electronics

MDPI

Article

An Extra-Contrast Affinity Network for Facial Expression Recognition in the Wild

Jiaqi Zhu [1], Shuaishi Liu [1,*], Siyang Yu [2] and Yihu Song [1]

[1] Department of Control Engineering, Changchun University of Technology, Changchun 130012, China; zhujiaqi2587@163.com (J.Z.); songyihu12138@163.com (Y.S.)

[2] Department of Digital Media, Changchun University of Technology, Changchun 130012, China; yusiyang@ccut.edu.cn

* Correspondence: liushuaishi@ccut.edu.cn

Abstract: Learning discriminative features for facial expression recognition (FER) in the wild is a challenging task due to the significant intra-class variations, inter-class similarities, and extreme class imbalances. In order to solve these issues, a contrastive-learning-based extra-contrast affinity network (ECAN) method is proposed. The ECAN consists of a feature processing network and two proposed loss functions, namely extra negative supervised contrastive loss (ENSC loss) and multi-view affinity loss (MVA loss). The feature processing network provides current and historical deep features to satisfy the necessary conditions for these loss functions. Specifically, the ENSC loss function simultaneously considers many positive samples and extra negative samples from other minibatches to maximize intra-class similarity and the inter-class separation of deep features, while also automatically turning the attention of the model to majority and minority classes to alleviate the class imbalance issue. The MVA loss function improves upon the center loss function by leveraging additional deep feature groups from other minibatches to dynamically learn more accurate class centers and further enhance the intra-class compactness of deep features. The numerical results obtained using two public wild FER datasets (RAFDB and FER2013) indicate that the proposed method outperforms most state-of-the-art models in FER.

Keywords: contrastive learning; deep metric learning; imbalance classification; facial expression recognition

check for updates

Citation: Zhu, J.; Liu, Z.; Yu, S.; Song, Y. An Extra-Contrast Affinity Network for Facial Expression Recognition in the Wild. *Electronics* **2022**, *11*, 2288. https://doi.org/10.3390/electronics11152288

Academic Editor: Byung-Gyu Kim

Received: 29 June 2022
Accepted: 19 July 2022
Published: 22 July 2022

1. Introduction

Facial expression is one of the most direct, fundamental, and universal signals in non-verbal communication. The analysis of facial expressions is an active area of computer vision research. Furthermore, FER is widely used in numerous aspects of modern society, such as human–computer interaction (HCI), facial expression synthesis, education, psychotherapy, and social robotics. Ekman et al. defined six basic expressions: anger, disgust, fear, happiness, sadness, and surprise [1]; neutrality and contempt have been recently included in FER datasets as additional categories. In recent years, methods based on deep neural networks have not only outperformed traditional methods of FER but also achieved remarkable success in many other fields [2–6].

FER applications in the real world require wild FER datasets that contain a massive number of annotated images acquired in an unconstrained environment. Two obstacles hinder the ability of a convolutional neural network (CNN) to learn from wild FER datasets. One is the significant intra-class variation and inter-class similarity, and the other is the extreme class imbalance. A contrastive learning method named ECAN is proposed in this paper to solve these issues. It includes three main components: the feature processing network, the ENSC loss function, and the MVA loss function. The first component uses the classic Siamese network and memory-bank-based architecture along with contrastive learning methods, as used by MOCO [7]. The other two components are proposed for the first time in this paper; they are introduced briefly in this section, and more details can be found in Section 3.

Intra-class variation and inter-class similarity are noteworthy features of wild FER datasets; as a result, discriminative features are easily confusable. An effective way to solve this issue is to use deep metric learning (DML) to enhance the discrimination of confusable facial features. One typical idea that is widely employed is utilizing pairs of positive and negative samples; the loss functions bring anchors and positives closer and push away negatives in the deep feature space. The triplet loss function [8] was designed to consider an anchor with a sample that had the same label (i.e., a positive sample) in the current minibatch as a positive pair and a sample that had a different label (i.e., a negative sample) as a negative pair; it learned by decreasing the distance between positive pairs and increasing the distance between negative pairs. The N-pair loss function [9] improved upon the learning effect of the triplet loss function by expanding the number of negative pairs. Recently, with contrastive learning methods, a novel selection strategy has been proposed that refers to positive and negative pairs. Since no labels are available in self-supervised contrastive learning methods, the positive pair usually consists of a pair of data augmentations (also known as "views") of the same sample, and the negative pair consists of anchors and other views in the minibatch. In the case of supervised learning, the supervised contrastive (SupCon) loss function proposed by Prannay et al. reasonably utilized labels to add multiple positive pairs. The ENSC loss function considers not only the negative samples in the current minibatch but also a large number of extra negative samples from other minibatches to improve upon the SupCon loss function. This idea more effectively promotes the inter-class separation of deep features. It also greatly expands negative pairs, which is another benefit.

Furthermore, another commonly used loss function in DML is the center loss function [10], which aims to increase the intra-class compactness of deep features. There are many ideas inspired by the center loss function, such as the island loss function [11], which trained class centers while also pushing them away from each other; DACL [12], which improved upon the center loss function by adaptively selecting a subset of important feature elements; and DAN [13], which promoted the separation of class centers using a simple idea that barely increased the computational cost. The above-mentioned methods improved upon the center loss function from different perspectives; however, they only focused on the deep features of the current minibatch to optimize feature distribution and learn the class centers. In contrast, in our proposal, the MVA loss function simultaneously considers the deep features of multiple minibatches to dynamically and more accurately learn the class centers. Moreover, our method can further improve the compactness of intra-class features using the MVA loss function.

Another problem from which wild FER datasets suffer is extreme class imbalance. Usually, classes such as fear and disgust are in the minority due to a lack of representative data. Other expressions, such as neutral, happy, sad, surprised, and angry, are in the majority [14]. Deep neural networks tend to focus more attention on training the majority classes than the minority classes. The proposed negative sample screening strategy for the ENSC loss function implicitly addresses this issue. It regulates the attention paid by the deep neural network to the majority and minority classes based on their contribution percentages while calculating loss, which means that the majority classes provide less of a contribution and the minority classes provide more of a contribution. This strategy alleviates the negative impact of the class imbalance issue on feature learning. As shown in Figure 1, the ECAN method does not take a two-step training approach (i.e., first train the feature extraction network and then train the classifier) like the conventional scheme for self-supervised contrastive learning; in contrast, it employs a one-step training approach combined with cross-entropy loss (CE loss) to obtain classification results directly.

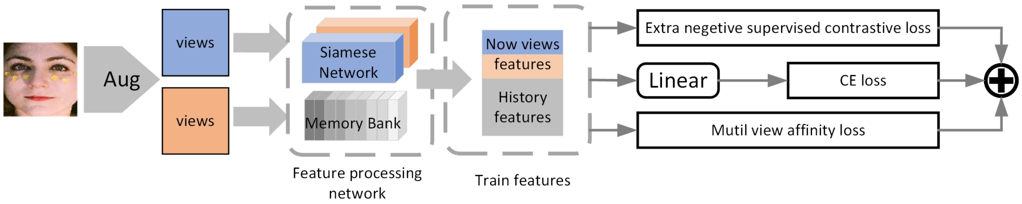


Figure 1. Overview of our proposed ECAN method.

The main contributions of this paper are summarized as follows:

- This paper proposes the ENSC loss function, which can utilize multiple minibatches of extra negative samples to increase inter-class separation and alleviate the effects of class imbalance.
- To further increase intra-class compactness, the MVA loss function is proposed to dynamically and more accurately learn class centers with the help of multiple minibatches of deep features.
- This paper presents the results of extensive experiments on two standard wild FER datasets. Experimental results show that the proposed ECAN method achieved 89.77% accuracy and 83.14% average accuracy on the RAFDB dataset and 73.73% accuracy on the FER2013 dataset; these results are consistent with current state-of-the-art FER performance.

The rest of this paper is organized as follows. Section 2 reviews related work, with a particular focus on deep metric learning and contrastive learning. Section 3 elaborates the proposed method. Evaluation and experimental results are demonstrated in Section 4. Discussion and analysis of the proposal can be found in Section 4.7. Finally, Section 5 summarizes the paper.

2. Related Works

2.1. DML for FER

DML can strongly modulate the distribution of deep features to tackle large intra-class variations and inter-class similarities. Most existing DML methods were developed for facial recognition applications, but FER applications can also benefit from DML. Meng et al. proposed an identity-aware convolutional neural network (IACNN) [15] that could distinguish expression-related features and identity-related features. Liu et al. proposed an (N+M)-tuplet clusters loss function [16] that could utilize tuples to form deep feature clusters on the same expression and then separate them from each other. The foundation for this approach was laid by the N-pair [9] and triplet [8] loss functions. Wen et al. proposed a center loss function [10] to learn the center distribution of each class; it penalized the distance between deep features and their corresponding class centers. Farzaneh et al. proposed a discriminant distribution agnostic (DDA) loss function [16] that could utilize the center loss function to cluster features of the same class and the softmax loss function to separate adjacent classes. Cai et al. improved on the center loss function by adding an extra objective function called the island loss function [11] that learned the cosine distance between class centers. The separate loss function [17] proposed by Li et al. further extended the island loss function; it simultaneously maximized the cosine similarity between deep features and their corresponding class centers and minimized the cosine similarity between class centers. Farzaneh et al. proposed a deep attentive center loss function [12] that optimized the center loss function by learning the relationship of each class center. Most of the above methods did not adopt a special strategy for training class centers, so it was necessary to improve upon the center loss function using a novel strategy.

2.2. Contrastive Learning

Contrastive learning is a DML method, and its core idea recently inspired a series of self-supervised contrastive learning methods [18]. Specifically, it maximizes the consistency between two "views" of the same image while repelling "views" from different images. These views can be obtained through various approaches, including data augmentation [19], color decomposition [20], patch cropping [21], and image segmentation [22]. The tactic chosen for the selection of views can greatly affect the performance of contrastive learning. In the work of Chen et al., an in-depth study of views obtained through data augmentation was performed [19]. A special multi-crop method of obtaining views was adopted in the work of Mathilde et al. [23]; they introduced clustering ideas to improve contrastive learning. Similarly, Li et al. proposed a clustering-based method called PCL [24] that was suitable for large-scale classification tasks due to the large number of cluster centers. He et al. proposed the MOCO method [7], which only required small batches to achieve good learning results because it designed a memory bank to store historical features. Recently, some methods were proposed that did not fully follow the core idea of self-supervised contrastive learning. For example, the BYOL method [25] only optimized the consistency of positive pairs without considering negative pairs. Prannay et al. proposed supervised contrastive learning [26], which was different from self-supervised contrastive learning because it simultaneously narrowed the distances between many positive pairs in the deep feature space. Supervised contrastive learning did not improve on negative pairs, although it made full use of positive pairs. In addition, it also did not take into account the possible impact of the class imbalance issue on training. Therefore, it was necessary to design a more comprehensive loss function.

3. Method

This explanation of the ECAN method is organized into 3 subsections that correspond to its three 3 main components: the feature processing network, the ENSC loss function, and the MVA loss function. Figure 2 shows a detailed overview of the ECAN method. First, random data augmentation is twice performed on the same expression image to generate two views. Then, the feature processing network extracts the deep features from the views. The features used for training are composed of the extracted deep features of the current views and the historical deep feature groups from the memory bank. Finally, the deep features of the current minibatch are linked with a linear classification layer, and the classification loss is calculated using the CE loss function. At the same time, the proposed ENSC and MVA loss functions are computed at the deep-feature level.

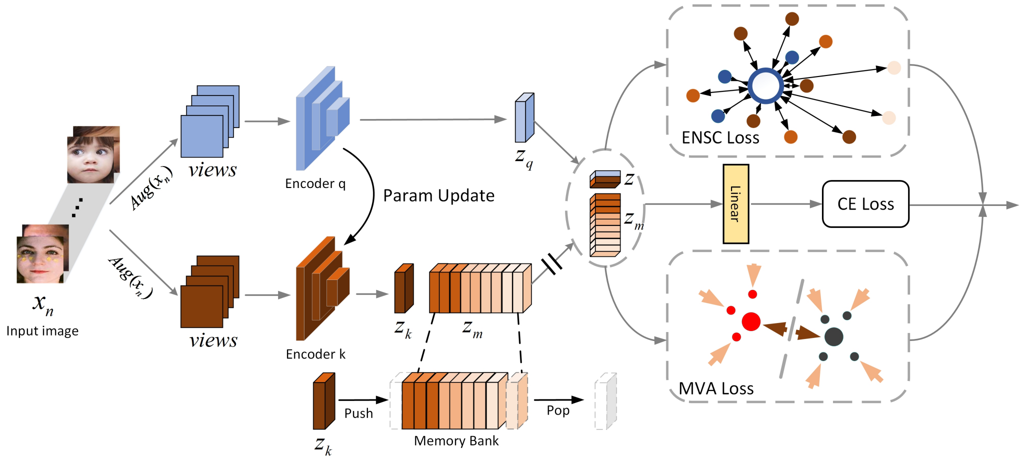


Figure 2. A detailed representation of the process of our proposed ECAN method. The hollow circle in the middle of the ENSC loss function diagram represents the anchor, the three dark blue circles represent the positives of the current minibatch, and the other colored circles represent the negatives. The symbol "||" indicates the blocked gradient update. In addition, the yellow linear layer is connected to z_q alone and not to any other feature groups.

3.1. Feature Processing Network

The feature processing network uses a similar architecture to MOCO [7], consisting of a Siamese network and a memory bank. The memory bank provides the necessary deep feature groups for the two loss functions. The deep features belonging to a single minibatch are called one deep feature group in this paper. For a set of N random sample/label pairs, $\{x_n, y_n\}_{n=1\ldots N}$, the data used for minibatch training consist of $2N$ pairs, $\{\tilde{x}_\ell, \tilde{y}_\ell\}_{\ell=1\ldots 2N}$, where $\tilde{x}_{2n}$ and $\tilde{x}_{2n-1}$ are two random data augmentations (also known as "views") of $\{x_n\}_{n=1\ldots N}$ and the labels $\tilde{y}_{2n-1} = \tilde{y}_{2n} = y_n$, which is called the "two-view minibatch" in this paper. Let the random data augmentation be $Aug(\cdot)$; then, $\tilde{x}_{2n} = Aug(x_n)$ and $\tilde{x}_{2n-1} = Aug(x_n)$. Since the base of contrastive learning is feature invariance (i.e., image transformation does not change the discriminative features of the image), it requires not only differences between views but also fully preserved discriminative features. So, the choice of the data augmentation approach can affect the performance of contrastive learning considerably. In this research, three levels of data augmentation (none, weak, and strong) are deployed to study their impacts on model performance.

The Siamese network extracts the deep features of these views. Let two encoders of the Siamese network be $Enc_q(\cdot)$ and $Enc_k(\cdot)$, and the deep features extracted by the Siamese network are $z_q = Enc_q(\tilde{x}_{2n})$ and $z_k = Enc_k(\tilde{x}_{2n-1})$ for a two-view minibatch. The total feature of the two-view minibatch is $z = z_q \cup z_k$. The Siamese network does not train both encoders at the same time for the purpose of accelerating model training and saving computing resources. If we set the parameter of $Enc_q(\cdot)$ to θ_q and the parameter of $Enc_k(\cdot)$ to θ_k, only θ_q is updated during back-propagation, and θ_k is updated according to the following formula after each step of training.

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q \tag{1}$$

where $m \in [0, 1)$ represents the hyper-parameters of the momentum coefficient, which is set to 0.999 based on the experience of MOCO [7].

z_k and the corresponding label y_k of the current and historical minibatches are stored in the memory bank. Since the quality of z_k improves along with updates to the Siamese network, the oldest historical z_k is the most outdated. The memory bank needs to update dynamically to remove outdated feature groups. The update process of the memory bank can be expressed as the following, if we denote a memory bank that stores K groups of z_k as z_m.

$$z_m^{new} \leftarrow \mathrm{Push}\Big([z_k, y_k], \mathrm{Pop}\Big(\Big[z_{m(K)}, y_{m(K)}\Big]\Big)\Big) \tag{2}$$

where z_m^{new} is the new memory bank after updating, $\mathrm{Push}(z, m)$ means to insert the given feature group z into the m queue, $[z_k, y_k]$ is the feature group/label pair of the current minibatch, $\mathrm{Pop}(z_{m(K)})$ means to dequeue the oldest feature group/label pair $[z_{m(K)}, y_{m(K)}]$ in the memory bank, and the memory bank is updated synchronously with θ_k after each step. An initialization process that freezes all parameter updates is performed to populate the memory bank at the beginning of training since the memory bank has no features. The acquired total feature z of the current two-view minibatch and the memory bank z_m are used for subsequent loss calculations.

3.2. The Extra Negative Supervised Contrastive Loss Function

In this subsection, we review the SupCon loss function and discuss the improvements made to it by the proposed ENSC loss function.

3.2.1. Review of the SupCon Loss Function

The contrastive loss approaches are divided into two types, namely self-supervised learning and supervised learning. In the case of self-supervised learning, the contrastive learning method treats each sample as an independent positive and the rest as negatives due to the lack of labels. Negatives that are the same as true positives are called false negatives. Obviously, false negatives are an obstacle to feature learning, but a self-supervised contrastive learning method that completely eliminates the influence of false negatives has not yet been developed.

Within a two-view minibatch, let $i \in I = \{1 \ldots 2N\}$ be the index of a view, and let $j(i)$ be the index of the other view from the same source sample. In self-supervised contrastive learning, the loss takes the following form:

$$\mathcal{L}^{selfcon} = \sum_{i \in I} \mathcal{L}_i^{self} = -\sum_{i \in I} \log \frac{\exp\left(z_i \cdot z_{j(i)} / \tau\right)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)} \tag{3}$$

where z_i represents the deep features of the view, called the "anchor", $z_{j(i)}$ is the positive, and $A(i) = I \backslash \{i\}$ due to the fact that it is meaningless to measure the distance between the same samples. z_a contains both the positive and remaining samples (i.e., the negatives), the symbol $\cdot$ denotes the inner product, and $\tau \in \mathcal{R}^+$ is a temperature parameter. Note that for each anchor, there is 1 positive pair and 2N-2 negative pairs. The denominator has a total of 2N-1 terms (the positive and negatives).

In the case of supervised contrastive learning, each anchor has multiple positive pairs since labels can be used. The SupCon loss function takes the following form:

$$\mathcal{L}^{supcon} = \sum_{i \in I} \mathcal{L}_i^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)} \tag{4}$$

where $P(i) \equiv \{p \in A(i) : \tilde{y}_p = \tilde{y}_i\}$ is the set of all positive indices except itself in the two-view minibatch, and $|P(i)|$ is its cardinality. Note that each anchor has more than one positive pair and the others are negative pairs.

3.2.2. Improvement by the Extra Negative Supervised Contrastive Loss Function

The SupCon loss function is reviewed in Equation (4). Like with the end-to-end contrastive learning method [19], batch size can significantly affect training performance [26]. In the literature, the best results were achieved with a batch size of 6144, which was very demanding on computational resources. Our experiments in Section 4.4.3 show that with a batch size limit of 256, FER models trained with the SupCon loss function still have huge room for improvement. According to our analysis, this is because the SupCon loss function considers too few positives and negatives. In addition, empirically, negative pairs play a larger role than positive pairs; positive pairs consider feature learning within a class and, consequently, fewer positive pairs can usually meet the demand, whereas negative pairs consider the feature distinctions between classes, which involves learning more classes of discriminative features. This is more complicated; therefore, more negative samples are needed to produce an effect. From a macro point of view, directly increasing the batch size does not change the ratio of positives to negatives in a minibatch. So, even if the batch size is increased, the negatives are still relatively limited.

In order to obtain more negatives, the ECAN utilizes the memory bank to provide extra features. One simple extension method is to directly use all memory bank feature groups as negatives; in this paper, that approach is called "extra supervised contrastive loss" (ESC loss) and takes the following form:

$$\mathcal{L}^{esc} = \sum_{i \in I} \mathcal{L}_i^{esc} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot (z_a + z_m) / \tau)} \tag{5}$$

where z_m represents all feature groups in the memory bank. In fact, this loss has the same flaw as the self-supervised contrastive loss function even though it adds negative samples. There is still a non-negligible number of false negatives in z_m. Training with false negatives is essentially equivalent to promoting inter-class separation, which is a negative optimization. The ENSC loss function is a further improvement over the ESC loss function because it only retains true negatives to ensure that negative optimization does not occur. The ENSC loss function takes the following form:

$$\mathcal{L}^{ensc} = \sum_{i \in I} \mathcal{L}_i^{ensc} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(z_i \cdot z_p / \tau\right)}{\sum_{a \in A(i)} \exp(z_i \cdot (z_a + z_{mn}) / \tau)} \tag{6}$$

where $z_{mn} \in z_m|_{y_m \neq y_i}$ represents all view features in z_m that are different from the class of z_i (i.e., extra negative samples). Each anchor of the ENSC loss function can accurately push away more negatives to elevate deep feature separation between different expressions. In addition, the same intra-class learning effect obtained with the SupCon loss function can be achieved because z_p is not reduced.

On the other hand, our negative sample screening strategy also implies the suppression of class imbalance. The number of z_{mn} is significantly different for each expression class. When z_i belongs to the majority class, z_{mn} provides more negatives, which makes it more difficult and slower to reduce this part of the loss. After a period of training, the main contribution of the loss comes from the minority class, which can make the network refocus on those minority classes that were not valued originally. This alleviates the negative impact on the feature processing network caused by the class imbalance issue of wild FER datasets.

3.3. Multi-View Affinity Loss

The center loss function is a widely adopted DML method in which the similarities between the deep features and their corresponding class centers are measured. It aims to partition the deep feature space into K clusters for a K-class classification problem. The objective function in the center loss function minimizes the within-cluster sum of squares between the deep features and their corresponding class centers. Assuming that a minibatch has m samples, the deep features of each sample are represented by $x_i \in \mathcal{R}^{m \times d}$, and the corresponding class center is $c_{yi} \in \mathcal{R}^{m \times d}$. The center loss function can be written as follows:

$$\mathcal{L}^{\text{center}} = \frac{1}{2m} \sum_{i=1}^{m} \left\| x_i - c_{y_i} \right\|_2^2 \tag{7}$$

In recent years, many variants of the center loss function have been proposed to enhance its training effect, such as DAN [13], which proposes an affinity loss function for the FER task. It can be written as follows:

$$\mathcal{L}^{af} = \frac{\sum_{i=1}^{m} \left\| x_i - c_{y_i} \right\|_2^2}{\sigma_c^2} \tag{8}$$

where σ_c^2 represents the variance between class centers. Like most center loss function variants, the affinity loss function only measures the distance between the deep features of the current minibatch and their corresponding class centers. The authors of this work believe that not only the current minibatch deep features but also the deep features of the other minibatches have the potential to contribute to learning class centers. To test this idea, this paper presents the newly designed MVA loss function, which takes full advantage of the historical features of the memory bank.

Notably, since historical feature groups cannot directly optimize the network weights through back-propagation, we treat them as an adjustment coefficient α_v in order to weight the affinity loss. If we denote $v \in \{1 \ldots V\}$ as the index of z_q and consider $V - 1$ historical feature groups, the α_v takes the following form:

$$\alpha_v = 1 + \frac{\sum_{v=1}^{V} \left(\mathcal{L}_v^{af} - \overline{\mathcal{L}^{af}} \right)^2}{V} \tag{9}$$

where $\mathcal{L}_v^{af}$ represents the v-th feature group value of the affinity loss, and $\overline{\mathcal{L}^{af}}$ represents the mean of the $\mathcal{L}_v^{af}$ of all V feature groups. The MVA loss function takes the following form:

$$\mathcal{L}^{mva} = \alpha_v \cdot \frac{\sum_{i=1}^{m} \left\| x_i - c_{y_i} \right\|_2^2}{\sigma_c^2} \tag{10}$$

In fact, α_v is dynamically adjusted following the $\mathcal{L}_v^{af}$ of all V feature groups. When α_v is large, it indicates that the current class center has seriously deviated from the true class center; thus, the MVA loss function punishes the model more severely to learn the true class center. On the contrary, when all values of $\mathcal{L}_v^{af}$ are nearly equal, it indicates that the current class center is close enough to the true value. The value of α_v approaches 1 in this case, and the MVA loss function degenerates into the affinity loss function. Optimizing the MVA loss function can increase the compactness of intra-class features and eventually leads to learning more accurate class centers.

In the end, the loss functions employed in this paper are the ENSC loss function, the MVA loss function, and the cross-entropy loss function for classification. The total loss can be summarized in the following form:

$$\mathcal{L} = (1 - \lambda_1)\mathcal{L}^{cls} + \lambda_1 \mathcal{L}^{ensc} + \lambda_2 \mathcal{L}^{mva} \tag{11}$$

where λ_1 represents the ratio between $\mathcal{L}^{ensc}$ and $\mathcal{L}^{cls}$, and λ_2 represents the individual weight coefficient of $\mathcal{L}^{mva}$.

4. Experiment and Results

In this section, we describe the experimental evaluation results of the proposed method in detail. The data augmentation experiments were designed first to study the impact of data augmentation on the ECAN. Ablation experiments were then designed to evaluate the impact of hyper-parameters and analyze the improvements offered by the ENSC and MVA loss functions compared to the SupCon and center loss functions, respectively. Next, the effect of ECAN on the class imbalance issue was visually demonstrated using a confusion matrix. Finally, the performance of the proposed method compared to other state-of-the-art methods was evaluated on two widely used wild FER datasets.

4.1. Datasets

RAFDB. The RAFDB dataset [27] is a real-world FER dataset with over 29,670 face images downloaded from the Internet. The dataset adopts labeling methods such as the manual labeling of single-label and double-label subsets, as well as the automatic labeling of landmark locations. There are a total of 15,339 images that can be used for seven classes of expression classification; each image is cropped to 100 × 100 fixed pixels, and the images are divided into 12,271 training images and 3068 test images. This dataset suffers from an extreme class imbalance issue.

FER2013. The FER2013 dataset [28] was released for the FER competition of the ICML2013 Expression Learning Challenge. The dataset consists of 35,886 facial expression images collected outside the laboratory, with different light intensities and shooting backgrounds. The dataset consists of 28,708 training set images, 3589 public validation set images, and 3589 public test set images. It has the same classes of expressions as the RAFDB dataset and also has the issue of class imbalance. We used the training set for training and reported the recognition accuracy obtained on the public test set.

4.2. Implementation Details

For the RAFDB dataset, we used the official aligned samples, and the input images were reshaped to 224 × 224 pixels. For the FER2013 dataset, we used the original 48 × 48 pixel grayscale samples. The three levels of data augmentation were not exactly the same for these two datasets due to the large differences in samples. The baseline of this paper consists of the ResNet-18 network and the cross-entropy loss function. We used a model pretrained on the MS-Celeb-1M facial recognition dataset as the backbone of the Siamese network for the RAFDB dataset.

We used Pytorch to implement the experimental code and trained it on an Nvidia RTX-3090 GPU. For all tasks, the weight hyper-parameters of total loss were set to $\lambda_1 = 0.2$ and $\lambda_2 = 0.7$ by default. The number of historical feature group hyper-parameters (v) of the proposed MVA loss function was set to 5 by default.

For the training strategy, we trained the model using the SGD optimizer with the same configuration as GAF [29]. The batch size was set to 256, and the initial learning rate was 0.1. We trained 40 epochs on the RAFDB dataset because the RAFDB dataset uses a pretrained model. For the FER2013 dataset, we trained 200 epochs to achieve the best results.

4.3. Data Augmentation Experiment

Research by SIMCLR [19] in the context of self-supervision shows that data augmentation can have a significant impact on contrastive learning. In this paper, data augmentation was divided into three levels (none, weak, and strong) to evaluate the performance gap between the baseline and ECAN methods. Table 1 shows the data augmentation configurations for the RAFDB and FER2013 datasets. Table 2 shows the corresponding accuracy results.

Table 1. Configurations of different data augmentation levels for the RAFDB and FER2013 datasets.

		RAFDB			FER2013	
Augment Level	**No**	**Weak**	**Strong**	**No**	**Weak**	**Strong**
RandomCrop	-	-	✓	-	✓	✓
RandomColorJitter	-	-	✓	-	-	✓
RandomHorizontalFlip	-	✓	✓	-	✓	✓
RandomRotation	-	-	✓	-	✓	✓
RandomErasing	-	✓	✓	-	✓	✓
RandomGrayscale	-	✓	✓	-	✓	✓
RandomGaussianBlur	-	-	✓	-	-	-
RandomAffine	-	-	-	-	-	✓
FiveCrop	-	-	-	-	-	✓

Table 2. The accuracy results of the baseline and ECAN methods at different data augmentation levels.

		RAFDB			FER2013	
Method	**No**	**Weak**	**Strong**	**No**	**Weak**	**Strong**
Baseline	85.69%	86.70%	87.97%	62.02%	70.60%	72.33%
ECAN	86.80%	87.97%	89.77%	63.53%	72.28%	73.73%

As can be seen from the results, our ECAN method achieved the highest accuracy with a strong level of data augmentation. In addition, ECAN was more accurate than the baseline at all levels of data augmentation. This suggests that ECAN still works even when the level of data augmentation is inappropriate. This paper uses the strong data augmentation level as the default setting for the following experiments.

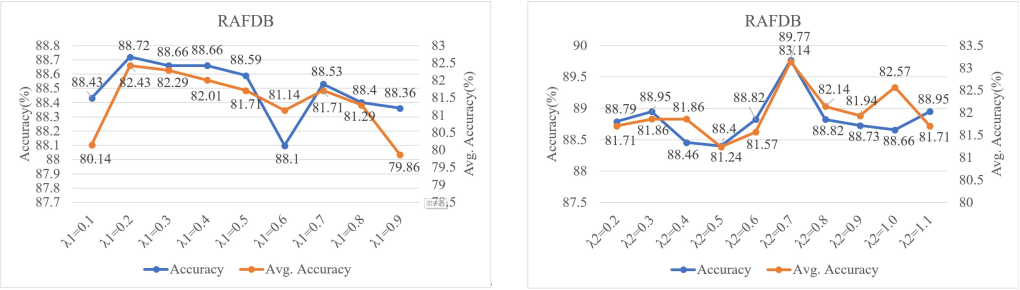
4.4. Ablation Studies

To verify the effectiveness of the proposed ECAN, we conducted an ablation study on the RAFDB dataset. This paper provides data not only on the accuracy but also the average accuracy, which is a more precise metric for the imbalanced samples in the RAFDB dataset.

4.4.1. Different Values of the Weight Hyper-Parameters λ_1 and λ_2

We first investigated the ratio between the proposed ENSC loss function and the cross-entropy loss function. The value of λ_1 controls how much they contribute; a larger λ_1 makes the total loss more concentrated in the ENSC loss function. Figure 3a shows the accuracy and average accuracy data obtained for the RAFDB dataset with different values of λ_1. It can be seen that the ENSC loss function significantly improved average accuracy. When $\lambda_1 = 0.2$, the average accuracy was significantly improved, but accuracy did not increase proportionally. This result shows that the ENSC loss function effectively improves the prediction accuracy of the minority class, resulting in an increase in average accuracy. However, since the majority class contributes more to the accuracy, there was no significant increase in accuracy. When $\lambda_1 > 0.2$, both the accuracy and average accuracy started to decrease steadily. This was because the linear classification layer could not be trained efficiently due to the reduced contribution of the cross-entropy loss function.

Next, we studied the effect of different values of λ_2 on FER performance when λ_1 = 0.2. Figure 3b shows the accuracy and average accuracy data obtained for the RAFDB dataset with different values of λ_2. It can be seen that the best result was obtained at $\lambda_2 = 0.7$ and that a larger value $(1.1 > \lambda_2 > 0.7)$ was better than a smaller value $(0.2 < \lambda_2 < 0.7)$. This was because the MVA loss function requires the training of class centers, and a small value of λ_2 caused the training of class centers to be slower.



(**a**)

(**b**)

Figure 3. Ablation studies using the RAFDB dataset. (**a**) The impact of the weight hyper-parameter λ_1. (**b**) The impact of the weight hyper-parameter λ_2.

4.4.2. Different Values of the Historical Feature Group Hyper-Parameter v for the MVA Loss Function

Figure 4 shows the effect of using different values of v on the MVA loss function. It can be seen that the training results were not very stable when v was small $(v < 4)$. This was because the MVA loss function considered fewer historical feature groups, which led to insufficient training robustness. The training results started to gradually decline when $v > 5$, which was due to the outdated historical feature groups that could no longer accurately represent the current model and thus impaired feature learning. In summary, the hyper-parameter v needs to be adjusted according to the actual situation to achieve the best effect.

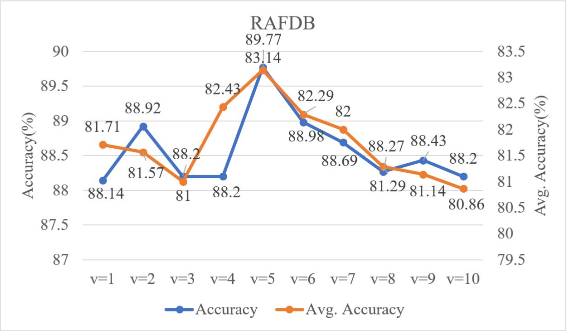


Figure 4. Ablation studies for the historical feature group hyper-parameter v of the proposed MVA loss function using the RAFDB dataset.

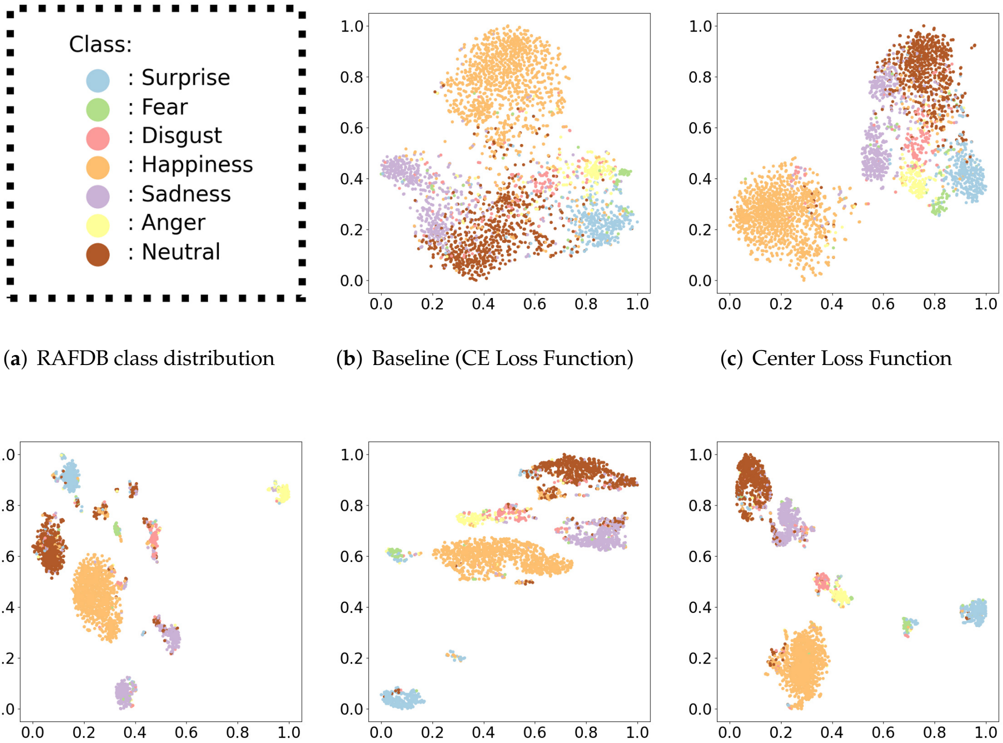
4.4.3. Effects of the ENSC and MVA Loss Functions

Table 3 presents the results of the performance comparison of our ENSC and MVA loss functions with other loss functions. The average accuracy is the mean of the diagonal values in the confusion matrix, which is often used to evaluate the true performance of a model on class-imbalanced datasets. It can be seen that our ENSC loss function displayed a large improvement in average accuracy compared to the SupCon loss function. Likewise, the MVA loss function also played a crucial role in the final performance.

Table 3. Performance comparison of different loss functions using the RAFDB dataset.

Loss Function	Accuracy (%)	Avg. Accuracy (%)
Cross-Entropy	87.97	80.43
Center	88.62	81.86
SupCon	88.01	79.45
ENSC	88.72	82.43
MVA + ENSC	89.77	83.14

Figure 5 shows T-SNE visualizations of different loss functions to more clearly assess the impact of these loss functions on deep features. Figure 5b is the result of the baseline; its inter-class boundaries are blurred, and the distribution of feature points is very chaotic.



(**a**) RAFDB class distribution (**b**) Baseline (CE Loss Function) (**c**) Center Loss Function

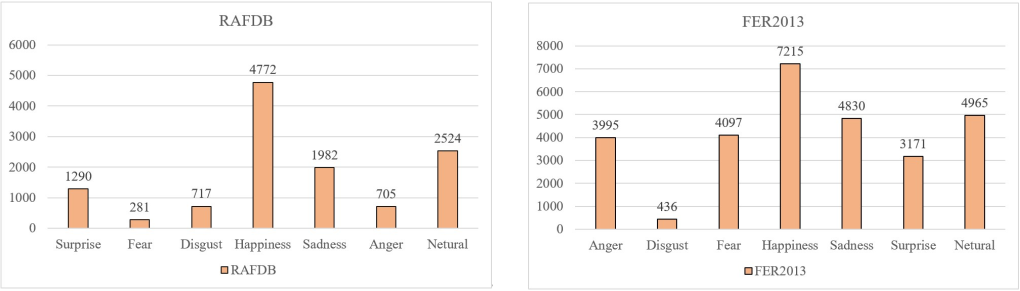
(**d**) SupCon Loss Function (**e**) ENSC Loss Function (**f**) ECAN Method

Figure 5. The T-SNE visualization results for different loss functions applied to the RAFDB test set. The axes reflect the positions of all feature points on the two-dimensional plane. (**a**) Distribution of classes for the RAFDB dataset. (**b**) Cross-entropy loss function (CE loss function). (**c**) Center loss function. (**d**) SupCon loss function. (**e**) ENSC loss function. (**f**) ECAN (ENSC and MVA loss functions).

From Figure 5c, it can be seen that the center loss function aggregated the “happy” class, which had the largest sample size, into a single cluster while clustering other classes into a cluster of similar size; this was the result of the class imbalance issue. In addition, the boundaries between different classes were also very blurred. Figure 5d is the best result of the SupCon loss function at a batch size of 256. The SupCon loss function made the feature points within the class dense by virtue of having multiple positive pairs. However, it could not accurately cluster the feature points of each class. Not only are the feature points of "surprise", "disgust", "happiness", "neutral", etc. clustered in the middle to form false clusters, but the "sadness" class is also wrongly separated. Figure 5e is the result of our ENSC loss function. It improved upon the results of the SupCon loss function because it had enough negative pairs. It can be seen that most of the feature points are correctly clustered, and the inter-class boundaries of each cluster are obvious. In addition, the improvement of the ENSC loss function in terms of the class imbalance issue can be clearly seen when compared with the results for the center loss function. Figure 5f is the final result of our ECAN method. The MVA loss function further improved the compactness of each cluster when compared to the results for the ENSC loss function alone. The near absence of discrete feature points that do not belong to any cluster indicates that the MVA loss function learned more accurate class centers.

4.5. Confusion Matrix Analysis

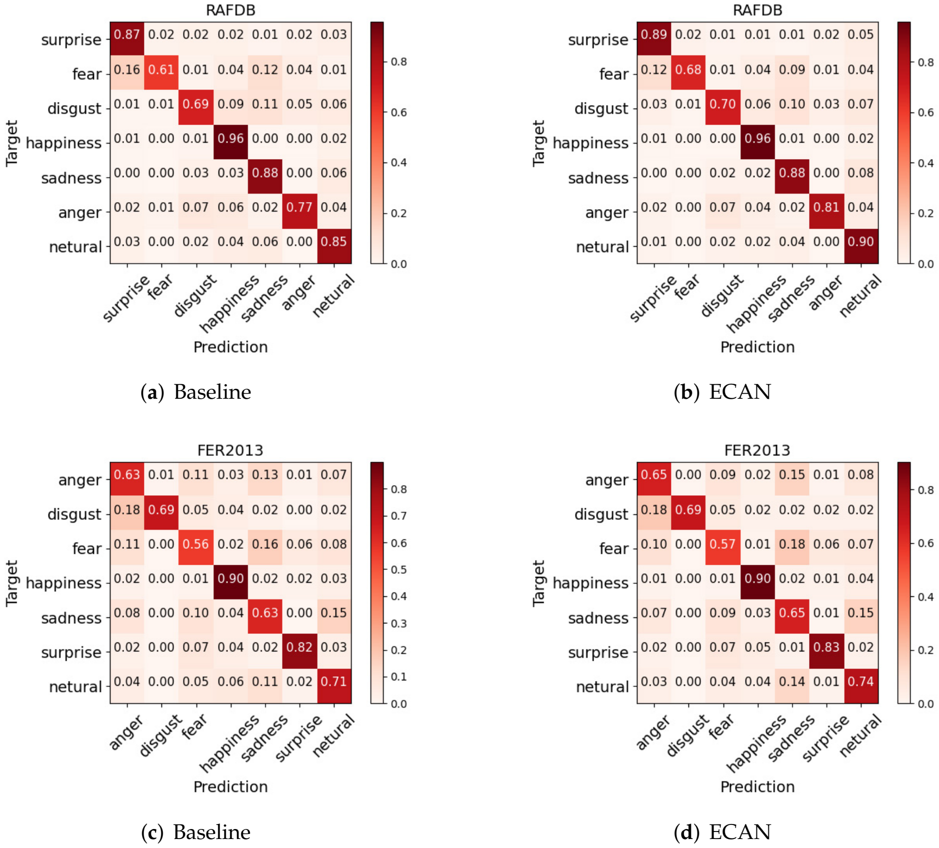
We analyzed the accuracy results for each expression class using a confusion matrix. Figure 6 shows the distribution of the training sets for the RAFDB and FER2013 datasets. Figure 7 shows the confusion matrix results for the baseline and ECAN methods when applied to the RAFDB and FER2013 datasets.



(**a**) (**b**)

Figure 6. The distribution of training set samples in the RAFDB (**a**) and FER2013 (**b**) datasets. It can be seen that surprise, fear, disgust, and anger belong to the minority class in the RAFDB dataset, and disgust belongs to the minority class in the FER2013 dataset.

Figure 7a,b reflects the results of the baseline and ECAN methods for the RAFDB test set. It is observed that the accuracy of the ECAN method improved for the minority class. The categories with the smallest sample sizes, such as fear and anger, had the largest improvement, reaching 7% and 4%, respectively. These results fully reflect the effect of the ECAN method on the class imbalance issue. Not only that, but the accuracy rates for all other expression classes except the happy class were also improved, which shows that our ECAN method effectively distinguished confusing expression features. Figure 7c,d shows the results for the FER2013 dataset. The ECAN method improved accuracy for all classes except disgust and pleasure. The authors believe that the ECAN method did not improve the accuracy rate for disgust because the discriminative features of the disgust class are so obvious that our baseline was sufficient.



(**a**) Baseline

(**b**) ECAN

(**c**) Baseline

(**d**) ECAN

Figure 7. Confusion matrix results for the baseline and ECAN methods on the RAFDB and FER2013 test sets.

4.6. Comparison with Previous Results

We compared our ECAN method and recent state-of-the-art methods using the RAFDB and FER2013 datasets, and the results are presented in Tables 4 and 5. This paper presents not only the macro-f1 score results but also the average accuracy results for the RAFDB dataset, given that some recent methods use average accuracy as a more refined evaluation metric.

Table 4. Performance comparison for the RAFDB dataset.

Method	Accuracy (%)	Avg. Accuracy (%)	Macro-F1 Score
DDA-loss [16]	86.90	79.71	-
Ad-Corre [30]	86.96	-	0.8057 *
SCN [31]	87.03	-	-
DACL [12]	87.78	80.44	-
PSR [32]	88.98	80.78	-
EfficientFace [33]	88.36	-	-
MviT [34]	88.62	-	-
DAN [13]	89.70	82.75 *	-
Baseline (Strong Augment)	87.97	80.43	0.8153
ECAN (ours)	89.77	83.14	0.8429

* Note: The average accuracy given in the original paper that presented DAN was 85.32%, but the true value should be 82.75%. The authors of DAN acknowledged the true average accuracy value in their github issues. The macro-f1 score of the Ad-Corre method was obtained by taking the mean value of the f1 score for each expression category given in the original paper.

Table 5. Performance comparison for the FER2013 dataset.

Method	Accuracy (%)	Macro-F1 Score
Attentional ConvNet [35]	70.02	-
Shao et al. [36]	71.14	-
MBCC [37]	71.52	0.7029 *
BreG-NeXt [38]	71.53	0.7100
Ad-Corre [30]	72.03	0.7123 *
VGG [39]	73.28	-
Baseline (Strong Augment)	72.33	0.7157
ECAN (ours)	73.73	0.7325

* Note: The macro-f1 score of the Ad-Corre and MBCC methods were obtained by taking the mean value of the f1 score for each expression category given in the original papers.

The proposed ECAN method achieved 89.77% accuracy and 83.14% average accuracy on the RAFDB dataset and 73.73% accuracy on the FER2013 dataset, which is better than the current state-of-the-art methods based on the ResNet-18 network architecture. These results clearly demonstrate the robust learning effect of the proposed ECAN for wild FER datasets.

4.7. Discussion

In this section, we discuss some valuable information discovered during the development of the ECAN method. When we conducted the data augmentation experiments, we found that data augmentation had a very significant effect on our baseline; these effects sometimes even exceeded the results of some of the latest methods. Such results suggest that data augmentation may be a key factor in improving performance not only for contrastive learning but also for other methods. We did not consider the class imbalance issue in the initial design of the ENSC loss function; the experimental results proved that the ENSC loss function was effective at addressing the issue, but after consideration, we believe that it may be a feasible idea to use historical minibatch samples to augment those classes with insufficient sample sizes. Experiments have fully demonstrated that our ECAN method is highly competitive with respect to wild FER datasets, but it still has some problems and should be improved. Experience shows that when the number of historical feature groups considered by the MVA loss function was larger, the training of the model was more stable; however, at the same time, this led to performance degradation. The question of how to solve this problem is one future research direction. Moreover, other annotations, such as multi-label annotations and landmarks, provided richer supervised information. In addition, if a scheme is designed to associate deep features with such supervised information, the model should be able to learn more accurate feature clusters.

5. Conclusions

This paper proposes a contrastive-learning-based ECAN method, i.e., an extra-contrast affinity network. It is used to effectively learn highly discriminative features in wild FER scenes. ECAN contains two newly proposed metric loss functions to optimize deep feature output using a feature processing network. The proposed ENSC loss function simultaneously considers multiple positive pairs and more extra negative pairs to effectively improve inter-class separation and intra-class compactness in cases of extreme class imbalance. The proposed MVA loss function utilizes historical feature groups to dynamically learn more accurate class centers and can further improve intra-class compactness. Experiments on two widely used wild FER datasets demonstrated the improvements of the ECAN over other methods.

Author Contributions: Conceptualization, J.Z. and Y.S.; Data curation, J.Z.; Formal analysis, J.Z.; Funding acquisition, S.L.; Investigation, J.Z.; Methodology, J.Z.; Project administration, S.L.; Resources, S.L.; Software, J.Z.; Supervision, S.L.; Validation, J.Z. and S.L.; Visualization, Y.S.; Writing—original draft, J.Z. and S.Y.; Writing—review & editing, S.L. and S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Project of the National Natural Science Foundation of China under Grant No. 62106023.

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

References

1. Ekman, P. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol*. **1971**, *17*, 124–129. [CrossRef] [PubMed]
2. Sun, Z.; Wang, G.; Jin, L.; Cheng, C; Zhang, B.; Yu, J. Noise-suppressing zeroing neural network for online solving time-varying matrix square roots problems: A control-theoretic approach. *Expert Syst. Appl.* **2022**, *192*, 116272. [CrossRef]
3. Jin, L.; Wei, L.; Li, S. Gradient-based differential neural-solution to time-dependent nonlinear optimization. *IEEE Trans. Autom. Control* **2022**, 1. [CrossRef]
4. Jin, L.; Li, J.; Sun, Z.; Lu, J.; Wang, F. Neural dynamics for computing perturbed nonlinear equations applied to ACP-based lower limb motion intention recognition. *IEEE Trans. Syst. Man Cybern*. **2022**, *52*, 5105–5113. [CrossRef]
5. Sun, Z.; Shi, T.; Jin, L.; Zhang, B.; Pang, Z.; Yu, J. Discrete-time zeroing neural network of O(τ^4) pattern for online time-varying nonlinear optimization: Application to manipulator motion generation. *J. Frankl. Inst. Eng. Appl. Math.* **2021**, *358*, 7203–7220. [CrossRef]
6. Liu, K.; Liu, Y.; Zhang, Y.; Wei, L.; Sun, Z.; Jin, L. Five-step discrete-time noise-tolerant zeroing neural network model for time-varying matrix inversion: Application to manipulator motion generation. *Eng. Appl. Artif. Intell.* **2021**, *103*, 104306. [CrossRef]
7. He, K.; Fan, H.; Wu, Y.; Xie, S.; Ross G. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
8. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823.
9. Sohn, K. Improved deep metric learning with multi-class N-pair loss objective. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 1857–1865.
10. Wen,Y.; Zhang, K.; Li, Z.; Yu, Q. A discriminative feature learning approach for deep face recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 499–515.
11. Cai, J.; Meng, Z.; Khan, A.; Li, Z.; O'Reilly, J.; Tong, Y. Island loss for learning discriminative features in facial expression recognition. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 302–309.
12. Farzaneh, A.; Qi, X. Facial expression recognition in the wild via deep attentive center loss. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 2401–2410.
13. Wen, Z.; Lin, W.; Wang, T.; Xu, G. Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv* **2021**, arXiv:2109.07270v3.
14. Liu, X.; Vijaya Kumar, B.V.K.; You, J.; Jia, P. Adaptive deep metric learning for identity-aware facial expression recognition. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 522–531.
15. Meng, Z.; Liu, P.; Cai, J.; Han, S.; Tong, Y. Identity-aware convolutional neural network for facial expression recognition. *J. Syst. Eng. Electron.* **2017**, *28*, 784–792.
16. Farzaneh, A.; Qi, X. Discriminant distribution-agnostic loss for facial expression recognition in the wild. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 1631–1639.
17. Li, Y.; Lu, Y.; Li, J.; Lu. G. Separate loss for basic and compound facial expression recognition in the wild. In Proceedings of the 11th Asian Conference on Machine Learning (ACML), Nagoya, Japan, 17–19 November 2019; pp. 897–911.
18. Tian, Y.; Hénaff, O.; Oord, A. Divide and contrast: Self-supervised learning from uncurated data. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10043–10054.
19. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv* **2020**, arXiv:2006.10029.
20. Tian, Y.; Krishnan, D.; Isola, P. Contrastive multiview coding. *arXiv* **2019**, arXiv:1906.05849.
21. Bachman, P.; Hjelm, R.; Buchwalter, W. Learning representations by maximizing mutual information across views. *arXiv* **2019**, arXiv:1906.00910.
22. Hénaff, O.; Koppula, S.; Alayrac, J.; Oord, A.; Vinyals, O.; Carreira, J. Efficient visual pretraining with contrastive detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10066–10076.

23. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv* **2020**, arXiv:2006.09882.
24. Li, J.; Zhou, P.; Xiong, C.; Socher, R.; Hoi, S. Prototypical contrastive learning of unsupervised representations. *arXiv* **2020**, arXiv:2005.04966.
25. Grill, J.; Strub, F.; Altché, F. Bootstrap your own latent: A new approach to self-supervised Learning. *arXiv* **2020**, arXiv:2006.07733.
26. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised Contrastive Learning. *arXiv* **2020**, arXiv:2004.11362.
27. Li, S.; Deng, W. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Trans. Image Process.* **2019**, *28*, 356–370. [CrossRef] [PubMed]
28. Goodfellow, I.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H. Challenges in representation learning: A report on three machine learning contests . *Off. J. Int. Neural Netw. Soc.* **2015**, *64*, 59–63. [CrossRef]
29. Liu, M.; Chen, X.; Du, X.; Jin, L. Activated gradients for deep neural networks. *IEEE Trans. Neural Netw. Learn. Syst*. **2021**, *99*, 1–13. [CrossRef]
30. Fard, A.; Mahoor, M. Ad-Corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access* **2022**, *10*, 26756–26768. [CrossRef]
31. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6896–6905.
32. Vo, T.; Lee, G.; Yang, H.; Kim, S. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access* **2020**, *8*, 131988–132001. [CrossRef]
33. Zhao, Z.; Liu, Q.; Zhou, F. Robust lightweight facial expression recognition network with label distribution training. *AAAI Conf. Artif. Intell.* **2021**, *35*, 3510–3519.
34. Li, H.; Sui, M.; Zhao, F.; Zha, Z.; Wu, F. MViT: Mask vision transformer for facial expression recognition in the wild. *arXiv* **2021**, arXiv:2106.04520.
35. Minaee, S.; Abdolrashidi, A. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors* **2021**, *21*, 3046. [CrossRef] [PubMed]
36. Shao, J.; Qian, Y. Three convolutional neural network models for facial expression recognition in the wild. *Neurocomputing* **2019**, *355*, 82–92. [CrossRef]
37. Shi, C.; Tan, C.; Wang, L. A facial expression recognition method based on a multibranch cross-connection convolutional neural network. *IEEE Access* **2021**, *9*, 39255–39274. [CrossRef]
38. Hasani, B.; Negi, P.; Mahoor, M. BReG-NeXt: Facial affect computing using adaptive residual networks with bounded gradient. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1023–1036. [CrossRef]
39. Khaireddin, Y.; Chen, Z. Facial emotion recognition: State of the art performance on FER2013. *arXiv* **2021**, arXiv:2105.03588v1.