Facial Expression Recognition in the Wild via Deep Attentive Center Loss

Amir Hossein Farzaneh and Xiaojun Qi
Department of Computer Science
Utah State University
Logan, UT 84322, USA
farzaneh@aggiemail.usu.edu, xiaojun.qi@usu.edu

Abstract

Learning discriminative features for Facial Expression Recognition (FER) in the wild using Convolutional Neural Networks (CNNs) is a non-trivial task due to the significant intra-class variations and inter-class similarities. Deep Metric Learning (DML) approaches such as center loss and its variants jointly optimized with softmax loss have been adopted in many FER methods to enhance the discriminative power of learned features in the embedding space. However, equally supervising all features with the metric learning method might include irrelevant features and ultimately degrade the generalization ability of the learning algorithm. We propose a Deep Attentive Center Loss (DACL) method to adaptively select a subset of significant feature elements for enhanced discrimination. The proposed DACL integrates an attention mechanism to estimate attention weights correlated with feature importance using the intermediate spatial feature maps in CNN as context. The estimated weights accommodate the sparse formulation of center loss to selectively achieve intra-class compactness and inter-class separation for the relevant information in the embedding space. An extensive study on two widely used wild FER datasets demonstrates the superiority of the proposed DACL method compared to state-of-the-art methods.

1. Introduction

Analyzing facial expressions is an active field of research in computer vision. Facial Expression Recognition (FER) is an important visual recognition technology to detect emotions given the input to the intelligent system is a facial image. FER is widely used in Human-Computer Interaction (HCI), driver monitoring for autonomous driving, education, healthcare, and psychological treatments. Recently, Deep Neural Network (DNN) approaches have demonstrated significant performance in visual recognition tasks. Notably, Convolutional Neural Network (CNN) methods [4, 12, 31, 23], as prominent deep learning techniques that automatically extract deep feature representations, have significantly outperformed conventional methods in FER [28, 37, 39, 40].

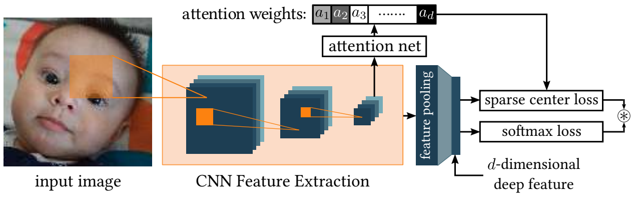

Figure 1. The high-level overview of our proposed Deep Attentive Center Loss (DACL) method: A Convolutional Neural Network (CNN) yields spatial convolutional features and a feature pooling layer extracts the final d-dimensional deep feature vector for softmax loss and sparse center loss. The last convolutional features are fed to an attention network as context to estimate the attention weights. The estimated weights guide the sparse center loss module to achieve intra-class compactness and inter-class separation for an adaptively selected subset of feature elements. $\circledast$ indicates a linear combination of softmax loss and sparse center loss.

For any visual recognition system with a fixed set of classes, the input space (*i.e.*, a 2D image) is mapped to a high-dimensional feature representation vector that captures the input image's semantics. Deep CNN-based methods extract spatial features that capture the input image's abstract semantics by composing features from lower levels to higher levels. A pooling layer then converts the spatial features into a single deep feature vector. In practice, a softmax loss estimates a probability distribution over all classes in the final stage.

Intuitively, a better recognition system is built on an efficiently discriminated space of embedded deep features. On the other hand, real-world FER applications require a massive corpus of annotated images acquired in an unconstrained environment, namely wild FER datasets [24, 14]. Accordingly, for the task of FER in the wild, where the images exhibit significant intra-class variation and inter-class similarity, feature discrimination is a critical supervision step. However, the widely used softmax loss is incapable of yielding discriminative features in wild scenarios. To address this shortcoming, Deep Metric Learning (DML) approaches constrain the embedding space to obtain well-discriminated deep features. Specifically, DML methods achieve intra-class compactness and inter-class separation by maximizing the similarity between deep features and their corresponding class prototypes in the embedding space.

In a typical DML problem, the deep feature equally contributes to the DML's objective function along all dimensions. Therefore, DML methods are prone to discriminate redundant and noisy information along with important information encoded in the deep feature vector. This leads to over-fitting and hinders the generalization ability of the learning algorithm.

To address the aforementioned shortcomings, we design a modular attention-based DML approach, called Deep Attentive Center Loss (DACL), to selectively learn to discriminate exclusively the relevant information in the embedding space. Our method is inspired by visual attention described in cognitive neuroscience as the perception of the most relevant subset of sensory data. As shown in Figure 1, given the last convolutional spatial feature map as a context, our attention network produces attention weights to guide the DML objective function with the most relevant information. A reformulation of the center loss [35], called sparse center loss, is further proposed as the DML objective function with the advantages of simplicity and straightforward computation. Since our proposed method is designed to be modular, it can be easily developed and integrated with other DML approaches.

The main contributions of our work are summarized as follows:

- We propose a novel attention mechanism that yields context-based attention weights to estimate the weighted contribution of each dimension in the DML's objective function.
- We propose the sparse center loss as the DML's objective function that uses the estimated weights obtained by the attention mechanism to selectively discriminate deep features along its dimensions in the embedding space. Sparse center loss is jointly optimized with softmax loss and can be trained using the standard Stochastic Gradient Descent (SGD).
- We show that the modular DACL method, which consists of the attention network and the sparse center loss, can be trained using the standard SGD algorithm and can therefore be promptly applied to any state-of-the-art network architectures and DML methods with minimal intervention.
- We conduct extensive experiments on two popular large-scale wild FER datasets (RAF-DB and AffectNet) to show the improved generalization ability and the superiority of the proposed modular DACL method compared to other state-of-the-art methods.

2. Related Work

In this section, we review the methods in Facial Expression Recognition (FER) from two perspectives: 1. Methods that particularly enhance FER with Deep Metric Learning (DML) and 2. FER methods that tackle the wild dataset challenges.

2.1. FER with DML

DML enhances the discrimination power of softmax loss function to tackle the large intra-class variation and inter-class similarity. Although most of the existing DML methods are developed for Face Recognition applications, FER has also enjoyed the DML benefits. Meng *et al*. [22] develop an Identity-Aware Convolutional Neural Network (IACNN) that jointly discriminates expression-related and identity-related features. Contrastive loss [8] is applied to the extracted deep features to pull those with similar labels together and push those with different labels away from each other. Similarly, Liu *et al*. [20] propose (N+M)-tuplet clusters loss function adapted from (N+1)-tuplet loss [29] and Coupled Clusters Loss (CCL) [19] to address the the difficulty of anchor selection in triplet loss [6]. Particularly, inputs are mined as a set of N positive samples and a set of M negative samples. During training, the samples in the negative set are moved away from the center of positive samples, and the positive samples are simultaneously clustered around their corresponding center. Locality-Preserving loss (LP-loss) [13], inspired by center loss [35], is embedded in a Deep Locality-Preserving CNN (DLP-CNN) to enforce intra-class compactness by locally clustering deep features using the k-nearest neighbor algorithm. Cai *et al*. [1] improve on center loss by adding an extra objective function called Island loss to achieve intra-class compactness and inter-class separation simultaneously. Island loss maximizes the cosine distance between the class centers in the embedding space. Similarly, Li *et al*. [15] propose separate loss as a cosine version of center loss and Island loss. The intra loss and inter loss in separate loss maximize the cosine similarity between the features belonging to a class and minimize the cosine similarity between the class centers in the embedding space. Li *et al*. [18] propose a multi-scale CNN with an attention mechanism to learn the importance of different convolutional receptive fields in the network. Additionally, softmax loss is jointly supervised with a regularized version of the center loss to incorporate a distance margin while discriminating features in the embedding space. Farzaneh and Qi [3] propose a discriminant distribution-agnostic loss (DDA loss) to implicitly enforce inter-class separation for both majority and minority classes under extreme class imbalance scenarios. Specifically, DDA loss regulates the Euclidean distance of a sample among all classes in the embedding space during forward propagation.

2.2. FER in the Wild

Methods that are developed for real-world FER applications use a large-scale dataset with a wild attribute that exhibit a diverse spectrum of subjects in an unconstrained environment. Li *et al*. [16, 17] propose CNN methods with attention mechanism, namely patch-based Attention CNN (pACNN) and global-local-based Attention (gACNN), to tackle the face occlusion challenge associated with wild FER datasets. The attention mechanism estimates a weight for each local patch in the feature space correlating to their obstructed-ness and a global weight for the whole feature map. Intuitively, occluded patches are assigned with small weights. pACNN concatenates only the weighted local patches while gACNN also incorporates the weighted global feature in concatenation to represent the input image.

Alternatively, Zhao *et al*. [38] introduce a Feature Selection Network (FSN) that automatically filters out irrelevant features in the network. FSN calculates the local influence of features to yield a filter mask. Additionally, a face mask that filters out the features corresponding to the areas beyond the face is generated. The two generated masks adjust the final feature to represent the input image. Pan *et al*. [25] tackle occlusion by training a CNN on non-occluded images to guide the output of an identical CNN on their corresponding occluded image. The output of the former network guides the latter network's output using the joint supervision of four different loss functions in the label space and the feature space. Wang *et al*. [34] design a Region Attention Network (RAN) to address pose and occlusion in wild FER datasets by passing regions around facial landmarks for a single image to a CNN. The final feature vector is obtained by combining weighted feature vectors of cropped regions using a self-attention module.

Florea *et al*. [5] combine semi-supervised learning and inductive transfer learning into an Annealed Label Transfer (ALT) framework to tackle the label scarcity issue. ALT transfers a learner's knowledge on a labeled wild FER dataset to an unlabeled face dataset to generate pseudo labels. The pseudo label's confidence is increased to enhance the primary learner's performance in recognition. Zeng *et al*. [36] propose Inconsistent Pseudo Annotations to Latent Truth (IPA2LT) to address label noise issue and alleviate prediction bias to a specific wild dataset. IPA2LT trains a Latent Truth Network (LTNet) to extract the true latent label for a sample using the inconsistency between the labels generated with a prediction model and manual labels. Wang *et al*. [33] address label uncertainty by proposing a Self-Cure Network (SCN) to re-label the mis-labeled samples. A self-attention mechanism estimates a weight for each sample in a batch based on the network's prediction and identifies label uncertainty using a margin-based loss function.

3. Proposed Method

In this section, we briefly review the necessary preliminaries related to our work. We then introduce the two building blocks of our proposed Deep Attentive Center Loss (DACL) method, namely, the sparse center loss and the attention network. Finally, we discuss how DACL is trained and optimized with the standard Stochastic Gradient Descent (SGD).

3.1. Preliminaries

Given a training mini-batch of m samples $D_m = \{(X_i, y_i)|i = 1, ..., m\}$, where X_i is the input, and $y_i \in \{1, ..., K\}$ is its corresponding label for a K-class classification problem, let the spatial feature map $x_i^* \in \mathbb{R}^{N_C \times N_H \times N_W}$ be the output of a Convolutional Neural Network (CNN). A pooling layer $\mathcal{P}$ (*e.g*., fully-connected layer or average pooling layer) takes x_i^* as input and extracts a d-dimensional deep feature $x_i \in \mathbb{R}^d$.

The conventional softmax loss combines a fully-connected layer, softmax function, and the cross-entropy loss to estimate a probability distribution over all classes and measures the prediction error. The deep feature x_i as input to the fully-connected layer is mapped to a raw score vector $z_i = [z_{i1}, ..., z_{iK}]^T \in \mathbb{R}^{K \times 1}$ through a linear transformation as follows:

$$z_i = W^T x_i + B \tag{1}$$

where $W = [w_1, ..., w_K] \in \mathbb{R}^{d \times K}$ and $B = [b_1, ..., b_K] \in \mathbb{R}^{K \times 1}$ are the class weights and bias parameters for the fully-connected layer, respectively. A probability distribution $p(y = j|x_i)$ is then calculated over all classes using the softmax function. Finally, the cross-entropy loss function computes the discrepancy between prediction and the true label y_i to formulate the softmax loss function $\mathcal{L}_S$ as follows:

$$\begin{aligned} \mathcal{L}_S &= -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{K} y_i \log p(y = j|x_i) \\ &= -\frac{1}{m} \sum_{i=1}^{m} \log \frac{e^{w_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{K} e^{w_j^T x_i + b_j}} \end{aligned} \tag{2}$$

3.2. Sparse Center Loss

Center loss is a widely adopted DML method where the similarity is measured between the deep features and their

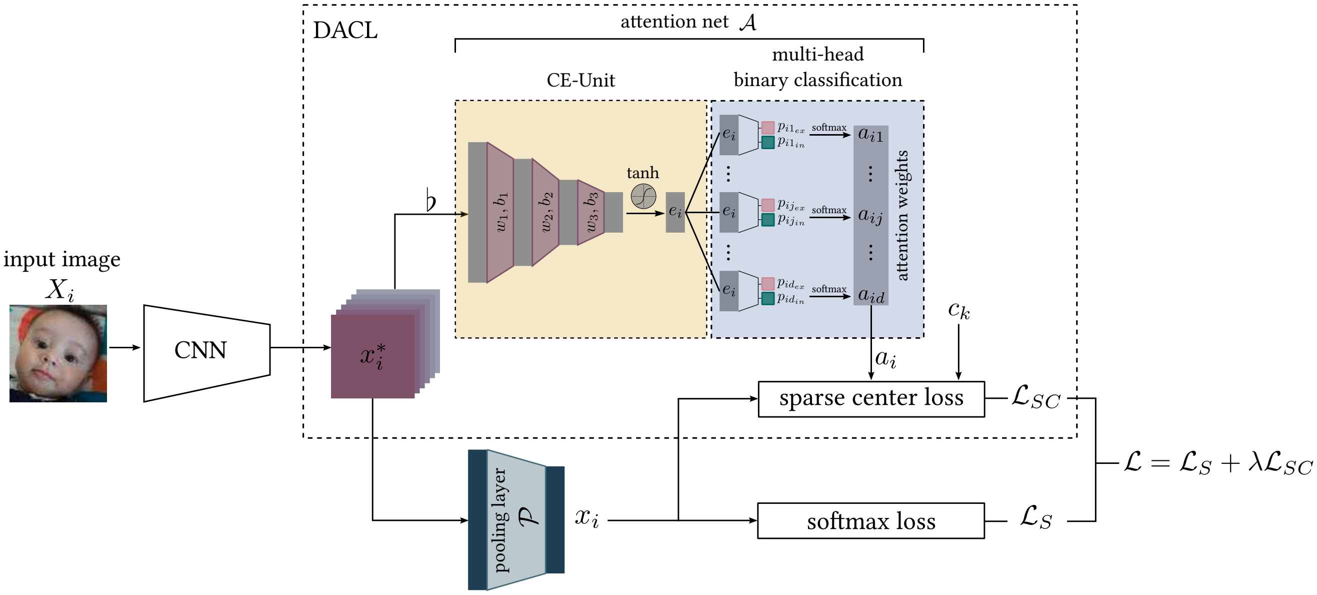

Figure 2. The illustration of the proposed DACL method. An input image X_i is fed to the CNN to yield the convolutional spatial feature map x_i^*. DACL is a hybrid combination of an attention network $\mathcal{A}$ and a sparse center loss. The CE-Unit in DACL's attention mechanism takes the spatial feature map as a context and yields an encoded latent feature vector e_i to eliminate noise and irrelevant information. A multi-head binary classification module then calculates the attention weight a_{ij} corresponding to the j-th dimension in the deep feature x_i. Finally, the sparse center loss $\mathcal{L}_{SC}$ calculates a weighted WCSS and is fractionally accumulated with the softmax loss $\mathcal{L}_S$ to compose the final loss $\mathcal{L}$.

corresponding class centers (class prototypes). The objective function in center loss minimizes the Within Cluster Sum of Squares (WCSS) between the deep features and their corresponding class centers. That is, it aims to partition the embedding space into K clusters for a K-class classification problem. Given a training mini-batch of m samples, let $\mathbf{x_i} = [x_{i1}, x_{i2}, ..., x_{id}]^T \in \mathbb{R}^d$ be the i-th sample deep feature vector belonging to the y_i-th class, where $y_i \in \{1, ..., K\}$ and $\mathbf{c_{y_i}} = [c_{y_i1}, ..., c_{y_id}]^T \in \mathbb{R}^d$ be its corresponding class center. Center loss minimizes the following criterion defined as:

$$\mathcal{L}_C = \frac{1}{2m} \sum_{i=1}^{m} \sum_{j=1}^{d} \|x_{ij} - c_{y_ij}\|_2^2 \tag{3}$$

where WCSS is minimized by equally penalizing the Euclidean distance between the deep features and their corresponding class centers in the embedding space.

We argue that not all the elements in a feature vector are relevant to discrimination. Our goal is to select only a subset of elements in a deep feature vector to contribute in the discrimination. Accordingly, to filter out irrelevant features in the discrimination process, we weight the calculated Euclidean distance at each dimension in Eq. 3 and develop a a sparse center loss method as follows:

$$\mathcal{L}_{SC} = \frac{1}{2m} \sum_{i=1}^{m} \sum_{j=1}^{d} a_{ij} \odot \|x_{ij} - c_{y_ij}\|_2^2 \tag{4}$$
$$\text{subject to} \quad 0 < a_{ij} \leq 1 \quad \forall j, \quad (j = 1, ..., d).$$

where $\odot$ indicates element-wise multiplication and a_{ij} denotes the weight of the i-th deep feature along the dimension $j \in \{1, ..., d\}$ in the embedding space. Intuitively, the sparse center loss calculates a weighted WCSS. It should be noted that Eq. 4 reduces to the standard center loss in Eq. 3 if $a_{i1} = ... = a_{id}$.

3.3. Attention Network

We design an auxiliary attention network attached to the CNN to dynamically estimate the weights $a_i \in \mathbb{R}^d$ for the sparse center loss based on the input. Specifically, we seek an adaptive and flexible approach to estimate the weights for the sparse center loss that adjusts to the task and the input data. Ideally, we require the weights to be determined by a neural network. For this purpose, we propose an attention network $\mathcal{A}$ that adaptively computes an attention weight vector to govern the contribution of deep feature x_i along the j-th dimension in Eq. 4. This attention network together with the sparse center loss comprises the two building blocks of the proposed DACL method. Figure 2 presents the proposed attention network in DACL. It has two major components: 1. The Context Encoder Unit (CE-Unit), which takes the spatial feature map from the CNN as input (context) and generates a latent representation and 2. The multi-head binary classification module that takes the latent representation and estimates the attention weights. It should be emphasized that the context for the attention network is at the convolutional feature-level to preserve the spatial information.

We build a dense CE-Unit by stacking three trainable fully-connected linear layers to extract exclusively relevant information from the context as follows:

$$\begin{aligned} e_i = \tanh(\text{BN}(W_3^T \text{relu}(\text{BN}(W_2^T \text{relu}(\text{BN}(...\\ ...W_1^T \flat(x_i^*) + b_1)) + b_2)) + b_3)) \end{aligned} \tag{5}$$

where x_i^* is the last convolutional feature map in the CNN *i.e.*, the context feature for the i-th sample, the operator $\flat : \mathbb{R}^{1\times N_C \times N_H \times N_W} \rightarrow \mathbb{R}^{1\times \mathcal{N}_C \mathcal{N}_\mathcal{H} \mathcal{N}_\mathcal{W}}$ flattens the convolutional feature map, W_l and b_l are respectively the weights and biases for l-th linear layer in the attention network where $l = 1, 2, 3$. Layers are interjected with batch normalization BN$(.)$ [11] and rectified linear units relu$(.)$ to capture non-linear relationships between layers. The final hyperbolic tangent function $\tanh(.)$ as element-wise nonlinearity preserves both positive and negative activation values for a smoother gradient flow in the network. We initialize the linear layer weights using the *He* initialization method [9], and the biases are initialized to 0. The CE-Unit defined in Eq. 5 extracts an encoded latent feature vector $e_i \in \mathbb{R}^{d' \ll d}$ for the i-th sample in a lower dimension to eliminate irrelevant information while keeping the important information. The CE-Unit is adjustable in terms of layer parameters to match a specific task.

To estimate the attention weight of the j-th dimension correlating to the d-dimensional deep feature x_i at dimension j, we attach a multi-head binary classification (inclusion/exclusion) module to the CE-Unit. The latent d'-dimensional feature vector e_i is shared among d linear units, *i.e.*, heads with two outputs each, to calculate two raw scores for the deep feature x_i along dimension j as follows:

$$\begin{aligned} p_{ij_{in}} &= A_{j_{in}}^T e_i + b_{j_{in}} \\ p_{ij_{ex}} &= A_{j_{ex}}^T e_i + b_{j_{ex}} \end{aligned} \tag{6}$$

where $A_j \in \mathbb{R}^{d' \times 2}$ and $b_j \in \mathbb{R}^2$ are the learnable weights and biases for each classification head with subscript in representing inclusion and subscript ex representing exclusion, and $p_{ij_{in}}$ and $p_{ij_{ex}}$ denote the inclusion and exclusion scores for the j-th dimension in x_i, respectively. A softmax function is applied on each head's output to normalize the scores subject to the constraint in Eq. 4. Finally, the corresponding attention weight a_{ij} is calculated as follows:

$$a_{ij} = \frac{\exp(p_{ij_{in}})}{\exp(p_{ij_{in}}) + \exp(p_{ij_{ex}})} \tag{7}$$

The differentiable softmax function employed on the raw scores limits the value of the estimated attention weights in the range $(0, 1]$.

3.4. Training and Optimization

Our proposed DACL method as illustrated in Figure 2 is trained in an end-to-end manner, where the sparse center loss is jointly supervised with softmax loss to compose the final loss as follows:

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_{SC} \tag{8}$$

where λ controls the contribution of the sparse center loss $\mathcal{L}_{SC}$ to the total loss $\mathcal{L}$. The parameters associated with DACL can be optimized using the standard SGD algorithm. The gradient of the sparse center loss with respect to the deep features are obtained as follows:

$$\frac{\partial \mathcal{L}_{SC}}{\partial x_i} = \frac{1}{m} a_i \odot (x_i - c_{y_i}) \tag{9}$$

and the gradient of the sparse center loss with respect to the attention weights are obtained as follows:

$$\frac{\partial \mathcal{L}_{SC}}{\partial a_i} = \frac{1}{2m} \|x_i - c_{y_i}\|_2^2 \tag{10}$$

The centers c_k are initialized using the *He* initialization method and are updated according to a moving average strategy as follows:

$$\Delta_{c_k} = \frac{\sum_{i=1}^m \delta_{y_i j} a_i \odot (c_j - x_i)}{\epsilon + \sum_{i=1}^m \delta_{y_i j}} \tag{11}$$

where the Kronecker delta function is defined as $\delta_{ij} = 1$ for $i = j$ and 0 otherwise. The gradients with respect to the context feature x_i^* is trivially calculated according to the chain rule. We summarize training a supervised learning algorithm (*e.g.*, prediction model) with DACL in Algorithm 1.

4. Experiments

In this section, we first describe two publicly available wild FER datasets, *i.e.*, Affect from the Internet (AffectNet) [24] and Real-world Affective Face Database (RAF-DB) [14]. Then, we conduct extensive experiments on these two widely used wild Facial Expression Recognition (FER) datasets to demonstrate the superior performance of our proposed Deep Attentive Center Loss (DACL). We evaluate our method on the wild FER datasets compared with two baselines (softmax loss and center loss) and various state-of-the-art methods. Finally, we visualize the learned attention weights to interpret our model intuitively.

Algorithm 1 Training a supervised learning algorithm (*e.g.*, prediction model) with DACL.

```
Input: Training dataset $D = \{(X_i, y_i)|i = 1, ..., N\}$; Initialized CNN parameters $\theta_C$, pooling layer parameters $\theta_P$, attention network parameters $\theta_A$, softmax loss FC layer $\theta_S$, and centers $C = \{c_k|k = 1, ..., K\}$; Hyper-parameters $\alpha$, $\lambda$, and learning rate $\mu$; The number of iterations $t \leftarrow 0$.
Output: Updated parameters $\theta_C$, $\theta_P$, $\theta_A$, $\theta_S$, and $C$.
while not converged do
    Sample a mini-batch of size $m$ from the training dataset $D_m = \{(X_i, y_i)|i = 1, ..., m\}$;
    Compute the context features $\{x_i^*|i = 1, ..., m\}$ using the CNN.
    Compute the deep features $\{x_i|i = 1, ..., m\}$ with the pooling layer.
    Compute the attention weights $\{a_i|i = 1, ..., m\}$ by Eq. 5 - 7.
    Compute the softmax loss $\mathcal{L}_S^t$ by Eq. 2.
    Compute the sparse center loss $\mathcal{L}_{SC}^t$ by Eq. 4.
    Compute the total loss by Eq. 8: $\mathcal{L}^t = \mathcal{L}_S^t + \lambda \mathcal{L}_{SC}^t$.
    Compute the softmax loss gradients:
        $\hat{g}_S^t \leftarrow \frac{\partial \mathcal{L}_S^t}{\partial \theta_S}$
    Compute the pooling layer gradients:
        $\hat{g}_P^t \leftarrow \frac{1}{m} \sum_{i=1}^{m} \frac{\partial x_i^t}{\partial \theta_P} \frac{\partial \mathcal{L}_S^t}{\partial x_i^t} + \lambda \frac{\partial \mathcal{L}_{SC}^t}{\partial x_i^t}$.
    Compute the attention network gradients:
        $\hat{g}_A^t \leftarrow \frac{1}{m} \sum_{i=1}^{m} \frac{\partial a_i^t}{\partial \theta_A} (\frac{\partial \mathcal{L}_{SC}^t}{\partial a_i^t})$.
    Compute the CNN gradients:
        $\hat{g}_C^t \leftarrow \frac{1}{m} \sum_{i=1}^{m} \frac{\partial x_i^{*t}}{\partial \theta_C} (\frac{\partial \mathcal{A}^t}{\partial x_i^{*t}} + \frac{\partial \mathcal{P}^t}{\partial x_i^{*t}})$.
    Compute $\Delta c_k$ by Eq. 11.
    $t \leftarrow t + 1$.
    Update $c_k$ for each $k$: $c_k^{t+1} = c_k^t - \alpha \Delta c_k$.
    Update the model parameters:
        $\theta_{\{C,P,A,S\}}^{t+1} = \theta_{\{C,P,A,S\}}^t - \mu^t \hat{g}_{\{C,P,A,S\}}^t$
end while
```

4.1. Datasets

Compared to lab-controlled datasets such as CK+ [21], MMI [26], and Oulu-CASIA [30], wild FER datasets are acquired in an unconstrained setting offering a broad diversity across pose, gender, age, demography, image quality, and illumination. RAF-DB and AffectNet are two widely used wild FER datasets in research.

RAF-DB contains 30,000 facial images acquired using crowd-sourcing techniques. Images are annotated with categorical and compound expressions by 30 trained human annotators. For our RAF-DB experiments, we only use the 12,271 training images and 3,068 images in the test set with six discrete basic expressions identified by Ekman and Friesen [2] (*i.e.*, *happy*, *sad*, *surprise*, *anger*, *fear*, and *disgust*) and *neutral* expression.

AffectNet is the largest publicly available wild FER dataset with 450,000 facial images acquired from the internet and manually annotated with categorical expressions and dimensional affect (valence and arousal). For our experiments, we use 280,000 training images and 3,500 images in the validation set annotated with six basic expressions and *neutral* expression. Since the test set is not released by the authors, we use the validation set for our evaluations. Following state-of-the-art FER methods, we exclude the *contempt* expression in our experiments.

4.2. Implementation Details

We use ResNet-18 [10], a standard Convolutional Neural Network (CNN), as our backbone architecture in our experiments. Since FER's domain is close to the Face Recognition task, we pre-train ResNet-18 on MS-CELEB-1M [7], a face dataset with 10 million images of nearly 100,000 subjects. We use the standard Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. We augment the input images on-the-fly by extracting random crops (one central, and one for each corner and their horizontal flips). At test time, we use the central crop of the input image. Crops of size 224×224 are extracted from the input images with size 256×256. We train ResNet-18 on RAF-DB for 60 epochs with an initial learning rate of 0.01 decayed by a factor of 10 every 20 epochs. Alternatively, we train ResNet-18 on AffectNet for 20 epochs with an initial learning rate of 0.01 decayed by a factor of 5 every five epochs. We use a batch size of 128 for both datasets. The hyper-parameters α and λ are empirically set as 0.5 and 0.01, respectively.

With our specific backbone architecture setup, the deep feature x_i is 512-dimensional, the last convolutional feature map x_i^* is of size $512 \times 7 \times 7$ and the pooling layer is the standard 2D average pooling layer in ResNet-18. The CE-Unit in DACL is designed by stacking three fully-connected layers with 3,584, 512, and 64 channels, respectively. Hence, the latent feature vector e_i is 64-dimensional. Accordingly, we have 512 heads in our multi-head binary classification module that yields a 512-dimensional attention weight vector. We train our models using the PyTorch deep learning framework [27] on an NVIDIA 2080Ti GPU with 11GB of V-RAM. The source code is publicly available at https://github.com/amirhfarzaneh/dacl.

4.3. Recognition Results

We present wild FER results in Table 1 and Table 2 for RAF-DB and AffectNet, respectively. Unlike AffectNet, RAF-DB's test set is imbalanced. Therefore, we report the average accuracy, which is the mean of diagonal values in the confusion matrix alongside the standard accuracy across all classes for RAF-DB.

Table 1. Expression recognition performance of various methods on RAF-DB test set in terms of standard accuracy and average accuracy.

Method	Acc. (%)	Avg. Acc. (%)
FSN [38]	81.10	72.46
pACNN [16]	83.27	-
DLP-CNN [14]	84.13	74.20
ALT [5]	84.50	76.50
gACNN [17]	85.07	-
separate loss [15]	86.38	77.25
IPA2LT [36]	86.77	-
RAN [34]	86.90	-
DDA loss [3]	86.90	79.71
SCN [33]	87.03	-
softmax loss	86.54	79.43
center loss [35]	87.06	79.71
DACL	**87.78**	**80.44**

Our DACL method outperforms our baseline methods and other state-of-the-art methods and achieves a recognition accuracy of 87.78% and an average recognition accuracy of 80.44% on RAF-DB. Similarly, DACL outperforms the baseline methods and other-state-of-the-art methods on AffectNet with an accuracy of 65.20%. We also notice that DACL improves both baseline methods by a larger margin compared to the margin of improvement by center loss over softmax loss. In other words, center loss improves on softmax loss, but the generalization ability is sub-optimal. However, our proposed DACL significantly improves the generalization ability of the center loss. We depict some correctly classified and misclassified sample images from both wild FER datasets by the DACL method in Figure 3.

We present the confusion matrices obtained by the baseline methods (softmax loss and center loss) and our proposed DACL framework on both wild FER datasets in Figure 4 to evaluate the recognition accuracy of individual classes. DACL boosts the recognition accuracy of all classes in RAF-DB's test set except for *surprise* and *disgust* when comparing with softmax loss. The overall performance of DACL on RAF-DB is better since the recognition accuracy of *surprise*, *fear*, and *disgust* is significantly higher than center loss. We notice that DACL outperforms the baseline methods on AffectNet except for the *angry* class while the recognition accuracy of *sad* and *disgust* classes are significantly higher than both baselines. Overall, DACL outperforms baseline methods across all classes in RAF-DB and AffectNet.

4.4. Attention Weights Visualization

To demonstrate the interpretability of our proposed approach, we illustrate the 512-dimensional attention weights in Figure 5. We randomly select two learned attention weight vectors from the *neutral* class, and three learned attention weights from the *surprise* class. It is clear that the learned attention weights from the same classes follow very similar patterns, and the attention weights from different classes are not similar. For instance, both *neutral* samples exhibit attention weights that are filtered out around dimensions 0, 150, 190, 480, and 500. On the other hand, all samples from the *surprise* class depict attention weights that are filtered out around dimensions 50, 140, 220, and 480. Evidently, the *surprise* 2 and *surprise* 3 samples have learned almost identical attention weights. Consequently, we can verify that DACL adaptively learns the contribution

Table 2. Expression recognition performance of various methods on AffectNet validation set in terms of accuracy.

Method	Accuracy (%)
pACNN [16]	55.33
IPA2LT [36]	57.31
IPFR [32]	57.40
gACNN [17]	58.78
separate loss [15]	58.89
DDA loss [3]	62.34
softmax loss	63.86
center loss [35]	64.09
DACL	**65.20**

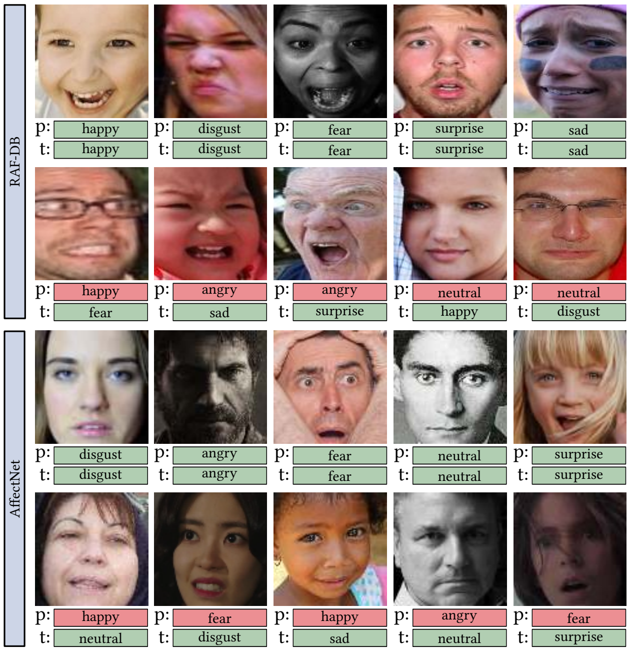

Figure 3. Sample correctly classified and misclassified images from RAF-DB and AffectNet from the model trained with DACL method. ”p” is for prediction and ”t” is for true label.

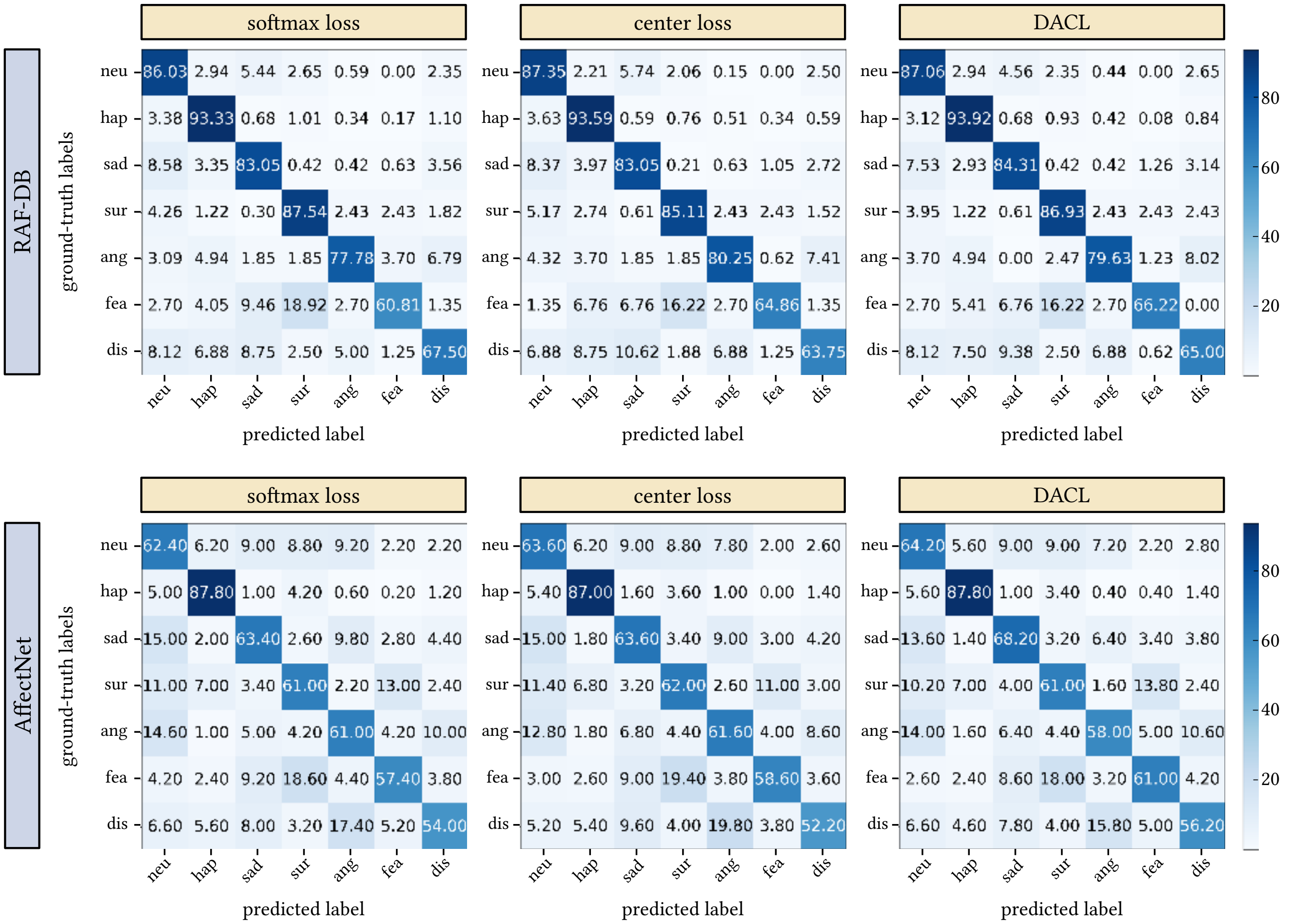

Figure 4. Confusion matrices obtained by baseline methods (softmax loss and center loss) and the proposed DACL framework on : ***top row:*** RAF-DB's test set, and ***bottom row:*** AffectNet's validation set.

of features along each dimension in the DML's objective function.

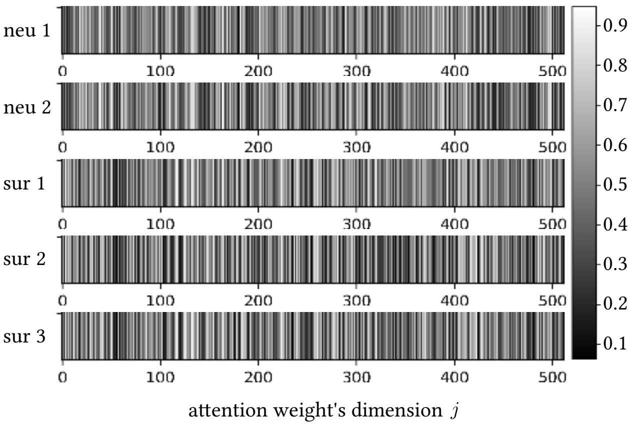

Figure 5. Illustration of learned attention weights for five different samples.

5. Conclusions

In this paper, we propose a flexible method called Deep Attentive Center Loss (DACL) for Facial Expression Recognition (FER) under wild scenarios. Our hybrid approach takes advantage of a sparse re-formulation of center loss to adaptively control the contribution of the deep feature representations in the Deep Metric Learning's objective function. Additionally, an attention mechanism that is fully parameterized by a customizable neural network estimates the probability of contribution along all dimensions by providing attention weights to the sparse center loss. We empirically show that DACL outperforms our baseline methods (softmax loss and center loss) and other state-of-the-art methods on two wild FER datasets, namely RAF-DB and AffectNet.

DACL can be easily customized to solve other classification tasks to increase feature discrimination. Moreover, the proposed approach is easily extensible with other deep metric learning (DML) objective functions.

References

[1] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island Loss for Learning Discriminative Features in Facial Expression Recognition. In *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 302–309, 2018.

[2] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.

[3] Amir Hossein Farzaneh and Xiaojun Qi. Discriminant distribution-agnostic loss for facial expression recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1631–1639, 2020.

[4] Beat Fasel. Robust face analysis using convolutional neural networks. In *Object Recognition Supported by User Interaction for Service Robots*, volume 2, pages 40–43 vol.2, 2002.

[5] Corneliu Florea, Laura Florea, Mihai Alexandru Badea, and Constantin Vertan. Annealed label transfer for face expression recognition. In *British Machine Vision Conference (BMVC)*, 2019.

[6] Yanan Guo, Dapeng Tao, Jun Yu, Hao Xiong, Yaotang Li, and Dacheng Tao. Deep Neural Networks with Relativity Learning for facial expression recognition. In *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–6, 2016.

[7] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision (ECCV)*, 2016.

[8] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality Reduction by Learning an Invariant Mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1735–1742, 2006.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, Santiago, Chile, 2015. IEEE.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[11] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[12] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, et al. Combining modality specific deep neural networks for emotion recognition in video. In *ACM International Conference on Multimodal Interaction (ICMI)*, pages 543–550, 2013.

[13] Shan Li and Weihong Deng. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019.

[14] Shan Li, Weihong Deng, and JunPing Du. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017.

[15] Yingjian Li, Yao Lu, Jinxing Li, and Guangming Lu. Separate loss for basic and compound facial expression recognition in the wild. In *Asian Conference on Machine Learning (ACML)*, volume 101, pages 897–911, 2019.

[16] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Patch-Gated CNN for Occlusion-aware Facial Expression Recognition. In *International Conference on Pattern Recognition (ICPR)*, pages 2209–2214, 2018.

[17] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2019.

[18] Zhenghao Li, Song Wu, and Guoqiang Xiao. Facial Expression Recognition by Multi-Scale CNN with Regularized Center Loss. In *International Conference on Pattern Recognition (ICPR)*, pages 3384–3389, 2018.

[19] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep Relative Distance Learning: Tell the Difference between Similar Vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2167–2175, 2016.

[20] Xiaofeng Liu, B. V. K. Vijaya Kumar, Jane You, and Ping Jia. Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 522–531, 2017.

[21] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 94–101, 2010.

[22] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-Aware Convolutional Neural Network for Facial Expression Recognition. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 558–565, 2017.

[23] Ali Mollahosseini, David Chan, and Mohammad H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016.

[24] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019.

[25] Bowen Pan, Shangfei Wang, and Bin Xia. Occluded facial expression recognition enhanced through privileged information. In *ACM International Conference on Multimedia*, pages 566–573, 2019.

[26] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo*, pages 5 pp.–, 2005.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8026–8037. 2019.

[28] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[29] Kihyuk Sohn. Improved deep metric learning with multi-class N-pair loss objective. In *International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1857–1865, 2016.

[30] Matti Taini, Guoying Zhao, Stan Z. Li, and Matti Pietikainen. Facial expression recognition from near-infrared video sequences. In *International Conference on Pattern Recognition*, pages 1–4, 2008.

[31] Yichuan Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

[32] Can Wang, Shangfei Wang, and Guang Liang. Identity- and pose-robust facial expression recognition through adversarial feature learning. In *ACM International Conference on Multimedia*, pages 238–246, 2019.

[33] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6896–6905, 2020.

[34] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020.

[35] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A Discriminative Feature Learning Approach for Deep Face Recognition. In *European Conference on Computer Vision (ECCV)*, pages 499–515, 2016.

[36] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial Expression Recognition with Inconsistently Annotated Datasets. In *European Conference on Computer Vision (ECCV)*, pages 227–243, Cham, 2018.

[37] Guoying Zhao and Matti Pietikainen. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.

[38] Shuwen Zhao, Haibin Cai, Honghai Liu, Jianhua Zhang, and Shengyong Chen. Feature selection mechanism in CNNs for facial expression recognition. In *British Machine Vision Conference (BMVC)*, 2018.

[39] Ruicong Zhi, Markus Flierl, Qiuqi Ruan, and W. Bastiaan Kleijn. Graph-Preserving Sparse Nonnegative Matrix Factorization With Application to Facial Expression Recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(1):38–52, 2011.

[40] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N. Metaxas. Learning active facial patches for expression analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2562–2569, 2012.