0.32
Fe	0.60	0.07	0.02	0.05	0.17	0.00	0.10
Ha	0.03	0.02	0.00	0.90	0.00	0.00	0.05
Sa	0.27	0.10	0.02	0.13	0.23	0.04	0.21
Su	0.52	0.07	0.09	0.07	0.00	0.17	0.09
Ne	0.18	0.02	0.05	0.10	0.08	0.03	0.54

(b) CK+

	An	Di	Fe	Ha	Sa	Su	Ne
An	0.93	0.02	0.00	0.00	0.02	0.00	0.02
Di	0.00	0.98	0.00	0.00	0.00	0.00	0.02
Fe	0.00	0.00	0.88	0.00	0.04	0.04	0.04
Ha	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Sa	0.00	0.00	0.00	0.00	0.93	0.00	0.07
Su	0.00	0.00	0.00	0.00	0.00	0.99	0.01
Ne	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(c) AffectNet

	An	Di	Fe	Ha	Sa	Su	Ne
An	0.64	0.04	0.01	0.04	0.02	0.01	0.23
Di	0.28	0.18	0.04	0.18	0.23	0.08	0.01
Fe	0.03	0.00	0.40	0.08	0.21	0.22	0.06
Ha	0.00	0.00	0.00	0.96	0.00	0.02	0.03
Sa	0.04	0.00	0.03	0.05	0.62	0.01	0.25
Su	0.06	0.00	0.06	0.12	0.01	0.62	0.14
Ne	0.04	0.00	0.00	0.12	0.04	0.01	0.79

(d) Oulu-CASIA

	An	Di	Fe	Ha	Sa	Su
An	0.81	0.11	0.03	0.01	0.04	0.00
Di	0.10	0.85	0.03	0.00	0.03	0.00
Fe	0.01	0.01	0.82	0.07	0.01	0.06
Ha	0.01	0.00	0.03	0.96	0.00	0.00
Sa	0.09	0.04	0.01	0.01	0.85	0.00
Su	0.01	0.00	0.01	0.00	0.00	0.97

(e) MMI

	An	Di	Fe	Ha	Sa	Su
An	0.80	0.07	0.00	0.00	0.10	0.03
Di	0.03	0.88	0.00	0.06	0.00	0.03
Fe	0.14	0.04	0.61	0.04	0.04	0.14
Ha	0.00	0.07	0.07	0.86	0.00	0.00
Sa	0.12	0.09	0.09	0.00	0.69	0.00
Su	0.05	0.00	0.12	0.00	0.00	0.83

(f) BU-3DFE

	An	Di	Fe	Ha	Sa	Su
An	0.83	0.04	0.06	0.00	0.07	0.00
Di	0.07	0.80	0.07	0.00	0.02	0.04
Fe	0.04	0.10	0.75	0.07	0.02	0.02
Ha	0.00	0.01	0.08	0.91	0.00	0.00
Sa	0.12	0.02	0.02	0.00	0.83	0.01
Su	0.00	0.00	0.04	0.01	0.00	0.95

Fig. 9. The confusion matrices of the AFEW (a), CK+ (b), AffectNet (c), Oulu-CASIA (d), MMI (e) and BU-3DFE (f) databases.

While the mean margins of the top 15 smallest (blue) and largest (red) nonzero margins are 0.66 and 1.53, the mean confusion probabilities corresponding to the small and large nonzero margins are 3.93% and 5.07% for the FER2013 database. Larger confusion probability of a class pair implies the finer difference scale. In this case, a larger margin is consequently more beneficial since more triplets can be selected in the training to boost the discriminative ability for this class pair. The large variance among the pairwise margins corresponding to all the expression pairs also illustrates the usefulness of the class-pair margin strategy.

The evolution curves of six class-pair margins, i.e. $\alpha_{Ne,An}$, $\alpha_{Ne,Di}$, $\alpha_{Ne,Fe}$, $\alpha_{Ne,Ha}$, $\alpha_{Ne,Sa}$ and $\alpha_{Ne,Su}$ for FER2013 dataset are visualized in Fig. 8(d). Fig. 8(d) shows that the margins can gradually evolve to the corresponding values in Fig. 8(c).

To study the outliers detected for the proposed triplet loss, the largely posed and occluded faces detected for the FER2013 database are demonstrated in Fig. 10. The figure shows that largely posed and occluded faces are properly detected by the introduced face-pose regressor and multi-task CNN.

Fig. 11 further demonstrates six example outliers before (the 2nd row) and after (the 3rd row) the removal of largely posed and occluded expressions. As shown in Figs. 11(a)-11(c), the largely posed and occluded expression images can

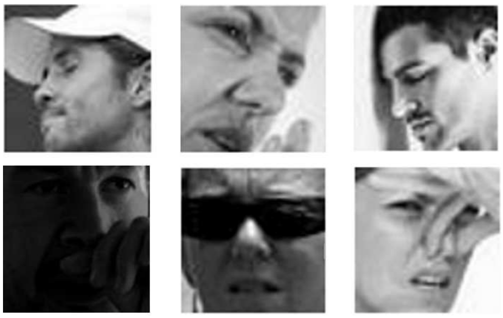


Fig. 10. The posed (the 1st row) and occluded (the 2nd row) faces detected with face-pose regressor and multi-task CNN for the FER2013 database.

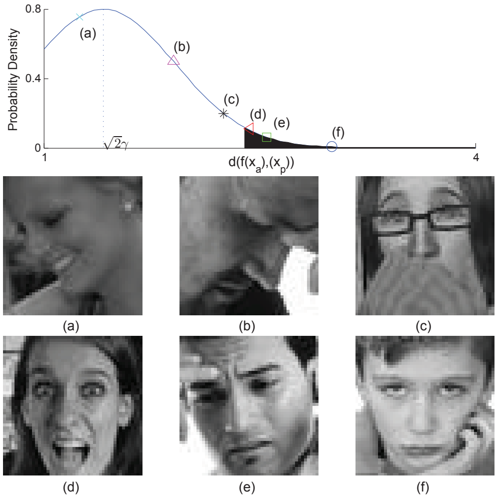


Fig. 11. Largely posed and occluded faces ((a)-(c)) and confusing outliers ((d)-(f)) that lie in the non-rejection and rejection regions during positive sample selection.

lie on the acceptance region of the feature distance distribution, and correctly recognized as 'happy', 'sad' and 'surprise' with large probabilities. However, these outliers may misguide the network to correlate the expression features with the pose and occlusion, and decrease the generalization ability on the frontal and un-occluded faces. Thus, the exclusion of the largely posed and occluded faces is beneficial for the triplet loss learning.

The last row of Fig. 11 presents three example outliers detected with feature distance distribution during positive sample selection, whose anchor-positive distances lie in the rejection region of the normal sample hypothesis. One can observe that the expressions labeled with 'angry', 'fear' and 'neutral' in Figs. 11(d)-11(f) can be easily confused with 'surprise', 'sad' and 'sad', respectively. Thus, the exclusion of these confusing expressions from the triplet loss can reduce misguidance during network training and help the network to generalize well to normal expressions.

TABLE II
The performances (%) with different triplet loss settings for the seven databases. 'Basic setting' denotes the combination of the softmax, center losses and the feature normalization; 'Tri1' or 'Tri15' denotes the setting of 1 or 15 margins; 'RandTri', 'HardTri', 'DisSupTri', 'PosSupTri', 'OccSupTri' and 'AllSupTri' denote the selection strategies of random, the hardest, the outlier-suppressed triplets according to feature distribution, pose, occlusion and their combination, respectively. All the algorithm settings except 'Base line' include 'Basic setting'.

Database	Algorithm Description	Recognition rates (%)
FER2013	Softmax Only	68.91
	Basic setting	71.66
	Tri1 + RandTri	71.80
	Tri1 + HardTri	71.41
	Tri21 + HardTri	72.14
	Tri42 + HardTri	71.83
	Tri1 + DisSupTri	71.86
	Tri21 + DisSupTri	72.64
	Tri42 + DisSupTri	72.22
	Tri21 + DisSupTri + OccSupTri	73.28
	Tri21 + DisSupTri + PosSupTri	**73.78**
	Tri21 + AllSupTri	73.56
	Tri42 + AllSupTri	73.27
AFEW	Softmax Only	41.64
	Basic setting	42.19
	Tri1 + RandTri	40.82
	Tri21 + HardTri	43.01
	Tri42 + HardTri	44.12
	Tri21 + DisSupTri + OccSupTri	43.01
	Tri21 + DisSupTri + PosSupTri	44.93
	Tri21 + AllSupTri	46.30
	Tri42 + AllSupTri	**46.84**
CK+	Softmax Only	95.17
	Basic setting	95.87
	Tri1 + RandTri	96.21
	Tri21 + HardTri	96.64
	Tri42 + HardTri	96.48
	Tri21 + DisSupTri	**97.61**
	Tri42 + DisSupTri	97.13
AffectNet	Softmax Only	58.03
	Basic setting	58.63
	Tri1 + RandTri	59.74
	Tri21 + HardTri	59.50
	Tri42 + HardTri	59.8
	Tri21 + DisSupTri	**60.12**
	Tri42 + DisSupTri	59.92
Oulu-CASIA	Softmax Only	84.79
	Basic setting	85.90
	Tri1 + RandTri	85.97
	Tri15 + HardTri	86.04
	Tri30 + HardTri	87.13
	Tri15 + DisSupTri	87.29
	Tri30 + DisSupTri	**87.94**
MMI	Softmax Only	75.12
	Basic setting	76.10
	Tri1 + RandTri	76.10
	Tri15 + HardTri	76.59
	Tri30 + HardTri	77.07
	Tri15 + DisSupTri	78.05
	Tri30 + DisSupTri	**78.53**
BU-3DFE	Softmax Only	80.91
	Basic setting	82.33
	Tri1 + RandTri	82.15
	Tri15 + HardTri	83.16
	Tri30 + HardTri	83.74
	Tri15 + DisSupTri	**84.50**
	Tri30 + DisSupTri	84.41

C. Ablation Study of the Proposed Triplet Loss

To test the performance of the proposed triplet loss, the ResNet network is trained with different loss strategies, i.e. the class-pair margins, different outlier suppressions and their combinations, then their performances on the seven databases are presented in Table II.

In Table II, the performance with the softmax loss is listed as the benchmark, while the setting with the combination of the softmax, center losses and the feature normalization is used as the basic training setting. To compare the performances of the proposed loss with other variants of the triplet loss, the combination of $Tri1$ and $RandTri$ is used as the original triplet loss. Table II shows that both class-pair margins and outlier-suppressed hard triplet selection improve the performance of the original triplet loss on all of the seven databases.

For the FER2013 database, while the proposed loss achieves an improvement of 0.84% over the original loss (the 3rd and 8th rows), the proposed loss with 21 margins achieves an improvement of 0.78% over the variant with single margin (the 7th and 8th rows), which justifies the effectiveness of the class-pair margins. Meanwhile, benefited from the exclusion of abnormal samples with feature distance distribution, the proposed loss achieves an improvement of 0.5% over the variant with the hardest triplet selection (the 5th and 8th rows). When the detection of largely posed faces is embedded, the proposed triplet loss achieves the best performance of 73.78%, i.e. 4.87% above the baseline, and about 2.12% over the basic loss setting. Similar improvements by the triplet loss with $\#class(\#class-1)/2$ margins over the Softmax-only loss and the basic setting are also observed for AFEW, where improvements of 4.66% and 4.11% are achieved. Meanwhile, the triplet loss with $\#class(\#class-1)$ margins achieves an improvement of 0.54% over the loss with $\#class(\#class-1)/2$ margins.

For the other five databases with less posed and occluded faces, the outlier suppression based on only feature distance distribution is employed, which achieves rather competitive performances when $\#class(\#class-1)/2$ margins are used. Meanwhile, improvements of 0.65% and 0.48% are achieved by the setting of $\#class(\#class-1)$ over the setting of $\#class(\#class-1)/2$ margins for Oulu-CASIA and MMI databases, respectively. Improvements of 2.44%, 2.09%, 3.15%, 3.41% and 3.59% over the softmax loss, and 1.74%, 1.49%, 2.04%, 2.43% and 2.17% over the best basic setting are achieved for the CK+, AffectNet, Oulu-CASIA, MMI and BU-3DFE databases, respectively.

D. Comparison with State of the Arts

To compare the performance of the proposed algorithm with other algorithms, Table III lists the performances of the proposed approach and state-of-the-art algorithms on the seven databases. For the FER2013 database, our algorithm achieved as high as 73.78% accuracy, which is even 2.58% higher than that of the challenge winner [46], i.e. 71.2%. It is worthwhile that the algorithm [23] also employed the triplet loss for FER, while our triplet loss achieves an improvement of 0.96% over them [23]. For four of the employed seven databases, the proposed algorithm achieved the best performances among the algorithms for comparison, where improvements of 0.96%, 1.41%, 0.26% and 0.62%, are achieved for the FER2013, AFEW, CK+ and AffectNet databases, respectively. The proposed algorithm also ranks the second on the MMI and BU-3DFE databases.

TABLE III
COMPARISON OF DIFFERENT RELATED ALGORITHMS ON THE SEVEN EXPRESSION DATABASES. 'PROT.' DENOTES THE EMPLOYED PROTOCOL AND ITS VALUE '3' DENOTES '3-FOLD'. 'DATA.', 'SUB.', 'FER.', 'AF.', 'A.N.', 'OULU.' AND 'BU.' ARE THE ABBREVIATIONS OF 'DATABASE', 'SUBJECT', 'FER2013', 'AFEW', 'AFFECTNET', 'OULU-CASIA' AND 'BU-3DFE'.

Data.	Algorithm	Sub.	Prot.	Recog. rate (%)
FER.	Deeper DNN [45]	-	-	66.4
	DNN with SVM [46]	-	-	71.2
	Multi-scale CNN [23]	-	-	72.82
	Adaptive Feature Losses [38]	-	-	72.67
	Ours (21 margins)	-	-	**73.78**
AF.	Multi-clue Fusion [47]	-	3	44.46
	Score-level Classifier Fusion [48]	-	3	44.47
	3D CNN [49]	-	-	39.69
	Single CNN-RNN [49]	-	-	45.43
	Ours (42 margins)	-	-	**46.84**
CK+	Deeper DNN [45]	106	5	93.2
	Salient Facial Parts [50]	-	10	94.09
	DeRL [44]	118	10	97.30
	Adaptive Feature Losses [38]	106	10	97.35
	Ours (21 margins)	106	10	**97.61**
A.N.	PG-CNN [51]	-	-	55.33
	gACNN [52]	-	-	58.78
	IPA2LT [53]	-	-	57.31
	RAN [54]	-	-	59.5
	Ours (21 margins)	-	-	**60.12**
Oulu.	STM-Explet [55]	80	10	74.59
	Atlases [56]	80	10	75.52
	DTAGN-Joint [57]	80	10	81.46
	DeRL [44]	80	10	88.0
	Adaptive Feature Losses [38]	80	10	85.83
	Augmentation with GAN [58]	80	10	**88.25**
	Ours (30 margins)	80	10	87.94
MMI	STM-Explet [55]	205	10	75.12
	DTAGN-Joint [57]	-	-	70.24
	IACNN [59]	208	-	71.55
	DeRL [44]	208	10	73.23
	Augmentation with GAN [58]	208	10	**81.13**
	Ours (30 margins)	205	10	78.53
BU.	Geometric Scattering [60]	-	-	**84.80**
	Sample Order [61]	64	-	72.89
	DeRL [44]	100	10	84.17
	Ours (15 margins)	100	10	84.50

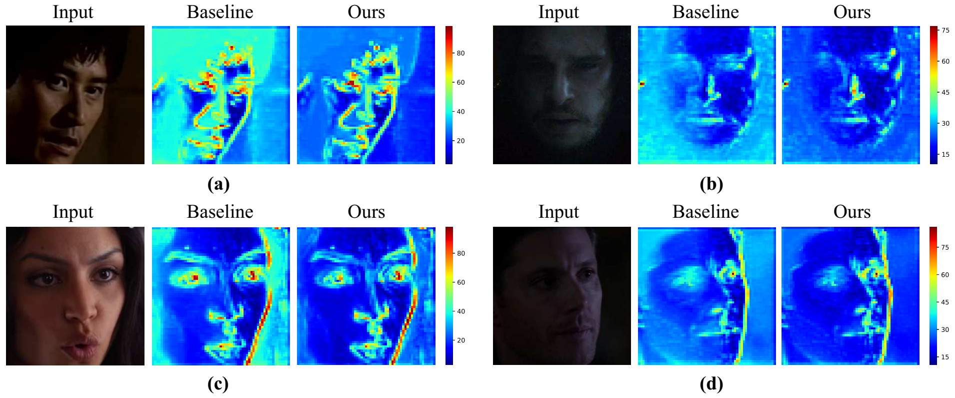


Fig. 12. The average of the feature maps generated by the baseline and the proposed algorithm.

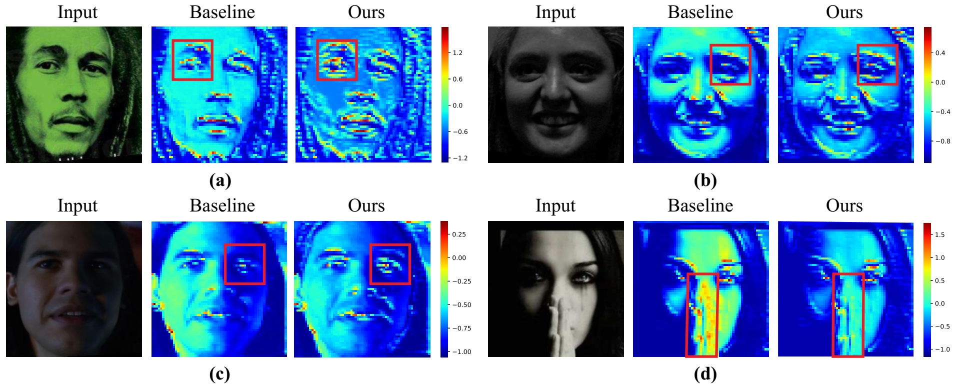


Fig. 13. The 25th feature map generated by the baseline and the proposed algorithm.

For the AFEW database, although only the peak frames are used (the algorithms [48], [49] employed the sequential expression video), the proposed algorithm achieves the best performance, i.e. 46.84%, among five state of the arts.

For the AffectNet database, the proposed algorithm achieved better performance, i.e. 60.12%, than the occlusion-aware methods [51], [52], the latent truth discovering method [53] and region attention network [54], i.e. an improvement of 0.62% is achieved.

For the Oulu-CASIA database, the proposed algorithm achieves a rather competitive performance, i.e. 87.94%, compared with the performance (88.0%) achieved by de-expression residual learning [44]. Though the proposed triplet loss is constructed with only the FC layers, rather than multiple intermediate layers in the study [44], our algorithm achieves better performances for the CK+ and MMI databases, and significantly better performance for the BU-3DFE database.

For the MMI database, the proposed algorithm ranks the 2nd, i.e. 78.53% among six state-of-the-art approaches. While the proposed algorithm employed single network, the runtime cost of the additional construction of the proposed triplet loss is marginal. Compared with the approach [58] that achieves the best performance on MMI, i.e. 81.13%, the proposed algorithm does not require an additional GAN training during hard triplet generation.

For the BU-3DFE database, the proposed algorithm achieved a competitive performance, i.e. 84.50%, compared with the best performance, i.e. 84.80%, of geometric scattering representation [60]. While the study [60] used the 3D data for the recognition, our algorithm only employed 2D images.

Though our proposed triplet loss does not achieve the best performances for all of the databases, it balances the performances for these databases and yields competitive accuracies, compared with the best performances.

E. Visualization of the Feature Maps

In order to explore the work mechanism of the proposed algorithm, feature maps from the outputs of the second residual block are visualized. Figs. 12 and Fig. 13 present the original image, and the feature maps after the training with ‘Softmax

Only' (baseline) and 'Tri21 + DisSupTri' (ours), corresponding to the average and 25th feature maps, respectively.

Compared with the baseline, Figs. 12(a)-(d) show that expression insensitive regions, such as the hair and background of the non-face region in Figs. 12(a)-(d) and the cheek and forehead of the face region in Figs. 12(b)-(d), are better suppressed by the proposed algorithm. Fig. 13 shows that the feature maps generated with the proposed algorithm display larger responses on the expression sensitive regions, such as the eyes and lips in Figs. 13(a)-(c), while smaller responses on the occlusion region in Fig. 13(d).

Thus, benefit from the suppression of outlier samples during training, the proposed algorithm is able to de-activate the expression insensitive regions, highlight the expression sensitive regions, and help to improve the generalization performance of FER.

F. Cross-database Experiments

To study the generalization performance of the proposed algorithm, cross-database experiment is performed among six databases, i.e. FER2013, CK+, Oulu-CASIA, MMI, BU-3DFE and AFEW. Instead of conducting all the cross-database experiments, the training datasets with the top two competitive performances for each testing dataset are used for the evaluation and comparison. The results of the proposed algorithm, together with that of the sparse feature loss [62] are presented in Table IV.

Table IV shows that the proposed algorithm outperforms the sparse feature loss [62] in most cases, where the proposed algorithm achieves improvements of 1.99% or 2.25% over sparse feature loss [62] for FER2013 when CK+ or MMI is used for training. When MMI and Oulu-CASIA are used for training and testing, the proposed algorithm achieves an improvement of 14.37% over the sparse feature loss [62]. By suppressing outlier samples during training, the proposed algorithm is able to reduce the abnormal features, which can consequently improve its cross-database performances of the learned features.

G. Computational Efficiency

To study the runtime cost of the proposed algorithm, theoretical analysis and actual runtime costs of network training with the baseline and the proposed algorithm on the AFEW6.0 dataset, are presented in Table V. The comparison related with network testing is not considered since the same network architecture is used for inference and the detection of outlier occlusion and pose is not demanded. For the comparison, the online training and offline processing are studied separately.

For deep network training, the floating point operations (FLOPs) related with the convolution blocks is $T_{conv} \sim O(\sum_{l=1}^{d_{conv}} n_{map,l}^2 \cdot n_{ker,l}^2 \cdot n_{cha,l-1} \cdot n_{cha,l})$, where d_{conv} is the number of convolution layers, $n_{map,l}$, $n_{ker,l}$ and $n_{cha,l}$ are the feature map size, the kernel size and the channel number in the l-th layer. While the time complexities for the layers of pooling, ReLU, batch normalization and FC layer are negligible compared with that of the convolution layers.

Table V shows that the theoretical runtime costs of the baseline and the proposed algorithms are about the same. While the actual runtime cost of the proposed algorithm exceeds that of the baseline due to the CPU executions of margin updating and hard triplet selection, they can be speeded up in GPU for real application. Table V also shows that the theoretical and actual runtime costs of offline computation, i.e. the detection of largely occluded and posed faces, are almost negligible compared with the online training, as the offline operations run for only one epoch.

TABLE V
THEORETICAL AND ACTUAL RUNTIME COSTS OF THE BASELINE AND THE PROPOSED METHOD. THE ACTUAL RUNTIME IS EVALUATED ON THE AFEW6.0 DATASET.

Runtime Cost	Baseline (Softmax Only, Online)	Ours (the Proposed Loss, Online)	Ours (Occlusion Outlier and Pose Selection, Offline)
Theoretical Runtime Cost (FLOPs)	1.7047×10^9	1.7047×10^9 $+ 0.2457 \times 10^6$	1.2185×10^7
Actual Runtime Cost in Seconds (s)	155 s/epoch × 120 epochs	261 s/epoch × 120 epochs	576 s/epoch × 1 epoch

IV. CONCLUSIONS AND FUTURE WORKS

To take into account the deform intensity inconsistency among expression pairs and the outliers that potentially impair the network generalization performance for facial expression recognition (FER), this work proposed a triplet loss based on class-pair margins and multistage outlier suppression. To address the class-pair inconsistency, class order information and self-adaptive model are used in the construction and renewal of the class-pair margins. To reduce the misguidance introduced by abnormal hard triplets, the training samples are screened based on multi-stage detections of outliers, i.e. largely posed, occluded expressions or faces with abnormal offset from the mean feature presentation. Extensive experiments on seven public databases, including ablation study, comparison with the state of the arts, feature map visualization, cross-database evaluation and computational efficiency analysis, show that the network with the proposed triplet loss achieved better performance than that without the proposed pairwise margins and outlier-suppression strategies. Compared with state-of-the-art approaches, competitive and balanced performances have been observed.

TABLE IV

CROSS-DATABASE PERFORMANCES (%) ON SIX DATABASES. THE DATABASE IN THE BRACKET IS USED FOR TRAINING.

Methods	FER2013	CK+	Oulu-CASIA	MMI	BU-3DFE	AFEW
Sparse Feature Loss [62]	39.72 (CK+) 60.19 (MMI)	77.02 (MMI) **84.47** (Oulu-CASIA)	42.08 (CK+) 50.83 (MMI)	60.0 (CK+) 61.46 (Oulu-CASIA)	-	-
Ours	**41.71** (CK+) **62.44** (MMI)	76.05 (MMI) 82.42 (Oulu-CASIA)	**61.25** (CK+) **65.20** (MMI)	**62.68** (CK+) 58.53 (Oulu-CASIA)	62.10 (CK+)	44.9 (FER)

However, improvements or future works still require further exploration. First, the region for hard expression selection can be roughly reduced with a pre-trained model to reduce the time complexity of hard triplet selection. Second, several additional hyper-parameters are introduced in equations (4) and (6), whose best settings, sensitivity analysis and the self-adaptive model can be further explored. Third, the class-pair margin should be extended to group-pair margin for recognition tasks with a larger number of categories, where each group shall include multiple classes. Fourth, the complementarity between the proposed algorithm and the algorithms that achieve better performances on the considered databases, e.g. [25], [58] and [54], needs further exploitation. Lastly, the proposed algorithm is general, which can be used for more recognition or verification tasks.

V. APPENDIX

Lemma 1. *[37] Given the surface of an n-dimensional hypersphere ($n > 3$) with the radius of γ, we use d to denote the Euclidean distance between any two points randomly sampled from the surface. The random variable d obeys the normal distribution with mean $\langle d \rangle$ and variance σ_d^2 as follows*

$$\begin{cases} \langle d \rangle = \gamma\sqrt{2}(1 - \frac{1}{8n} + O(\frac{1}{n^2})), \\ \sigma_d^2 = \gamma^2(\frac{1}{2n} + \frac{7}{16n^2} + O(\frac{1}{n^3})). \end{cases} \tag{15}$$

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their helpful comments and constructive suggestions. The work was supported by Natural Science Foundation of China under grants no. 61602315, 61672357 and U1713214, the Science and Technology Project of Guangdong Province under grant no. 2020A1515010707 and 2018A050501014, the Science and Technology Innovation Commission of Shenzhen under grant no. JCYJ20190808165203670, the SZU Branch of Shenzhen Institute of Artificial Intelligence and Robotics for Society.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778.

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

[3] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 1320-1329.

[4] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 4004-4012.

[5] Hermans A, Beyer L, and Leibe B, "In defense of the triplet loss for person re-identification," in *arXiv:1703.07737*, 2017.

[6] Y. Yuan, W. Chen, Y. Yang, and Z. Wang, "In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation," in *Proc. Comput. Vis. Pattern Recognit.*, 2020, pp. 354-355.

[7] Q. Xiao, H. Luo, and C. Zhang, "Margin sample mining loss: a deep learning based method for person re-identification," in *arXiv:1710.00478*, 2017.

[8] C. Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2859-2867.

[9] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai, "Hard-aware point-to-set deep metric for person re-identification," in *The European Conference on Computer Vision*, 2018, pp. 188-204.

[10] S. Zhou, F. Wang, Z. Huang, and J. Wang, "Discriminative feature learning with consistent attention regularization for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, 2019, pp. 8040-8049.

[11] W. Ge, W. Huang, D. Dong, and M. R. Scott, "Deep metric learning with hierarchical triplet loss," in *The European Conference on Computer Vision*, 2018, pp. 269-285.

[12] B. Yu, T. Liu, M. Gong, C. Ding, and D. Tao, "Correcting the triplet selection bias for triplet loss," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 1-17.

[13] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 3741-3750.

[14] Y. Yuan, K. Yang, and C. Zhang, "Hard-aware deeply cascaded embedding," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 814-823.

[15] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1857-1865.

[16] R. Hadsell, S. Chopra, and Y. Lecun, "Dimensionality reduction by learning an invariant mapping," in *Proc. Comput. Vis. Pattern Recognit.*, 2006, pp. 1735-1742.

[17] J. Wang, Z. Wang, C. Gao, N. Sang, and R. Huang, "DeepList: Learning deep features with adaptive listwise constraint for person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 513-524, 2017.

[18] Y. Li, Z. Jia, J. Zhang, K. Huang, and T. Tan, "Deep semantic structural constraints for zero-shot learning," in *Proc. Conf. AAAI Artif. Intell.*, 2018, pp. 7049-7056.

[19] J. Wang, S. Zhou, J. Wang, and Q. Hou, "Deep ranking model by large adaptive margin learning for person re-identification," *Pattern Recog.*, vol. 74, no., pp. 241-252, 2017.

[20] B. Chen and W. Deng, "Deep embedding learning with adaptive large margin N-pair loss for image retrieval and clustering," *Pattern Recog.*, vol. 93, no., pp. 353-364, 2019.

[21] S. Zhang, L. Xie, W. Wu, H. Yu, and Z. Zhu, "Identity-enhanced

network for facial expression recognition," in *Asian Conference on Computer Vision*, 2018, pp. 534-550.

[22] Y. Kim, B. I. Yoo, Y. Kwak, C. Choi, and J. Kim, "Deep generative-contrastive networks for facial expression recognition," in *arXiv:1703.07140*, 2017.

[23] J. Wang and C. Yuan, "Facial expression recognition with multi-scale convolution neural network," in *Proc. Pacific-Rim Conf. Multimed.*, 2016, pp. 376-385.

[24] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proc. Comput. Vis. Pattern Recognit. Workshop*, 2017, pp. 522-531.

[25] S. Li and W. Deng, "Deep facial expression recognition: a survey," *IEEE Trans. Affect. Comput.*, doi: 10.1109/TAFFC.2020.2981446.

[26] L. Wu, Y. Wang, J. Gao, and X. Li, "Deep adaptive feature embedding with local sample distributions for person re-identification," *Pattern Recog.*, vol. 73, no., pp. 275-288, 2018.

[27] J. Hu, J. Lu, Y. Tan, J. Yuan and J. Zhou, "Local large-margin multi-metric learning for face and kinship verification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1875-1891, 2018.

[28] M. Kyperountas, A. Tefas, and I. Pitas, "Salient feature and reliable classifier selection for facial expression classification," *Pattern Recog.*, vol. 43, no. 3, pp. 972-986, 2010.

[29] Y. Tian, Z. Wen, W. Xie, X. Zhang, L. Shen, and J. Duan, "Outlier-suppressed triplet loss with adaptive class-aware margins for facial expression recognition," in *Proc. Int. Conf. Image Process.*, 2019, pp. 46-50.

[30] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. Int. Conf. Multimed. Expo*, 2005, pp. 5.

[31] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," in *arXiv:1703.09507*, 2017.

[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.*, 2016, pp. 499-515.

[33] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial features," in *Proc. Int. Conf. Pattern Recog. Workshop*, 2004, pp. 17-25.

[34] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1021-1030.

[35] Y. Xia, B. Zhang, and F. Coenen, "Face occlusion detection based on multi-task convolution neural network," in *Proc. Int. Conf. Fuzzy Syst. Knowl. Discov.*, 2015, pp. 375-379.

[36] V. K. B. G, G. Carneiro, and I. Reid, "Learning local image descriptors with deep Siamese and triplet convolutional networks by minimising global loss functions," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 5385-5394.

[37] The sphere game in n dimensions, http://faculty.madisoncollege.edu /alehnen/sphere/hypers.htm, 2018.

[38] W. Xie, L. Shen, and J. Duan, "Adaptive weighting of handcrafted feature losses for facial expression recognition," *IEEE Trans. Cybern.*, doi: 10.1109/TCYB.2019.2925095.

[39] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. Int. Conf. Comput. Vis. Workshop*, 2011, pp. 2106-2112.

[40] T. Kanade, Y. Tian, and J. F. Cohn, "Comprehensive database for facial expression analysis," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2000, pp. 46-53.

[41] 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, 2017.

[42] G. Zhao, X. Huang, M. Taini, and S. Z. Li, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607-619, 2011.

[43] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 211-216.

[44] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 2168-2177.

[45] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1-10.

[46] Y. Tang, "Deep learning using linear support vector machines," in *arXiv:1306.0239*, 2013.

[47] J. Yan, W. Zheng, C. Zhen, C. Tang, and S. Ning, "Multi-clue fusion for emotion recognition in the wild," in *Int. Conf. Multimodal Interact.*, 2016, pp. 458-463..

[48] W. Ding, M. Xu, D. Huang, W. Lin, M. Dong, X. Yu, and H. Li, "Audio and face video emotion recognition in the wild using deep neural networks and small datasets," in *Int. Conf. Multimodal Interact.*, 2016, pp. 506-513.

[49] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Int. Conf. Multimodal Interact.*, 2016, pp. 445-450.

[50] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1-12, 2015.

[51] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-Gated CNN for occlusion-aware facial expression recognition," in *Proc. ICPR, Aug.*, 2018, pp. 2209-2214.

[52] 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, 2019.

[53] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European conference on computer vision*, 2018, pp. 222-237.

[54] 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.

[55] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 1749-1756.

[56] Y. Guo, G. Zhao, and M. Pietikainen, "Dynamic facial expression recognition using longitudinal facial expression atlases," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 631-644.

[57] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 2983-2991.

[58] X. Liu, Kumar, P. Jia and J. You, "Hard negative generation for identity-disentangled facial expression recognition", *Pattern Recog.*, vol. 88, no., pp. 1-12, 2020.

[59] Z. Meng, L. Ping, C. Jie, S. Han, and T. Yan, "Identity-aware convolutional neural network for facial expression recognition," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 558-565.

[60] X. Yang, H. Di, Y. Wang, and L. Chen, "Automatic 3D facial expression recognition using geometric scattering representation," in *Proc. Int. Conf. Autom. Face Gesture Recognit. Workshop*, 2015, pp. 1-6.

[61] A. T. Lopes, E. De Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," *Pattern Recog.*, vol. 61, no., pp. 610-628, 2017.

[62] W. Xie, X. Jia, L. Shen, and M. Yang, "Sparse deep feature learning for facial expression recognition," *Pattern Recog.*, vol. 96, no., pp. 106966, 2019.

Weicheng Xie is currently an assistant professor at School of Computer Science & Software Engineering, Shenzhen University, China. He received the B.S. degree in statistics from Central China Normal University in 2008, the M.S. degree in probability and mathematical statistics and Ph.D. degree in computational mathematics from Wuhan University, China in 2010 and 2013. He has been a visiting research fellow with School of Computer Science, University of Nottingham, UK. His current researches focus on image processing and facial expression synthesis and recognition.

Haoqian Wu received the B.Sc. degree in college of Computer Science & Technology from Southwest University of Science and Technology in 2019. Now he is working toward the M.Sc. degree in School of Computer Science & Software Engineering, Shenzhen University. His research interests include facial expression recognition and transfer learning.

Yi Tian received the B.Sc. degree in college of Communications and Information Engineering from University of Electronic Science and Technology of China. She received M.Sc. degree in computer science from School of Computer Science & Software Engineering, Shenzhen University. Her research interests include facial expression recognition and image data augmentation.

Mengchao Bai received the B.Sc. degree in college of Electronic and Information Engineering from Shenzhen University. He received M.Sc. degree in computer science from School of Computer Science & Software Engineering, Shenzhen University. His research interests include facial expression recognition and image translation.

Linlin Shen is currently a professor at School of Computer Science & Software Engineering, Shenzhen University, China. He received the B.Sc. degree from Shanghai Jiaotong University, Shanghai, China, and the Ph.D. degree from University of Nottingham, Nottingham, U.K., in 2005. He was a Research Fellow with the Medical School, University of Nottingham, researching brain image processing of magnetic resonance imaging. His research interest covers pattern recognition, medical image processing and deep learning.