

Article

Symmetric Multi-Scale Residual Network Ensemble with Weighted Evidence Fusion Strategy for Facial Expression Recognition

Juan Liu ^{1,2}, Min Hu ² , Ying Wang ¹, Zhong Huang ^{1,*} and Julang Jiang ¹

¹ School of Electronic Engineering and Intelligent Manufacturing, Anqing Normal University, Anqing 246133, China

² Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, School of Computer Science and Information, Hefei University of Technology, Hefei 230009, China

* Correspondence: huangzh@aqnu.edu.cn

Abstract: To extract facial features with different receptive fields and improve the decision fusion performance of network ensemble, a symmetric multi-scale residual network (SMResNet) ensemble with a weighted evidence fusion (WEF) strategy for facial expression recognition (FER) was proposed. Firstly, aiming at the defect of connecting different filter groups of Res2Net only from one direction in a hierarchical residual-like style, a symmetric multi-scale residual (SMR) block, which can symmetrically extract the features from two directions, was improved. Secondly, to highlight the role of different facial regions, a network ensemble was constructed based on three networks of SMResNet to extract the decision-level semantic of the whole face, eyes, and mouth regions, respectively. Meanwhile, the decision-level semantics of three regions were regarded as different pieces of evidence for decision-level fusion based on the Dempster-Shafer (D-S) evidence theory. Finally, to fuse the different regional expression evidence of the network ensemble, which has ambiguity and uncertainty, a WEF strategy was introduced to overcome conflicts within evidence based on the support degree adjustment. The experimental results showed that the facial expression recognition rates achieved 88.73%, 88.46%, and 88.52% on FERPlus, RAF-DB, and CAER-S datasets, respectively. Compared with other state-of-the-art methods on three datasets, the proposed network ensemble, which not only focuses the decision-level semantics of key regions, but also addresses to the whole face for the absence of regional semantics under occlusion and posture variations, improved the performance of facial expression recognition in the wild.

Keywords: facial expression recognition; machine learning; symmetric multi-scale residual network; network ensemble; D-S evidence theory; symmetry and asymmetry



Citation: Liu, J.; Hu, M.; Wang, Y.; Huang, Z.; Jiang, J. Symmetric Multi-Scale Residual Network Ensemble with Weighted Evidence Fusion Strategy for Facial Expression Recognition. *Symmetry* **2023**, *15*, 1228. <https://doi.org/10.3390/sym15061228>

Academic Editors: Antonio Palacios, Yiming Tang and Alexander Zaslavski

Received: 12 May 2023

Revised: 2 June 2023

Accepted: 5 June 2023

Published: 8 June 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 expression plays an important role in daily communication as it is a natural and universal way for human to convey emotional states [1,2]. Currently, automatic facial expression recognition (FER) has many applications, such as social robots, safe driving, intelligent medicine, and other human–computer interaction fields [3,4]. In recent years, numerous novel methods have been proposed for FER in experimental environments. However, it is challenging to cope with the diversity and complexity of FER in uncontrolled environments (e.g., non-frontal faces, fuzzy faces, partially occluded faces, and spontaneous expressions) [5–9].

At present, FER is divided into feature engineering-based methods and end-to-end deep learning-based methods. Feature engineering-based methods, which include feature extraction and feature classification, have a common difficulty of selecting robust facial features for expression classification [10], and lack compensation and fault-tolerance mechanisms for occlusions and posture variations in the wild. Recently, to reduce the

interference of human factors and the errors in manual feature selection, end-to-end deep learning-based methods, especially convolutional neural networks (CNNs), have been rapidly applied to FER [11]. Owing to their rich semantic representation capabilities, CNN-based methods [7,12–14] have achieved an advanced recognition rate and outperformed the previous feature engineering-based methods. Research shows that deeper convolution has a wider receptive field for rich semantic features, but is easily affected by occlusions and varying postures [15]. Therefore, a multi-scale Res2Net block was designed to extract features with different receptive fields [16]. However, Res2Net obtains multi-scale features only from left to right in the basic block, which results in a limited range of receptive fields, and how to expand receptive fields for facial features extraction is a problem to be solved in the paper.

In addition, the current end-to-end deep learning-based methods, which only take a whole face image as input, are difficult to accurately capture the subtle changes in key regions of the face and ignore the importance of different facial regions. Psychological research shows that facial expression details focus on the eyebrows, eyes, and mouth regions [17]. Furthermore, human beings can effectively use local regions and the whole face to perceive the semantics transmitted by incomplete faces [18]. Hence, how to highlight the role of key regions and maintain the interoperability of the whole face is the other problem to be solved in this paper.

Recent studies have shown that the performance of a set of multiple networks is better than that of a single network [19,20], and the decision-level-based network ensemble strategy becomes mainstream [21]. In the decision-level network ensemble, the Dempster-Shafer (D-S) evidence theory, which has unique advantages in terms of flexibility and effectiveness of modeling uncertainty and imprecision, has been one of the most competitive fusion strategies [22]. However, the expression semantics of different regions have a certain degree of ambiguity and uncertainty, and the traditional D-S combination rules (DCRs) fail when there are conflicts in evidence [23]. How to reduce the conflicts between the decision-level semantic of different regions is a third problem to be solved.

To solve the above problems, a FER framework based on a symmetric multi-scale residual network (SMResNet) ensemble with a weighted evidence fusion (WEF) strategy was proposed. The overall structure of the proposed framework is shown in Figure 1. It mainly includes a preprocessing module for face alignment and facial key region location, a decision-level semantic extraction based on SMResNet ensemble, and a decision-level semantic fusion with the WEF strategy. In the preprocessing module, due to facial images in unconstrained scenes having complex backgrounds, illuminations, head poses, and local occlusions, a pipeline for face alignment and the location of facial key regions was designed. In the decision-level semantic extraction, to overcome the defect of connecting different filter groups of Res2Net only from one direction, a symmetric multi-scale residual (SMR) block was firstly improved to symmetrically extract the features of different receptive fields from two directions. Secondly, a network ensemble, composed of three networks of SMResNet, was constructed to extract the decision-level semantics of the whole face, eyes, and mouth. In the decision-level semantic fusion, the outputs of three SMResNets were regarded as different evidence, and the D-S evidence theory was addressed to decision-level fusion of these three pieces of evidence. Given the ambiguity and uncertainty of different expression evidence, a WEF strategy was introduced to overcome the conflicts in evidence based on the support degree adjustment. The main contributions of this paper are as follows.

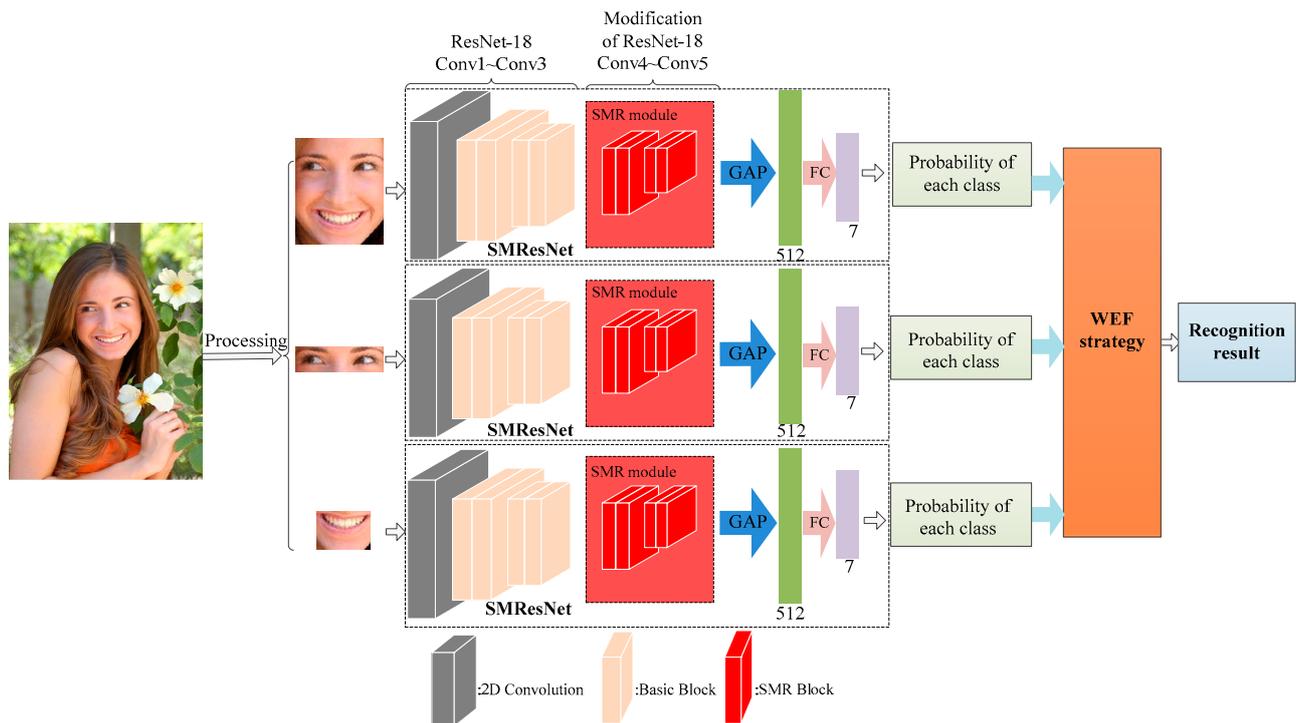


Figure 1. The overall framework of the proposed method.

- (1) To further extract features in a wider range of receptive fields, an SMR block, which symmetrically constructs hierarchical residual-like connections from two directions in a basic block, was improved. It represents the multi-scale feature of fine-grained level and can further expand the receptive field range of each network layer.
- (2) To establish the complementarity of global facial features and regional detail features, a network ensemble of SMResNet, which consists of one 2D convolution layer, four basic blocks, one SMR module composed of four cascaded SMR blocks, a global average pooling (GAP) layer, and a full connection (FC) layer, was constructed. The constructed ensemble framework not only focuses on the decision-level semantics of the eye and mouth regions, but also addresses the whole face for the absence of regional semantics under occlusions and posture variations.
- (3) To restrain the ambiguity and uncertainty of decision-level semantics from the constructed SMResNet ensemble, a decision-level semantic fusion with WEF strategy was proposed based on the D-S evidence theory. The proposed strategy, which overcomes the conflicts in evidence by the support degree adjustment, is helpful to minimize the influence of evidence with small weight on the decision-making judgment, and reduce conflicting information between evidence to satisfy the DCRs for the decision-level fusion of three pieces of regional evidence.

The rest of this paper is organized as follows: Section 2 puts forward the review of previous work, Section 3 introduces the proposed framework of FER system, Section 4 discusses the experiments and results, and Section 5 concludes this paper.

2. Related Work

This paper addresses facial expression recognition based on network ensemble; hence, current facial expression recognition methods, network ensemble structures, and decision-level fusion strategies were reviewed.

2.1. Facial Expression Recognition Methods

Due to the important role of facial expression recognition in the field of computer vision, research on facial expression recognition methods has received wide public concern.

Currently, FER methods can be divided into traditional feature engineering-based methods and end-to-end deep learning-based methods.

The traditional feature engineering-based methods, composed of facial feature extraction and facial expression classification, mainly explore facial expression variations using geometric and appearance features [11], and classify facial expression categories based on support vector machine (SVM) [24] or principal component analysis (PCA) [25]. In general, geometric feature-based methods have the advantages of low dimension and being insensitive to illumination variations, but their ability for local detail description is weak. Appearance feature-based methods, such as Gabor [26], local binary pattern (LBP) [27], and histogram of oriented gradient (HOG) features [28], contain a large amount of expression information and have a relatively stable extraction process, but the extracted features have high dimension and are susceptible to illumination variations, and even mixed with inference data. Hence, the traditional feature engineering-based methods not only have difficulty selecting robust features for expression classification, but also lack compensation and fault-tolerance mechanisms for occlusion and posture variations in the wild. Furthermore, the separate processes of feature extraction and classification cannot be integrated into an end-to-end model.

Due to the increasing amount of data used to train deep models and the improvement in GPU technology, an important part of the advancement in recognition performance is the advent of deep learning methods. The end-to-end deep learning-based methods attempt to capture high-level semantics through multiple hierarchical structures of non-linear transformations and representations. In addition to subject identity bias, variations in posture, illumination, and occlusions are also common in unconstrained facial expression scenes, which are nonlinearly confused with facial expressions, reinforcing the need for deep networks to address large intra-class variability and learn effective specific expression representations. A survey of the research on deep learning FER can be found in [21,29,30]. Depending on different network structures, the end-to-end deep learning-based methods can be further divided into convolutional neural networks [7,12–14], deep belief networks [31], deep autoencoders [32], recurrent neural networks [33], and generative adversarial networks [34]. Compared with other neural networks, CNNs possess unique superiority in facial spatial semantic extraction due to their own convolution and pooling operations, and have better performance for FER in the wild [7,12–14].

Li et al. [12] proposed a CNN with an attention mechanism named as ACNN based on VGG-16 [35], which could perceive the occlusion regions of the face and focus on the most discriminative un-occluded regions. Each region of interest (ROI) was weighed via a gate unit that computed a weight from the region itself. To capture the importance of facial region information, and make a reasonable trade-off between the region and global features, Wang et al. [13] proposed the new regional attention network (RAN) using VGG-16 [35] or ResNet-18 [36] as the backbone network. The weight distribution of facial regions was realized by the self-attention module and the region bias loss. The influence of occlusion and irrelevant regions was reduced or eliminated by increasing the attention weight of important regions. The VGGNet increases the network depth and uses filters with a smaller kernel size. A deeper structure can expand the receptive fields, which is very useful for extracting features from a larger scale. However, as the depth of the network increases, the gradient may vanish or explode. Different from VGGNet, the ResNet block introduces a short connection to neural networks to alleviate these problems and can obtain much deeper network structures [36]. Short connections in ResNet allow different combinations of convolution operators to obtain a large number of equivalent feature scales. Recently, hierarchical residual-like connections are introduced into the Res2Net block [16] to enable the variation of receptive fields at a finer granularity for capturing details and global features. The related experimental results have shown that Res2Net is better than ResNet in the context of several representative computer vision tasks. However, Res2Net extracts multi-scale features only from left to right in the basic block, which results in a limited range of receptive fields.

In addition, Zhao et al. [14] extracted global features using multi-scale modules, which reduced the sensitivity of deep convolution to occlusions and posture variations, and then obtained attention semantics from the four regions of middle-level facial features. However, it was necessary to design hyperparameters to balance global features and local features, resulting in achieving decision fusion for expression classification.

2.2. Network Ensemble Structures

Recent advances in deep learning have shown that combining multiple deep learning models can considerably outperform the approach of using only a single deep learning model for challenging recognition problems [19,37]. Cho et al. proposed a novel deep convolutional neural network (DCNN) ensemble for FER in the wild [37]. Karnati [38] designed a texture-based feature-level ensemble parallel network (FLEPNet) for FER, and addressed insufficient training data and intra-class facial appearance variations.

When implementing a network ensemble, two key factors should be considered: sufficient diversity of networks to ensure complementarity and an appropriate ensemble strategy that can effectively aggregate the committee networks [21]. There are many methods to generate network diversity, such as different training data, several preprocessing methods, varying network models, various parameters, and so on. The committee networks that extract local features and global features in our paper are used to construct the network ensemble. Each member of the committee network can be assembled at two different levels: feature-level and decision-level [39]. For the feature-level ensemble, the most commonly used strategy is to concatenate the features learned from different networks. However, the feature-level fusion method has the problem of dimension catastrophe. Therefore, the decision fusion method is chosen in our paper.

2.3. Decision Fusion Strategies

Decision fusion strategies mainly include majority vote and D-S evidence theory. The majority voting strategy is a simple and effective method for decision-level data fusion [40]. However, the majority voting method does not consider the importance and confidence of each individual, and is prone to controversial decision results.

D-S evidence theory, which has the flexibility and effectiveness to model uncertainty and imprecision without considering prior information, has been widely used in various fields of information fusion [41], such as decision-making, pattern recognition, risk analysis, supplier selection, fault diagnosis, and so on. Specifically, the fusion results generated through the DCRs exhibit fault tolerance, enabling them to better support decision-making [42]. However, the D-S evidence theory may lead to counterintuitive results when fusing highly conflicting evidence [43]. To solve this problem, it has been proposed to modify DCRs and preprocess the evidence [44]. Considering the advantages of D-S in decision fusion, and the ambiguity and uncertainty of facial semantics, this paper proposed a decision-level semantic fusion with WEF strategy that integrates D-S evidence theory into a network ensemble framework for FER.

3. Methodology

The recognition framework proposed in this paper mainly includes a preprocessing module for face alignment and facial key region location, a decision-level semantic extraction based on the SMResNet ensemble, and a decision-level semantic fusion with the WEF strategy. At last, to solve the parameters of SMResNets and improve the convergence speed of the network ensemble, the optimization for the SMResNet ensemble was designed.

3.1. A Preprocessing Module for Face Alignment and Facial Key Region Location

Unlike frontal facial images collected in laboratory settings, facial images in the wild have complex backgrounds, head posture deviation, non-uniform illumination, and local occlusion. To suppress redundant information in input images and improve the anti-interference

ability of facial regions, a preprocessing module was designed. The pipeline of the proposed preprocessing module mainly included face alignment and facial key region location.

3.1.1. Face Alignment

To learn meaningful features for training deep neural networks, preprocessing is usually required to align and standardize the visual semantic information conveyed by the human face [45]. The face alignment algorithm [46], which is robust to face occlusion, was addressed to the calibrated position of landmarks in our paper. Firstly, 68 facial landmark points, each of which can be defined by coordinates, were detected. The positional relationship of 68 facial landmark points is shown in Figure 2.

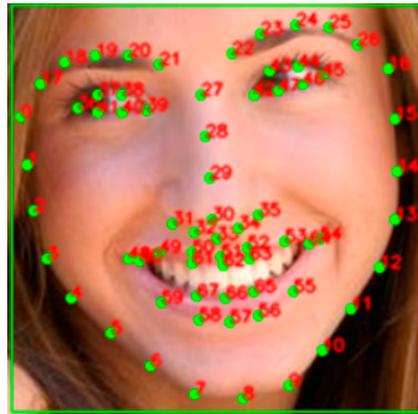


Figure 2. Positional relationship of 68 facial landmark points.

Secondly, twelve landmarks (No. 36 to 47) around the left and right eyes were selected to achieve facial image alignment, as shown in Figure 3. The angle θ between the horizontal axis and the line segment, running from one eye center to the other eye center, and the inter-eye distance d between the centers of the two eyes, are calculated as:

$$\theta = \arctan \frac{\sum_{i=42}^{47} y_i - \sum_{i=36}^{41} y_i}{\sum_{i=42}^{47} x_i - \sum_{i=36}^{41} x_i}, \quad (1)$$

$$d = \frac{\sum_{i=42}^{47} x_i - \sum_{i=36}^{41} x_i}{6}, \quad (2)$$

where x_i, y_i represent the abscissa and ordinate of the i -th landmark, respectively.



Figure 3. Face alignment and ROI_{face} cropping.

Finally, according to facial proportion structure and human experience, a vertical factor and horizontal factor for ROI of face are delimited 2.2 (considering 0.6 for the region above the eyes and 1.6 for the region below) and 1.8, respectively. The final facial region ROI_{face} is cropped from the aligned facial image:

$$ROI_{face} = Image_{\theta}(P_{LU}, W_F, H_F), \quad (3)$$

where $Image_{\theta}$ represents the aligned facial image after rotating the θ angle, P_{LU} , W_F , and H_F represent the left upper vertex, the width and height of the cropped facial ROI region, respectively. Specifically,

$$P_{LU} : x = P_{27} : x - 0.6d, \quad P_{LU} : y = P_{27} : y - 0.9d, \quad (4)$$

$$W_F = 1.8d, \quad H_F = 2.2d, \quad (5)$$

where $P : x$ and $P : y$ are the abscissa and ordinate of landmark P , respectively.

3.1.2. Facial Key Region Location

Human facial expressions are not only closely related to the global information presented on the face, but also to the local details of key facial regions such as the eyes and mouth. Therefore, to highlight the important role of different facial regions in FER, six landmarks in ROI_{face} are used to achieve localization and cropping of the eye and mouth regions, as shown in Figure 4. The six landmarks in Figure 4 are represented as $P = \{P_{LEBU}, P_{REBU}, P_{NM}, P_{NL}, P_{MLC}, P_{MRC}\}$, where the landmarks represent upper of left eyebrow, upper of right eyebrow, middle of nose, lower of nose, left corner of mouth, and right corner of mouth, respectively.

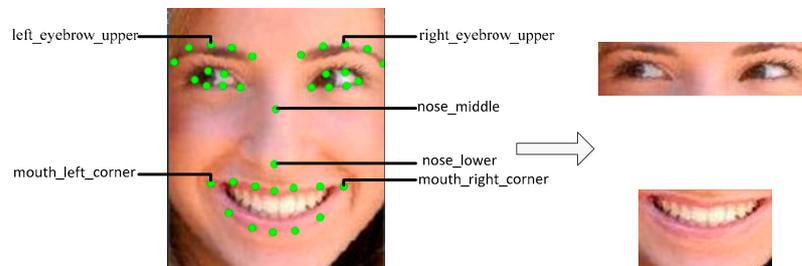


Figure 4. Selection of key regions.

Firstly, the eye region ROI_{eyes} is calculated from the cropped image ROI_{face} :

$$ROI_{eyes} = ROI_{face}(P_{LUER}, W_E, H_E), \quad (6)$$

where P_{LUER} , W_E , and H_E represent the left upper vertex, the width and height of the eye region, respectively. Specifically,

$$P_{LUER} : x = 0, \quad P_{LUER} : y = \min(P_{LEBU} : y, P_{REBU} : y), \quad (7)$$

$$W_E = W_F, \quad H_E = P_{NM} : y - P_{LUER} : y, \quad (8)$$

where \min is the function taking the minimum value.

Similarly, the mouth region ROI_{mouth} is calculated from the cropped image ROI_{face} :

$$ROI_{mouth} = ROI_{face}(P_{LUMR}, W_M, H_M), \quad (9)$$

$$P_{LUMR} : x = P_{MLC} : x, \quad P_{LUER} : y = P_{NL} : y, \quad (10)$$

$$W_M = P_{MRC} : x - P_{MLC} : x, H_M = H_F - P_{LUER} : y, \tag{11}$$

where P_{LUMR} , W_M , and H_M represent the left upper vertex, the width and height of the mouth region, respectively.

Through the above pipeline for face alignment and facial key region location, the whole face, eye region, and mouth region can be cropped from the input image. The designed pipeline can suppress redundant information of the input image and improve the anti-interference ability of facial regions.

3.2. A Decision-Level Semantic Extraction Based on SMResNet Ensemble

In this section, we introduce the construction of SMResNet ensemble. Firstly, to extract facial features of different receptive fields with a wider range, an SMR block was proposed. Then, the SMResNet, which embeds SMR blocks, was designed. Finally, the SMResNet ensemble, composed of three SMResNets, was constructed to extract decision-level semantic features of the whole facial region and key facial expression regions such as the eyes and mouth.

3.2.1. SMR Block

To extract multi-scale facial features, the basic block, composed of a group of a 3×3 filter, was embedded into ResNet-18 and ResNet-34 [36], as shown in Figure 5a. However, most existing methods based on the basic block represent the multi-scale features in a layer-wise manner. The Res2Net block sought smaller groups of filters, and connected different filter groups in a hierarchical residual-like style [16], as shown in Figure 5b. The Res2Net block represents multi-scale features at a granular level. In the Res2Net block, the hierarchical residual-like connections within a single basic block can increase the range of receptive fields for each network layer. However, the direction of the filter 3×3 residual connection is a single direction, thus causing the range of receptive fields for each network layer limited. To extract facial features of different receptive fields with a wider range, an SMR block, which symmetrically learns the multi-scale features from both directions, was proposed, as shown in Figure 5c.

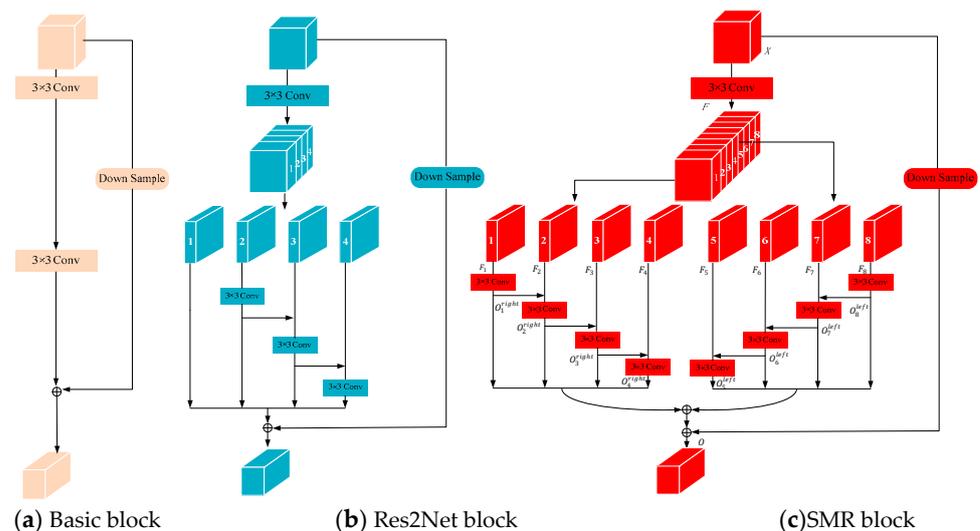


Figure 5. Comparison of basic block, Res2Net block, and SMR block, (a) Basic block; (b) Res2Net block; (c) SMR block.

Firstly, the output of the previous modules is regarded as the input X of SMR block. After the 3×3 convolution, the feature map F can split into s feature map subsets:

$$F = Conv(X), \tag{12}$$

$$\{F_i\}_{i=1}^s = \text{split}(F), \quad (13)$$

where F_i , which has $1/s$ number of channels of the input feature map, represents the i -th feature map subset; s is the number of feature map subsets.

Secondly, F_i ($1 \leq i \leq s$) is carried out a series of 3×3 convolutions, and the symmetrical multi-scale features O_i^{left} and O_j^{right} of F_i can be expressed as:

$$O_i^{\text{left}} = \begin{cases} \text{Conv}_i^{\text{left}}(F_i) & i = s \\ \text{Conv}_i^{\text{left}}(F_i + O_{i+1}^{\text{left}}) & s/2 + 1 \leq i < s' \end{cases} \quad (14)$$

$$O_j^{\text{right}} = \begin{cases} \text{Conv}_j^{\text{right}}(F_j) & j = 1 \\ \text{Conv}_j^{\text{right}}(F_j + O_{j-1}^{\text{right}}) & 1 < j \leq s/2' \end{cases} \quad (15)$$

where $\text{Conv}_i^{\text{left}}(\cdot)$, $\text{Conv}_j^{\text{right}}(\cdot)$ represent the 3×3 convolution operation in right branch and the left branch, respectively. Equation (14) shows that the $\text{Conv}_i^{\text{left}}(\cdot)$ operation can capture all the features from the subset $\{F_k, i \leq k \leq s\}$, while Equation (15) shows that the $\text{Conv}_j^{\text{right}}(\cdot)$ operation can capture all the features from the subset $\{F_k, k \leq j \leq s/2\}$. Since each F_k goes through a 3×3 convolution processing, O_i^{left} and O_j^{right} include a different number and proportion of feature subsets. To obtain more diverse multi-scale features, all O_i^{left} and O_j^{right} along the channel axis are concatenated. The final output of SMR block can be expressed as:

$$O = \text{Concat}\{O_i^{\text{left}}\}_{i=s/2+1}^s + \text{Concat}\{O_j^{\text{right}}\}_{j=1}^{s/2} + \text{DownSample}(X), \quad (16)$$

where $\text{Concat}(\cdot)$ indicates the concatenation operation along the channel axis; $\text{DownSample}(\cdot)$ represents downsampling operation. O is the output feature map of the input X . When the value of s is large, the learned feature contains more scale information, but it will increase the computational cost. In the experiment, set $s = 8$, which is a trade-off between performance and calculation.

Compared with Res2Net, the designed SMR block overcomes the problem of residual network connections being only one-way from left to right, and can extract the facial features from left to right and from right to left, resulting in a wider range of receptive fields.

3.2.2. SMResNet Ensemble for Decision-Level Semantic Extraction

To expand the receptive fields and extract regional facial expressions information, the SMResNet, which embeds SMR blocks, was designed, as shown in Figure 1. The designed SMResNet consisted of one 2D convolution layer, four basic blocks, one SMR module composed of four cascaded SMR blocks, a GAP layer, and an FC layer. The structural parameters of SMResNet are listed in Table 1.

Table 1. Structure parameters of SMResNet.

Layer	Structure Parameter	Output Size
Conv1	$7 \times 7, 64$, stride 2	$64 \times 112 \times 112$
Max pool	3×3 max pool, stride 2	$64 \times 56 \times 56$
Conv2	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$64 \times 56 \times 56$
Conv3	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$128 \times 28 \times 28$

Table 1. Cont.

Layer	Structure Parameter	Output Size
Modified Conv4	$\left[\begin{array}{c} (3 \times 3, 256) \\ (3 \times 3, 32) \times 4, (3 \times 3, 32) \times 4 \end{array} \right] \times 2$	$256 \times 14 \times 14$
Modified Conv5	$\left[\begin{array}{c} (3 \times 3, 512) \\ (3 \times 3, 64) \times 4, (3 \times 3, 64) \times 4 \end{array} \right] \times 2$	$512 \times 7 \times 7$
GAP	7×7 average pool	$512 \times 1 \times 1$
FC	$(512, 7)$	7

As opposed to directly taking global facial semantics from the facial ROI region as the input [11], the SMResNet ensemble, composed of three SMResNets, was constructed to extract decision-level semantic features of the whole facial region and key facial expression regions such as the eyes and mouth, as shown in Figure 1. Set the decision-level semantic of the SMResNet from i -th region is Y_i :

$$Y_i = (y_{i1}, \dots, y_{ij}, \dots, y_{iN}) \quad \left(\sum_{j=1}^N y_{ij} = 1, \quad i = 1, 2, 3 \right), \quad (17)$$

where Y_1 , Y_2 , and Y_3 represent the decision-level semantic of the whole face, eyes region and mouth region, respectively. y_{ij} represents the j th output signal in the softmax output layer of the i -th SMResNet. $N = 7$ represents seven classes of facial expressions.

3.3. A Decision-Level Semantic Fusion with WEF Strategy

The decision-level semantics, outputted from the softmax layer of individual SMResNet, not only have a certain degree of ambiguity and uncertainty in the same region, but also have conflicts between different regions. How to integrate decision-level semantic across different regions is the focus of this section. The D-S evidence theory, which has unique advantages in terms of flexibility and effectiveness of modeling uncertainty and imprecision, is widely used in various fields of information decision fusion. Hence, the D-S evidence theory was used to achieve multi-region decision-level semantics fusion. In addition, considering the semantic conflicts among different regions, a decision-level semantic fusion with WEF strategy was proposed for FER.

3.3.1. Basic Probability Assignment in D-S Evidence Theory

Let Θ be a finite set of mutually exclusive and exhaustive hypotheses on a problem domain, which is referred to as the framework of discernment. If $A \subseteq 2^\Theta$, then $2^\Theta \rightarrow [0, 1]$, and the conditions defined in Equation (18) are satisfied.

$$\begin{cases} \sum m(A) = 1 \\ m(\emptyset) = 0 \end{cases}, \quad (18)$$

where m is the basic probability assignment (BPA) function on Θ , and $m(A)$ can be interpreted as a measure of the belief that one is willing to commit exactly to A . If $m(A) > 0$, then A is called a focal element. Let two BPAs m_1 and m_2 on the frame of discernment of Θ and assuming that these BPAs are independent. The DCRs [43] are defined as follows:

$$\begin{cases} m(A) = \frac{\sum_{B \cap C = A} m_1(B)m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B)m_2(C)}, A \neq \emptyset, \\ m(\emptyset) = 0 \end{cases}, \quad (19)$$

where B and C are the elements of 2^Θ , $\sum_{B \cap C = \emptyset} m_1(B)m_2(C)$ is the conflict coefficient between two pieces of evidence.

The FER framework $\Theta = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7\}$ is a 7-tuple containing seven different expressions, where $c_1, c_2, c_3, c_4, c_5, c_6$ and c_7 represent seven different expressions of happiness, sadness, surprise, disgust, anger, fear and neutral, respectively. At the same time, the decision-level semantics $\{Y_i\}_{i=1}^3$ from the whole face region, eyes region and

mouth region are regarded as three pieces of evidence. The BPA of evidence m_i for the category c_j is given as:

$$m_i(c_j) = y_{ij} \quad (i = 1, 2, 3; j = 1, 2, \dots, 7). \quad (20)$$

3.3.2. WEF Strategy Using Support Degree of Evidence

The decision-level semantics from the whole face, eyes, and mouth regions have significant conflicts, and the traditional DCRs fail when there are conflicts between evidences [43,44]. Hence, the WEF is proposed. The flowchart of WEF is shown in Figure 6. Firstly, belief Jensen-Shannon (BJS) divergence [43] was used to describe the degree of conflict between each piece of evidence, which was then transformed into the degree of support between each piece of evidence. The weighting coefficient, which represents the degree of importance of the evidence, was determined by the degree of support. Secondly, the weighted coefficient was used to adjust the BPA based on the idea of the discount rate. Finally, the adjusted BPAs were synthesized using the DCRs.

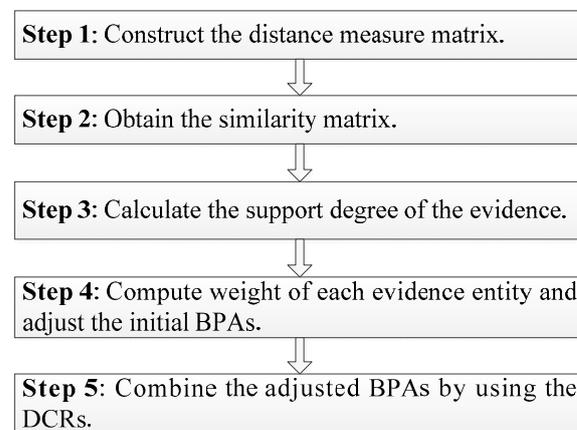


Figure 6. Flowchart of WEF strategy.

The decision-level semantics $\{m_i\}_{i=1}^3$ of the face, eyes, and mouth regions were regarded as the evidence set $E = \{E_i\}_{i=1}^3$. The distance metric d_{ij} of two pieces of evidence m_i and m_j can be calculated by the BJS divergence [43]:

$$d_{ij} = \frac{1}{2} \left[\sum_k m_i(A_k) \log \left(\frac{2m_i(A_k)}{m_i(A_k) + m_j(A_k)} \right) + \sum_k m_j(A_k) \log \left(\frac{2m_j(A_k)}{m_i(A_k) + m_j(A_k)} \right) \right], \quad (21)$$

where A_k ($k = 1, 2, \dots, N$) is a hypothesis of belief function, and m_i ($i = 1, 2, \dots, n$) and m_j ($j = 1, 2, \dots, n$) are two BPAs on the same frame of discernment Θ , containing N mutually exclusive and exhaustive hypotheses, where $n = 3$.

According to (21), a conflict matrix M_c of the evidence set E can be expressed as:

$$M_c = \begin{bmatrix} 0 & \cdots & d_{1i} & \cdots & d_{1n} \\ \vdots & \cdots & \vdots & \vdots & \vdots \\ d_{i1} & \cdots & 0 & \cdots & d_{in} \\ \vdots & \cdots & \vdots & \vdots & \vdots \\ d_{n1} & \cdots & d_{ni} & \cdots & 0 \end{bmatrix}. \quad (22)$$

Given that $d_{ij} = d_{ji}$ and $d_{ii} = 0$, the conflict matrix M_c is a symmetric matrix. If there was a smaller conflict between any two pieces of evidence, the similarity was higher between them. Thus, the conflict matrix M_c was transformed into a similarity matrix M_X :

$$M_X = \begin{bmatrix} 1 & \cdots & x_{1i} & \cdots & x_{1n} \\ \vdots & \cdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & 1 & \cdots & x_{in} \\ \vdots & \cdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{ni} & \cdots & 1 \end{bmatrix}, \quad (23)$$

where $x_{ij} = 1 - d_{ij}$ represents the degree of similarity between the evidence m_i and m_j . The confidence degree of the evidence m_i supported by other evidence is given as:

$$Sup(m_i) = \sum_{j=1, j \neq i}^n x_{ij} (i = 1, 2, \dots, n), \quad (24)$$

$Sup(m_i)$ computes the sum of all other elements of each row, except for its own similarity in the similarity matrix M_X , and it reflects the extent to which m_i is supported by other evidence. As is well known, if the similarity between one evidence and other evidence is high, they are considered mutually supportive; conversely, they are assumed to have a low degree of mutual support. Hence, the weight w_i of the evidence m_i in the fusion system can be calculated as:

$$w_i = \frac{Sup(m_i)}{\max_{1 \leq i \leq n} (Sup(m_i))} (i = 1, \dots, n). \quad (25)$$

To highlight the importance of different regional evidence and improve the reliability and fault tolerance of the fusion results, the mutual support was used to modify the BPA of each piece of evidence. The weight $\{w_i\}_{i=1}^3$ was transformed into the initial BPAs of evidence based on the idea of discount rate.

$$m'_i(A_k) = w_i m_i(A_k), \forall A \in 2^\Theta, A_k \neq \Theta, \quad (26)$$

$$m'_i(\Theta) = (1 - w_i) + w_i m_i(\Theta), \quad (27)$$

where m'_i represents the BPA after adjustment of m_i . From Equation (26), the deterministic information provided by the evidence element A with low mutual support is reduced, and Equation (27) increases the uncertainty information provided by the uncertainty element of the evidence Θ . This phenomenon reduces the impact of evidence with low mutual support on the overall fusion result.

Finally, the adjusted BPAs $\{m'_i\}_{i=1}^3$ are synthesized using the DCRs.

$$m(A) = \begin{cases} \frac{\sum_{\cap A_k = A} \prod_{i=1}^3 m'_i(A_k)}{1 - \sum_{\cap A_k = \emptyset} \prod_{i=1}^3 m'_i(A_k)}, & A \neq \emptyset, \\ 0, & A = \emptyset \end{cases}, \quad (28)$$

where $\sum_{\cap A_k = \emptyset} \prod_{i=1}^3 m'_i(A_k)$ is the coefficient of revaluation. Based on the proposed WEF strategy, the final facial expression category of decision-level fusion can be expressed as:

$$C_{result} = \arg \max_{1 \leq j \leq 7} m(c_j). \quad (29)$$

3.4. Optimization for SMResNet Ensemble

To solve the parameters of SMResNets and improve the convergence speed of the ensemble optimization, the strategies of independent optimization within branches and

joint optimization between branches were carried out. In the independent optimization, the cross-entropy loss is regarded as the optimization function for the whole face branch, eyes region branch, and mouth region branch, respectively.

$$L(Y_i, Y'_i) = - \sum_{j=1}^N y_{ij} \log y'_{ij} (i = 1, 2, 3; N = 7), \quad (30)$$

where Y_i and Y'_i represent the ground truth and predicted decision-level semantic of the i -th branch of SMResNet.

In the joint optimization, taking the independent optimization parameters as the initial values, the total loss of three branches was regarded as the objective function for the fine-tuning of each branch parameter:

$$L(Y, Y') = - \sum_{i=1}^3 \sum_{j=1}^N y_{ij} \log y'_{ij} (N = 7), \quad (31)$$

where Y and Y' represent the ground truth and the final output of SMResNet ensemble with the WEF strategy. This coarse-to-fine tuning strategy, which maximizes the information interaction among branches, can obtain better performance for FER.

4. Experimental Results and Analysis

In this section, the experimental evaluations were presented. Before showing the results, we firstly describe the datasets and experimental settings. Then, the experimental results on different datasets, an ablation analysis, and a comparison with the state-of-the-art FER methods are provided. Finally, we conclude the experiment results and analysis.

4.1. Datasets and Experimental Settings

To verify the proposed SMResNet with the WEF strategy, experiments are carried out on three facial expression datasets, including FERPlus, RAF-DB, and CAER-S. The FERPlus dataset contains 35,887 facial expression images [47] and eight classes of expressions, including 28,709 images as the training set, 3589 images as verification set and 3589 images as test set. Due to its ease of comparison with other methods and analysis of generalization ability, the proposed expression recognition framework focuses on the classification of seven classes of expressions. Therefore, the contempt expression in the FERPlus dataset was not considered in the experiment. The RAF-DB dataset [48] contains 30,000 facial images annotated with basic or compound expressions. In our experiment, only seven basic expressions were used, and the training set and test set include 12,271 and 3068 images, respectively. CAER-S dataset [49], created by selecting static frames from CAER dataset, contains 65,983 images, and is divided into two sets: training set (44,996 samples) and test set (20,987 samples). Each image was labeled to one of seven expressions.

For all datasets, face images were preprocessed based on the pipeline for face alignment and facial key region location, and resized to 224×224 pixels. The experiments were implemented in the server environment of Tesla T4 based on Python 3.7, Pytorch 1.3.0, and Cuda10.2. The related parameter settings are shown in Table 2.

Table 2. Related parameter settings.

Items	Settings
Size of input image	224×224
Learning rate	0.01
Weight decay	1×10^{-3}
Batch size	32
Optimizer	SGD
Momentum	0.9
Iterations	100

4.2. Experimental Results on Different Datasets

The FER rates of different regions and the proposed network ensemble are shown in Table 3. On the one hand, Table 3 shows that the performance of SMResNet based on three regions is excellent, which explains that each region retains most of the information for FER. On the other hand, compared to the individual sub-network, decision-level fusion increases the recognition rate by 2–4%, which illustrates the effectiveness and feasibility of the proposed WEF strategy.

Table 3. Experimental results on different datasets.

Datasets	Facial Region	Eyes Region	Mouth Region	Network Ensemble
FERPlus	85.36	85.15	84.45	88.73
RAF-DB	85.98	84.56	84.17	88.46
CAER-S	85.32	84.74	84.11	88.52

Figure 7 shows the FER rate of SMResNet ensemble on FERlus, RAF-DB, and CAER-S. For simplicity, the expressions of happiness, sadness, surprise, disgust, anger, fear, and neutral are expressed as Ha, Sa, Su, Di, An, Fe, and Ne. On all kinds of datasets, the proposed method not only has a high recognition rate for expressions such as happiness and surprise with rich details, but also maintains a high recognition rate for expressions such as sadness and disgust with insignificant local features and mutual confusion.

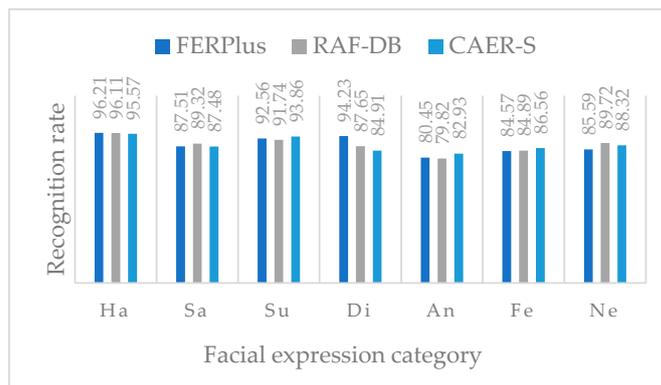


Figure 7. FER rate of SMResNet ensemble on FERlus, RAF-DB, and CAER-S.

Specially, the confusion matrix of the seven expressions on the FERPlus dataset is shown in Table 4. As can be seen from Table 4, we found that the result of happiness expression was the best (96.21%), the result of anger expression was the lowest (80.45%), and the larger inter-class error rate occurred between anger and sadness (8.84%).

Table 4. Confusion matrix of the FERPlus dataset.

	Ha	Sa	Su	Di	An	Fe	Ne
Ha	96.21	0	0	3.79	0	0	0
Sa	0	87.51	0	0	3.57	8.25	0.67
Su	0	0	92.56	0	0	5.42	2.02
Di	2.78	0	0	94.23	0.73	1.25	1.01
An	0	8.84	0	5.05	80.45	2.57	3.09
Fe	0	8.32	0	0.83	4.23	84.57	2.05
Ne	0	6.94	0.29	3.01	2.78	1.39	85.59

The confusion matrix of the seven expressions on the RAF-DB dataset is shown in Table 5. As can be seen from Table 5, we found that the result of happiness expression was the best (96.11%), the result of anger expression was the lowest (79.82%), and the larger inter-class error rate occurred between fear and anger (9.72%).

Table 5. Confusion matrix of the RAF-DB dataset.

	Ha	Sa	Su	Di	An	Fe	Ne
Ha	96.11	0	0	3.37	0	0	0.52
Sa	0	89.32	0	1.24	3.65	4.25	1.54
Su	0	0.12	91.74	0	3.21	3.71	1.22
Di	6.12	1.25	2.21	87.65	1.29	0.67	0.81
An	2.13	6.25	1.54	7.64	79.82	1.86	0.76
Fe	0	2.54	0	1.47	9.72	84.89	1.38
Ne	0	5.78	0	1.02	2.19	1.29	89.72

The confusion matrix of the seven expressions on the CAER-S dataset is shown in Table 6. As can be seen from Table 6, we found that the result of happiness expression was the best (95.57%), the result of anger expression was the lowest (82.93%), and the larger inter-class error rate occurred between fear and sadness (7.94%).

Table 6. Confusion matrix of the CAER-S dataset.

	Ha	Sa	Su	Di	An	Fe	Ne
Ha	95.57	0	0	3.01	1.29	0	0.13
Sa	0	87.48	0	1.39	3.68	5.78	1.67
Su	0	0	93.86	1.05	4.23	0.42	0.44
Di	3.78	2.17	0	84.91	3.79	3.61	1.74
An	0	6.10	0.41	5.58	82.93	4.22	0.76
Fe	0	7.94	0	2.11	2.84	86.56	0.55
Ne	0	5.63	0	2.08	1.82	2.15	88.32

After analyzing the data in Tables 4–6, it can be concluded that the result of happiness expression is the best, which can be attributed to the fact that it has the characteristics of upturned corners of mouth and narrowed eyes. The expression with the lowest recognition rate was anger, which is due to certain similarities between anger expression and fear as well as sadness. It can also be seen that the error recognition rate between various expressions did not exceed 10%. This also indicates that the proposed weighted fusion strategy based on support degree can suppress evidence conflicts, improve the credibility of regional evidence, and enhance inter-class discrimination.

4.3. Ablation Analysis

To validate the effectiveness of each component in our recognition framework, we conducted an ablation analysis. In our experiments, different network structures and different decision-level fusion methods were studied, respectively.

4.3.1. Comparison of Different Network Structures

To prove the effectiveness for feature extraction, three network structures, the baseline network of ResNet-18, the multi-scale Res2Net-18, and symmetric multi-scale SMResNet, were carried out on the FERPlus dataset, the RAF-DB dataset, and the CAER-S dataset, respectively. The comparison of FER with different network structures for the whole facial region is shown in Table 7. From Table 7, we can see that the experimental results of SMResNet for the whole facial region were the best.

Table 7. Comparison of FER with different network structures for the whole facial region.

Network	FERPlus	RAF-DB	CAER-S
Baseline [36]	83.63	82.59	84.67
Res2Net-18 [16]	84.75	83.86	85.01
SMResNet	85.36	85.98	85.32

The comparison of FER with different network structures for the network ensemble is shown in Table 8. In Table 8, compared with the baseline network of ResNet-18, the hierarchical residual-like connections are integrated into SMR block, thus improving by 1.70%, 1.90%, and 2.89% on FERPlus, RAF-DB, and CAER-S datasets, respectively. Compared with Res2Net-18, the designed SMR block overcame the problem of residual network connections being only one-way from left to right, and could extract the facial features from left to right and from right to left, resulting in an increase of 0.88%, 0.73%, and 1.67% on the FERPlus, RAF-DB, and CAER-S datasets, respectively. This indicates that the proposed SMR block can effectively extract facial expression features from different regions, thereby improving the FER rate.

Table 8. Comparison of FER with different network structures for the network ensemble.

Network	FERPlus	RAF-DB	CAER-S
Baseline [36]	87.03	86.56	85.63
Res2Net-18 [16]	87.85	87.73	86.85
SMResNet	88.73	88.46	88.52

To better explain the effect of the SMR module with the basic ResNet, Res2Net, we further conducted visualization of the proposed module through gradient-weighted class activation mapping (Grad-CAM) [50]. A comparison of Grad-CAM with different modules is shown in Figure 8. In Figure 8, SMR can pay attention to specific regions which are beneficial to facial expression recognition, even if there are occlusion and non-frontal pose issues in the facial images. For one thing, the SMR module added shallow geometric features to deep semantic information, for another, the range of receptive fields was wider since the multi-scale direction of SMR module is bidirectional. Hence, the constructed SMR module, which reduces the sensitivity of the deeper convolutions towards occlusion and variant pose, can obtain a more complete representation of the features.



Figure 8. Comparison of Grad-CAM with different network modules.

4.3.2. Comparison of Different Decision-Level Fusion Methods

To verify the superiority of the proposed decision-level fusion strategy, the WEF was compared with other decision-level fusion methods, as shown in Table 9. In Table 9, compared with the majority vote, WEF overcame the shortcomings of not reflecting the credibility of evidence, thus improving by 0.78%, 1.74%, and 2.38% on the FERPlus, RAF-DB, and CAER-S datasets, respectively. Compared with D-S evidence fusion, WEF suppresses conflicts between evidence and effectively solves the ambiguity and uncertainty of regional expressions, thus improving by 0.63%, 0.82%, and 1.29% on the FERPlus, RAF-DB and CAER-S datasets, respectively.

Table 9. Comparison of different decision-level fusion methods.

Fusion Method	FERPlus	RAF-DB	CAER-S
Majority vote	87.95	86.72	86.14
D-S evidence fusion	88.10	87.64	87.23
WEF	88.73	88.46	88.52

Specially, to illustrate the effect of the proposed WEF strategy, the disgust expression of RAF-DB dataset is chosen as an example. The BPAs and the modified BPAs are shown in Table 10. Table 10 illustrates that the calculated conflict coefficient between two pieces of evidence of the facial region and the eyes region is close to 1, so in this case data fusion applying D-S theory is not feasible. The results obtained by different regions are inconsistent, but the validity of the WEF in recognizing facial expressions can be seen from the fusion results.

Table 10. Results of data fusion using the WEF.

Expression Image	Region	Basic Probability Assignment								Recognition Result
		$m(c_1)$	$m(c_2)$	$m(c_3)$	$m(c_4)$	$m(c_5)$	$m(c_6)$	$m(c_7)$	$m(\Theta)$	
Disgust	Facial	0.1562	0.0004	0.0001	0.0011	0.8381	0.0041	0	0	Anger
	Eyes	0.0082	0.0004	0.1417	0.8483	0.0004	0.0010	0	0	Disgust
	Mouth	0.0039	0.0005	0.4411	0.5505	0.0002	0.0038	0	0	Disgust
		Modified Basic Probability Assignment								
		$m'(c_1)$	$m'(c_2)$	$m'(c_3)$	$m'(c_4)$	$m'(c_5)$	$m'(c_6)$	$m'(c_7)$	$m'(\Theta)$	
Disgust	Facial	0.0804	0.0002	0.0001	0.0006	0.4313	0.0021	0	0.4853	uncertain
	Eyes	0.0082	0.0004	0.1417	0.8483	0.0004	0.0010	0	0	Disgust
	Mouth	0.0039	0.0005	0.4389	0.5477	0.0002	0.0038	0	0.0050	Disgust
	WEF	0.0001	0	0.1123	0.8874	0.0001	0.0001	0	0	Disgust

$c_1, c_2, c_3, c_4, c_5, c_6$ and c_7 represent seven different expressions of happiness, sadness, surprise, disgust, anger, fear and neutral, respectively.

4.4. Comparison with State-of-the-Art Methods

To illustrate the performance of FER, we compared the proposed method with the state-of-art methods on the three datasets described above. The experimental results are shown in Tables 11–13.

Table 11. Comparison with state-of-the-art methods on FERPlus dataset.

Method	Years	Recognition Rate (%)
CSLD [47]	2016	85.10
gACNN [12]	2019	84.86
RAN [13]	2020	88.55
SCN [51]	2020	88.01
Ours	2023	88.73

Table 12. Comparison with state-of-the-art methods on RAF-DB dataset.

Method	Years	Recognition Rate (%)
IPA2LT [52]	2018	86.77
Separate Loss [53]	2019	86.38
gACNN [12]	2019	85.07
RAN [13]	2020	86.90
LDL-ALSG [54]	2020	85.53
CVT [7]	2021	88.14
Ours	2023	88.46

Table 13. Comparison with state-of-the-art methods on CAER-S dataset.

Method	Years	Recognition Rate (%)
ResNet-18 [35]	2016	84.67
ResNet-50 [35]	2016	84.81
CAER-Net-S [49]	2019	73.51
Res2Net-18 [16]	2021	85.01
Res2Net-50 [16]	2021	85.35
MA-Net [14]	2021	88.42
Ours	2023	88.52

Table 11 shows the comparison of the proposed method with several state-of-the-art methods on FERPlus dataset. The results in Table 11 show that the proposed SMResNet ensemble reaches 88.73%, which achieves a higher recognition rate and outperforms other recent state-of-the-art methods.

We also performed a comparison with advanced methods on RAF-DB dataset, as shown in Table 12. Consistent with other methods, we verified the effectiveness of SMResNet ensemble by recognizing seven basic expressions. The results in Table 12 show that the proposed method achieves recognition rate of 88.46% on RAF-DB, and obtained a higher recognition rate than the other six methods.

The FER recognition results of our proposed method and some state-of-the-art methods on CAER-S dataset are shown in Table 13. Since CAER-S dataset was released recently, only [49] has evaluated method on it. This paper has conducted several experiments using some state-of-the-art networks on it (such as ResNet-18, ResNet-50, Res2Net-18 and Res2Net-50). From Table 13, we can see that the proposed method achieves higher recognition rate even compared with the deeper network such as Res2Net-50.

4.5. Summary of Experiment

The effectiveness of the proposed method was verified on the FERPlus, RAF-DB, and CAER-S datasets. The experimental results demonstrate that the FER rates achieved 88.73%, 88.46%, and 88.52% on three datasets, respectively. In particular, the confusion matrix of the seven facial expressions, the FER results of different face regions, and the network ensemble were discussed. The comparative results indicated the proposed network ensemble not only had a high recognition rate for expressions such as happiness and surprise with rich details, but also maintained good recognition results for expressions such as sadness and disgust with insignificant local features and mutual confusion. This also showed that the WEF strategy based on the support degree can improve the credibility of regional evidence, thus enhancing the inter-class discrimination of different facial expression categories. Subsequently, the ablation experiments, which introduce the ensemble of different network structures, the visualization of different network modules, and the different decision-level fusion strategies, indicated that the proposed ensemble framework can focus the decision-level semantics of key regions and address the whole face for the absence of regional semantics under occlusions and posture variations. Particularly, the modified BPAs of the disgust expression showed the WEF strategy effectively solves the ambiguity and uncertainty of regional expressions, thus further boosting the effectiveness of FER. Finally, comparisons with state-of-the-art methods were made on the three databases. The experimental results illustrated that the proposed scheme improved the performance of FER in the wild.

5. Conclusions

To highlight the role of different facial regions with different receptive fields and implement the decision-level fusion of the network ensemble, the FER method based on the SMResNet ensemble with WEF strategy was proposed. Firstly, the pipeline for face alignment and facial key region location was designed to suppress redundant information of the input image and improve the anti-interference ability of facial regions. Secondly, to extract

features with a wider range of different receptive fields, the SMR block, which learned symmetrically the multi-scale features from both directions, was improved. Meanwhile, the SMResNet, which consisted of one 2D convolution layer, four basic blocks, and one SMR module composed of four cascaded SMR blocks, a GAP layer, and a FC layer, was designed. Instead of directly feeding global facial semantics of the facial ROI region into a deep neural network, the SMResNet ensemble, composed of three SMResNets, was constructed to extract decision-level semantic features of the whole face, eye, and mouth regions. Finally, the decision-level semantics extracted from the three regions were regarded as three pieces of evidence to realize decision fusion. Meanwhile, the WEF strategy, which overcame the conflicts between evidence by the support degree adjustment, was proposed to restrain the ambiguity and uncertainty of regional semantics based on D-S evidence theory.

The effectiveness of the proposed network ensemble was verified by experiments on the FERPlus, RAF-DB, and CAER-S datasets. The impact of different network structures and decision-level fusion strategies were discussed. The experimental results demonstrated that the FER rates achieved 88.73%, 88.46%, and 88.52% on three datasets, respectively. Compared with other state-of-the-art methods on three datasets, the proposed ensemble framework had a higher recognition rate, and improved the performance of facial expression recognition in the wild. The ablation experiments and visualization results based on Grad-CAM indicate that the proposed ensemble framework not only focused the decision-level semantics of key regions, but also addressed the whole face for the absence of regional semantics under occlusions and posture variations.

The advantages of our proposed method are as follows:

- (1) The pipeline for face alignment and facial key region location was designed. The designed pipeline can suppress redundant information in the input image and improve the anti-interference ability of facial regions.
- (2) The SMR block, which symmetrically constructs a hierarchical residual-like connection from two directions in a basic block, was improved. The proposed block represents the multi-scale feature of fine-grained level and can further expand the receptive field range of each network layer.
- (3) The SMResNet ensemble, which is composed of three SMResNets, is constructed for decision-level semantics extraction of the whole face, eye, and mouth regions. The ensemble framework not only focuses on the decision-level semantics of key regions, but also addresses the whole face, thus perceiving the semantics transmitted by incomplete faces under occlusions and posture variations.
- (4) The WEF strategy for decision-level semantic fusion was proposed based on D-S evidence theory. The proposed strategy minimizes the influence of evidence with a small weight on the decision-making judgment, and reduces conflicting information between evidence to satisfy the decision-level fusion of three regional pieces of evidence.

Although the proposed network ensemble focused on comprehensive information by cropping facial and two key regions to achieve a good recognition rate, the current network ensemble, which extracts facial features with different receptive fields and fuses regional decision-level semantics with WEF strategy, was mainly addressed to the image-based FER. With the emergence of multimedia technology, exploring multi-modal compensation mechanisms for robustness in the interference environment and improving SMResNet with contextual temporal attention for video-based FER is our next step.

Author Contributions: Conceptualization, J.L., M.H., Y.W. and Z.H.; methodology and investigation, J.L., Y.W., Z.H. and J.J.; resources, J.L., Y.W. and M.H.; writing—original draft preparation, J.L., M.H., Y.W. and Z.H.; writing—review and editing, J.L., M.H., Z.H. and J.J.; supervision, M.H. and J.J.; project administration, J.L., Y.W. and M.H.; funding acquisition, J.L., M.H., Z.H. and J.J. 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 under grant 62176084, the Natural Science Foundation of Anhui Province of China under Grant 1908085MF195

and 2008085MF193, and the Natural Science Research Project of the Education Department of Anhui Province under Grant 2022AH051038 and KJ2020A0508.

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China under grant 62176084, the Natural Science Foundation of Anhui Province of China under Grant 1908085MF195 and 2008085MF193, and the Natural Science Research Project of the Education Department of Anhui Province under Grant 2022AH051038 and KJ2020A0508. We acknowledge the use of the facilities and equipment provided by the Hefei University of Technology. We would like to thank every party stated above for providing help and assistance in this research.

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

Abbreviations

A list of all acronyms:

Basic probability assignment	BPA
Convolutional neural networks	CNNs
Deep convolutional neural network	DCNN
Dempster–Shafer	D-S
D-S combination rules	DCRs
Facial expression recognition	FER
Feature-level ensemble parallel network	FLEPNet
Full connection	FC
Global average pooling	GAP
Gradient-weighted class activation mapping	Grad-CAM
Histogram of oriented gradient	HOG
Local binary pattern	LBP
Principal component analysis	PCA
Region of interest	ROI
Regional attention network	RAN
Support vector machine	SVM
Symmetric multi-scale residual network	SMResNet
Weighted evidence fusion	WEF

References

1. Tong, X.Y.; Sun, S.L.; Fu, M.X. Adaptive Weight based on Overlapping Blocks Network for Facial Expression Recognition. *Image Vision Comput.* **2022**, *120*, 104399. [[CrossRef](#)]
2. Khan, S.; Chen, L.; Yan, H. Co-clustering to Reveal Salient Facial Features for Expression Recognition. *IEEE Trans. Affect. Comput.* **2020**, *11*, 348–360. [[CrossRef](#)]
3. Zhao, Y.; Xu, J. An Improved Micro-Expression Recognition Method Based on Necessary Morphological Patches. *Symmetry* **2019**, *11*, 497. [[CrossRef](#)]
4. Zhang, Z.Y.; Sun, X.; Li, J. MAN: Mining Ambiguity and Noise for Facial Expression Recognition in the Wild. *Pattern. Recognit. Lett.* **2022**, *164*, 23–29. [[CrossRef](#)]
5. Tang, Y.; Pan, Z.; Pedrycz, W.; Ren, F.; Song, X. Viewpoint-based Kernel Fuzzy Clustering with Weight Information Granules. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *7*, 342–356. [[CrossRef](#)]
6. Tang, Y.; Huang, J.; Pedrycz, W.; Li, B.; Ren, F. A Fuzzy Cluster Validity Index Induced by Triple Center Relation. *IEEE Trans. Cybern.* **2023**. [[CrossRef](#)]
7. Ma, F.; Sun, B.; Li, S. Robust Facial Expression Recognition with Convolutional Visual Transformers. *arXiv* **2021**, arXiv:2103.16854. [[CrossRef](#)]
8. Liu, C.; Hirota, K.; Dai, Y.P. Patch Attention Convolutional Vision Transformer for Facial Expression Recognition with Occlusion. *Inform. Sci.* **2023**, *619*, 781–794. [[CrossRef](#)]
9. Jiang, B.; Zhang, Q.W.; Li, Z.H.; Wu, Q.G.; Zhang, H.L. Non-frontal Facial Expression Recognition based on Salient Facial Patches. *EURASIP J. Image Video Process.* **2021**, *2021*, 15. [[CrossRef](#)]
10. Majumder, A.; Behera, L.; Subramanian, V.K. Automatic Facial Expression Recognition System using Deep Network-based Data Fusion. *IEEE Trans. Cybern.* **2018**, *48*, 103–114. [[CrossRef](#)]
11. Lopes, A.T.; Aguiar, E.D.; Souza, A.F.D.; Oliveira-Santos, T. Facial Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order. *Pattern. Recogn.* **2017**, *61*, 610–628. [[CrossRef](#)]

12. 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](#)]
13. Wang, K.; Peng, X.; Yang, J. Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4057–4069. [[CrossRef](#)]
14. 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*, 6544–6554. [[CrossRef](#)]
15. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 5–12 September 2014; pp. 818–833.
16. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Philip, T. Res2Net: A New Multi-scale Backbone Architecture. *IEEE Trans. Patten. Anal.* **2021**, *43*, 652–662. [[CrossRef](#)]
17. Lin, Z.; Liu, Q.S.; Peng, Y.; Liu, B.; Metaxas, D.N. Learning Active Facial Patches for Expression Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2562–2569. [[CrossRef](#)]
18. Yovel, G.; Duchaine, B. Specialized Face Perception Mechanisms Extract both Part and Spacing Information: Evidence from Developmental Prosopagnosia. *J. Cogn. Neurosci.* **2006**, *18*, 580–593. [[CrossRef](#)]
19. Pons, G.; Masip, D. Supervised Committee of Convolutional Neural Networks in Automated Facial Expression Analysis. *IEEE Trans. Affect. Comput.* **2018**, *9*, 343–350. [[CrossRef](#)]
20. Wen, G.; Zhi, H.; Li, H.; Li, D.; Xun, E. Ensemble of Deep Neural Networks with Probability-based Fusion for Facial Expression Recognition. *Cogn. Comput.* **2017**, *9*, 5155. [[CrossRef](#)]
21. Li, S.; Deng, W. Deep Facial Expression Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2022**, *13*, 1195–1215. [[CrossRef](#)]
22. Pan, Y.; Zhang, L.; Li, Z.; Ding, L. Improved Fuzzy Bayesian Network-Based Risk Analysis with Interval-Valued Fuzzy Sets and D–S Evidence Theory. *IEEE Trans. Fuzzy. Syst.* **2020**, *18*, 2063–2077. [[CrossRef](#)]
23. Liu, X.; Liu, S.; Xiang, J.; Sun, R. A Conflict Evidence Fusion Method based on the Composite Discount Factor and the Game Theory. *Inform. Fusion.* **2023**, *94*, 281–296. [[CrossRef](#)]
24. Turan, C.; Neergaard, K.D.; Lam, K.M. Facial Expressions of Comprehension (FEC). *IEEE Trans. Affect. Comput.* **2022**, *13*, 335–346. [[CrossRef](#)]
25. Saurav, S.; Saini, R.; Singh, S. Facial Expression Recognition using Dynamic Local Ternary Patterns with Kernel Extreme Learning Machine Classifier. *IEEE Access* **2021**, *9*, 120844–120868. [[CrossRef](#)]
26. Verma, K.; Khunteta, A. Facial Expression Recognition using Gabor Filter and Multi-layer Artificial Neural Network. In Proceedings of the International Conference on Information, Communication, Instrumentation and Control, Indore, India, 17–19 August 2017; pp. 1–5.
27. He, Y.; Chen, S. Person-Independent Facial Expression Recognition based on Improved Local Binary Pattern and Higher-Order Singular Value Decomposition. *IEEE Access* **2020**, *8*, 190184–190193. [[CrossRef](#)]
28. Wang, H.; Wei, S.; Fang, B. Facial Expression Recognition using Iterative Fusion of MO-HOG and Deep Features. *J. Supercomput.* **2020**, *76*, 3211–3221. [[CrossRef](#)]
29. Rouast, P.V.; Adam, M.; Chiong, R. Deep Learning for Human Affect Recognition: Insights and New Developments. *IEEE Trans. Affect. Comput.* **2021**, *12*, 524–543. [[CrossRef](#)]
30. Fan, Y.R.; Li, V.O.K.; Lam, J.C.K. Facial Expression Recognition with Deeply-Supervised Attention Network. *IEEE Trans. Affect. Comput.* **2022**, *13*, 1057–1071. [[CrossRef](#)]
31. Vasudha; Kakkar, D. Facial Expression Recognition with LDPP & LTP using Deep Belief Network. In Proceedings of the International Conference on Signal Processing and Integrated Networks, Noida, India, 22–23 February 2018; pp. 503–508. [[CrossRef](#)]
32. Chen, L.; Su, W.; Wu, M.; Pedrycz, W.; Hirota, K. A Fuzzy Deep Neural Network with Sparse Autoencoder for Emotional Intention Understanding in Human–Robot Interaction. *IEEE Trans. Fuzzy Syst.* **2020**, *28*, 1252–1264. [[CrossRef](#)]
33. Lee, J.; Kim, S.; Kim, S.; Sohn, K. Multi-Modal Recurrent Attention Networks for Facial Expression Recognition. *IEEE Trans. Image Process.* **2020**, *29*, 6977–6991. [[CrossRef](#)]
34. Hajarolasvadi, N.; Ramírez, M.A.; Beccaro, W.; Demirel, H. Generative Adversarial Networks in Human Emotion Synthesis: A Review. *IEEE Access* **2020**, *8*, 218499–218529. [[CrossRef](#)]
35. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–14. [[CrossRef](#)]
36. 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, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. [[CrossRef](#)]
37. Choi, J.Y.; Lee, B. Combining Deep Convolutional Neural Networks with Stochastic Ensemble Weight Optimization for Facial Expression Recognition in the Wild. *IEEE Trans. Multimed.* **2023**, *25*, 100–111. [[CrossRef](#)]
38. Karnati, M.; Seal, A.; Yazidi, A.; Krejcar, O. FLEPNet: Feature Level Ensemble Parallel Network for Facial Expression Recognition. *IEEE Trans. Affect. Comput.* **2022**, *13*, 2058–2070. [[CrossRef](#)]
39. 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](#)]
40. Baumgartner, S.; Huemer, M.; Lunglmayr, M. Efficient Majority Voting in Digital Hardware. *IEEE Trans. Circuits-II: Express Briefs* **2022**, *69*, 2266–2270. [[CrossRef](#)]

41. Zhao, G.; Chen, A.; Lu, G.; Liu, W. Data Fusion Algorithm based on Fuzzy Sets and D-S Theory of Evidence. *Tsinghua Sci. Technol.* **2020**, *25*, 12–19. [[CrossRef](#)]
42. Gao, S.; Deng, Y. An Evidential Evaluation of Nuclear Safeguards. *Int. J. Distrib. Sens. Netw.* **2019**, *15*, 12–19. [[CrossRef](#)]
43. Xiao, F. Multi-sensor Data Fusion based on the Belief Divergence Measure of Evidences and the Belief Entropy. *Inform. Fusion.* **2019**, *46*, 23–32. [[CrossRef](#)]
44. Jiang, W.; Wang, S. An Uncertainty Measure for Interval-Valued Evidences. *Int. J. Comput. Commun.* **2017**, *12*, 631–644. [[CrossRef](#)]
45. Martínez, B.; Valstar, M.F.; Jiang, B.; Pantic, M. Automatic Analysis of Facial Actions: A Survey. *IEEE Trans. Affect. Comput.* **2019**, *10*, 325–347. [[CrossRef](#)]
46. Zhang, J.; Kan, M.; Shan, S. Occlusion-free Face Alignment: Deep Regression Networks Coupled with De-corrupt Autoencoders. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3428–3437. [[CrossRef](#)]
47. Barsoum, E.; Zhang, C.; Ferrer, C. Training Deep Networks for Facial Expression Recognition with Crowd-sourced Label Distribution. In Proceedings of the International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 279–283. [[CrossRef](#)]
48. 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](#)]
49. Lee, J.; Kim, S.; Park, J. Context-aware Emotion Recognition Networks. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10143–10152. [[CrossRef](#)]
50. Selvaraju, R.R.; Cogswell, M.; Das, A. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2983–2991. [[CrossRef](#)]
51. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing Uncertainties for Large-scale Facial Expression Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, DC, USA, 13–19 June 2020; pp. 6897–6906. [[CrossRef](#)]
52. Zeng, J.; Shan, S.; Chen, X. Facial Expression Recognition with Inconsistently Annotated Datasets. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 222–237. [[CrossRef](#)]
53. Li, Y.; Lu, Y.; Li, J.; Lu, G. Separate Loss for Basic and Compound Facial Expression Recognition in the Wild. In Proceedings of the Asian Conference on Machine Learning, Nagoya, Japan, 17–19 November 2019; pp. 897–911.
54. 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 Conference on Computer Vision and Pattern Recognition, Seattle, DC, USA, 13–19 June 2020; pp. 13984–13993. [[CrossRef](#)]

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.