electronics

MDPI

Article

Facial Expression Recognition in the Wild for Low-Resolution Images Using Voting Residual Network

José L. Gómez-Sirvent [1,2], Francisco López de la Rosa [1,3], María T. López [1,2] and Antonio Fernández-Caballero [1,2,4,*]

[1] Instituto de Investigación en Informática de Albacete, Universidad de Castilla-La Mancha, 02071 Albacete, Spain; jose.gomez@uclm.es (J.L.G.-S.); francisco.lopezrosa@uclm.es (F.L.d.l.R.); maria.lbonal@uclm.es (M.T.L.)
[2] Departamento de Sistemas Informáticos, Universidad de Castilla-La Mancha, 02071 Albacete, Spain
[3] Departamento de Ingeniería Eléctrica, Electrónica, Automática y Comunicaciones, Universidad de Castilla-La Mancha, 02071 Albacete, Spain
[4] CIBERSAM-ISCIII (Biomedical Research Networking Center in Mental Health, Instituto de Salud Carlos III), 28016 Madrid, Spain
* Correspondence: antonio.fdez@uclm.es

Abstract: Facial expression recognition (FER) in the wild has attracted much attention in recent years due to its wide range of applications. Most current approaches use deep learning models trained on relatively large images, which significantly reduces their accuracy when they have to infer low-resolution images. In this paper, a residual voting network is proposed for the classification of low-resolution facial expression images. Specifically, the network consists of a modified ResNet-18, which divides each sample into multiple overlapping crops, makes a prediction of the class to which each of the crops belongs, and by soft-voting the predictions of all the crops, the network determines the class of the sample. A novel aspect of this work is that the image splitting is not performed before entering the network, but at an intermediate point in the network, which significantly reduces the resource consumption. The proposed approach was evaluated on two popular benchmark datasets (AffectNet and RAF-DB) by scaling the images to a network input size of 48 × 48. The proposed model reported an accuracy of 63.06% on AffectNet and 85.69% on RAF-DB with seven classes in both cases, which are values comparable to those provided by other current approaches using much larger images.

Keywords: facial expression recognition; emotions; low resolution; AffectNet; RAF-DB; voting

Citation: Gómez-Sirvent, J.L.; López de la Rosa, F.; López, M.T.; Fernández-Caballero, A. Facial Expression Recognition in the Wild for Low-Resolution Images Using Voting Residual Network. *Electronics* **2023**, *12*, 3837. https://doi.org/10.3390/electronics12183837

Academic Editor: Byung-Gyu Kim

Received: 25 July 2023
Revised: 6 September 2023
Accepted: 8 September 2023
Published: 11 September 2023

1. Introduction

In recent years, the need to recognize a person's emotions has increased, and there has been a growing interest in human emotion recognition across various fields, including brain–computer interfaces [1,2], assistance [3], medicine [4], psychology [5,6], and marketing [7]. Facial expressions are one of the primary nonverbal means of conveying emotion and play an important role in everyday human communication. According to a seminal paper [8], more than half of the messages related to feelings and attitudes are contained in facial expressions. Emotions are continuous in nature. However, it is common to measure them on a discrete scale. Ekman and Friesen [9] identified six universal emotions based on a study of people from different cultures. This study showed that people, regardless of their culture, perceive some basic emotions in the same way. These basic emotions are happiness, anger, disgust, sadness, surprise, and fear. Over time, critics of this model have emerged [10], arguing that emotions are not universal and have a high cultural component; nevertheless, the model of the six basic emotions continues to be widely used in emotion recognition [11].

In the last few decades, facial expression recognition (FER) has come a long way thanks to advances in computer vision and machine learning [12]. Traditionally, feature-extraction algorithms such as scale-invariant feature transform or local binary patterns and classification algorithms such as support vector machines or artificial neural networks have been used for this task [13–15]. However, the current trend is to use convolutional neural networks (CNNs), which perform feature extraction and classification at the same time [16]. FER datasets can be divided into two main categories, depending on how the samples were obtained: laboratory-controlled or wild. In lab-controlled datasets, all images are taken under the same conditions. Therefore, there are no variations in illumination, occlusion, or pose. Under these conditions, it is relatively easy to achieve high classification accuracy without resorting to complex models; in fact, in some datasets, such as CK+ or JAFFE, 100% of the images can be correctly classified [17,18]. On the other hand, in-the-wild datasets contain images taken under uncontrolled conditions, such as those found in the real world. In this scenario, the classification accuracy is significantly lower than that obtained in laboratory-controlled datasets [19,20].

FER in-the-wild datasets typically contain images of many different sizes [21], which are typically scaled to 224×224 to feed a neural network. The main drawback of training a CNN on images of this size is that, as the network infers lower-resolution images, classification accuracy drops significantly [20]. There are many applications where obtaining high-resolution images of human faces is not feasible, for example when trying to determine the emotional state of multiple people simultaneously in large spaces such as shopping malls, parks, or airports. In these situations, each person is at a different distance from the camera, resulting in images of different resolutions. As highlighted in a survey [22], these circumstances present a variety of challenges, including occlusion, pose, low resolution, scale variations, and variations in illumination levels. In addition, they underscore the importance of the efficiency of FER models when processing images of multiple people in real-time.

In these scenarios, a network trained on low-resolution images can be more robust because the network is less dependent on fine details that are not present in low-resolution images, thus increasing its ability to generalize. In addition, working with smaller images reduces the computational cost and bandwidth required to transfer the images from the cameras to the computer where they are processed. CNNs that work with low-resolution images are lighter because the features occupy less memory, making it possible to use methods such as ensemble learning, which is not widely used in deep learning due to its high computational complexity [23]. Ensemble learning methods combine the results of multiple machine learning estimators with the goal of obtaining a model that generalizes better than the estimators of which it is composed [24]. Assembling n CNNs means multiplying by n the number of trainable parameters of the model and the size of the features of each image in the network.

To illustrate the problem more clearly, let us consider a possible application scenario, such as a smart campus proposal, where FER is used to quantify the degree of user comfort [25]. This environment uses a network of wireless cameras that send video to servers that recognize people's faces and predict their emotions in real-time. In this environment, it is not possible to control the distance of building occupants from the camera, their position, or possible occlusions. In addition, there can be hundreds or even thousands of people in this type of environment, which means managing a huge flow of data. In this context, the resolution of the images is a fundamental variable: the lower the resolution of the images, the lower the workload of the system is. For this reason, the use of lightweight models specifically designed for processing low-resolution images in real-world conditions is of interest.

This paper proposes a model for facial expression recognition from low-resolution images based on a residual network and a voting strategy. This model tries to improve the generalization capability of the neural network by combining different classification results, as is done in ensemble methods, but without increasing the complexity of the model

excessively. For this purpose, instead of using n networks to obtain n votes (predictions), n patches of each image are taken and fed into a single network, which provides the n votes based on which the class of the image is determined. Thus, the number of trainable parameters of the model remains constant regardless of the number of votes. If the image is partitioned into multiple patches before entering the network, the size of the image features is multiplied by n. To address this problem, our work explored different intermediate points in the network where partitioning can be performed. The closer the splitting point is to the network exit, the smaller the dimensionality of the image features and, consequently, the smaller the space required to store them. In this paper, the proposed approach was implemented on a residual network, although theoretically, this approach could be used with virtually any CNN.

The main contributions of this paper can be summarized as follows:

- Integration of a voting mechanism: A voting mechanism is integrated into a residual network architecture to improve the accuracy of facial expression recognition in low-resolution images.
- Retention of model size: The proposed voting approach does not increase the number of trainable parameters regardless of where the image is split. This differs from conventional ensemble model implementations, which tend to increase the number of trainable parameters.
- Examination of split points: The method is evaluated by performing image splitting at different points in the network to determine the impact on network accuracy and forward pass size after applying the model. The closer the split point is to the network output, the lower the dimensionality of the image features and, thus, the smaller the space required to store them.
- Experimental evaluation: Extensive experiments were performed on two public benchmark datasets (AffectNet and RAF-DB) to validate the effectiveness of the proposed method. The images of these datasets were resized to 48 $\times$ 48 px, which makes the proposed method very robust to variations in image resolution. The results showed that it is possible to reduce the resolution of the images used to train FER models without significant loss of accuracy.

The rest of this paper is organized as follows. In Section 2, several existing algorithms for FER in the wild are discussed. The materials and methods used in this work are presented in Section 3. Section 4 contains the experimental results. Finally, Section 5 draws the conclusions of the study.

2. Related Work

2.1. Facial Expression Recognition in the Wild

Many of today's FER methods use a conventional CNN as the backbone and add attention modules to it to improve classification accuracy [26–28]. These modules are used to force the CNN to learn and focus more on important information instead of learning useless background information. For example, an occlusion-aware FER system using a CNN with an attention mechanism has been developed [19]. The authors divided the resulting feature vector of the last convolutional layer of a CNN into 24 regions of interest using a region decomposition scheme and trained an attention mechanism module capable of learning a low weight for a blocked region and a high weight for an unblocked and informative region with them. In a recent paper, a multi-headed cross-attention network was proposed that achieved state-of-the-art performance on three public FER in-the-wild datasets [29]. In this network, the attention module implements spatial and channel attention, which allows capturing higher-order interactions between local features. A CNN was also trained with multiple patches of the same image, and an attention module was added to the network output [30]. This model achieved state-of-the-art results on four public FER in-the-wild datasets. Our method has in common that each sample is split into multiple patches in both methods. However, in our approach, the splitting is

performed within the CNN, which significantly reduces the size of the image features within the network.

Nevertheless, there are other approaches that achieve high performance in FER in the wild without relying on attentional mechanisms. For example, a novel multitask learning framework that exploits the dependencies between these two models using a graph convolutional network has recently been proposed [31]. The results of their experiments showed that their method can improve the performance over different datasets and backbone architectures. In addition, a paper proposed three novel CNN models with different architectures and performed extensive evaluations on three popular datasets, demonstrating that these models are competitive and representative in the field of FER in the wild research [32]. A very recent paper introduced a few-shot learning model called the convolutional relation network for FER in the wild, which was trained by exploiting a feature similarity comparison among the sufficient samples of the emotion categories to identify new classes with few samples [33].

Another approach that has proven to be state-of-the-art in FER is transformer-based methods. Inspired by transformers used in natural language processing, vision transformers (ViTs) have been proposed as an alternative to CNNs for various computer vision problems such as image generation [34] or classification [35]. In this approach, images are divided into multiple samples, and this sequence of images is used as the input to the model. Compared to CNNs, ViTs are more robust for classifying images that have noise or are magnified, but these models are generally more computationally expensive [36].

However, the pure structure of ViTs, which does not reflect local features, is not suitable for detecting subtle changes between different facial expressions, and therefore, the performance of these models for FER may be inferior to those based on CNNs. In order to exploit the advantages and minimize the limitations of both approaches, hybrid models combining ViTs and CNNs for FER have been developed in recent years. Huang et al. [37] proposed a novel framework with two attention mechanisms for CNN-based models and a token-based visual transformer technique for image classification of facial expressions in the wild. With this model, they achieved state-of-the-art performance on different datasets without the need for additional training data. With the same goal, Kim et al. [38] proposed a hybrid approach with a ViT as the backbone. By introducing a squeeze module, they were able to reduce the computational complexity by reducing the number of feature dimensions, while increasing the FER performance by simultaneously combining global and local features. Similarly, Liu et al. [39] used a CNN to extract local image features and fed a ViT with a positional embedding generator to model correlations between different visual features from a global perspective. With this hybrid model of only 28.4 million parameters, they surpassed the state-of-the-art in FER with occlusion and outperformed other models with hundreds of millions of parameters.

In addition, Ma et al. [40] proposed a model using two ResNet-18 for parallel feature extraction from the original image and an image obtained by local binary patterns. To model the relationships between features, they used a visual transformer, where features from both networks are merged. However, the improvement in accuracy provided by this model implies a significant increase in computational load compared to other approaches, since this model uses between 51.8 and 108.5 million parameters in the different implementations described in this work. What our proposal has in common with transformer-based methods is that both approaches divide the images into several patches in order to improve the accuracy of emotion recognition. However, in transformer-based models, each patch contains only a small part of the face, which forces working with a large number of samples and, consequently, increases the computational complexity. In contrast, in our approach, almost the entire face is contained in each sample, which reduces the number of samples required and makes the model robust to image translations.

2.2. FER in the Wild from Low-Resolution Images

In real-world applications, some or all of the images may be low-resolution. Under these conditions, the accuracy of models trained on high-resolution images is significantly reduced. FER in the wild from low-resolution or variable-resolution images is a less-explored area. However, in recent years, several approaches have been proposed to address this problem. Yan et al. [41] proposed a filter-based subspace learning method that outperformed the state-of-the-art on posed facial expression datasets, but the results were significantly worse on in-the-wild image datasets. The authors argued that this method has a learning capacity superior to some CNN-based methods. However, it requires the image to be converted to gray scale, which can result in the loss of valuable information. Another approach that has proven effective is the use of denoising techniques on low-resolution images to increase classification accuracy [42].

On the other hand, super-resolution-based methods have shown promising results in this field [43,44]. Super-resolution algorithms are generative models used to obtain high-resolution images from small images. The output of these models can be used as the input to any CNN that works with high-resolution images, such as those described in the previous subsection. The main drawback of these methods is the computational cost, since adding a generative model before the classifier means increasing the memory needed to process the image and the number of floating-point operations. While conventional CNNs typically require less than 10 giga floating-point operations per second (GFLOPs) to process an image, super-resolution-based models can require thousands of GFLOPs [43].

Another approach that has shown promising results in this area is knowledge distillation [45], which basically consists of transferring knowledge from heavy models trained on high-resolution images to lighter models operating on low-resolution images. For example, Ma et al. [46] obtained high accuracy rates in FER of resolution-degraded images with a model in which they used the knowledge of different levels of features of a teacher network to transfer it to a lighter student network. O. Huang et al. [47] proposed a feature-map-distillation (FMD) framework in which the size of the feature map of the teacher and learner networks was different. With this approach, they achieved better results on several recognition tasks than with twenty state-of-the-art knowledge-distillation methods.

3. Materials and Methods

The notations and symbols used in this article are shown in Table 1.

Table 1. Notations and symbols used in this article.

Notation	Description
conv	2D convolutional layer; applies sliding convolutional filters to 2D input.
adaptive avg pool	2D adaptive average pooling layer; computes the kernel size required to generate a specified output dimensionality from the given input.
b	Batch size; the number of input samples processed simultaneously during training or inference.
fc	Fully connected layer; applies a linear transformation to the input vector through a weights matrix.
num_channels	Number of output channels; representing the depth of feature maps produced by a layer.
x	Batch of input images; a multi-dimensional array containing the input image data.
x_{patches}	List that stores patches extracted from the images; used for further processing.
img	Individual image in the batch of images.
$h_{\text{old}}, w_{\text{old}}$	Original height and width of *img*.
$h_{\text{new}}, w_{\text{new}}$	Height and width of the image after calculating new dimensions based on a specific ratio.
$x_{\text{start}}, x_{\text{end}}, y_{\text{start}}, y_{\text{end}}$	Coordinates representing the position of a patch in the image.
patch	Extracted image patch with dimensions h_{new} by w_{new}, representing a smaller region of the input image.
mirrored_img	Image horizontally mirrored.
x_{final}	The final result contains all patches of the original and mirrored images, concatenated along the batch size dimension.

3.1. Network Architecture

Our work used the ResNet-18 architecture for ImageNet [48] as a baseline, with slight modifications to adapt it to smaller images (see Figure 1). ResNet-18 is a lightweight and robust architecture that has demonstrated its superiority over other heavier architectures in computer vision tasks [48]. This model has been widely used as the backbone for various models that have reached the state-of-the-art in FER in the wild [49–53]. The skip connections in ResNet-18 allow faster convergence during training by providing shortcuts for gradient flow. This leads to faster training and often better generalization, especially when dealing with complex datasets. In addition, this model has a relatively simple and easily manipulated structure that allows the splitting and voting strategy presented in this paper to be easily incorporated at various points in the architecture.

In the original architecture, designed for images of 224 × 224 px, the first convolutional layer used a kernel of size 7 × 7 and a stride of two, followed by a max-pooling layer to reduce the dimensionality quickly. In our network, the input images had a size of 48 × 48 px, so it was not necessary to reduce the dimensionality so quickly. Therefore, we chose a 3 × 3 kernel in the first layer with a stride of one and removed the max-pooling layer. Otherwise, the baseline architecture was identical to the original.

We believe that the 7 × 7 kernel size in the original architecture may be excessive for 48 × 48 px images and may result in a loss of fine-grained features. The use of a 3 × 3 kernel is congruent with the ResNet approach for CIFAR, which is designed to work with 32 × 32 px images. On the other hand, the decision to eliminate the max-pooling layer was due to the fact that, for images with an input size of 48 × 48, the dimensionality at this point of the network is already low without the need to perform any pooling operation. In the original architecture, this layer allows reducing the dimensionality from 112 × 112 to 56 × 56. In our case, however, the dimensionality is 48 × 48.

From the baseline architecture, three variants were developed by introducing a voting strategy. For this purpose, each image entering the network was cut into five overlapping square patches; the side dimension of each patch was 5/6 of the image side dimension, and one patch was taken at the center of the image and one at each corner of the image (see Figure 2). Once the patches were obtained, a copy of the patches was made and mirrored horizontally. In this way, 10 sub-images were obtained from each image. The pseudocode of the image division process is shown in Algorithm 1.

Accurately recognizing facial expressions in real-world situations is challenging due to the variety of lighting conditions, poses, backgrounds, and occlusions. By dividing the image into multiple overlapping slices, we sought to increase the robustness and stability of the model by forcing it to work with incomplete and off-center images. This approach yields different predictions from the same image, as if it were an assembler model, but without the need for ten different neural networks. The choice of a 5/6 ratio for image cropping is based on the observation that this ratio preserves most of the facial area relevant to facial expressions in most images, while eliminating less-informative regions. In addition, the dimensions of the images produced by each module within the baseline architecture are all divisible by six, which simplifies the cropping procedure.

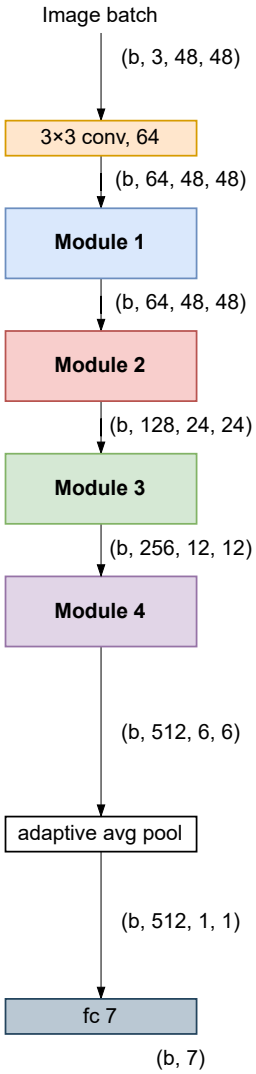

Figure 1. Modified ResNet-18. Note: "b" denotes batch size.

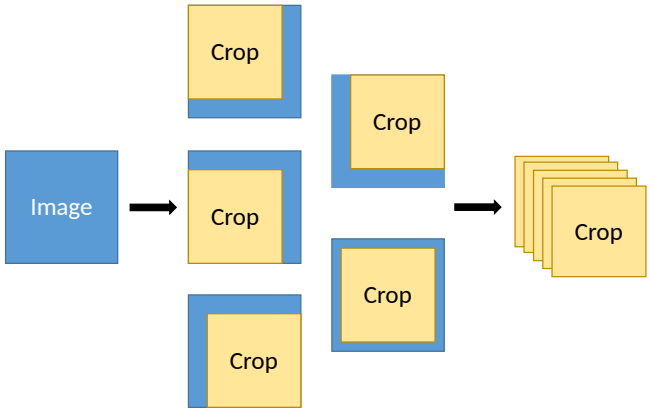

Figure 2. Image cropping.

Algorithm 1 Image patch generation

Input: Batch of images: x of shape $(\text{batch\_size}, \text{num\_channels}, \text{height}, \text{width})$
Output: Patches: x_{patches} of shape (batch_size $\times$ 10, num_channels, new_height, new_width)
Initialize an empty list x_{patches}
for each image img in batch x **do**
    Get dimensions of img: $h_{\text{old}} = \text{height}(img)$, $w_{\text{old}} = \text{width}(img)$
    Calculate new dimensions: $h_{\text{new}} = \frac{5}{6} \times h_{\text{old}}$, $w_{\text{new}} = \frac{5}{6} \times w_{\text{old}}$
    for each patch position **do**
        Calculate patch coordinates: x_{start}, x_{end}, y_{start}, y_{end} for the current position
        Extract patch: $patch = img[:, :, y_{\text{start}} : y_{\text{end}}, x_{\text{start}} : x_{\text{end}}]$
        Add $patch$ to x_{patches}
    end for
    Mirror img horizontally: $mirrored\_img = \text{horizontal\_flip}(img)$
    for each patch position **do**
        Calculate patch coordinates for mirrored image
        Extract patch from mirrored image
        Add mirrored $patch$ to x_{patches}
    end for
end for
Concatenate patches along the batch dimension: $x_{\text{final}} = \text{concatenate}(x_{\text{patches}}, \text{dim} = 0)$

Cropping before entering the network would significantly increase the size of the image features within the network. For this reason, in this work, cropping was performed at an intermediate point in the network where the dimensionality is lower. Specifically, three points were examined (see Figure 3), namely the output of Modules 1, 2, and 3. The output of the fully connected (FC) layer of the network provided the predictions of each of the 10 crops into which the image had been divided. The final prediction of the class to which the image belonged was obtained by calculating the average of these 10 predictions (soft voting). The closer the cropping point was to the input of the network, the larger the size of the images within the network (forward pass) and, consequently, the smaller the number of images that could be processed per batch in both training and inference. Table 2 shows the estimated forward pass sizes per image of the proposed architecture, depending on where in the network the image was divided into 10 crops. Included also in the first row is the forward pass size that the baseline architecture would have if the image were split before entering the network. In this case, the effective forward pass size would be 70 MB because it would have to be multiplied by 10 since 10 sub-images would be used to predict each sample.

The forward pass size was measured by calculating the space occupied by the tensors on the graphics card before and after inserting a batch of images and dividing this value by the number of images in each batch. The occupied space was measured using the *torch.cuda.memory_allocated*() function of PyTorch [54].

Table 2. Estimated forward pass size and giga floating-point operations per second (GFLOPs) of different ResNet-18 architectures.

Architecture	Input Size	Forward Pass Size (MB)	GFLOPs
Baseline	40 × 40	7.00	0.87
Baseline	48 × 48	10.08	1.26
Cropping after Module 1	48 × 48	40.34	6.66
Cropping after Module 2	48 × 48	23.34	4.85
Cropping after Module 3	48 × 48	14.84	3.05

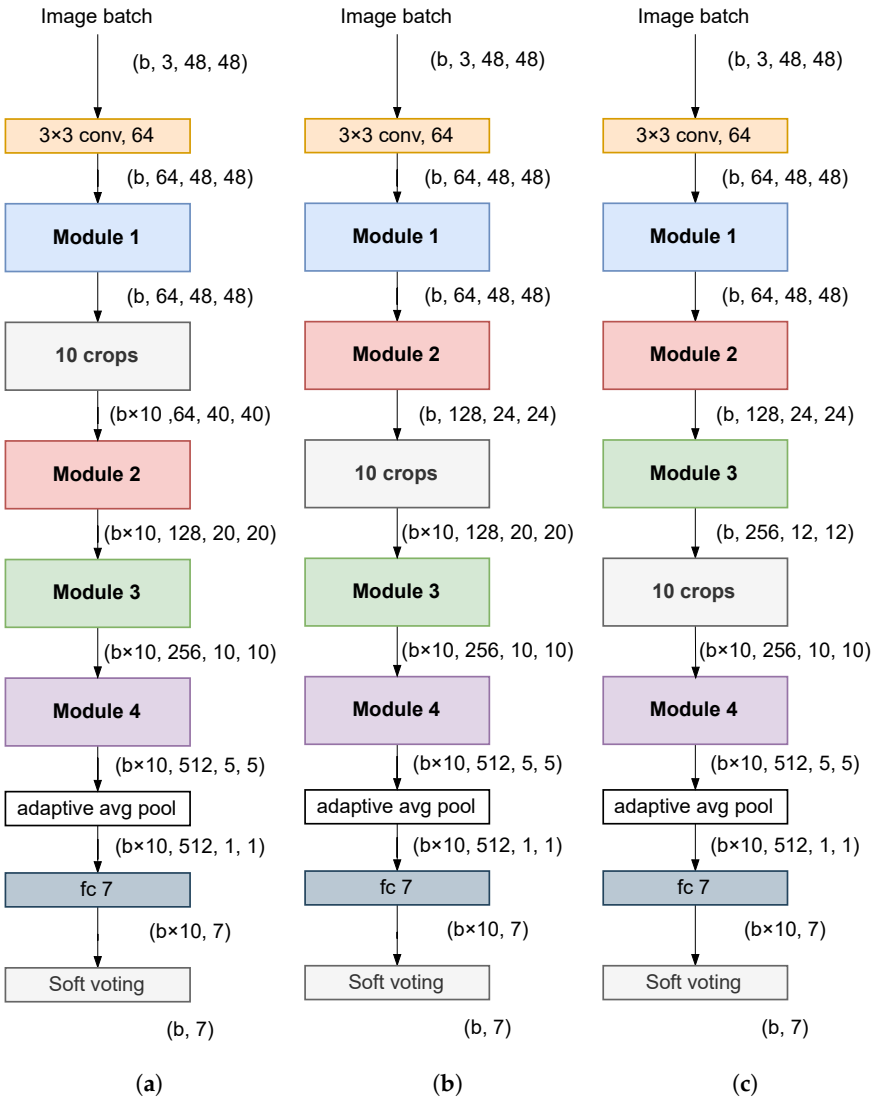

Figure 3. Variants of the baseline ResNet-18 with the addition of the voting strategy. (**a**) Cropping after Module 1. (**b**) Cropping after Module 2. (**c**) Cropping after Module 3. Note: "b" denotes batch size.

3.2. Datasets

To evaluate the proposed method, two popular FER datasets in the wild were used, namely AffectNet [55] and RAF-DB [56]. These datasets contain images of different sizes taken under challenging conditions with occlusions and variations in illumination and pose. Both datasets are highly imbalanced. Figure 4 shows the distribution of images in the original training set of the two datasets above.

3.2.1. AffectNet

AffectNet contains more than 440,000 facial images collected from the Internet by querying various search engines with 1250 emotion-related keywords in six different languages. Approximately 291,651 of these images were manually annotated for the presence of eight facial expressions (fear, happiness, sadness, neutral, anger, surprise, disgust, and contempt). In this work, the contempt expression was not used, so there was a total of 287,401 images, divided into two subsets, and 500 images of each class belonged to the validation set and the rest to the training set. Since the test set was not published by the authors, the validation set was used as the test set, and we created a new validation set by randomly taking 500 images from each class of the training set.

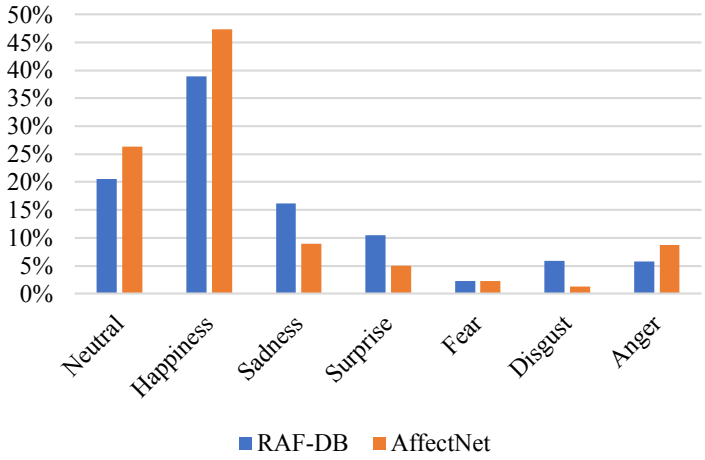

Figure 4. Class distribution of the training sets of the RAF-DB and AffectNet datasets.

3.2.2. RAF-DB

The RAF-DB dataset contains 29,672 manually annotated images collected from the Internet, which are divided into two subsets according to the type of annotation: (i) basic expressions and (ii) compound expressions. Here, only the subset of basic expressions was used, which includes six basic emotions (fear, happiness, sadness, anger, surprise, and disgust) and the neutral emotion. The subset of basic expressions was further divided into two subsets: training and testing, with 12,271 and 3068 images, respectively. Since we did not have a validation set, we created one by taking 10% of the images from the training set.

3.3. Implementation Details

The experiment was conducted on a workstation with the following hardware specifications: Intel(R) Core(TM) i7-10700KF processor, 32 GB DDR4 3600 MHz RAM, NVIDIA GeForce RXT 2070 Super graphics card. All networks were implemented using the PyTorch library [54]. All images were resized to 48 × 48 before entering the network. During training, online data augmentation was performed to increase the generalization capacity of the network and to avoid overfitting. The following transformations were randomly applied to the images in the training phase:

- *Resized crop:* A part of the image was cropped and resized to its original size.
- *Color jitter*: The brightness, contrast, and/or saturation of the image were randomly altered.
- *Translation*: The image was scrolled horizontally and/or vertically.
- *Rotation*: The image was randomly rotated with respect to its center.
- *Erase*: A random rectangular area of the image was erased.

The AffectNet dataset is highly imbalanced. However, the validation set of this dataset (which we used as the test set) is balanced. Therefore, in training, we downsampled the majority classes and oversampled the minority classes to improve the average accuracy of the model. In the case of RAF-DB, both the training and test sets are unbalanced, although, for a real application, it would be more useful to train the model by balancing the dataset, as was performed in AffectNet. It was decided not to do this because most of the work we compared our model to used overall accuracy rather than average accuracy as a metric to evaluate their models. Balancing the dataset for training can reduce the overall accuracy of the classifier, so we feel that comparing it to other works that did not balance it would not be fair.

Training the networks with AffectNet was performed from scratch, while training with RAF-DB, which contains approximately 18-times fewer images, was performed by initializing the networks with the AffectNet weights. The Adam [57] optimizer was used to train all networks in both datasets. The initial learning rate was 0.0002 and was reduced after each iteration by multiplying it by 0.65 for AffectNet and by 0.95 for RAF-DB. The

cross-entropy between the FC layer output and the image labels was used as the loss function in the model optimization. The FC layer output was used instead of the soft voting result to increase the robustness of the model, as it was forced to learn to correctly classify the different crops into which the image had been divided, rather than just the whole image. All models were trained with a batch size of 135. For the AffectNet dataset, the model was trained on twenty epochs and, for the RAF-DB dataset, on forty epochs.

For both datasets, the best model was selected based on the classification accuracy on the validation set (created by taking images from the original training set). Once the best model was determined, it was re-trained using the images from the training and validation sets and then evaluated using the images from the test set to obtain the final classification results.

3.4. Evaluation of Results

Accuracy is defined as the ratio of the number of samples correctly classified to the total number of samples (see Equation (1)).

$$\text{accuracy} = \frac{\text{number of correct predictions}}{\text{total number of predictions}} \quad (1)$$

Overall accuracy is not the most-appropriate metric to evaluate a classifier on an imbalanced dataset, since its value depends mainly on the results obtained on the majority classes. However, it is the most-commonly used metric to evaluate the performance of a classifier in the field of FER. In this work, it was decided to use this metric to evaluate the model, since it allowed us to compare the results with almost all the works that used the same datasets as here. In addition to the overall accuracy, confusion matrices were used to evaluate the performance of the classifiers. Confusion matrices allowed us to visualize when one class was confused with another, allowing us to work with different types of errors separately. Specifically, normalized confusion matrices were used, i.e., matrices in which the columns of the actual classes were divided by the total number of images contained in each class. In this way, the classification accuracy of each class was obtained on the diagonal of the matrix.

4. Results and Discussion

Table 3 shows the classification results on the validation set. The introduction of the voting strategy improved the classification results with respect to the baseline architecture in both datasets. The best results were obtained by cropping after the first module of the network. However, similar results were obtained by cropping after the second module, and in this case, the size of the forward pass was much smaller. The results suggest that the closer the image splitting was performed to the network input, the better the classification accuracy. Furthermore, it was observed that the soft voting strategy gave better results than the hard voting strategy in all cases.

Table 3. Classification accuracy of the validation set of the different architectures studied (highest accuracy is in bold).

Architecture	AffectNet		RAF-DB	
	Soft Voting	**Hard Voting**	**Soft Voting**	**Hard Voting**
Cropping after Module 1	**63.69%**	63.14%	**85.88%**	85.06%
Cropping after Module 2	63.51%	62.34%	85.31%	84.90%
Cropping after Module 3	62.80%	62.34%	85.31%	84.73%
Baseline	62.29%		83.76%	

Once it was determined that cropping after Module 1 produced the best classification results, the final model was retrained using the images from the training and validation sets. The accuracy was 63.06% for AffectNet and 85.69% for RAF-DB. Figures 5 and 6 show the confusion matrices of our model evaluated on the test set of the above datasets.

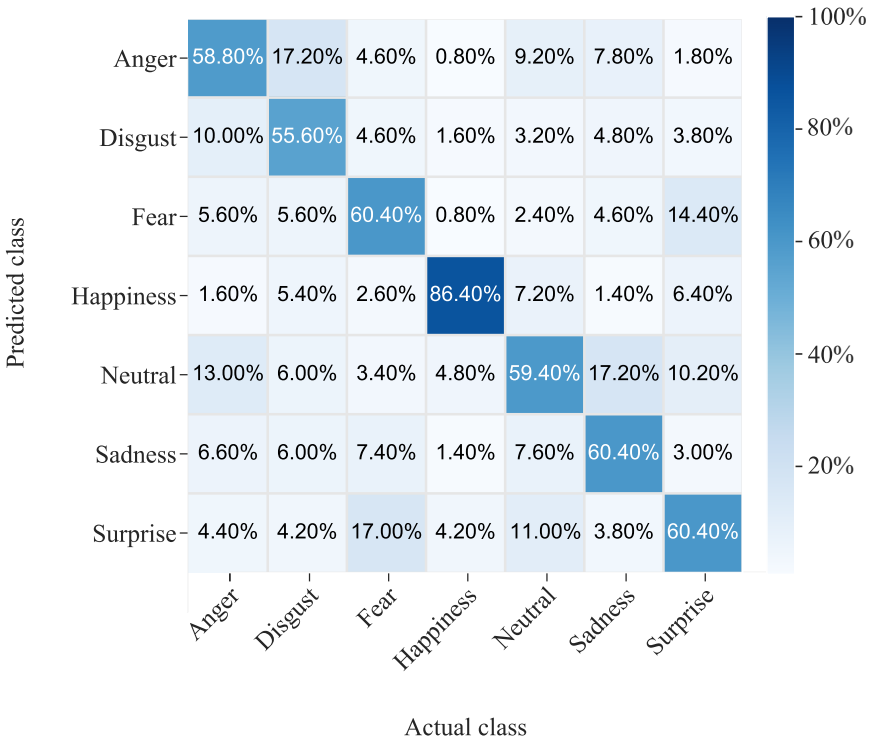

Figure 5. Confusion matrix of the original validation set of the AffectNet database.

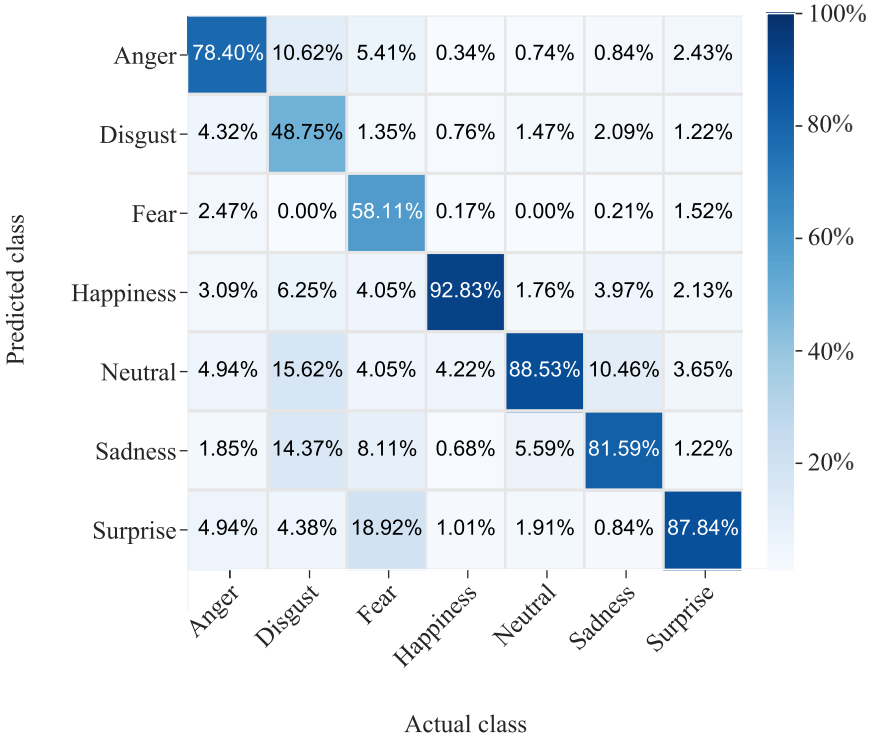

Figure 6. Confusion matrix of the test set of RAF-DB.

In the AffectNet dataset, the best-classified emotion was happiness with an accuracy of 86.40%. The accuracy of the other classes ranged from 55.60% to 60.40%. As in the previous dataset, the class with the highest accuracy in RAF-DB was happiness, in this case with 92.93%. In this dataset, the three classes with the worst accuracy coincided with the three classes with the lowest number of images (see Figure 4). This is probably because the RAF-DB dataset was not balanced during the training of the network, as was performed with AffectNet.

In both datasets, the worst-classified class was disgust, with 17.2% of the AffectNet images with this label being mistaken for anger and 10% of those with anger being assigned to disgust. The difficulty in recognizing disgust may be due to the fact that this facial emotion often shares visual features with other expressions, such as anger. In RAF-DB, on the other hand, there are not many false positives for disgust, but there are many false negatives; about 30% of the disgust images were assigned to neutral or sad. This pattern was also observed for fear, where the false positives did not exceed 5%, while the false negatives were more than 40%.

These results are consistent with those obtained in most of the works listed in Tables 4 and 5, where happiness was always the best-classified emotion and disgust was usually the worst-classified. Considering that happiness is the easiest facial expression for humans to recognize, it is expected that images with this facial expression will be better labeled and, therefore, easier for the machine to recognize. When interpreting the data, it is important to note that the images in these datasets were obtained from Internet search engines, and it is not possible to know for certain what emotions the people in these images were experiencing. These datasets were labeled by human annotators, and there were discrepancies between the annotators' responses for quite a few images, especially in the case of AffectNet, where the degree of inter-annotator agreement for eight facial expressions was 60.7%.

Next, we compared the best results obtained with several state-of-the-art methods on AffectNet and RAF-DB. For AffectNet with seven classes, we did not find any work using an image size like ours, so the comparison with other works is not entirely fair. As shown in Table 4, a classification accuracy of 63.06% was obtained, which is 3.4% lower than the state-of-the-art, but higher than the accuracy obtained by some recent models using much larger images than ours.

Table 4. Comparison with the state-of-the-art methods on AffectNet dataset (seven classes) in terms of accuracy (highest accuracy is in bold).

Method	Years	Image Size	Accuracy
CNNs and BoVW [58]	2019	224 × 224	63.31%
gACNN [19]	2019	224 × 224	58.78%
HERO [49]	2019	224 × 224	62.11%
SNA-DFER [59]	2020	112 × 112	62.70%
ResNet-50 [60]	2020	100 × 100	61.57%
SAANet [61]	2020	224 × 224	63.71%
GCN [31]	2021	227 × 227	**66.46%**
ACSI-Net [50]	2022	256 × 256	65.83%
MAFT [51]	2022	224 × 224	65.17%
FG-AGR [62]	2023	224 × 224	64.91%
Baseline (ours)	2023	48 × 48	61.97%
Voting (ours)	2023	48 × 48	63.06%

On the other hand, for RAF-DB, the E-FCNN [44] and RCAN [43] methods were evaluated on images of a similar size to ours and reported similar results to those obtained with the proposed approach (see Table 5). However, these two methods are based on super-resolution, which means that the neural networks were trained on images larger than those used for inference. In contrast, in the proposed method, the training images are the same size as the test images.

Table 5. Comparison with the state-of-the-art methods on the RAF-DB dataset in terms of accuracy (highest accuracy is in bold).

Method	Year	Image Size	Accuracy
gACNN [19]	2019	224 × 224	85.07%
E-FCNN [44]	2020	50 × 50	84.62%
IFSL (SVM) [41]	2020	32 × 32	76.90%
ResNet-50 [60]	2020	100 × 100	87.00%
SCAN and CCI [63]	2021	224 × 224	89.02%
ACSI-Net [50]	2022	256 × 256	86.86%
MAFT [51]	2022	224 × 224	88.75%
RCAN [43]	2022	50 × 50	85.76%
MATF [52]	2022	100 × 100	88.52%
EAC [53]	2022	224 × 224	89.99%
FG-AGR [62]	2023	224 × 224	**90.81%**
Baseline (ours)	2023	48 × 48	84.32%
Voting (ours)	2023	48 × 48	85.69%

Regarding the number of trainable parameters, our model is one of the lightest with 11.17 million parameters, both for the baseline architecture and for the architectures where the splitting and voting strategy was implemented. Most of the papers listed in Tables 4 and 5 did not report the number of trainable parameters or the GFLOPs of the model. Only a recent paper reported 35.74 million parameters and 3686 GFLOPs [43]. However, considering only the network used as the backbone in each paper, it can be observed that most of the models used more trainable parameters than those used in our approach.

For example, some approaches [19,49,58,61] used VGG architectures for feature extraction with significantly more trainable parameters than the modified ResNet-18 we used. On the other hand, other works [60,63] used a ResNet-50 as their backbone, an architecture that has about 25 million trainable parameters in its usual implementation. The ResNet-18 architecture is one of the most-widely used [49–53]. These approaches use this architecture as a feature extractor. ResNet-18 uses 11.69 million parameters in its usual implementation, but these papers used this architecture as a feature extractor and added other elements at the end of the network, so the total number of parameters of the complete model can be significantly higher.

5. Conclusions

In this paper, a residual voting network was proposed for the classification of low-resolution facial expression images. The introduction of the voting strategy into the network improved the accuracy of the baseline model without significantly increasing the number of trainable network parameters. Different intermediate points in the network at which images could be cropped were examined, and it was observed that the closer the point was to the input of the network, the greater the improvement in accuracy compared to the baseline architecture. However, the size of the forward pass also increased. Based on the above, the best point for cropping will depend on the application. It will be a matter of selecting the closest possible point to the network input, depending on the available computational resources.

In addition, the proposed method was compared with some of the more-recent approaches. The experimental results obtained showed that our method was able to achieve classification accuracies similar to those reported by other methods using images larger than ours. Therefore, we concluded that it is possible to use low-resolution images for training FER models without a significant reduction in classification accuracy.

The method presented in this paper, with slight modifications to adapt it to the output dimensions of the different layers of the network, can be applied to almost any CNN. Therefore, it will be necessary in the future to study the usefulness of this method in other network architectures and in other classification or even regression tasks.

A possible future research direction is to explore the feasibility of applying this method to networks operating at the standard 224 × 224 image size. Although the proposed method

does not increase the number of trainable model parameters, it does increase the memory requirements. Investigating its performance on larger images could provide valuable information about its versatility and feasibility for other applications. Another promising research direction is the integration of this model into a ViT architecture. Both approaches share the partitioning of the image into multiple patches, suggesting that the creation of a hybrid model may be feasible.

Although the proposed model is intended for FER, it may also be interesting to explore whether it can improve the performance of the base model in other image-processing tasks. This research could help determine the true potential and usefulness of the model. In conclusion, the adaptability and performance of the proposed method gives rise to several avenues for future research.

Author Contributions: Conceptualization, J.L.G.-S. and F.L.d.l.R.; methodology, J.L.G.-S.; software, J.L.G.-S.; validation, F.L.d.l.R., M.T.L. and A.F.-C.; writing—original draft preparation, J.L.G.-S.; writing—review and editing, A.F.-C.; funding acquisition, A.F.-C. All authors have read and agreed to the published version of the manuscript.

Funding: Grants PID2020-115220RB-C21 and EQC2019-006063-P funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way to make Europe". Grant BES-2021-097834 funded by MCIN/AEI/10.13039/501100011033 and by "ESF Investing in your future". Grant 2023-PRED-21291 funded by Universidad de Castilla-La Mancha and by "ESF Investing in your future". Grant 2022-GRIN-34436 funded by Universidad de Castilla-La Mancha and by "ERDF A way of making Europe". This research was also partially funded by CIBERSAM, Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación, and co-funded by "ERDF A way to make Europe".

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional neural network
FC	Fully connected
FER	Facial expression recognition
FMD	Feature map distillation
GFLOPs	Giga floating-point operations per second
ViT	Vision Transformer

References

1. García-Martínez, B.; Fernández-Caballero, A.; Martínez-Rodrigo, A.; Novais, P. Analysis of Electroencephalographic Signals from a Brain-Computer Interface for Emotions Detection. In Proceedings of the Advances in Computational Intelligence, Berlin, Germany, 16–18 December 2021; pp. 219—229 [CrossRef]
2. Sánchez-Reolid, R.; García, A.S.; Vicente-Querol, M.A.; Fernández-Aguilar, L.; López, M.T.; Fernández-Caballero, A.; González, P. Artificial Neural Networks to Assess Emotional States from Brain-Computer Interface. *Electronics* **2018**, *7*, 384. [CrossRef]
3. Martínez, A.; Belmonte, L.M.; García, A.S.; Fernández-Caballero, A.; Morales, R. Facial Emotion Recognition from an Unmanned Flying Social Robot for Home Care of Dependent People. *Electronics* **2021**, *10*, 868. [CrossRef]
4. Kumfor, F.; Piguet, O. Emotion recognition in the dementias: Brain correlates and patient implications. *Neurodegener. Dis. Manag.* **2013**, *3*, 277–288. [CrossRef]
5. Monferrer, M.; García, A.S.; Ricarte, J.J.; Montes, M.J.; Fernández-Caballero, A.; Fernández-Sotos, P. Facial emotion recognition in patients with depression compared to healthy controls when using human avatars. *Sci. Rep.* **2023**, *13*, 6007. [CrossRef]
6. Monferrer, M.; García, A.S.; Ricarte, J.J.; Montes, M.J.; Fernández-Sotos, P.; Fernández-Caballero, A. Facial Affect Recognition in Depression Using Human Avatars. *Appl. Sci.* **2023**, *13*, 1609. [CrossRef]
7. Consoli, D. A new concept of marketing: The emotional marketing. *Broad Res. Account. Negot. Distrib.* **2010**, *1*, 52–59.
8. Mehrabian, A.; Russell, J.A. *An Approach to Environmental Psychology*; The MIT Press: Cambridge, MA, USA, 1974.
9. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef]
10. Russell, J.A. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychol. Bull.* **1994**, *115*, 102–141. [CrossRef]

11. Upadhyay, A.; Dewangan, A.K. Facial expression recognition: A review. *Int. J. Latest Trends Eng. Technol.* **2016**, *3*, 237–243.
12. Li, S.; Deng, W. Deep Facial Expression Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2022**, *13*, 1195–1215. [CrossRef]
13. Ryumina, E.; Dresvyanskiy, D.; Karpov, A. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing* **2022**, *514*, 435–450. [CrossRef]
14. Lozano-Monasor, E.; López, M.; Vigo-Bustos, F.; Fernández-Caballero, A. Facial expression recognition in ageing adults: From lab to ambient assisted living. *J. Ambient Intell. Humaniz. Comput.* **2017**, *8*, 567–578. [CrossRef]
15. Lozano-Monasor, E.; López, M.T.; Fernández-Caballero, A.; Vigo-Bustos, F. Facial Expression Recognition from Webcam Based on Active Shape Models and Support Vector Machines. In Proceedings of the Ambient Assisted Living and Daily Activities, Belfast, UK, 2–5 December 2014; Pecchia, L., Chen, L.L., Nugent, C., Bravo, J., Eds.; Springer: Cham, Switzerland, 2014; pp. 147–154.
16. Revina, I.M.; Emmanuel, W.S. A Survey on Human Face Expression Recognition Techniques. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *33*, 619–628. [CrossRef]
17. Kandeel, A.; Rahmanian, M.; Zulkernine, F.; Abbas, H.M.; Hassanein, H. Facial Expression Recognition Using a Simplified Convolutional Neural Network Model. In Proceedings of the 2020 International Conference on Communications, Signal Processing, and their Applications, Sharjah, United Arab Emirates, 16–18 March 2021; pp. 1–6. [CrossRef]
18. Taee, E.J.A.; Jasim, Q.M. Blurred Facial Expression Recognition System by Using Convolution Neural Network. *Webology* **2020**, *17*, 804–816. [CrossRef]
19. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism. *IEEE Trans. Image Process.* **2019**, *28*, 2439–2450. [CrossRef]
20. Zhao, Z.; Liu, Q.; Wang, S. Learning Deep Global Multi-Scale and Local Attention Features for Facial Expression Recognition in the Wild. *IEEE Trans. Image Process.* **2021**, *30*, 6544–6556. [CrossRef]
21. Patel, K.; Mehta, D.; Mistry, C.; Gupta, R.; Tanwar, S.; Kumar, N.; Alazab, M. Facial Sentiment Analysis Using AI Techniques: State-of-the-Art, Taxonomies, and Challenges. *IEEE Access* **2020**, *8*, 90495–90519. [CrossRef]
22. Deshmukh, S.; Patwardhan, M.; Mahajan, A. Survey on real-time facial expression recognition techniques. *IET Biom.* **2016**, *5*, 155–163. [CrossRef]
23. Pham, L.; Vu, T.H.; Tran, T.A. Facial Expression Recognition Using Residual Masking Network. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: New York, NY, USA, 2021; pp. 4513–4519. [CrossRef]
24. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2019**, *14*, 241–258. [CrossRef]
25. Zaballos, A.; Briones, A.; Massa, A.; Centelles, P.; Caballero, V. A Smart Campus' Digital Twin for Sustainable Comfort Monitoring. *Sustainability* **2020**, *12*, 9196. [CrossRef]
26. Li, J.; Jin, K.; Zhou, D.; Kubota, N.; Ju, Z. Attention mechanism-based CNN for facial expression recognition. *Neurocomputing* **2020**, *411*, 340–350. [CrossRef]
27. Sun, W.; Zhao, H.; Jin, Z. A visual attention based ROI detection method for facial expression recognition. *Neurocomputing* **2018**, *296*, 12–22. [CrossRef]
28. Wang, Z.; Zeng, F.; Liu, S.; Zeng, B. OAENet: Oriented attention ensemble for accurate facial expression recognition. *Pattern Recognit.* **2021**, *112*, 107694. [CrossRef]
29. Wen, Z.; Lin, W.; Wang, T.; Xu, G. Distract Your Attention: Multi-Head Cross Attention Network for Facial Expression Recognition. *Biomimetics* **2023**, *8*, 199. [CrossRef] [PubMed]
30. 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]
31. 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, Jodhpur, India, 15–18 December 2021; IEEE: New York, NY, USA, 2021; pp. 1–8. [CrossRef]
32. Shao, J.; Qian, Y. Three convolutional neural network models for facial expression recognition in the wild. *Neurocomputing* **2019**, *355*, 82–92. [CrossRef]
33. Zhu, Q.; Mao, Q.; Jia, H.; Noi, O.E.N.; Tu, J. Convolutional relation network for facial expression recognition in the wild with few-shot learning. *Expert Syst. Appl.* **2022**, *189*, 116046. [CrossRef]
34. Dubey, S.R.; Singh, S.K. Transformer-based Generative Adversarial Networks in Computer Vision: A Comprehensive Survey. *arXiv* **2023**, arXiv:2302.08641. https://doi.org/10.48550/ARXIV.2302.08641.
35. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. https://doi.org/10.48550/ARXIV.2010.11929.
36. Maurício, J.; Domingues, I.; Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Appl. Sci.* **2023**, *13*, 5521. [CrossRef]
37. Huang, Q.; Huang, C.; Wang, X.; Jiang, F. Facial expression recognition with grid-wise attention and visual transformer. *Inf. Sci.* **2021**, *580*, 35–54. [CrossRef]
38. Kim, S.; Nam, J.; Ko, B.C. Facial Expression Recognition Based on Squeeze Vision Transformer. *Sensors* **2022**, *22*, 3729. [CrossRef] [PubMed]

39. Liu, C.; Hirota, K.; Dai, Y. Patch attention convolutional vision transformer for facial expression recognition with occlusion. *Inf. Sci.* **2023**, *619*, 781–794. [CrossRef]
40. Ma, F.; Sun, B.; Li, S. Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion. *IEEE Trans. Affect. Comput.* **2023**, *14*, 1236–1248. [CrossRef]
41. Yan, Y.; Zhang, Z.; Chen, S.; Wang, H. Low-resolution facial expression recognition: A filter learning perspective. *Signal Process.* **2020**, *169*, 107370. [CrossRef]
42. Bodavarapu, P.N.R.; Srinivas, P.V.V.S. Facial expression recognition for low resolution images using convolutional neural networks and denoising techniques. *Indian J. Sci. Technol.* **2021**, *14*, 971–983. [CrossRef]
43. Nan, F.; Jing, W.; Tian, F.; Zhang, J.; Chao, K.M.; Hong, Z.; Zheng, Q. Feature super-resolution based Facial Expression Recognition for multi-scale low-resolution images. *Knowl. Based Syst.* **2022**, *236*, 107678. [CrossRef]
44. Shao, J.; Cheng, Q. E-FCNN for tiny facial expression recognition. *Appl. Intell.* **2020**, *51*, 549–559. [CrossRef]
45. Lee, K.; Kim, S.; Lee, E.C. Fast and Accurate Facial Expression Image Classification and Regression Method Based on Knowledge Distillation. *Appl. Sci.* **2023**, *13*, 6409. [CrossRef]
46. Ma, T.; Tian, W.; Xie, Y. Multi-level knowledge distillation for low-resolution object detection and facial expression recognition. *Knowl-Based Syst.* **2022**, *240*, 108136. [CrossRef]
47. Huang, Z.; Yang, S.; Zhou, M.; Li, Z.; Gong, Z.; Chen, Y. Feature Map Distillation of Thin Nets for Low-Resolution Object Recognition. *IEEE Trans. Image Process.* **2022**, *31*, 1364–1379. [CrossRef] [PubMed]
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 27–30 June 2016, Las Vegas, NV, USA; IEEE: New York, NY, USA, 2016; pp. 770–778. [CrossRef]
49. Hua, W.; Dai, F.; Huang, L.; Xiong, J.; Gui, G. HERO: Human Emotions Recognition for Realizing Intelligent Internet of Things. *IEEE Access* **2019**, *7*, 24321–24332. [CrossRef]
50. Li, X.; Zhu, C.; Zhou, F. Facial Expression Recognition: One Attention-Modulated Contextual Spatial Information Network. *Entropy* **2022**, *24*, 882. [CrossRef] [PubMed]
51. Fu, B.; Mao, Y.; Fu, S.; Ren, Y.; Luo, Z. Blindfold Attention: Novel Mask Strategy for Facial Expression Recognition. In Proceedings of the 2022 International Conference on Multimedia Retrieval, Newark, NJ, USA, 27–30 June 2022; ACM: Frisco, TX, USA, 2022; pp. 624–630. [CrossRef]
52. Guo, Y.; Huang, J.; Xiong, M.; Wang, Z.; Hu, X.; Wang, J.; Hijji, M. Facial expressions recognition with multi-region divided attention networks for smart education cloud applications. *Neurocomputing* **2022**, *493*, 119–128. [CrossRef]
53. Zhang, Y.; Wang, C.; Ling, X.; Deng, W. Learn from All: Erasing Attention Consistency for Noisy Label Facial Expression Recognition. In *Lecture Notes in Computer Science*; Springer Nature: Cham, Switzerland, 2022; pp. 418–434. [CrossRef]
54. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
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.* **2019**, *10*, 18–31. [CrossRef]
56. 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]
57. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980. [CrossRef]
58. Georgescu, M.I.; Ionescu, R.T.; Popescu, M. Local Learning with Deep and Handcrafted Features for Facial Expression Recognition. *IEEE Access* **2019**, *7*, 64827–64836. [CrossRef]
59. Fu, Y.; Wu, X.; Li, X.; Pan, Z.; Luo, D. Semantic Neighborhood-Aware Deep Facial Expression Recognition. *IEEE Trans. Image Process.* **2020**, *29*, 6535–6548. [CrossRef]
60. Han, B.; Yun, W.H.; Yoo, J.H.; Kim, W.H. Toward Unbiased Facial Expression Recognition in the Wild via Cross-Dataset Adaptation. *IEEE Access* **2020**, *8*, 159172–159181. [CrossRef]
61. Liu, D.; Ouyang, X.; Xu, S.; Zhou, P.; He, K.; Wen, S. SAANet: Siamese action-units attention network for improving dynamic facial expression recognition. *Neurocomputing* **2020**, *413*, 145–157. [CrossRef]
62. Li, C.; Li, X.; Wang, X.; Huang, D.; Liu, Z.; Liao, L. FG-AGR: Fine-Grained Associative Graph Representation for Facial Expression Recognition in the Wild. *IEEE Trans. Circuits Syst. Video Technol.* 2023, *early access*. [CrossRef]
63. Gera, D.; Balasubramanian, S. Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. *Pattern Recognit. Lett.* **2021**, *145*, 58–66. [CrossRef]