information

Article

Mixing Global and Local Features for Long-Tailed Expression Recognition

Jiaxiong Zhou [1], Jian Li [2], Yubo Yan [2], Lei Wu [3] and Hao Xu [1,2,*]

[1] School of Artificial Intelligence, Jilin University, Changchun 130012, China
[2] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[3] College of Software, Jilin University, Changchun 130012, China
* Correspondence: xuhao@jlu.edu.cn

Abstract: Large-scale facial expression datasets are primarily composed of real-world facial expressions. Expression occlusion and large-angle faces are two important problems affecting the accuracy of expression recognition. Moreover, because facial expression data in natural scenes commonly follow a long-tailed distribution, trained models tend to recognize the majority classes while recognizing the minority classes with low accuracies. To improve the robustness and accuracy of expression recognition networks in an uncontrolled environment, this paper proposes an efficient network structure based on an attention mechanism that fuses global and local features (AM-FGL). We use a channel spatial model and local feature convolutional neural networks to perceive the global and local features of the human face, respectively. Because the distribution of real-world scene field expression datasets commonly follows a long-tail distribution, where neutral and happy expressions account for the tail expressions, a trained model exhibits low recognition accuracy for tail expressions such as fear and disgust. CutMix is a novel data enhancement method proposed in other fields; thus, based on the CutMix concept, a simple and effective data-balancing method is proposed (BC-EDB). The key idea is to paste key pixels (around eyes, mouths, and noses), which reduces the influence of overfitting. Our proposed method is more focused on the recognition of tail expression, occluded expression, and large-angle faces, and we achieved the most advanced results in occlusion-RAF-DB, 30° pose-RAF-DB, and 45° pose-RAF-DB with accuracies of 86.96%, 89.74%, and 88.53%.

Keywords: expression recognition; attention mechanism; long-tailed distributions

check for updates

Citation: Zhou, J.; Li, J.; Yan, Y.; Wu, L.; Xu, H. Mixing Global and Local Features for Long-Tailed Expression Recognition. *Information* **2023**, *14*, 83. https://doi.org/10.3390/info14020083

Academic Editor: Alessandra Lumini

Received: 26 December 2022
Revised: 25 January 2023
Accepted: 29 January 2023
Published: 1 February 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Facial expressions play a vital role in communication, and the automatic recognition of facial expressions is crucial in various fields. In the field of human–computer interaction, an environment adaptation system that detects a user's emotional state can be developed: for example, an intelligent classroom system that helps teachers identify and address students' emotional states while also fostering a positive learning environment [1]. In the medical field, pain detection is used to detect the progress of patients. Facial expression recognition is also frequently used in the field of intelligent driving to help detect whether a driver is sleepy or inattentive. Six basic expressions and neutral expressions are proposed according to psychological research [2], and the facial emotion recognition (FER) task is for classifying the input face image into seven categories: disgust, anger, happiness, fear, surprise, sadness, and neutral. Currently, the accuracy of the FER task in the laboratory scene significantly improved [3,4]. The faces in datasets in the laboratory scene face the camera with clear expressions, balanced lighting, and correct labels. However, the expression recognition of datasets containing large-scale real-world scenes faces significant difficulties. The main difficulties come from the occlusion of real faces and uneven facial postures; the data distribution is unbalanced, and the trained model is biased toward recognizing the majority class, whereas the recognition accuracy rate for the minority class is very low. The key

challenge is figuring out how to share visual knowledge between the head class and tail class and how to reduce confusion between the tail and head classes [5].

Problems such as occlusion and large-angle faces affect the expression recognition of real-world scenes [6,7], as shown in Figure 1. Deep learning has the potential to judge the current emotional state using facial data or even just by extracting information from the eyes [8]. To recognize the facial expression under occlusion, some researchers divide the face into different modules, giving high weights to the occluded module and low weights to the unobstructed module [9], enabling the trained model to focus on identifying the unobstructed module. In addition, some researchers use three-dimensional (3D) face models to extract human face information [10]. Generally, 3D face models can only use a single image to obtain the coordinate information of the key points of the face and recognize the facial expression state of adverse factors such as light scattering to obtain more abundant facial skeleton information, thereby obtaining good results. Some 3D models also use a generative adversarial network (GAN) to repair some image data that do not exist in two-dimensional (2D) space [11]. However, the former will increase the model's complexity and training time and make it difficult to converge. For large-scale expression recognition, some researchers used the model framework of federated learning [12], which uses a few labeled personal facial expression data samples to train the local model in each training round and aggregates all local model weights in the central server to obtain the global optimal model, which is the pioneering work in the few-shot task of facial expression recognition. Some researchers use the idea of a facial expression action unit [13] to learn the generated action mask [14]. First, they use the average difference between the neutral face and the corresponding expression face as the training guide and combine the prior domain knowledge. Some researchers, considering the uncertainty of fuzzy expression labels [15–17], adopt distributed labels or suppress the expression of uncertain labels to improve the accuracy of large-scale expression recognition [18]. However, some researchers proposed a new feature decomposition and reconstruction learning method for effective facial expression recognition [19]. They regard expression information as a combination of shared information about different expressions and unique information about each expression. Our proposed method is based on the lightweight shuffleNet V2 network [20], a model with fewer parameters and faster training; we use the attention mechanism to fuse the local and global feature information of the face (AM-FGL). The local feature information is primarily for face occlusion and large-angle faces; global feature information can obtain rich face information and improve recognition accuracies.



Figure 1. A large-scale facial expression dataset. In the real scene, the location of the face is not fixed. Some faces are at the center of the picture, and some are at the corner of the picture; the pose of the face is uncertain, including 30°, 45°, and 90° faces. Some faces are blocked, some facial features cannot be seen, and the image is unclear, making facial expression recognition difficult.

Real-world datasets are typically highly imbalanced [21], they have a long-tail distribution [22] and contain very few minority class data points and many majority class data points. For example, the numbers of neutral and happy expressions occupy the vast majority, whereas the proportion of expressions such as fear and disgust is very small, as shown in the two most commonly used large-scale facial expression datasets in Figure 2. Existing methods for solving such problems are oversampling and undersampling. Oversampling refers to the repeated sampling of a minority class to generate more minority class labels; however, this method lacks rich minority class context information, which can easily cause overfitting [23], and degrades the generalization ability of the classifier [24] because repeatedly chosen samples lack diversity but have highly similar image contexts. Undersampling is aimed at the majority class, ignoring part of the data according to the sample distribution. Although this method has no overfitting problem, many sample data points are ignored, making it difficult for the model to be fully trained [25]. We propose a novel minority class oversampling method, BC-EDB, for augmenting diverse minority class samples by pasting the key features of the face of the minority class such, as fear, into the rich context images [26] of the majority class, such as happiness, using them as background images, and fusing the labels of these two types of images to expand the number of minority groups. Unlike existing resampling methods that ignore majority class samples, our method aims to generate new minority class samples using the rich information of majority class samples, which is an interpolation of majority and minority class samples. Thus, in the decision boundary, the environment generates diverse data, which improves the generalization performance of the minority class. We increased the number of minority categories. The number of minority categories is not too small and minority categories are rich in background in order to reduce the influence of overfitting [27]. Our proposed method achieves advanced accuracy on large-scale FER datasets without increasing model complexity, and it achieves the most advanced accuracy in tail expression, occluded expression, and large-angle faces.

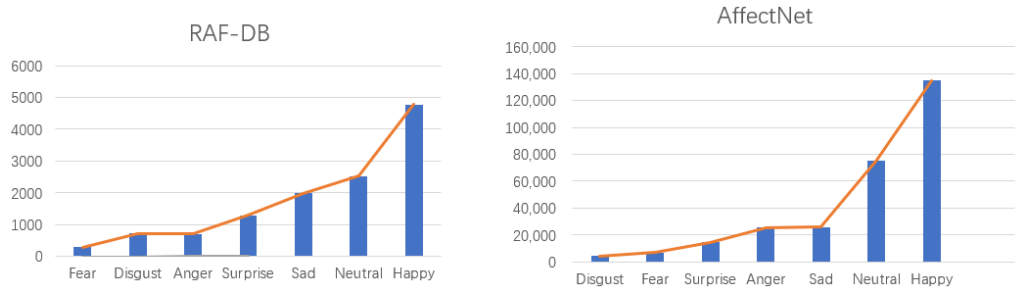


Figure 2. RAF-DB and AffectNet are large-scale facial expression datasets, where happy and neutral expressions occupy the vast majority of facial expression data, and other expressions only account for a small part, which conforms to a long-tail distribution. A model trained on such a dataset seldom refers to the characteristics of tail expression data.

Human expressions are complex, and there is complexity when classifying an expression into one category; some complex expressions may contain complex emotions. Occasionally, different expression recognition professionals have different interpretations of the same expression. Therefore, it is difficult to classify an expression into one category. To improve the recognition accuracy of expressions and better cooperate with CutMix [28] fusion tags, we employ distributed tags to deal with this problem [29].

Our contributions are as follows:

1. We propose an attention mechanism that combines global and local features to extract richer face information (AM-FGL).
2. Based on the CutMix data enhancement method for solving the problem of facial expression data imbalance in real-world scenes, we increased the diversity of the dataset without increasing the complexity of the model BC-EDB.

3. We improved the recognition accuracy of minority classes and achieved the most advanced results in occlusion-RAF-DB, 30° pose-RAF-DB, and 45° pose-RAF-DB with accuracies of 86.96%, 89.74%, and 88.53%.

The rest of this paper is organized as follows: Section 2 introduces related work on facial expression recognition. Section 3 briefly shows the details, which describe the overall architecture of the model. Section 4 presents the experimental results. In Section 5, we provide a summary of this research and explore potential avenues for further investigation.

2. Related Work

There are primarily two types of databases in the expression recognition field. The first dataset type includes datasets collected in the laboratory scene and obtained by researchers or professional actors under laboratory conditions, such as MMI, JAFFE, and CK+. Such facial expression datasets are very ideal even though they can easily deviate from the actual scene. The second dataset type includes datasets collected in the natural scene, such as RAF-DB and AffectNet. At this time, avoiding complex situations is difficult, such as occlusion and head posture changes. Moreover, facial expression datasets obtained under natural conditions will typically have a long-tail distribution, with many neutral and happy expressions and fewer fear and disgust expressions. In this case, the accuracy of the facial expression recognition task is not as good as that in the one observed in laboratory scenes. Furthermore, research on facial expression recognition-related algorithms in real-world scenes is more inclined toward improving facial expressions recognition robustness under more challenging real-world conditions [30].

2.1. Large-Scale Expression Data Research

In real scenes, the core areas (eyes and mouth) of human faces expressing emotions will be lost due to the interference of occlusion, face pose changes, and other issues [31], which will greatly affect such a scenario's accuracy with respect to model training and testing. For this reason, some methods are considered to normalize the face (frontal face and no occlusion) [32], and then the expression recognition model is trained. Early research on illumination normalization was mainly based on isotropic diffusion and based on discrete cosine transform, and histogram illumination normalization. Chieh-Ming Kuo proposed a weighted summation method that combines histogram equalization and linear mapping [33]; the most advanced effect was achieved at that time; some main content for pose normalization is to convert some side face images into frontal face images [34]. Zhipeng Bao proposed a single-image facial expression recognition method that is robust to facial orientation and care conditions [35], that is, to reconstruct a 3D face model from a single image. On this basis, a novel end-to-end deep neural network was proposed that utilizes both re-centered 3D models and FER task landmarks, such as the Rotate-and-Render method proposed by Hang Zhou [36]. First, the method can use a single picture for 3D reconstruction; secondly, the front face image of the original image is synthesized by back projection, and then the GAN network is used to repair the facial information that was lost when the side face was converted into a front face [37]. The scene under uncontrolled conditions—pose normalization—may lead to changes in facial expressions [38]. The occlusion problem is more complicated than the face pose change, mainly because the occlusion phenomenon is not very regular. The occlusion part may be any part of the face, and the occlusion may also be varied, such as masks and hair. In order to solve the facial expression recognition problem of people wearing masks, Pablo Barros modified the AffectNet dataset, proposed the MaskedAffect datasets for mask scenarios, and proposed the corresponding neural network structure—FaceChannel [39]—which is a lightweight convolutional neural network. It can adapt to different interaction scenarios, the number of network parameters is small, and it can produce a good effect in the scene of wearing a mask. We cannot treat the above problems in isolation because, in real-scene expression recognition, the above problems may appear at the same time. Kai Wang proposed a new regional attention network (RAN) to adaptively capture facial regions for occlusion

and pose [40]. Influenced by the changes in FER recognition, the author constructed six real-world occlusion and pose change model datasets from the existing large-scale datasets, FERPlus and AffectNet, and proposed a regional bias loss to encourage the most important regions to produce higher attention weights.

2.2. Research on Imbalanced Data

A widespread strategy for addressing long-tail distributions is to resample and rebalance the training data [41,42] either by sampling from rare classes more frequently [43] or by reducing the number of samples from common classes [44]. The former creates redundancy and quickly encounters the problem of overfitting rare classes since no complex background information is added, whereas the latter loses key information contained in large sample sets. Another approach is to introduce extra weights for different classes, which makes it very difficult to optimize the model in large-scale recognition scenarios. CutMix cuts out a part of the area but does not fill in 0 pixels, randomly fills in the area pixel values of other data in the concentrated training, and integrates the classified labels of the two according to the beta distribution. Based on CutMix, Seulki Park proposed Context-rich [27]. This method adopts the minority group oversampling method and diversifies the minority group samples by using the rich background of most categories as the background image pixels. The key idea is to paste the image of the minority category onto the rich background image of the majority category. This method proves the effectiveness of the oversampling method by conducting many experiments and ablation studies. The complexity of long-tail noisy face datasets has made traditional methods, such as resampling and reweighting, impractical. Yaoyao Zhong proposed a training strategy that treats head and tail data differently [45], which provides two training data streams, and the first training data stream uses head data to learn discrimination based on noise resistance, and the second data stream uses tail data to gradually mine out stable discriminant information from confused tail classes. These two data streams complement each other, share weights, and can save a substantial amount of GPU resources. Yu-Xiong Wang proposed a meta-learning network framework that uses the idea of transfer learning to transfer the learned knowledge of the rich head-class data to the poor tail-class data, and the final network can capture the concept of dynamic models [46]. Using the concept of data enhancement, Kangkang Zhu proposed a multi-modal solution based on a generative confrontation network, which combines the models of data enhancement and facial expression recognition [47]. This solution can process both 2D and 3D data, and the generated high-intensity expression data for recognition achieved good results. To achieve excellent performance, Hongxiang Gao, motivated by continuous learning, randomly selected samples from the head class and the up-sampled tail class, reconstructed multiple subsets, introduced a pre-trained backbone network [48], and learned weights by repeatedly training the target model.

3. Method

We propose a lightweight neural network with an attention mechanism that integrates global and local features. Different from previous research [49], we comprehensively consider the impact of local and global features on facial expression recognition and discuss their integration ratio in subsequent experiments. We assign the same weight to local and global features during training and then output the learned features via the attention module. To solve the long-tail distribution problem of datasets and improve the recognition accuracy of minority classes, we propose an oversampling method based on CutMix. The key idea of this method is to fuse the key pixels of minority classes as the foreground and the images of majority classes as the background in proportion to increase the number of minority classes. The overall architecture of the proposed method is shown in Figure 3.

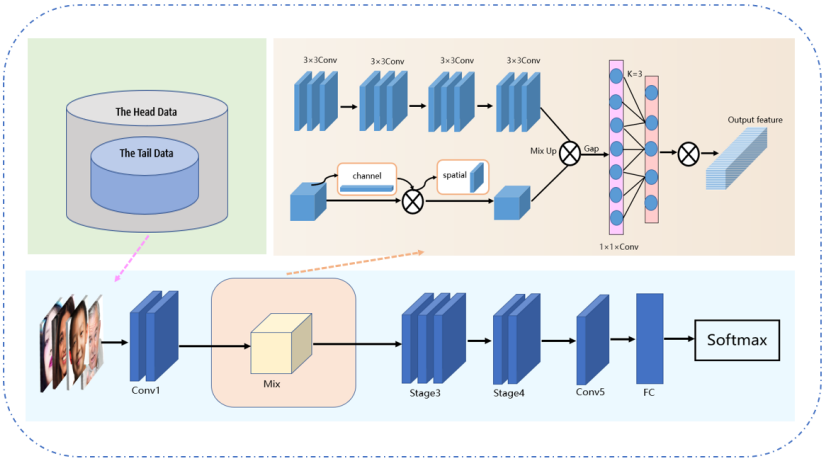


Figure 3. In the overall architecture framework diagram, the upper left corner is our proposed oversampling method, and the upper right corner for data enhancement is our proposed attention mechanism-based fusion of global and local features; that is, the mix layer. This part consists of three modules: channel space modulator, local feature layer, and attention mechanism module. At the bottom is our backbone network.

3.1. Backbone Network Based on ShuffleNet-v2

We adopt the concept of how to develop a lightweight network proposed by Ning Ma [20] and others to construct our network. In his paper, the author mentioned four principles: G1: equal channel width minimization memory access cost (MAC); G2: excessive group cooperation increases MAC; G3: network fragmentation reduces the degree of parallelism; G4: element-wise operations are non-neutral. We use ShuffleNet-V2, which was proposed by the same author as our backbone network, which is composed of Conv1, stage, mi (and our attention mechanism-based fusion of global and local feature layer), ConV5, full connection, and softmax layers. The size of our input image is $224 \times 224 \times 3$. After passing the Conv1 layer, the 2D global feature mapping will become $Fconv1 = R^{H \times W \times C}$. Here, $H = W = 56$, and $C = 29$. After entering the mix layer, the feature map will enter the global and local feature modules, respectively. The local feature module has four parallel patches, with each patch exhibiting $F_i^{local} = R^{H_i \times W_i \times C_i}$, where $H = W = 28$ and $C = 29$. After two layers of convolution, $F_i^{local} = R^{H_i \times W_i \times C_i}$, $i \in \{1, 2, 3, 4\}$. Here, $H = W = 14$ and $C = 116$, and then we will splice the characteristic graphs obtained by the four patches, $F_i^{local} = R^{H_i^{'} \times W_i^{'} \times C_i^{'}}$, where $H = W = 28$ and $C = 116$. We use the channel space modulator to remove the redundant global features of the original stage2 layer to obtain the key global feature information. In stage 3, $Fstage3 = R^{H_{stage3} \times W_{stage3} \times C_{stage3}}$, $H_{stage3} = W_{stage3} = 14, C_{stage3} = 232$, stage 4 $Fstage4 = R^{H_{stage4} \times W_{stage4} \times C_{stage4}}$, $H_{stage4} = W_{stage4} = 7$, and $C_{stage4} = 464$, and finally, we output our final classification results via the softmax layer.

3.2. Mixing Up Global Features and Local Features Based on Attention Mechanisms

Global features are used to perceive comprehensive image information, whereas local features learn the image information of different local positions of the face. The output weight ω can be formulated as follows:

$$\omega = \sigma(WY) \tag{1}$$

where Y represents the fusion of global and local features and does not require dimensionality reduction. W represents a matrix, as shown in Equation (2), including $K \times C$ parameters, which is a method for capturing local cross-channel interactions. It aims to

ensure efficiency and effectiveness, without increasing the complexity of the network, and to reduce training time. σ is a Sigmoid function.

$$\begin{bmatrix} \omega^{1,1} & \cdots & \omega^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \omega^{2,2} & \cdots & \omega^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \omega^{C,C-k+1} & \cdots & \omega^{C,C} \end{bmatrix} \tag{2}$$

We did not use the method of group calculation, avoiding the complete independence between different groups and allowing us to compromise between sharing weight and computing efficiency:

$$\omega_i = \sigma\left(\sum_{j=1}^{k} \omega^j y_i^j\right), y_i^j \in \Omega_i^k \tag{3}$$

where Ω_i^k denotes the collection of k-neighboring channels of y_i^j. The weight of y_i is determined by evaluating the interaction solely between y_i and its k neighboring channels, and ω^j implies that we let all channels share the same weight information:

$$\omega = \sigma(C1D_k(y)) \tag{4}$$

where $C1D$ represents $1D$ convolution, and this approach can be easily implemented using a high-speed $1D$ convolution with a kernel size of k:

$$Y = F_{global} + F_{local} \tag{5}$$

where F_{global} represents global information that is fused by stage2 and channel-spatial. After the global feature information of the high-dimensional stage2 is obtained, channel and spatial heat maps will be determined independently at two parallel branches [50] represented by $F_{channel}$ and $F_{spatial}$, respectively, which can be described as follows.

$$F_{global} = \sigma(F_{channel}(stage2) \times F_{spatial}(stage2)) \tag{6}$$

The channel and spatial modules have been proposed by Sanghyun Woo et al. [51]. The channel module aggregates the input information of the feature map via average and max pooling operations, sends the output from the two operations to the shared network (multilayer perception) to produce the channel attention map, and then applies element-wise summation, finally outputting the feature vector. The spatial module aggregates the channel module's information and generates two 2D maps via two pooling operations. Both modules are connected and convolved by a standard convolution layer to generate the final 2D spatial attention map. The acquisition of F_{local} (local features) is divided into four patches based on the face feature space. The patch is calculated by two 3×3 CNN, and after each convolution, the ReLU function is applied.

We then integrate global and local features, given that the aggregation feature is obtained via global average pooling (GAP). Then, a fast 1D convolution with the size of k is performed to generate the channel weight, where k is determined adaptively via the mapping of channel dimension C. The entire process can be expressed as follows.

$$\omega = \sigma\left(W\left(F_{global} + F_{local}\right)\right) \tag{7}$$

3.3. Data Enhancement Based on CutMix

For the long-tail problem of datasets, we propose a new oversampling method based on CutMix data enhancement. We use the context of the majority class sample to diversify the limited context of the minority class sample. We use a CAM [52] heat map to determine whether the key pixels of expression recognition are concentrated in the facial backbone area. CAM can map the output of the classification back to the original image and reveal

which part of the image has a significant impact on the final classification result, as shown in Figure 4. We paste a few key facial backbone images, including eyes, nose, and mouth, onto the majority of class samples. We use interpolation blur for the clipping region's boundary according to the distribution of most class samples.



Figure 4. Images from the RAF-DB dataset. The heat map results show that the key pixels are concentrated in the eyes, nose, and mouth, which proves the effectiveness of our key pixels.

We mix the images and ground truth labels of the training data. Moreover, we assume that $X \varepsilon R^{W \times H \times C}$ represents the training image and that Z represents the label. Our goal is to generate a new training sample, (X, Z), using two training samples (χ_A, z_A) and (χ_B, z_B), and the new training sample, (X, Z), is used to train the model with the original loss function. We define the merging operation as follows:

$$\begin{aligned} X &= M\Theta\chi_A + (1 - M)\Theta\chi_B \\ Z &= \lambda z_A + (1 - \lambda) z_B \end{aligned} \tag{8}$$

where M represents a binary mask, and if it is 1, the pixel needs to be filled; if it is 0, the location does not need to be filled. If a binary is set to 1, it is filled. Θ denotes element multiplication, and the combination ratio, λ, represents the sampling ratio between the two samples, which conforms to the $Beta(\alpha, \alpha)$ distribution. The equation does not directly represent the majority or minority of the sample. Considering the particularity of our expression recognition scene, we comprehensively consider the sampling data distribution of the foreground, (χ_A, z_A), and background samples, (χ_B, z_B), and the background samples should be biased toward the majority class. The foreground sample should be biased toward the minority class. Therefore, it is necessary to sample the background samples from the original data distribution. Foreground samples (χ_A, z_A) are sampled from the weighted distribution of secondary categories; background samples (χ_B, z_B) are sampled from most classes. Next, we use the above formula to combine the training datasets.

CAM is a method that can realize computer vision visualization. For classification tasks, CAM can determine which part of an image affects the final classification judgment. The calculation method is as follows: For a CNN model, GAP is applied to the last feature map to calculate the mean value of each channel, and then the maximum classification score is calculated by mapping the full connection layer to the classification score, which is the judgment result of the model's output. Then, the gradient of the largest category output relative to the last feature map is calculated, and the gradient is visualized on the original diagram as a thermodynamic diagram. Generally, this method judges which part of the high-level features extracted by the network has a greater impact on the final classification. Based on the above algorithm principle, before starting our data enhancement experiment, we first determine where the most influential pixels are mainly concentrated for the expression recognition tasks, known as the key pixels. We incorporate a CAM module into our proposed algorithm model. The incorporated module does not affect the operation of our algorithm. The purpose is to find the gradient of high-dimensional features that can influence the classification results when finding the key pixels. Our experimental results are as follows.

To obtain the key point position of the face, we use the Dlib algorithm [53] to mark key face landmarks. The Dlib algorithm first recognizes the position of the face and then recognizes the 68 key point positions of the face and, finally, the face landmark information. In total, 68 key point positions were marked, and each key point can be represented by a coordinate, marking the position of the key pixel for expression recognition. We calculate

the coordinates of the key points of the eye, nose, and mouth; calculate the convex hull; and generate a mask containing key pixels, as shown in Figure 5. In addition, the majority class face images are calculated to indicate where to paste. Considering the different sizes of the foreground and background images, we use the Platts analysis method to align the sizes and positions of the two masks. The formula is expressed as follows:

$$\sum_{i=1}^{68} \left\| sRp_i^T + T - q_i^T \right\|^2 \tag{9}$$

Here, R represents a 2×2 orthogonal matrix, s represents a scalar quantity, T represents a 2D vector, and p_i and q_i denote the rows of the matrices labeled foreground and background images, respectively.

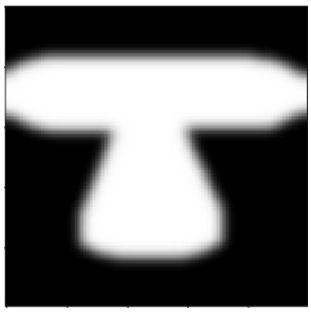


Figure 5. Large-scale facial expression dataset map. In a real-world scene, the position of the face is not fixed. Some faces are located in the center of the picture, whereas some faces are located at the corner of the picture. The pose of the face is uncertain, including 30°, 45°, and 90° faces. Some faces are blocked, not all facial features can be seen, and the image is unclear, which makes facial expression recognition difficult.

After we finish face alignment, in the final paste operation, we use the union of the masks of the two faces. To make the generated image more natural, we adjusted the color of the minority image to match the background image, as shown in Figure 6.

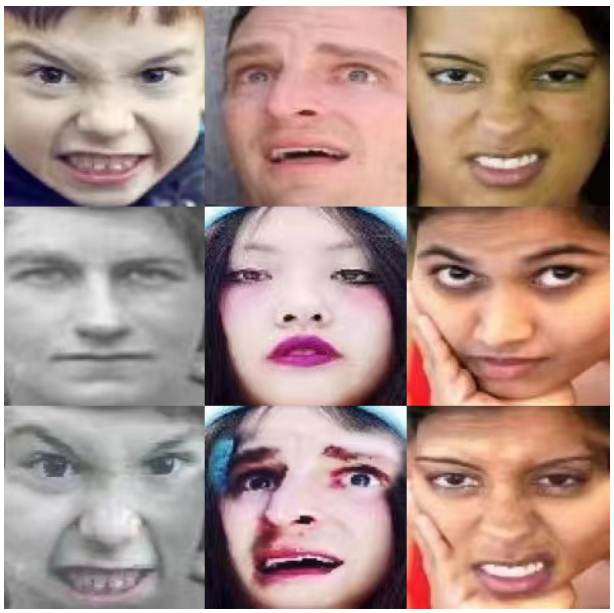


Figure 6. The image in the first row is the dataset that we need to expand, and we use it as the foreground image. The image in the second row is the head dataset, which will be used as the background image, and the image in the third row is our final fused new image, which we will use as the training model.

4. Experiment

4.1. Datasets

To verify the effectiveness of our method, we train our model on two unbalanced large-scale datasets—RAF-DB and AffectNet; restricted conditions—occlusion-RAF-DB; and pose change datasets—30° pose-RAF-DB and 45° pose-RAF-DB—in our experiment.

RAF-DB [54] is a large-scale expression dataset, marked by 40 staff members, and contains neutral, happy, sad, surprised, fear, disgust, and angry expressions (seven types of expressions). It contains more than 30,000 facial images, including 12,271 images as training data and 3068 images as test data. The number of various expressions is presented in Table 1.

Table 1. We classify them according to the labels given by RAF-DB and show the data of each category. The number of happy and neutral expressions is 4772 and 2524, respectively, which is significantly higher than those of fear, disgust, and anger expressions.

Expression	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger
numbers	2524	4772	1982	1290	281	716	704

AffectNet [55]: This dataset contains approximately 420,000 pictures, which are manually divided into 11 categories by labeling by professionals. Affectnet-7 is a category that comprises seven expression labels that we manually extracted according to the labels in the dataset. Compared with AffectNet-7, AffectNet-8 adds an eighth expression: contempt. We use AffectNet-7 and AffectNet-8 as our experimental datasets. The training dataset contains more than 280,000 images, and the test dataset contains about 3500 images. The distribution of the seven types of expressions of Affectnet-7 is shown in Table 2.

Table 2. Sample data from the AffectNet datasets, which originally had 11 classes. We extracted the seven expression classes that we needed from the datasets, as shown below. There are 134,915 happy and 75,374 neutral expressions, which are significantly more than fear, disgust, and anger expressions.

Expression	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger
numbers	75,374	134,915	25,959	14,590	6878	4303	25,383

Occlusion-RAF-DB and Pose-RAF-DB [56], which are subsets of RAF-DB: Occlusion-RAF-DB contain 735 images of different occlusion levels. Pose-RAF-DB includes 30° Pose-RAF-DB and 45° Pose-RAF-DB, which represent face poses greater than 30 degrees and face poses greater than 45 degrees. They contain 558 images and 1247 images, respectively.

4.2. Experimental Operation Details

For the images in all datasets, we use Retinaface [57] to detect the position of the face, as shown in Figure 7; mark the information of the five points of the eyes, nose, and mouth of the face; use affine transformation to align the face area [58,59]; cut out the center picture of the face; and crop and adjust the size to 224 × 224 pixels. In addition to face alignment, we use random cropping and random horizontal flipping to reduce the influence of overfitting. We use the distributed pretraining label and distributed label [29] and optimize the parameters via the stochastic gradient descent optimizer. Our experiments are conducted on a TITAN GPU running on 24 GB of memory. We trained 200 epochs, and the batch size is 128. The initial learning rate is 0.01, the Python version is 3.6, and the PyTorch version is 1.12.

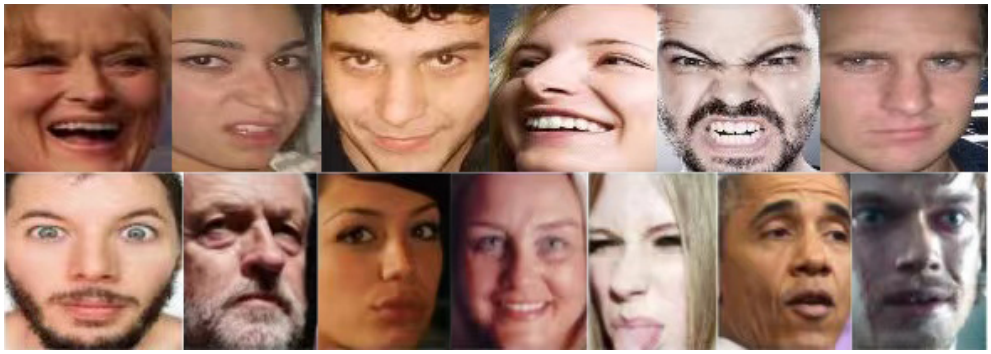


Figure 7. Sample images from RAF-DB and AffectNet datasets. We use the Retinaface algorithm to detect the location rectangle of the face and the location map of the five key points of the face. We calculate the angle to be converted via the positions of the center points of the left and right eyes and then align the face using affine transformations. The alignment effect is shown in the above figure.

We use our proposed data enhancement method (BC-EDB) proposed to expand the numbers of neutral, sad, surprise, fear, disgust, and anger expressions in RAF-DB to approximately 4500 images and also expand the categories of neutral, sad, surprise, fear, disgust, anger, and contempt expressions in the AffectNet dataset to approximately 90,000 to supplement the number of minority classes. To verify the identification effectiveness of our enhanced datasets in the minority classes, we compared the recognition accuracy of our proposed method for each class in the RAF-DB dataset with that of the EfficientFace network, RUL and EAC. The results are shown in Table 3, which can prove that our method significantly improved the classification accuracy of fear, disgust, and anger. Moreover, we proposed an attention mechanism that combines global and local features to extract richer face information (AM-FGL), which achieves state-of-the-art results on occlusion-RAF-DB and pose-RAF-DB.

Table 3. We compare the accuracy of our algorithm for each class in the RAF-DB dataset with that of the EfficientFace algorithm, RUL, and EAC, mainly comparing the recall and macro-f1. The results in the table show that our algorithm can significantly improve the recognition accuracy of fear, disgust, and anger minority classes.

	Classes	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger
Recall	Efficientface [29]	81.84	95.39	84.59	87.83	68.81	76.40	83.36
	RUL [60]	87.28	94.68	86.50	88.31	72.58	71.52	85.16
	EAC [61]	86.20	96.68	86.91	91.00	76.56	73.42	88.44
	Ours	85.88	93.64	85.43	92.73	78.95	77.62	89.33
Macro-f1	Efficientface [29]	84.92	94.25	83.79	85.44	57.24	57.24	81.97
	RUL [60]	87.54	95.35	87.77	87.49	66.18	69.45	83.28
	EAC [61]	88.06	96.27	88.79	90.00	68.70	72.96	84.14
	Ours	87.18	95.02	88.03	86.73	71.01	73.21	85.90

Evaluation metrics: We show the recall, precision, and macro-f1 to describe our experiment. We determine the recognition accuracy of each class. We present our recognition accuracy for each class when the overall recognition accuracy is the highest.

4.3. Experimental Results Compared with the State-of-the-Art Method

We compare our results with those of state-of-the-art expression recognition methods and conduct ablation studies under the same experimental environment. The experimental results are shown in the following table.

We compared the latest methods on the RAF-DB, AffectNet-7, and AffectNet-8 datasets, as shown in Table 4. All were trained in our environment according to the open-source code, and their accuracy was calculated. From the table, the recall accuracy of our method on the

RAF-DB, AffecNet-7, and AffecNet-8 datasets are 89.11%, 65.03%, and 61.11%, respectively, which is better than the accuracy of EfficientFace on RAF-DB (0.75%), AffecNet-7 (1.33%), and AffecNet-8 (1.22%). Our method is 0.97% better than SCN* and 2.21% higher than RAN on the RAF-DB dataset. Our method is 2.69% better than the DDA-Loss method on the AffecNet-7 dataset, 3.51% higher than FMPN, and 6.25% higher than gACNN. On the AffecNet-8 dataset, it is 3.11%, 1.61%, 0.59%, and 1.22% better than Weighted Ross, RAN, SCN, and EfficientFace.

Table 4. Comparison with state-of-the-art methods on RAF-DB, AffectNet-7, and AffectNet-8.

Datasets	Methods	Recall	Precision	Macro-f1
RAF-DB	RAN [40]	86.90	88.23	87.21
	SCN [56]	87.03	87.72	87.25
	SCN * [56]	88.14	88.38	88.17
	Efficient Face [29]	88.36	88.68	8845
	RUL [60]	88.98	89.21	89.07
	EAC [61]	90.12	90.22	90.14
	Ours	89.11	89.62	89.26
AffectNet-7	gACNN [26]	58.78	59.96	59.01
	FMPN [14]	61.52	63.01	62.02
	DDA-Loss [62]	62.34	63.33	62.42
	Efficient Face [29]	63.70	65.29	64.01
	DAN [63]	65.12	65.38	65.21
	EfficientNet-B2 [64]	66.13	67.65	67.34
	Ours	65.03	65.26	65.11
AffectNet-8	Weighted-Loss [55]	58.00	59.24	58.53
	RAN [40]	59.50	61.02	59.81
	SCN [56]	60.52	63.52	61.16
	Efficient Face [29]	59.89	62.30	59.43
	DAN [63]	61.82	62.01	61.89
	EfficientNet-B2 [64]	63.02	63.53	63.06
	Ours	61.11	64.62	61.59

* means that the algorithm is trained with RAF-DB and AffectNet-8.

Although our results are not as good as that of RUL and EAC on RAF-DB, as we said, we mainly focus on facial expression recognition in minority classes and restricted scenes. Thus, we show the identification accuracy of each class in RAF-DB in Table 3. No matter the accuracy of recall or macro-f1, the recognition accuracy of minority facial expressions, such as disgust, fear, and anger, is higher than RUL and EAC. Based on this, it can be proven that our proposed method can indeed improve the recognition accuracy of minority classes.

The method we proposed reached state-of-the-art expectations in limited scenes, such as occlusion and face posture change, as shown in Tables 5 and 6. To verify that our model can deal with occlusion and multi-pose problems, we conducted the following experiments on Occlusion-RAF-DB and Pose-RAF-DB datasets. In these experiments, we performed data enhancement processing on the datasets. From the data in the table, our method can obtain the most advanced accuracy on these datasets. The accuracy of our method on the occlusion-RAF-DB is 86.96%. The accuracy of our method on the 30° Pose-RAF-DB and 45° Pose-RAF-DB datasets is 89.74%, and 88.53%, respectively, which comprise the highest accuracy. The proposed method is 0.97% and 0.49% higher than RUL and EAC in occlusion-RAF-DB. Our proposed method surpasses RUL and EAC by 1.37% and 0.65%, respectively, in the 30° Pose-RAF-DB dataset, and it also demonstrates an improvement of 0.54% and 0.18% over RUL and EAC in the 45° Pose-RAF-DB datasets. These results prove that the proposed method can exhibit higher robustness on occlusion and multi-pose datasets.

Table 5. Comparison with state-of-the-art methods on occlusion-RAF-DB.

Datasets	Methods	Accuracy
Occlusion-RAF-DB	ResNet-18 [65]	80.19
	RAN [40]	82.72
	EfficientFace [29]	83.24
	RUL [60]	85.99
	EAC [61]	86.47
	Ours	86.96

Table 6. Comparison with state-of-the-art methods on pose-RAF-DB.

Datasets	Methods	Accuracy	
		Pose $\geqslant 30^{\circ}$	Pose $\geqslant 45^{\circ}$
Pose-RAF-DB	ResNet-18 [65]	84.04	83.15
	RAN [40]	86.74	85.20
	EfficientFace [29]	88.13	86.92
	RUL [60]	88.37	87.99
	EAC [61]	89.09	88.35
	Ours	89.74	88.53

To explore the fusion ratio of local and global features, we conducted the following experiments using the RAF-DB datasets. The results are shown in Table 7. Other configurations of the experiments were consistent with our formal experiments. To verify this fusion ratio, we also used our proposed data enhancement method to expand our datasets. The experimental results prove that the local features with a proportion of 0.5 and the global features with a proportion of 0.5 are optimal for the final result. An increase in the proportion of local or global features has no improved experimental effect. The experimental results are shown in the table. We believe that we must comprehensively analyze the local and global features. Biasing either side will result in a low feature fusion effect of the other side, and the feature with a low proportion has a low impact on the final experimental results, thereby reducing the final effect of the experiment. When we do not add local features or do not add global features, as the proportion of local features is 0 or the proportion of global features is 0, the result of the experiment is often lower than the result of adding both global and local features. Therefore, it is necessary to mix up global and local features.

Table 7. The fusion ratio of local and global features. The first column is the fusion factor of local features, the second column is the fusion factor of global features, and the third column is the experimental accuracy after they are fused and passed to the attention module.

Local Features	Global Features	Accuracy
0	1	87.68
0.2	0.8	88.12
0.3	0.7	88.61
0.4	0.6	88.21
0.5	0.5	89.11
0.6	0.4	88.38
0.7	0.3	87.98
0.8	0.2	87.92
1	0	87.78

To explore the partition granularity of the patch of the local features, we also conducted the following experiments. We divided the local features into 8, 12, and 16 patches to divide the local features of the face into different granularity and obtain more rich face information. The experimental results are as follows in Table 8. Our experimental results are run on the

enhanced RAF-DB datasets. By conducting an analysis, we believe that an increase in the number of patches does not necessarily result in more local feature information. In contrast, it may bring more redundant information and reduce the results of the experiment. If we introduce this redundant feature information into high-dimensional features, the model cannot distinguish the differences between different categories.

Table 8. The number of patches of local features. When we increased the number of patches, our experimental accuracy did not increase but decreased. Finally, we found that the effect was best when the number of patches was 4.

Patch	Patch Accuracy
4	89.11
8	88.26
12	87.38
16	86.20

According to the results, we found that the higher the number of patches, the lower the accuracy. We analyzed other feature information of the face that would be lost after dividing the number of patches into finer grains. Facial expressions cannot be simply and directly divided into more fine-grained areas. Because this may cause the blurring of boundary information, our accuracy will not improve after the number of patches increases.

4.4. Ablation Experiment

We verified the effectiveness of each component proposed via ablation experiments, as shown in Table 9. We conducted ablation research on our experiments using the RAF-DB, AffectNet-7, and AffectNet-8 datasets. In these experiments, we used EfficientFace as the baseline, with or without the AM-FGL module and BC-EDB module, to verify the impact of these two modules on the final experimental results and evaluate their effectiveness. We conducted our experiments on RAF-DB, AffectNet-7, and AffectNet-8 datasets.

Table 9. A checkmark, ✓, means that we have added this module, and no ✓means that it does not contain this module. Experiments can prove that the two added modules are effective.

Datasets	AM-FGL	BC-EDB	Accuracy
RAF-DB			88.36
	✓		88.60
		✓	88.91
	✓	✓	89.11
AffectNet-7			63.70
	✓		63.85
		✓	64.78
	✓	✓	65.03
AffectNet-8			59.89
	✓		60.03
		✓	60.52
	✓	✓	61.11

All experiments were conducted in the same environment. For the RAF-DB datasets, after adding only the AM-FGL module with fusion features, the experimental result is 0.24% better than that of the baseline network. After we only implemented the method of the BC-EDB module based on data enhancement, we improved the baseline network by 0.55%. For the AffectNet-7 dataset, our experimental result is 0.15% better than that of the baseline network after adding only the attention mechanism module with fusion features. After we implemented only the method of data enhancement based on oversampling, we improved the baseline network by 1.08%. For the AffectNet-8 dataset, our experimental

result is 0.14% better than that of the baseline network after adding only the attention mechanism module with fusion features. After we implemented only the method of data enhancement based on oversampling, we improved the baseline network by 1.22%. The ablation experiments prove that these two modules are effective and that the effect of the data enhancement method based on oversampling is better than that of the other module. Our fusion feature module, based on the attention mechanism, can fully perceive the global and local feature information of the face and will not add too much redundant feature information. Using the channel spatial module, we can extract global feature information; using the local feature extraction module, we can obtain more abundant local feature information without distortion. Our novel proposed oversampling method is based on CutMix data enhancement, which can fully supplement the number of datasets of minority classes. Because we use the majority classes as the background, our model does not overfit, and the recognition accuracy for minority classes is improved, making the model less biased toward identifying the majority classes.

Via the above experiments, we demonstrated the effects of the two proposed methods compared with the baseline. The proposed AM-FGL method has high robustness, verifying our ideas on the two large-scale datasets. Notably, our BC-EDB method can achieve a relatively large improvement on the baseline and sufficiently proves that the proposed data enhancement method can increase the number of minority classes. The accuracy of the test set shows that there is less overfitting phenomenon.

5. Conclusions

In this paper, we propose two methods: an efficient network structure based on an attention mechanism that fuses global and local features (AM-FGL) and a simple and effective data-balancing method based on CutMix (BC-EDB). The former can obtain specific facial information via the attention mechanism, and it is mainly used to solve the problem of occlusion and large-angle faces. The latter expands the number of tail datasets, and it is used to solve the long-tail distribution problem of the datasets. Our recognition accuracy reached the most advanced level in the case of minority classes, occlusion, and large-angle faces.

AM-FGL is a lightweight network that combines global and local features to solve the problems of face occlusion and large-angle faces in real-world scenes. It contains three modules: the attention mechanism module, the global features module, and the local features module. The global features module uses the channel space modulator to extract the global features of the face. By dividing the face into multiple patches, two groups of 3D CNN were designed for each patch. The ReLU activation function is used to extract the local features of the face. After that, the global and local features are fused; then, the proposed attention mechanism proposed is used. The proposed attention mechanism can share weight globally and can improve accuracies without reducing dimensions.

BC-EDB is a data enhancement method and can reduce the influence of overfitting. For the problem of long-tail distributions, we proposed our method based on CutMix. The method uses minority-class images key pixels as the foreground and majority-class images as the background. By conducting CAM experiments, we proved that in our facial expression recognition task, the positions around the eyes, nose, and mouth are key pixels. We fuse the foreground and background images, and we also fuse their tags according to the proportional distribution to expand the number of minority categories. By conducting many experiments, we proved that the proposed method has strong robustness under various conditions and exhibits advanced accuracy.

In the future, although the proposed data enhancement method can increase the number of a few classes by pasting the key pixels of a few classes to prevent uneven sampling, we used most areas where a few classes of faces generate expressions for pasting. Although we use the CAM algorithm to verify the effectiveness of our acquisition of key pixels, we do not know whether these pixels impact the results, and we performed the sampling proportion according to the Beta distribution. For the special scene in our

facial expression recognition task, we will explore whether other sampling methods are more effective.

Author Contributions: J.Z.: Conceptualization, methodology, software, data curation, writing—original draft preparation, and writing—reviewing and editing; J.L.: visualization, writing—reviewing and editing; Y.Y.: investigation; L.W.: validation; H.X.: supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (62077027), the Ministry of Science and Technology of the People's Republic of China (2018YFC2002500), the Jilin Province Development and Reform Commission, China (2019C053-1), the Education Department of Jilin Province, China (JJKH20200993K), and the Department of Science and Technology of Jilin Province, China (20200801002GH).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data underlying this article are available in the article.

Acknowledgments: The authors would like to thank all anonymous reviewers and editors for their helpful suggestions in the improvement of this paper.

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

References

1. Pabba, C.; Kumar, P. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Syst.* **2022**, *39*, e12839. [CrossRef]
2. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124. [CrossRef] [PubMed]
3. Zhan, C.; She, D.; Zhao, S.; Cheng, M.M.; Yang, J. Zero-shot emotion recognition via affective structural embedding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1151–1160.
4. Lei, J.; Liu, Z.; Zou, Z.; Li, T.; Juan, X.; Wang, S.; Yang, G.; Feng, Z. Mid-level Representation Enhancement and Graph Embedded Uncertainty Suppressing for Facial Expression Recognition. *arXiv* **2022**, arXiv:2207.13235.
5. Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; Yu, S.X. Large-scale long-tailed recognition in an open world. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2537–2546.
6. Cotter, S.F. Sparse representation for accurate classification of corrupted and occluded facial expressions. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 15–19 March 2010; pp. 838–841.
7. Kotsia, I.; Buciu, I.; Pitas, I. An analysis of facial expression recognition under partial facial image occlusion. *Image Vis. Comput.* **2008**, *26*, 1052–1067. [CrossRef]
8. Barros, P.; Sciutti, A. I Only Have Eyes for You: The Impact of Masks On Convolutional-Based Facial Expression Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1226–1231.
9. Bourel, F.; Chibelushi, C.C.; Low, A.A. Recognition of Facial Expressions in the Presence of Occlusion. In Proceedings of the BMVC, Manchester, UK, 10–13 September 2001; pp. 1–10.
10. Ly, S.T.; Do, N.T.; Lee, G.; Kim, S.H.; Yang, H.J. Multimodal 2D and 3D for In-The-Wild Facial Expression Recognition. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 2927–2934.
11. Li, K.; Zhao, Q. If-gan: Generative adversarial network for identity preserving facial image inpainting and frontalization. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 45–52.
12. Shome, D.; Kar, T. FedAffect: Few-shot federated learning for facial expression recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2021; pp. 4168–4175.
13. Cao, X.; Li, P. Dynamic facial expression recognition of sprinters based on multi-scale detail enhancement. *Int. J. Biom.* **2022**, *14*, 336–351.
14. Chen, Y.; Wang, J.; Chen, S.; Shi, Z.; Cai, J. Facial motion prior networks for facial expression recognition. In Proceedings of the 2019 IEEE Visual Communications and Image Processing (VCIP), Sydney, Australia, 1–4 December 2019; pp. 1–4.
15. She, J.; Hu, Y.; Shi, H.; Wang, J.; Shen, Q.; Mei, T. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6248–6257.
16. Chen, Y.; Joo, J. Understanding and mitigating annotation bias in facial expression recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 14980–14991.

17. Ahmady, M.; Mirkamali, S.S.; Pahlevanzadeh, B.; Pashaei, E.; Hosseinabadi, A.A.R.; Slowik, A. Facial expression recognition using fuzzified Pseudo Zernike Moments and structural features. *Fuzzy Sets Syst.* **2022**, *443*, 155–172. [CrossRef]
18. Chen, S.; Wang, J.; Chen, Y.; Shi, Z.; Geng, X.; Rui, Y. Label distribution learning on auxiliary label space graphs for facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13984–13993.
19. Gera, D.; Balasubramanian, S. Noisy Annotations Robust Consensual Collaborative Affect Expression Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3585–3592.
20. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
21. Wu, Y.; Liu, H.; Li, J.; Fu, Y. Deep face recognition with center invariant loss. In Proceedings of the on Thematic Workshops of ACM Multimedia 2017, Mountain View, CA, USA, 23–27 October 2017; pp. 408–414.
22. Wang, Y.X.; Ramanan, D.; Hebert, M. Learning to model the tail. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 7032–7042.
23. Yang, L.; Jiang, H.; Song, Q.; Guo, J. A Survey on Long-Tailed Visual Recognition. *Int. J. Comput. Vis.* **2022**, *130*, 1837–1872. [CrossRef]
24. Zhang, X.; Fang, Z.; Wen, Y.; Li, Z.; Qiao, Y. Range loss for deep face recognition with long-tailed training data. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5409–5418.
25. Mullick, S.S.; Datta, S.; Das, S. Generative adversarial minority oversampling. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1695–1704.
26. Lee, J.; Kim, S.; Kim, S.; Park, J.; Sohn, K. Context-aware emotion recognition networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10143–10152.
27. Park, S.; Hong, Y.; Heo, B.; Yun, S.; Choi, J.Y. The Majority Can Help The Minority: Context-rich Minority Oversampling for Long-tailed Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6887–6896.
28. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6023–6032.
29. Zhao, Z.; Liu, Q.; Zhou, F. Robust lightweight facial expression recognition network with label distribution training. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 3510–3519.
30. Antoniadis, P.; Pikoulis, I.; Filntisis, P.P.; Maragos, P. An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3645–3651.
31. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Trans. Image Process.* **2018**, *28*, 2439–2450. [CrossRef] [PubMed]
32. Xiong, W.; He, Y.; Zhang, Y.; Luo, W.; Ma, L.; Luo, J. Fine-grained image-to-image transformation towards visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5840–5849.
33. Kuo, C.M.; Lai, S.H.; Sarkis, M. A compact deep learning model for robust facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2121–2129.
34. Gecer, B.; Deng, J.; Zafeiriou, S. Ostec: One-shot texture completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7628–7638.
35. Bao, Z.; You, S.; Gu, L.; Yang, Z. Single-image facial expression recognition using deep 3d re-centralization. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
36. Zhou, H.; Liu, J.; Liu, Z.; Liu, Y.; Wang, X. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5911–5920.
37. Yang, H.; Zhu, K.; Huang, D.; Li, H.; Wang, Y.; Chen, L. Intensity enhancement via GAN for multimodal face expression recognition. *Neurocomputing* **2021**, *454*, 124–134. [CrossRef]
38. Bau, D.; Zhu, J.Y.; Wulff, J.; Peebles, W.; Strobelt, H.; Zhou, B.; Torralba, A. Seeing what a gan cannot generate. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4502–4511.
39. Barros, P.; Churamani, N.; Sciutti, A. The FaceChannel: A Light-weight Deep Neural Network for Facial Expression Recognition. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 652–656.
40. Wang, K.; Peng, X.; Yang, J.; Meng, D.; Qiao, Y. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4057–4069. [CrossRef] [PubMed]

41. Chu, P.; Bian, X.; Liu, S.; Ling, H. Feature space augmentation for long-tailed data. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 694–710.
42. Dong, Y.; Wang, X. A new over-sampling approach: Random-SMOTE for learning from imbalanced data sets. In Proceedings of the International Conference on Knowledge Science, Engineering and Management, Irvine, CA, USA, 12–14 December 2011; pp. 343–352.
43. Ando, S.; Huang, C.Y. Deep over-sampling framework for classifying imbalanced data. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Skopje, North Macedonia, 18–22 September 2017; pp. 770–785.
44. Hong, Y.; Han, S.; Choi, K.; Seo, S.; Kim, B.; Chang, B. Disentangling label distribution for long-tailed visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6626–6636.
45. Zhong, Y.; Deng, W.; Wang, M.; Hu, J.; Peng, J.; Tao, X.; Huang, Y. Unequal-training for deep face recognition with long-tailed noisy data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7812–7821.
46. Yin, X.; Yu, X.; Sohn, K.; Liu, X.; Chandraker, M. Feature transfer learning for face recognition with under-represented data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5704–5713.
47. Zhu, K.; Wang, Y.; Yang, H.; Huang, D.; Chen, L. Intensity enhancement via gan for multimodal facial expression recognition. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2020; pp. 1346–1350.
48. Gao, H.; An, S.; Li, J.; Liu, C. Deep balanced learning for long-tailed facial expressions recognition. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 11147–11153.
49. Zhang, H.; Su, W.; Yu, J.; Wang, Z. Weakly supervised local-global relation network for facial expression recognition. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021; pp. 1040–1046.
50. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
51. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
52. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
53. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
54. Li, S.; Deng, W.; Du, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2852–2861.
55. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [CrossRef]
56. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6897–6906.
57. Deng, J.; Guo, J.; Zhou, Y.; Yu, J.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-stage dense face localisation in the wild. *arXiv* **2019**, arXiv:1905.00641.
58. Xiong, X.; De la Torre, F. Supervised descent method and its applications to face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 532–539.
59. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
60. Zhang, Y.; Wang, C.; Deng, W. Relative Uncertainty Learning for Facial Expression Recognition. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17616–17627.
61. Zhang, Y.; Wang, C.; Ling, X.; Deng, W. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In Proceedings of the Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; pp. 418–434.
62. Farzaneh, A.H.; Qi, X. Discriminant distribution-agnostic loss for facial expression recognition in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 406–407.
63. Wen, Z.; Lin, W.; Wang, T.; Xu, G. Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv* **2021**, arXiv:2109.07270.

64. Antoniadis, P.; Filntisis, P.P.; Maragos, P. Exploiting Emotional Dependencies with Graph Convolutional Networks for Facial Expression Recognition. In Proceedings of the 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 15–18 December 2021; pp. 1–8. [CrossRef]
65. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 27–30 June 2016; pp. 770–778.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.