Discriminant Distribution-Agnostic Loss for Facial Expression Recognition in the Wild

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

Facial Expression Recognition (FER) has demonstrated remarkable progress due to the advancement of deep Convolutional Neural Networks (CNNs). FER's goal as a visual recognition problem is to learn a mapping from the facial embedding space to a set of fixed expression categories using a supervised learning algorithm. Softmax loss as the de facto standard in practice fails to learn discriminative features for efficient learning. Center loss and its variants as promising solutions increase deep feature discriminability in the embedding space and enable efficient learning. They fundamentally aim to maximize intra-class similarity and inter-class separation in the embedding space. However, center loss and its variants ignore the underlying extreme class imbalance in challenging wild FER datasets. As a result, they lead to a separation bias toward majority classes and leave minority classes overlapped in the embedding space. In this paper, we propose a novel Discriminant Distribution-Agnostic loss (DDA loss) to optimize the embedding space for extreme class imbalance scenarios. Specifically, DDA loss enforces inter-class separation of deep features for both majority and minority classes. Any CNN model can be trained with the DDA loss to yield well separated deep feature clusters in the embedding space. We conduct experiments on two popular large-scale wild FER datasets (RAF-DB and AffectNet) to show the discriminative power of the proposed loss function.

1. Introduction

Facial Expression Recognition (FER) has demonstrated substantial breakthrough results due to an explosion of research in computer vision tasks using Deep Neural Networks (DNNs). When the only piece of available information is a face, facial expressions are an essential visual channel to detect emotions. However, it is worthwhile to note that when additional contextual information is available, its analysis can be added to achieve more accurate emotion recognition [24]. Emotion recognition has been widely used in many aspects of modern society, such as healthcare, autonomous driving and driver safety, human-computer interaction, and education. The emergence of Convolutional Neural Networks (CNNs) [11] as a dominant deep learning technique offers an advanced tool for researchers to overcome the complications with large-scale FER datasets acquired in real-world scenarios [26, 14].

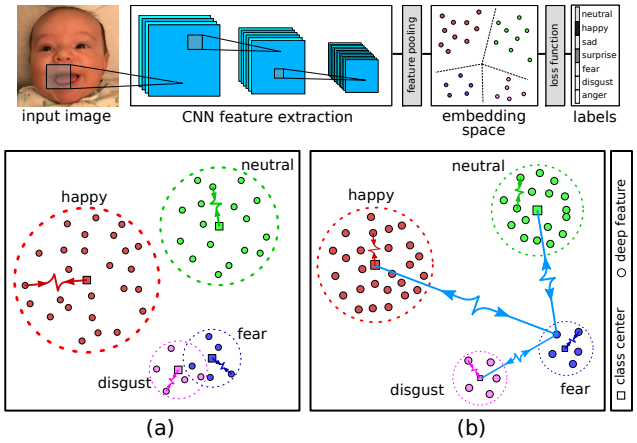


Figure 1. ***Top row:*** Illustration of the general pipeline for FER using a CNN model: CNN features are pooled in the embedding space and a loss function maps the deep features to expression labels. ***Bottom row:*** Example 2-D deep features in the embedding space learned by: (a) Center loss. (b) Discriminant Distribution-Agnostic (DDA) loss. DDA loss pushes the features of a class away from other class centers and pulls them toward their corresponding class centers to create compact and well separated feature clusters for both majority and minority classes.

While ubiquitous raw large-scale datasets have advanced research in FER, two major obstacles hinder the learning performance of deep CNNs applied in this setting: 1) Large intra-class variation and inter-class similarity, and 2) extreme class imbalance. Due to the in-the-wild attribute,

large-scale facial expression datasets acquired in an unconstrained environment inherently populate expression categories with significant variations in pose, gender, age, demography, image quality, and illumination. Additionally, facial expression categorization exhibits an intrinsic imbalance, a prevalent issue in many real-world data [7]. Commonly, categories such as *fear* and *disgust* are minority classes due to lack of representative data. Other expressions such as *neutral*, *happy*, *sad*, *surprise*, and *angry* are majority classes, which are represented with fair amount of data. The data complexity, along with extremely skewed class distribution, can severely degrade the performance of recognition models with deep CNNs.

Extracting discriminative facial features in the embedding space is a critical step towards solving the aforementioned issues. However, the widely used softmax loss is insufficient for delivering discriminant features for classification [20], [19]. Our work is motivated by Wen *et al*. [31], who pioneered center loss as a metric learning approach to yield discriminative deep features by clustering features in the embedding space. Empirically, as illustrated in Fig. 1 (a), when a CNN model is supervised by center loss in a wild dataset setting, minority classes tend to have overlapping feature clusters. Therefore, recognition performance for minority classes is sub-optimal when deep features learned by center loss are mapped to expression labels. Due to the inherent complex attributes of a wild FER dataset, optimal recognition of facial expressions requires designing new algorithms to translate raw data into an efficient representation for learning algorithms.

To learn discriminative features for FER in the wild, we propose a novel loss function, called Discriminant Distribution-Agnostic loss (DDA loss) to regulate deep features in the embedding space, where extreme class imbalance exits. The CNN models are trained under the joint supervision of softmax loss, center loss, and the proposed DDA loss. As shown in Fig. 1 (b), DDA loss creates distinctly segregated feature clusters and properly separates both majority and minority classes. Intuitively, DDA loss pushes the features of one class away from the centers of other classes and pulls them toward their corresponding class center. The discriminant deep features learned using the supervision of the DDA loss are compact and optimally separated in a d-dimensional embedding space. Consequently, the mapping from the embedding space to the label space is more efficient.

Our main contributions are summarized below:

1. We propose a novel loss function called Discriminant Distribution-Agnostic loss (DDA loss) to regulate the distribution of deep features in a d-dimensional embedding space. The proposed DDA loss implicitly maximizes the inter-class separation and minimizes intra-class variations of deep features for both majority and minority classes in extreme class imbalance scenarios. Deep CNNs trained with joint supervision of softmax loss, center loss, and DDA loss yield highly discriminant deep features for wild FER applications.
2. We show that DDA loss can be trained using the standard Stochastic Gradient Descent (SGD) algorithm and can therefore be promptly applied to any state-of-the-art network architectures with minimal intervention.
3. We conduct extensive experiments on a synthesized wild dataset and two popular large-scale wild FER datasets (AffectNet [26] and RAF-DB [14]) to demonstrate the improved recognition results with the proposed method.

2. Related Work

In this section, we review the related work on Facial Expression Recognition (FER) from two aspects: 1) FER using discriminative loss functions and 2) FER in the wild.

2.1. FER Using Discriminative Loss Functions

Metric learning has been adopted in many FER works to enhance the discrimination power of softmax loss. Identity-Aware CNN (IACNN) [25] utilizes an expression-sensitive contrastive loss on sample pairs to pull deep features with the same expression together and separate those with different expressions. Liu *et al*. [21] employ (N+M)-tuplet cluster loss on sample triplets to form clusters of deep features with the same expression and separate them from each other. However, these methods exhibit a slow convergence as the number of pairs and triplets grow in a quadratic and cubic way, respectively. Instead of sample mining, center loss [31] introduces an additional objective function coupled with softmax loss to create a compact representation of deep features by minimizing the distance between deep features to their corresponding class center. Locality-Preserving loss (LP-loss) [13], inspired by center loss, enforces intra-class compactness by locally clustering deep features using the k-nearest neighbor algorithm. Island Loss [2] introduces an additional objective function to center loss to maximize the cosine distance between the learned class centers. Separate loss [15] develops an intra-class loss and an inter-class loss to maximize the cosine similarity between deep features and their corresponding class center and minimize the cosine similarity between learned centers. Li *et al*. [18] design a multi-scale CNN to pool multiple feature vectors into a single feature embedding using the attention mechanism. Pooled features are classified using a regularized variant of center loss with a built-in margin.

2.2. FER in the Wild

FER in real-world settings requires large-scale unconstrained image datasets. Unlike lab-controlled datasets such

as CK+ [22], MMI [27], and JAFFE [23], wild FER datasets such as AffectNet [26] and RAF-DB [14] have been shown to yield more accurate models for real-world FER. Here, we review recent FER methods that tackle the wild setting.

Patch-based Attention-CNN (pA-CNN) [16] and global-local based Attention-CNN (gA-CNN) [17] integrate facial patch/region importance using an attention mechanism to recognize facial expressions in wild datasets (AffectNet and RAF-DB). Zhao *et al*. [33] propose Feature Selection Network (FSN) that employs a feature selection mechanism to automatically mask out the features with small influence for the subsequent layers in the network. FSN is evaluated on RAF-DB and FER2013 [5]. Florea *et al*. [3] develop Annealed Label Transform (ALT) algorithm to transfer a learner's knowledge on a labeled dataset (RAF-DB and FER+ [1]) to an unlabeled dataset (MegaFace [10]). By increasing the acquired pseudo label confidence, the original learner's performance in expression recognition is increased. Lee *et al*. [12] embed an attention mechanism in a Context-Aware Emotion Recognition Network (CAER-Net) to seek important visual clues in a video scene and combine them with facial features to recognize emotions in a TV show setting. Zeng *et al*. [32] propose an Inconsistent Pseudo Annotations to Latent Truth (IPA2LT) FER framework for RAF-DB and AffectNet. IPA2LT trains a Latent Truth Net (LT-Net) to discover latent true labels from inconsistency between pseudo-labels generated with a trained model and manual labels by maximizing the log-likelihood of inconsistent annotations. Georgescu *et al*. [4] combine deep features learned by three CNNs and the hand-crafted bag-of-visual-words (BOVW) features to boost the recognition performance on FER2013, FER+, and AffectNet. Siqueira *et al*. [29] propose a method called Ensemble with Shared Representation (ESR) to share the middle convolutional layers among an ensemble of CNNs to recognize facial expressions on FER+ and AffectNet.

3. Proposed Method

In this section, we first review necessary preliminaries. We then introduce the proposed Discriminant Distribution-Agnostic loss (DDA loss). Finally, we discuss DDA loss optimization and derive its corresponding gradients in back-propgation for Stochastic Gradient Descent (SGD) optimization.

3.1. Preliminaries

Given a training batch of m samples for a K-class image classification problem, let $x_i \in \mathbb{R}^d$ be the output d-dimensional deep feature of the i-th sample belonging to the y_i-th class, where $y_i \in \{1, ..., K\}$. The conventional softmax loss combines the last fully-connected layer, the softmax function, and the cross-entropy loss to measure the prediction error of the classifier. The last fully connected layer takes x_i and transforms it into a raw score vector (*i.e*., logits) $z_i = [z_{i1}, z_{i2}, ..., z_{iK}]^T \in \mathbb{R}^{K\times 1}$ through a linear transformation as follows:

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

where $W = [w_1, w_2, ..., w_K] \in \mathbb{R}^{d\times K}$ and $B = [b_1, b_2, ..., b_K] \in \mathbb{R}^{K\times 1}$ are the class weights and bias parameters for the last fully-connected layer, respectively. Each w_j is a d-dimensional vector and each b_j is a scalar where $j \in \{1, ..., K\}$. A probability distribution $p(y = j|x_i) = \frac{e^{z_{ij}}}{\sum_{j=1}^{K} e^{z_{ij}}}$ is then calculated over all classes using the softmax function. Finally, the cross-entropy computes the discrepancy between prediction and ground-truth to formulate the softmax loss function $\mathcal{L}_\mathcal{S}$ as follows:

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

where m is the total number of samples in a mini-batch. The softmax loss function is minimized by SGD to optimize the network parameters for a better classification. It also makes the learned features separated in an angular fashion in the embedding space since it calculates the vector dot product of $w \cdot x$ to minimize the angle between the deep feature x_i and its corresponding class weight w_{y_i} [20].

Center loss is jointly optimized with softmax loss to minimize the intra-class variations by minimizing the distance of the deep features to their corresponding class center in a d-dimensional embedding space. The center loss objective function penalizes the Euclidean distance between the deep feature vector of each sample $x_i \in \mathbb{R}^d$ and its corresponding class center $c_{y_i} \in \mathbb{R}^d$ as follows:

$$\mathcal{L}_C = \frac{1}{2m}\sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2 \tag{3}$$

where y_i is the class that x_i belongs to. Its joint optimization with softmax loss $\mathcal{L}_S$ is given as follows:

$$\mathcal{L} = \mathcal{L}_S + \lambda\mathcal{L}_C \tag{4}$$

where λ controls the contribution of $\mathcal{L}_C$ to the total loss $\mathcal{L}$. Individually, softmax loss $\mathcal{L}_S$ induces inter-class angular separation [30] and center loss $\mathcal{L}_C$ minimizes intra-class Euclidean distances to create compact clusters of features in the embedding space. The softmax loss in Eq. 2 is a special case of center loss with $\lambda = 0$.

3.2. Discriminant Distribution-Agnostic Loss

Training under the joint supervision of softmax loss and center loss creates compact clusters of deep features separated in an angular fashion. The softmax loss formulation incorporates all class weights to emphasize the angular separation of the deep feature x_i and class weights W. However, it has been proven to be unsuitable for a class imbalance setting [6]. On the other hand, center loss only penalizes the distance between a deep feature and its corresponding class center and disregards the contribution of other class centers. In an extreme class imbalance scenario, data points from minority classes and their corresponding class centers are minimally sampled in a training batch. The minimal learning impact from minority classes during mini-batch SGD optimization develops a bias towards majority classes. Thus, the efficiency of a learning algorithm supervised by center loss highly relies on the distribution of data among classes. Notably, in a wild setting, center loss delivers a sub-optimal classification performance for minority classes.

To circumvent this shortcoming, we aim to properly separate clustered deep feature vectors for both minority and majority classes in the embedding space. We argue that the Euclidean distance between the deep feature and all class centers should impact the forward propagation for a single sample to mitigate the bias toward majority classes as evidenced in center loss. To this end, we propose Discriminant Distribution-Agnostic loss (DDA loss) $\mathcal{L}_{DDA}$ as follows:

$$\begin{aligned}\mathcal{L}_{DDA} &= -\frac{1}{2m}\sum_{i=1}^{m}\sum_{k=1}^{N_k} y_{ik}\log p_C(x_i \in C_k|k) \\ &= -\frac{1}{2m}\sum_{i=1}^{m}\log\frac{e^{-\|x_i-c_{y_i}\|_2^2}}{\sum_{k=1}^{N_k} e^{-\|x_i-c_k\|_2^2}}\end{aligned} \tag{5}$$

where N_k is the number of classes, $y_{ik} = 1$ if x_i belongs to the k-th class and 0 otherwise, and C_k is the cluster for the k-th class in the embedding space. DDA loss estimates the probability of a deep feature x_i belonging to cluster k with its corresponding center c_k using a softmax function. Minimizing $\mathcal{L}_{DDA}$ is equivalent to maximizing the log-likelihood of the estimated probability $p_C(x_i \in C_k|k)$ over a batch of m samples. Compared to softmax loss in Eq. 2, which emphasizes the angular similarity, DDA loss separates the class features based on the Euclidean distance metric, which correlates with the feature vector's magnitude. Considering the magnitude difference between the learned features of a minority class and a majority class in a class imbalance setting is crucial in achieving intra-class compactness and inter-class separation.

Unlike center loss, DDA loss implicitly pushes the deep feature x_i away from any clusters C_k with $k \neq y_i$ and pulls itself towards its cluster C_k with $k = y_i$ in the embedding space with a single formulation. Intuitively, $\mathcal{L}_{DDA}$ considers the contribution from all majority and minority classes to update network parameters to achieve intra-class compactness and inter-class separability. The proposed DDA loss is distribution-agnostic and mitigates the bias towards majority classes.

DDA loss is jointly optimized with softmax loss and center loss to compose the total loss $\mathcal{L}$ by:

$$\mathcal{L} = \mathcal{L}_S + \lambda\mathcal{L}_C + \gamma\mathcal{L}_{DDA} \tag{6}$$

where the hyper-parameter γ controls the contribution of $\mathcal{L}_{DDA}$ to the total loss $\mathcal{L}$ and enables us to conduct quantitative analysis. The center loss defined in Eq. 4 is considered as a special case of this joint optimization when $\gamma = 0$.

3.3. Optimization

The proposed DDA loss is differentiable and can be optimized with the standard Stochastic Gradient Descent (SGD) algorithm. We study the SGD backpropagation optimization and the contribution of $\mathcal{L}_{DDA}$ gradients to the total loss $\mathcal{L}$. The joint optimization of $\mathcal{L}_{DDA}$ with softmax loss and center loss contributes to their gradients with respect to the deep feature x_i and centers c_k, respectively.

To simplify the derivative equations, we introduce the following intermediate notation:

$$p_{C_i} = \frac{e^{d_i}}{\sum_{k=1}^{N_k} e^{d_k}} \tag{7}$$

where $d_k = -\|x_i - c_k\|_2^2$. The gradient of DDA loss with respect to features x_i are computed according to the chain rule as follows:

$$\begin{aligned}\frac{\partial\mathcal{L}_{DDA}}{\partial x_i} &= \frac{\partial\mathcal{L}_{DDA}}{\partial d_j}\times\frac{\partial d_j}{\partial x_i} \\ &= -\frac{1}{2m}\times\frac{1}{p_{C_i}}\frac{\partial p_{C_i}}{\partial d_j}\times(-2)(x_i - c_{y_i}) \\ &= \frac{1}{m}(\delta_{ij} - p_{C_j})(x_i - c_{y_i})\end{aligned} \tag{8}$$

where the Kronecker delta function is defined as $\delta_{ij} = 1$ for $i = j$ and 0 otherwise.

Class centers are randomly initialized according to the *He* method [8]. We update the centers as follows:

$$c_k = c_k - \alpha\Delta c_k^* \tag{9}$$

where Δc_k^* is the combination of an average strategy (Δc_k) proposed in [31] and the gradients of DDA loss with respect to centers c_k as in:

$$\begin{aligned}\Delta c_k^* &= \Delta c_k + \frac{\partial\mathcal{L}_{DDA}}{\partial c_k} \\ &= \frac{\sum_{i=1}^{m}\delta_{y_ik}\cdot(c_k - x_i)}{1 + \sum_{i=1}^{m}\delta_{y_ik}} \\ &\quad + \frac{1}{m}\sum_{i=1}^{m}(\delta_{ij} - p_{C_j})(c_{y_i} - x_i)\end{aligned} \tag{10}$$

Algorithm 1 summarizes the major steps for training an end-to-end deep CNN model using DDA loss.

Algorithm 1 Training a supervised deep learning algorithm (*e.g.*, CNN) using DDA loss.

Input: Mini-batch features $\{x_i|i = 1, 2, ..., m\}$ extracted from a CNN model; Initialized parameters θ_C for convolutional filters in CNN; Initialized parameters $W = \{w_j|j = 1, 2, ..., N_k\}$ for the last FC layer and $C = \{c_k|k = 1, 2, ..., N_k\}$ for center loss and DDA loss; Hyper-parameters α, γ, λ, and learning rate μ; The number of iterations $t \leftarrow 0$.

Output: Updated parameters θ_C, W, and C.

while not converged **do**
  Compute the total joint loss: $\mathcal{L}^t = \mathcal{L}_S^t + \lambda\mathcal{L}_C^t + \gamma\mathcal{L}_{DDA}^t$.
  Compute the gradients: $\hat{g}_t \leftarrow \frac{1}{m}\sum_{i=1}^{m} \frac{\partial\mathcal{L}_S^t}{\partial x_i^t} + \lambda\frac{\partial\mathcal{L}_C^t}{\partial x_i^t} + \gamma\frac{\partial\mathcal{L}_{DDA}^t}{\partial x_i^t}$.
  Compute Δc_k^* by Eq. 10.
  $t \leftarrow t + 1$.
  Update w_j for each j: $w_j^{t+1} = w_j^t - \mu\frac{\partial\mathcal{L}_S^t}{\partial w_j^t}$.
  Update c_k for each k: $c_k^{t+1} = c_k^t - \alpha\Delta c_k^{t*}$.
  Update the CNN model parameters θ_C: $\theta_C^{t+1} = \theta_C^t - \mu^t\hat{g}_t$
end while

4. Experiments

We conduct extensive experiments to evaluate the performance of the proposed loss function and other state-of-the-art methods. We visually and quantitatively validate the superior performance of the proposed Discriminant Distribution-Aware loss (DDA loss) compared to the baseline loss functions, namely, softmax loss and center loss, on a wild toy dataset. We then evaluate the proposed DDA loss on two widely used wild FER datasets against the baseline loss functions and recent state-of-the-art methods that tackle the wild setting.

4.1. Wild MNIST Experiments

We present a toy experiment on the Wild MNIST (W-MNIST) dataset with ten classes, a subset of the MNIST dataset [11], to study the proposed method more intuitively. W-MNIST is comprised of randomly sampled image data (single hand-written digits) from the standard MNIST training set. To mimic the characteristics of a wild FER dataset, we drastically decrease the number of training data points in W-MNIST for a few categories by sampling only a few data points from MNIST. The distribution of data in W-MNIST is summarized in Fig. 2 (a). We illustrate two-dimensional (2-D) deep features learned by softmax loss and center loss in Fig. 2 (b) and (c), respectively, and the 2-D deep features learned by the proposed DDA loss with different γ values in Fig. 2 (d)-(f).

Method	λ / γ	Accuracy (%)
softmax loss	-	96.78
center loss [31]	0.01 / -	97.12
DDA loss	0.01 / 1.0	97.17
DDA loss	0.01 / 3.0	97.17
DDA loss	0.01 / 5.0	97.15
DDA loss	0.01 / 7.0	**97.34**

Table 1. Classification accuracy on the MNIST testing set by training the LeNets++ model with different losses on the W-MNIST training set.

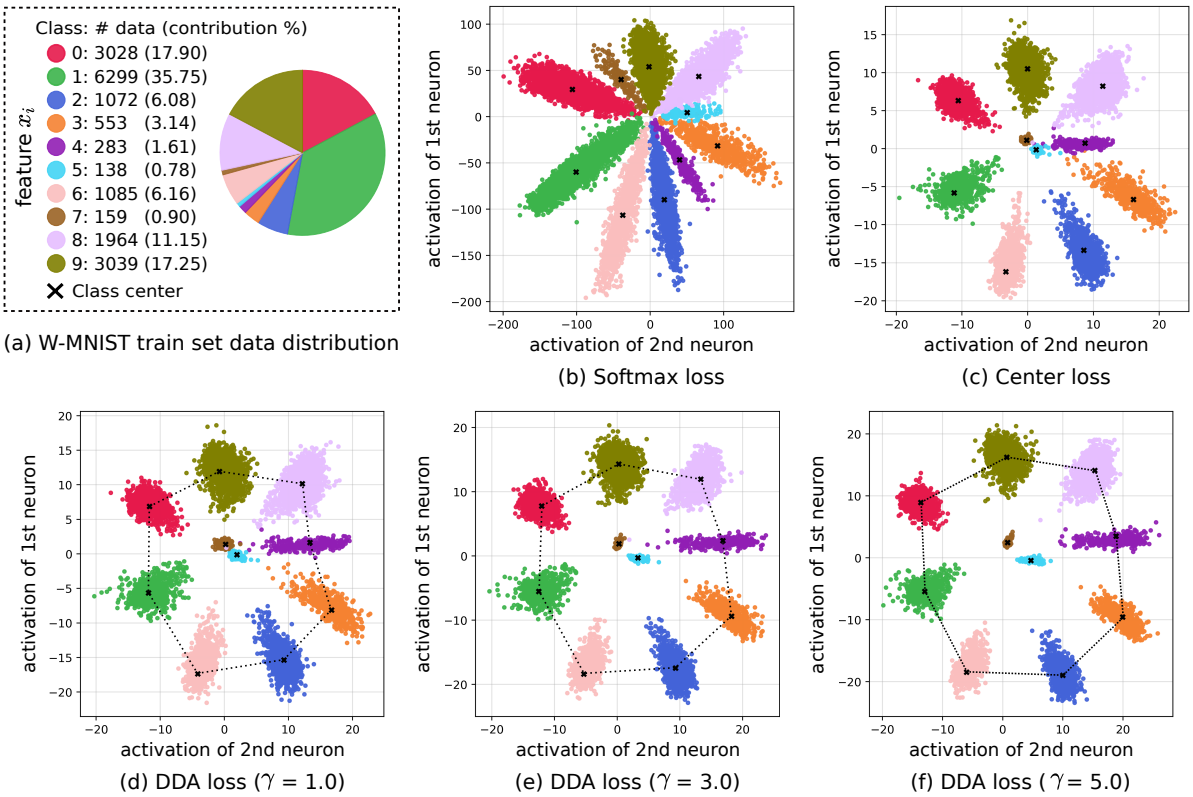
To display deep features on a 2-D plot, we use the CNN model LeNets++ [31] with six stacked convolutional layers and one fully-connected layer with two neurons. We train LeNets++ on the W-MNIST dataset using the standard stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 5×10^{-4} for 100 epochs. We use a batch size of 128 and set the initial learning rate as 0.001 with a decay factor of 1.25 every 20 epochs. We do not use any data augmentation on W-MNIST. We empirically set the hyper-parameter λ for center loss as 0.01 and experiment with different γ values for DDA loss.

As illustrated in Fig. 2, the deep features learned by center loss are more discriminative compared to the deep features learned by softmax loss. However, inter-class distances are optimized with a bias toward the majority classes in the embedding space. Consequently, minority classes are over-lapped, or their inter-class distances relative to majority classes are not optimized. On the other hand, DDA loss occupies the embedding space with compact and well-separated feature clusters for both majority and minority classes. As we increase the hyper-parameter γ, feature clusters tend to disperse further away from other clusters. Visualization of 2-D deep features verifies that the proposed DDA loss yields more discriminative features in a wild dataset setting since it achieves optimal inter-class separation and intra-class compactness for all classes.

To quantitatively evaluate the performance of the proposed DDA loss and the baseline loss functions, we train LeNets++ on the W-MNIST training set and test its recognition performance on the MNIST testing set. Table 1 summarizes the classification accuracy of the proposed DDA loss and two baseline loss functions (softmax loss and center loss) on the MNIST testing set. It clearly shows that the proposed DDA loss with $\gamma = 7.0$ outperforms both softmax loss and center loss by achieving an accuracy of 97.34%.

4.2. Wild FER Experiments

Real-world Affective Face Data-Base (RAF-DB) [14] and AffectNet [26] are the two largest and widely used wild Facial Expression Recognition (FER) datasets. RAF-



(a) W-MNIST train set data distribution

(b) Softmax loss

(c) Center loss

(d) DDA loss (γ = 1.0)

(e) DDA loss (γ = 3.0)

(f) DDA loss (γ = 5.0)

Figure 2. A wild toy experiment of training LeNets++ on the W-MNIST training set using different loss functions. (a) Distribution of data for the W-MNIST training set. Illustration of the distribution of 2-D deep features learned via: (b) Softmax loss, (c) Center loss, (d)-(f) DDA loss with different γ values.

DB contains 12,271 training images and 3,068 testing images aligned and annotated with six basic expressions (*i.e.*, *happy*, *sad*, *surprise*, *anger*, *disgust*, and *fear*) and *neutral* expression using crowd-sourcing techniques. AffectNet contains 280,000 training images and 3,500 testing images manually annotated with six basic expressions and *neutral* expression. Both datasets comprise of facial images in real world with various gender, age, demography, image quality, and illumination attributes. We first present the details of our implementations in terms of architecture, training, and hyper-parameters. We then analyze the recognition performance on both RAF-DB and AffectNet datasets and study the effect of hyper-parameter γ. Finally, we discuss our results and the limitations of the proposed method.

4.2.1 Implementation details

We fit ResNet-18 [9], a standard convolutional network, to both RAF-DB and AffectNet as the backbone architecture. We train and optimize ResNet-18 using SGD with an initial learning rate of 0.01, a momentum of 0.9, and a weight decay of 5×10^{-4}. We train ResNet-18 initialized with ImageNet weights on RAF-DB for 60 epochs with a batch size of 64 and decay the learning rate by a factor of 10 every 20 epochs. For AffectNet, we train ResNet-18 from scratch for 20 epochs with a batch size of 128 and decay the learning rate by a factor of 10 every five epochs. For both datasets, we augment the input images on-the-fly by extracting random crops (one central, and one for each corner and their horizontal flip). At test time, we use the central crop of the input image. Crops of size 90 (given images of size 100) and 224 (given images of size 256) are extracted from RAF-DB and AffectNet, respectively. Our models are trained using PyTorch deep learning framework [28] on a 2080Ti GPU with 11GBs of V-RAM.

4.2.2 Recognition Performance

Table 2 and Table 3 compare the expression recognition performance of the proposed DDA loss, the two baseline loss functions, and recent methods on RAF-DB and AffectNet, respectively. Since RAF-DB's testing set is imbalanced, we report both the standard accuracy and the average accuracy, which is the average of the main diagonal values in the confusion matrix. We empirically set the hyper-parameters for center loss as $\lambda = 0.01$ and $\alpha = 0.5$. To ensure a fair comparison, we use the same hyper-parameters in the proposed DDA loss. The proposed DDA loss, best optimized with $\gamma = 5.0$ outperforms other methods on RAF-DB by achieving the recognition accuracy of 86.99% and an aver-

Method	Acc. (%)	Avg. Acc. (%)
FSN [33]	81.10	72.46
PA-CNN [16]	83.27	-
DLP-CNN [13]	84.13	74.20
ALT [3]	84.50	76.50
GA-CNN [17]	85.07	-
SEP-LOSS [3]	86.38	77.25
IPA2LT [32]	86.77	-
softmax loss	85.56	77.28
center loss [31] ($\lambda = 0.01$)	86.25	77.81
DDA loss ($\lambda = 0.01, \gamma = 5.0$)	**86.90**	**79.71**

Table 2. Expression recognition performance of different methods on RAF-DB in terms of standard accuracy and average accuracy.

Method	Accuracy (%)
PA-CNN [16]	55.33
IPA2LT [32]	57.31
GA-CNN [17]	58.78
SEP-LOSS [15]	58.89
softmax loss	61.46
center loss [31]($\lambda = 0.01$)	61.69
DDA loss ($\lambda = 0.01, \gamma = 4.0$)	**62.34**

Table 3. Expression recognition performance of different methods on AffectNet in terms of accuracy.

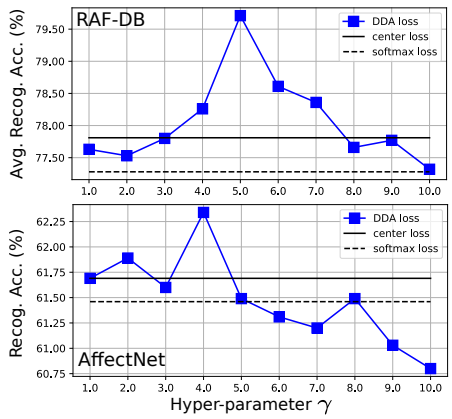


Figure 3. The effect of hyper-parameter γ for DDA loss on (top): The average recognition accuracy of RAF-DB and (bottom): The recognition accuracy of AffectNet.

age recognition accuracy of 79.71%. Similarly, DDA loss, best optimized with $\gamma = 4.0$, outperforms other methods on AffectNet by achieving the recognition accuracy of 62.34%.

4.2.3 The Effect of Hyper-parameter γ

Fig. 3 shows the effect of using different γ values for the proposed DDA loss on the FER performance for wild FER datasets. The contribution of DDA loss to the total loss is controlled by γ. Large γ values make the total loss focus more on DDA loss, and small γ values make the total loss focus more on softmax loss and center loss. Specifically, for large γ values, features either do not separate or do not exhibit compactness in the embedding space. Small γ values cannot separate the feature clusters efficiently to circumvent the issue with the learned features supervised by center loss. Our experiments on two datasets empirically show that softmax loss converges slower and cannot efficiently separate features in angular fashion when increasing the γ value to increase the contribution of DDA loss $\mathcal{L}_{DDA}$ in Eq. 6. Furthermore, the center loss objective function puts less emphasis on penalizing the distance between features and their corresponding class centers and achieves less intra-class compactness. Hence, the recognition rate starts to degrade after the peak performance with $\gamma = 5.0$ for RAF-DB and $\gamma = 4.0$ for AffectNet. Our experiments show that γ values larger than 10.0 will disrupt the balance between the three terms in the total loss $\mathcal{L}$ in Eq. 6 and significantly degrade the recognition rates.

4.2.4 Discussion

Although our method boosts the recognition performance when comparing with the two baseline methods, the results are not uniformly positive. For example, the proposed DDA loss tends to outperform center loss for RAF-DB dataset (Fig. 3 (a)). However, the performance of the proposed method quickly drops below center loss when $\gamma > 4.0$ for AffectNet dataset (Fig. 3 (b)). This behavior is mainly because DDA loss requires to be jointly optimized with center loss to maintain intra-class compactness. Furthermore, the relative size of majority classes to minority classes in AffectNet is significantly higher than the one in RAF-DB. Consequently, majority classes lose their intra-class structure and the recognition performance drops for all classes when the contribution of DDA loss increases.

We present the confusion matrices obtained by employing two baseline methods and the proposed method with DDA loss on RAF-DB and AffectNet in Fig. 4. It is clear that center loss boosts the recognition rates for most of the majority classes but degrades the recognition rates for minority classes. On the other hand, DDA loss boosts the recognition rates for minority classes and maintains the comparable recognition rates for majority classes. Specifically, we observe that the proposed method either maintains or boosts the recognition rates for majority classes except *neutral* and *surprise* for AffectNet. In Fig. 5, we provide sample correctly classified and misclassified images from RAF-DB and AffectNet predicted by our best models trained with DDA loss. Because AffectNet is much larger than RAF-DB, the human annotations are less accurate. This is a prevailing issue for large-scale datasets when resources are low and annotation can be subjective, which leads to more noisy ground-truth labels in AffectNet. Con-

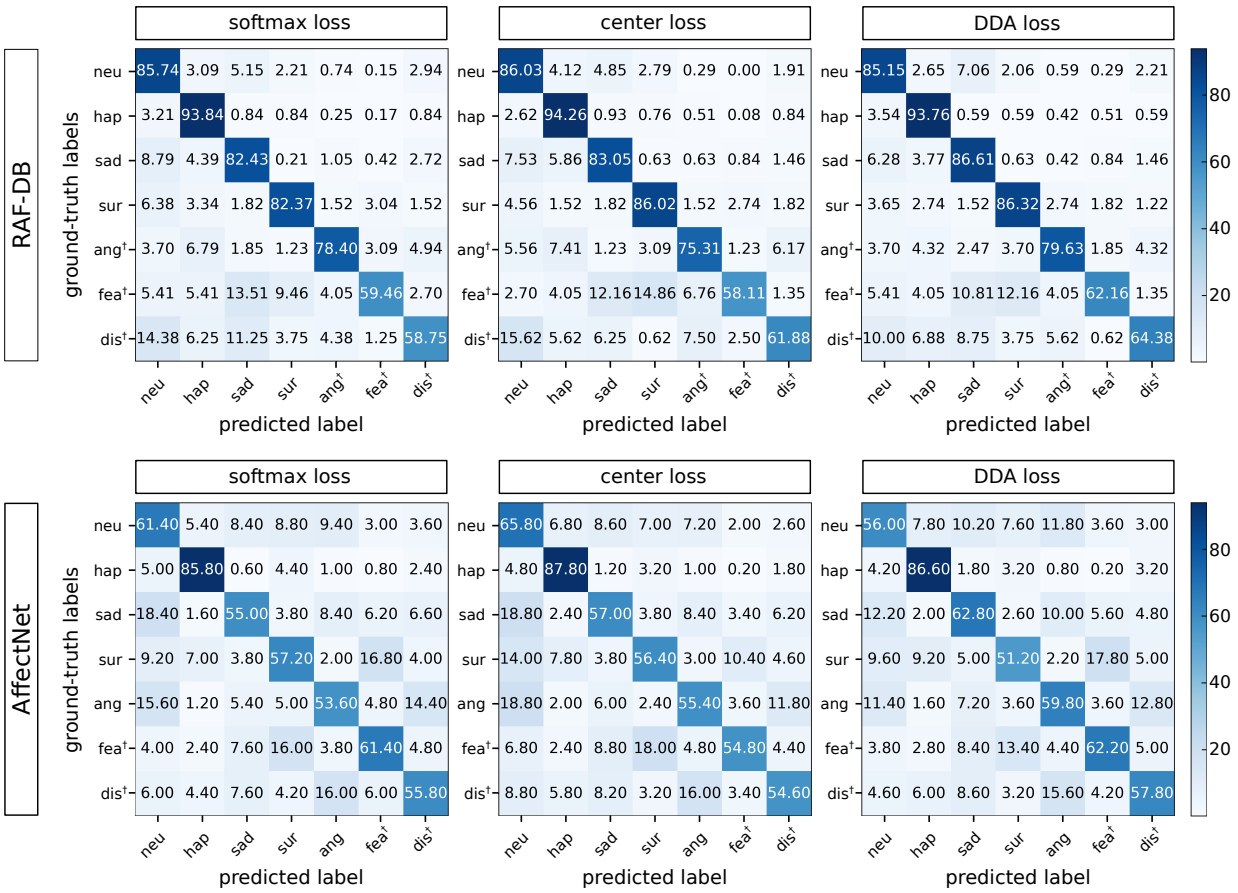


Figure 4. Confusion matrices for the recognition accuracy of: ***top row:*** RAF-DB, and ***bottom row:*** AffectNet using baseline methods and the proposed method. † Minority classes.

sequently, our models yield correct predictions that might contradict with the ground-turth labels.

5. Conclusions

We propose Discriminant Distribution-Agnostic loss (DDA loss) for Facial Expression Recognition (FER) in the wild settings. DDA loss implicitly pushes deep features of a class away from other classes and pulls them toward their corresponding class centers in the embedding space. Supervised jointly by softmax loss and center loss, DDA loss efficiently distributes feature clusters of both majority and minority classes in the embedding space where extremely imbalanced distribution of data exists. DDA loss can be optimized with the standard Stochastic Gradient Descent (SGD) algorithm and can be readily employed by any Convolutional Neural Network (CNN) to yield highly discriminative features that are efficient under wild scenarios. Experiments with a synthesized Wild MNIST (W-MNIST) dataset and two widely used wild FER datasets, RAF-DB and AffectNet, demonstrate the superior performance of DDA loss compared to other state-of-the-art methods.

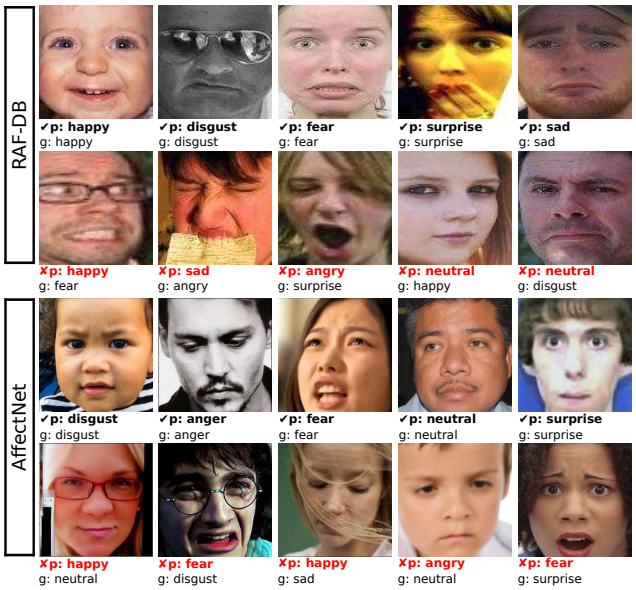


Figure 5. Sample correctly classified and misclassified images in ***top row:*** RAF-DB and ***bottom row:*** AffectNet. ✔ denotes correct classification and ✘ denotes misclassification.

References

[1] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ICMI*, pages 279–283, 2016.

[2] 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 *FG*, pages 302–309, 2018.

[3] Corneliu Florea, Laura Florea, Mihai Alexandru Badea, and Constantin Vertan. Annealed label transfer for face expression recognition. In *BMVC*, page 12, 2019.

[4] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, and Marius Popescu. Local Learning With Deep and Handcrafted Features for Facial Expression Recognition. *IEEE Access*, 7:64827–64836, 2019.

[5] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015.

[6] Munawar Hayat, Salman Khan, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Gaussian Affinity for Max-Margin Class Imbalanced Learning. In *ICCV*, pages 6468–6478, 2019.

[7] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *CVPR*, pages 1026–1034, 2015.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[10] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, pages 4873–4882, 2016.

[11] Yann LeCun, Léon Bottou, Yoshua Bengio, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[12] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *ICCV*, pages 10143–10152, 2019.

[13] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE TIP*, 28(1):356–370, 2018.

[14] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, pages 2852–2861, 2017.

[15] Yingjian Li, Yao Lu, Jinxing Li, and Guangming Lu. Separate loss for basic and compound facial expression recognition in the wild. In *ACML*, pages 897–911, 2019.

[16] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Patch-gated CNN for occlusion-aware facial expression recognition. In *ICPR*, pages 2209–2214, 2018.

[17] Y. Li, J. Zeng, S. Shan, and X. Chen. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE TIP*, 28(5):2439–2450, May 2019.

[18] Zhenghao Li, Song Wu, and Guoqiang Xiao. Facial expression recognition by multi-scale cnn with regularized center loss. In *ICPR*, pages 3384–3389, 2018.

[19] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pages 212–220, 2017.

[20] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7, 2016.

[21] Xiaofeng Liu, BVK Vijaya Kumar, Jane You, and Ping Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *CVPRW*, pages 20–29, 2017.

[22] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, et al. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, pages 94–101, 2010.

[23] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *FG*, pages 200–205, 1998.

[24] Aleix M. Martinez. Context may reveal how you feel. *Proceedings of the National Academy of Sciences*, 116(15):7169, 2019.

[25] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-aware convolutional neural network for facial expression recognition. In *FG*, pages 558–565, 2017.

[26] 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, 2017.

[27] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *ICME*, pages 5–8, 2005.

[28] Adam Paszke, Sam Gross, Francisco Massa, Lerer, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[29] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient facial feature learning with wide ensemble-based convolutional neural networks. In *AAAI*, 2020.

[30] Weitao Wan, Yuanyi Zhong, Tianpeng Li, and Jiansheng Chen. Rethinking feature distribution for loss functions in image classification. In *CVPR*, pages 9117–9126, 2018.

[31] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515, 2016.

[32] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *ECCV*, pages 222–237, 2018.

[33] Shuwen Zhao, Haibin Cai, Honghai Liu, Jianhua Zhang, and Shengyong Chen. Feature selection mechanism in cnns for facial expression recognition. In *BMVC*, 2018.