electronics

MDPI

Article

Three-Stream Convolutional Neural Network with Squeeze-and-Excitation Block for Near-Infrared Facial Expression Recognition

Ying Chen [1,2], Zhihao Zhang [1,2], Lei Zhong [1,2], Tong Chen [1,2,3,*], Juxiang Chen [1,2] and Yeda Yu [1,2]

1 Chongqing Key Laboratory of Nonlinear Circuit and Intelligent Information Processing, Southwest University, Chongqing 400715, China; chenyingly@email.swu.edu.cn (Y.C.); zzh085517@email.swu.edu.cn (Z.Z.); zl030610@email.swu.edu.cn (L.Z.); chenjuxiang@email.swu.edu.cn (J.C.); devil510@email.swu.edu.cn (Y.Y.)

2 Chongqing Key Laboratory of Artificial Intelligence and Service Robot Control Technology, Chongqing Institute of Green and Intelligent Technology, CAS, Chongqing 400715, China

3 Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China

* Correspondence: c_tong@swu.edu.cn; Tel.: +86-236-825-039

Received: 22 February 2019; Accepted: 26 March 2019; Published: 29 March 2019

Abstract: Near-infrared (NIR) facial expression recognition is resistant to illumination change. In this paper, we propose a three-stream three-dimensional convolution neural network with a squeeze-and-excitation (SE) block for NIR facial expression recognition. We fed each stream with different local regions, namely the eyes, nose, and mouth. By using an SE block, the network automatically allocated weights to different local features to further improve recognition accuracy. The experimental results on the Oulu-CASIA NIR facial expression database showed that the proposed method has a higher recognition rate than some state-of-the-art algorithms.

Keywords: NIR facial expression recognition; SE block; 3D CNN; adaptive feature weights calibration

1. Introduction

Facial expressions carry rich non-verbal information. Machines with the ability to understand facial expressions can better serve humans and fundamentally change the relationship between humans and machines. Therefore, automatic facial expression recognition has attracted attention from many fields, such as virtual reality [1,2], public security [3,4], and data-driven animation [5,6].

The effectiveness of facial expression recognition can be easily affected by environmental changes, such as changes of light, angle, and distance. Among these, the change of illumination conditions under visible light (VIS) (380–750 nm) has the largest influence [7,8]. To overcome this influence, an active near-infrared (NIR) illumination source (780–1100 nm) is used for the recognition. In this study, an NIR camera, together with the NIR illumination sources, were placed in front of the subjects. The intensity of the NIR illumination source was much higher than that of the ambient NIR light in indoor environments. Therefore, the ambient illumination problem could be solved as long as the active NIR illumination source is constant. The NIR recognition system is resistant to ambient illumination variations, and has been successfully applied to the field of face recognition [9]; it can perform well even in dark environments [10], in which normal imaging systems fail to perform recognition.

Facial expressions manifest themselves as movements of one or several discrete parts of the face, such as tightening the lips to express anger and raising the mouth to express happiness [11]. Some researchers use the features extracted from the entire face, which are called global features [12,13], for recognition, while other researchers use features extracted from specific parts, which are called local

features [14–17]. Many researchers have demonstrated that local features improve the performance of facial expression recognition compared with global features [18,19]. The main reason for this advancement is that the specific local regions contribute more accurate information of facial changes that help to distinguish the expressions, while the global region contains more identity information. Some researchers [20,21] have pointed out that the eyes, eyebrows, and mouth are the most expressive facial parts. However, it is unknown which part of the face should carry more weight in expression recognition or how the correct weight can be allocated to different parts of the face.

In earlier studies, many facial expression recognition systems used static images [22–24] that only contain spatial information as the input. However, facial expression can be a dynamic process, and the dynamic information of the face can better reflect the change of expression. Therefore, it is necessary to extract spatial and temporal information from the image sequences to facilitate recognition.

In the work reported in this paper, we designed a convolutional neural network (CNN) to complete NIR facial expression recognition. The CNN used is a three-stream three-dimensional (3D) CNN, which can learn spatio-temporal information from image sequences. In addition, the three inputs to the CNN are all local features, which not only reduce computational complexity, but also remove information not related to the expressions (such as identity information). A squeeze-and-excitation (SE) block is appended after the 3D CNN, which can automatically assign more weight to the local features that carry more expression information. To overcome the over-fitting problem caused by small data, features are extracted through three identical shallow networks. Finally, we add a global face stream to the local network, further increasing the recognition rate.

The main contributions of this paper are the following: (1) Three local regions of the face are used as the input of the network for the NIR expression recognition, which can not only accurately extract the facial expression information, but also reduce the computational complexity and dimensions; and (2) an SE block is added to model the dependencies between feature channels and adaptively learn the weight of the channel to gain efficient expression information and attenuate the useless information.

2. Related Work

Facial expressions can be decomposed into movement of one or more discrete facial action units (AUs). Inspired by this theory, Liu et al. [25] located common patches and unique patches of different expressions for recognition. However, this method could cause overlapping of located areas. Liu et al. [26] did further work and proposed a framework called FDM to select the active features of each expression without overlapping. Later, Liu et al. [27] proposed a 3D CNN with deformable action part constraints that can locate and code action units.

To extract temporal features while acquiring spatial features, Ji et al. [28] extended a CNN to a 3D CNN, which can extract the spatio-temporal information from image sequences. Szegedy et al. [29] utilized the 3D CNN to extract temporal information for video-based expression recognition. Chen et al. [30] proposed a new descriptor, the histogram of oriented gradients from three orthogonal planes (HOG-TOP), to extract the dynamic texture features from image sequences, which are fused with the geometric features to identify expressions. Fonnegra et al. [31] proposed a deep learning model and Yan et al. [32] presented collaborative-discriminative-multi-metric-learning (CDMML)-based image sequences for emotion recognition. To make the system more precise, Zia et al. [33] proposed a dynamic weight majority voting mechanism for the construction of ensemble systems. However, since these methods are all based on visible light, the impact of external illumination changes are not considered.

The NIR facial images/videos are hardly influenced by the ambient visible light change. Farokhi et al. [34] proposed a method of extracting global and local features by using Zernike moments (ZMs) and Hermite kernels (HKs), respectively, and then used the fused features to identify the NIR face. Taini et al. [35] assembled a near-infrared facial expression database and completed the first study based on NIR facial expression recognition. Zhao et al. [18] developed the database of NIR facial expressions, called the Oulu-CASIA NIR facial expression database, and used local binary patterns form three orthogonal planes (LBP-TOP) to extract dynamic local features. It was proved

in this work that NIR can overcome the influence of visible-light illumination changes on expression recognition. However, these methods must extract facial expression features manually. Jeni et al. [36] proposed a 3D-shape-information-based recognition technique and further proved that an NIR camera configuration is suitable for facial expressions under light-changing conditions. Wu et al. [37] proposed a three-stream 3D convolutional network for NIR facial expression recognition, using a combination of global and local features, but did not consider assigning different weights to local features.

3. Materials and Methods

3.1. 3D CNN

A 3D CNN is more suitable for spatial-temporal feature extraction. In [28], to process image sequences more efficiently, a 3D CNN approach is proposed to address action recognition problems. Through 3D convolution and pooling operations, a 3D CNN has the ability to learn temporal features.

A 3D CNN consists of an input layer, 3D convolution, 3D pooling (usually, each convolution layer is followed by the pooling layer), and a fully connected (FC) layer. The dimension of the input image sequences to the 3D CNN is represented as $d \times l \times h \times w$, where d is the number of the channels, l the number of frames of video clips, and h and w the height and width, respectively, of each frame. In addition, 3D convolution and pooling have a kernel size in $t \times k \times k$, where t is the temporal depth and k the spatial size.

3.2. Squeeze-and-Excitation Networks (SENets)

Hu et al. [38] proposed squeeze-and-excitation networks (SENets). The basic architectural unit of SENets is the SE building block, which is shown in Figure 1.

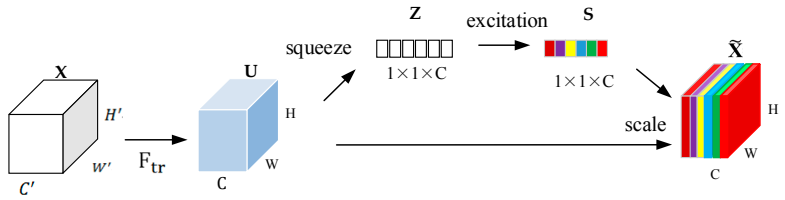

Figure 1. Squeeze-and-excitation (SE) block structure.

Before the SE block operation, input data X is transformed into features U through a series of convolution operations, i.e., $F_{tr} : X \rightarrow U$, $X \in R^{W' \times H' \times C'}$, $U \in R^{W \times H \times C}$, where F_{tr} represents the transformation from X to U, H (H') and W (W') are the frame height and width, respectively, and C (C') are the number channels.

The SE block mainly consists of two operations: Squeeze and excitation. Because the filter learned by each channel in the CNN operates on the local receptive field, each feature map in U cannot utilize the context information of other feature maps. The purpose of the squeeze operation is to have a global receptive field, so that the lower layers of the network can also use global information. The global average pooling operation is used to compress U (multiple feature maps) into Z, so that the C feature maps eventually become real columns of $1 \times 1 \times C$. The squeeze operation is performed by

$$z_m = F_{sq}(u_m) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} u_{m(i,j)} \tag{1}$$

where z_m represents the mth element of Z and u_m the mth element of U.

The excitation operation is a simple gating with a sigmoid activation. The purpose of this operation is to model the interdependence between feature channels by learning parameters to generate the weight of each feature channel. To meet these requirements and limit the model complexity and auxiliary generalization, two FC layers (1*1 conv layer) were introduced. One is the dimension

reduction layer, in which the parameter is W_1 and the dimension reduction ratio r; the other is a dimension increase layer with parameter W_2 followed by a Rectified linear unit (ReLU), $W_1 \in R^{\frac{C}{r} \times C}$ and $W_2 \in R^{C \times \frac{C}{r}}$. The excitation is performed by:

$$S = F_{ex}(Z, W) = \sigma(g(Z, W)) = \sigma(W_2\, \delta(W_1, Z)) \tag{2}$$

where S is the vector after excitation operation, and δ and σ refer to the ReLU function and the sigmoid function, respectively.

Finally, S is combined with U to obtain the final output by:

$$\tilde{x}_m = F_{scale}(u_m, s_m) = s_m \cdot u_m \tag{3}$$

where s_m is the mth element of S and $\tilde{x}_m$ the mth element of the final output $\tilde{X}$; F_{scale} refers to channel-wise multiplication.

The goal of the SE block is to greatly improve the expressiveness of the network; it adaptively recalibrates the feature weight by modeling the interdependencies between the channels. In more detail, it allows the network to use global information to selectively enhance the beneficial features of the channel and suppress the useless function channels.

3.3. Proposed System

In this paper, we propose a three-stream 3D CNN with an SE block called an SE three-stream fusion network (SETFNet). We took three local regions, the eyes (including eyebrows), nose, and mouth, from the facial expression image sequence as inputs to the three-stream network. After fusions of the three streams, an SE block was added to the network to adaptively learn the weight of each feature channel.

To avoid over-fitting problems, a deep CNN requires large amounts of data for training. However, the available database for NIR expression is small in size. To train a CNN model on a small database, researchers use a medium-size CNN [39,40]. Therefore, the SETFNet in this paper was also a medium-size CNN with four convolutional layers.

The structure of the proposed SETFNet is shown in Figure 2. It is a three-stream 3D CNN consisting of three identical sub-networks. Each sub-network consists of four convolutional layers and has the same parameters. The number of convolution kernels for the four convolution layers, first through fourth, is 16, 32, 64, and 128, respectively. The kernel size of the first convolution layer is 3×3×8, and a large temporal stride here is used to eliminate some useless information. The kernel size of the other three convolution layers is 3×3×3. The three streams were fused and followed by an SE block to recalibrate the weight of each stream. The details of each subnetwork are shown in Table 1.

Table 1. Configuration of each stream.

Layers	Kernel Parameter Settings	Number of Kernels	Output Size
Date			32 × 36 × 64
Conv	3 × 3 × 8	16	9 × 18 × 32
Pool1	2 × 2 × 1	16	9 × 18 × 32
Conv2	3 × 3 × 3	32	9 × 9 × 16
Pool2	2 × 2 × 2	32	8 × 8 × 15
Conv3	3 × 3 × 3	64	8 × 8 × 15
Pool3	2 × 2 × 2	64	4 × 4 × 8
Conv4	3 × 3 × 3	128	2 × 4 × 8
Pool4	2 × 2 × 1	128	2 × 2 × 4

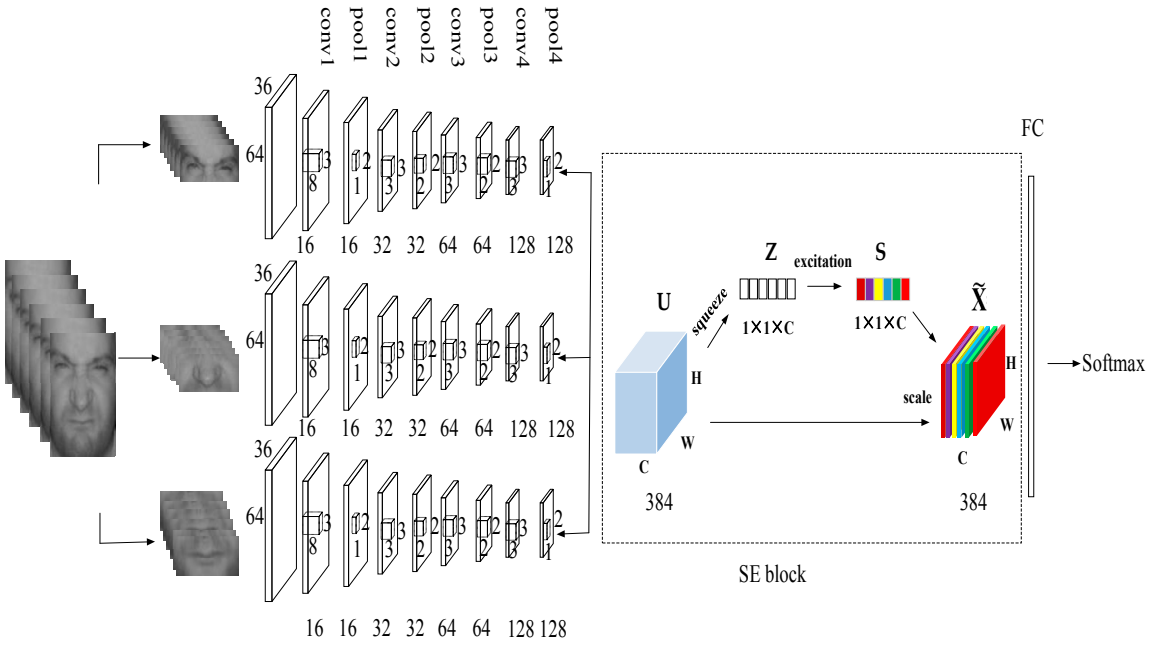

Figure 2. Overall structure of the proposed SE three-stream fusion network (SETFNet). The SE block is displayed in the dotted box.

Fusion Network

After extracting the features from the three regions (eyes, nose, and mouth), three stream features defined as T_1, T_2, and T_3 were obtained. The three stream features were then concatenated together to achieve better recognition by

$$T = T_1 \oplus T_2 \oplus T_3, \tag{4}$$

where T is the fused feature and $\oplus$ represents the concatenation operation. The concatenated features T were used as inputs to the next operation of the network.

3.4. Experiments

The proposed network was assessed on the Oulu-CASIA NIR facial expression database [18]. The network was implemented in the Caffe framework, which ran on a PC with a NVIDIA Geforce GTX 1080 graphical processing unit (GPU) (8 G). Training a model with the correct parameters is the key to achieving optimal performance, which has a direct impact on the experimental results. We trained the network from scratch using a batch size of 4, an initial learning rate of 10^{-3-3}, and a weight decay of 0.0005.

3.4.1. Database

Because the NIR facial expression database is not very common, the Oulu-CASIA NIR facial expression database is currently the only suitable one. It was collected in dark, weak, and normal light conditions, and consists of six kinds of facial expressions (anger, disgust, fear, happiness, sadness, and surprise) of 80 people between 23 and 58 years old, so each illumination condition has 480 image sequences. All expression sequences begin at the neutral emotion and end with the peak of the emotion. Each subject was asked to sit on a chair in the observation room in a way that they were in front of the camera. The distance between the face and camera was approximately 60 cm. Subjects made expressions according to the image sequences, while videos were captured by a USB 2.0 PC Camera (SN9C 201 & 202). Each clip was filmed by the camera at a frame rate of 25 fps. The image resolution was 320 × 240.

The aforementioned database has been used in many studies of facial expression recognition. It has been proved that the identification task under dark illumination conditions is the most difficult [18], because the facial image loses most of the texture features in dark light conditions. Therefore, we tested the proposed network on this most difficult sub-dataset (dark illumination condition).

We used the very popular method of tenfold cross-validation. All of the image sequences were divided into 10 groups. At each fold, nine groups were used to train the network and the rest were used for testing. During the entire experiment, there was no overlap between the training and testing sets.

3.4.2. Data Pre-Processing

In our experiment, a video sequence was pre-processed in the following three steps: (1) Frame-by-frame face detection; (2) locating eyes, nose, and mouth; and (3) cropping off the eyes, nose, and mouth areas. We found that step 2 had a significant effect on the performance of the network, so the choice of area to perform accurate spotting is crucial. To ensure that this was done accurately, the local areas were cropped based on the location of landmark points annotated by a robust landmark detector, discriminative response map fitting (DRMF) [41]. DRMF not only achieves good performance in landmark-detection methods [30], but also consumes very little computation time.

The cropping of these local areas was done by an automatic method. Since some of the cuts are inaccurate, manual cropping was used. Using the facial landmark points annotated earlier, the three regions were identified by using rectangular bounding boxes determined based on the eyes, nose, and mouth landmark points. We segmented the three local regions according to the following eleven points: E1 (x_1, y_1), E2 (x_2, y_2), E3 (x_3, y_3), E4 (x_4, y_4), E5 (x_5, y_5), N1 (x_6, y_6), N2 (x_7, y_7), M1 (x_8, y_8), M2 (x_9, y_9), M3 (x_{10}, y_{10}), and M4 (x_{11}, y_{11}) (shown in Figure 3). The center point of the rectangular bounding box of the eye region is L1 = E5 (x_5, y_5), and the length and width of the rectangle are $\frac{5}{3}|x_2 - x_1|$ and $\frac{4}{3}|y_4 - y_1|$, respectively. The center point of the rectangular bounding box of the nose region is L2 = $(x_5, \frac{y_7 - y_6}{2})$, and the length and width of the rectangle are $|y_7 - y_6|$ and $|x_3 - x_4|$, respectively. The center point of the rectangular bounding box of the mouth region is L3 = $(x_5, \frac{y_{11} - y_9}{2})$, and the length and width of the rectangle are $\frac{5}{3}|x_{10} - x_8|$ and $\frac{4}{3}|y_{11} - y_9|$, respectively.

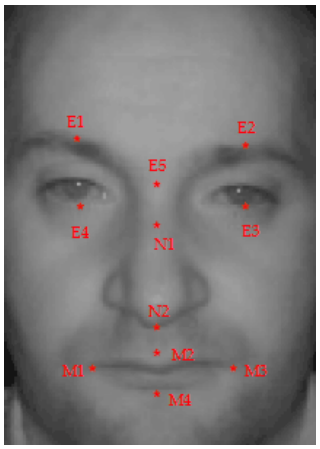

Figure 3. Positions of 11 points for segmenting three regions.

For the network input, each video sequence is normalized to 32 frames using the linear interpolation method [42]. Each frame of a global face (whole face) and local areas were resized to 88 × 108 and 36 × 64, respectively. To reduce the amount of calculation, all input images were converted to 8-bit grayscale.

4. Results and Discussion

4.1. Comparisons of Different Streams and Their Fusion

Table 2 shows the average results of tenfold cross-validation for each local region using a single sub-network (one stream) and a fused network. The feature information of the eye (including eyebrows), nose, and mouth regions is extracted by a single stream and the recognition rates are 35.37%, 42.76%, and 68.35%, respectively. The mouth region has the highest recognition rate, which may indicate that this part is the most expressive part in the database. The recognition rate of the eye region is the lowest among the three regions. This may be due to some of the participants wearing glasses. In the NIR face image, the NIR light reflected by the glasses removes the feature of the eyes, so that the frames with glasses have a great influence on recognition. At the same time, we can see that the performance of the recognition rate of the three-local-stream-fused networks (TFNets) reaches 78.68%, which is much higher than that of each single stream network (eye, 35.37%; nose, 42.76%; mouth, 68.35%). This indicates that our fusion is very effective in improving the recognition rate. After the network was fused, we added the SE block that automatically allocates weights to different streams. Since the SE block can make the entire network adaptively learn the weight of the feature channel, the SETFNet further improves the recognition rate, reaching a recognition rate of 80.34%.

Table 2. Comparison of different local and fused networks.

Architecture	Accuracy (%)	Time (s)
Eye	35.37	
Nose	42.76	0.515
Mouth	68.35	
TFNet	78.68	1.158
SETFNet	80.34	1.237
SETFNet + global	81.67	2.142

To investigate whether the SETFNet had extracted most of the expression features, we added one more stream to the SETFNet, which takes the frame of the global face as the input. Because each frame of the global face has larger spatial size than that of each local area, we added one more convolution pair to this added stream. The network structure is shown in Figure 4, with the fourth stream being the global face stream. When it is added to the SETFNet, the recognition rate becomes 81.67%. The SETFNet itself can achieve an 80.34% recognition rate. That is to say, after adding the entire face as input, the improvement of the recognition rate is still limited. This may indicate that the SETFNet has extracted most of the expression features.

Table 2 also shows the time consumption of various single sub-networks and fused networks. The time for a single sub-network to process an image sequence is 0.515 s, and the time for TFNet and SETFNet to process a sequence is 1.158 and 1.237 s, respectively. Considering the large improvement in recognition rate made by the TFNet and SETFNet, the increase of computation time is acceptable. However, when a global face stream is added to the SETFNet, the time for the network to process a sequence is 2.142 s. The slight increase in recognition rate (80.34% versus 81.67%) made by the global stream is at the expense of the processing time (1.237 s versus 2.142 s). However, all of the computation time may be within acceptable limits, since the input is 32 frames. Under the hardware settings used (NVIDIA Geforce GTX 1080 GPU (8G) for deep-learning acceleration), the SETFNet can process 32/1.237 = 25.87 frames every second. The frame rate of a normal imaging system is 25–30 fps, and 25.87 fps is within this range, which means that the SETFNet can give the recognition result just 1 s of lag in real-time imaging if the computation is performed in parallel with the imaging. With better hardware, the computation time can be further decreased to or to less than 1 s, which makes the processing a real-time process. Therefore, this network could be used in real applications.

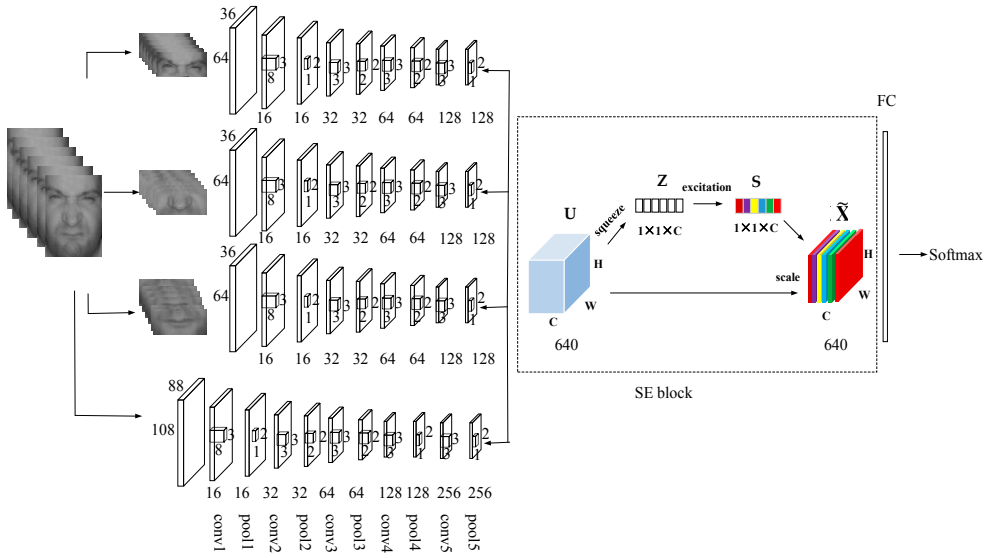

Figure 4. Structure of SETFNet plus global face stream.

The recognition rate of the eye region is the lowest among the three regions. One reason may be that the eyes have fewer features than the other parts; another reason could be that some of the subjects wear glasses. To verify the effect of glasses on the recognition rate, we input the eyes with and without glasses into the sub-network separately. The recognition results are shown in Table 3. It is seen that the recognition rate without glasses is better than that with glasses, which indicates that the glasses remove some features of the eyes. Since we divided the dataset into two parts, the recognition rates of wearing glasses and not wearing glasses are lower than that of the single sub-network with all data as the input.

Table 3. Comparison of recognition rate with and without glasses.

Category	Accuracy (%)
With glasses	30.13
Without glasses	31.45

4.2. Comparison of Embedded SE Block

The SE block was added to the network after the fusion so that the network could receive the information of the entire network and have a global receptive field. In the SE block, the reduction ratio r is an important parameter that can change the capacity and computational cost. We compared different reduction ratios r in our network model and the results are shown in the Table 4. When $r = 16$, the accuracy is the highest; therefore, r is set as 16.

Table 4. Comparison of different network reduction ratios.

Architecture		Accuracy (%)
SETFNet	r = 4	79.82
	r = 8	79.12
	r = 16	80.34
	r = 32	79.54
SETFNet + global	r = 4	80.57
	r = 8	81.25
	r = 16	81.67
	r = 32	80.38

4.3. Comparisons with Other Methods

Table 5 shows the different expression recognition rates of different methods on the Oulu-CASIA NIR facial expression database under dark-lighting conditions. For all of the methods, we used the tenfold cross-validation method to obtain an average recognition rate. The results of Deep Temporal Geometry Network (DTAGN), 3D CNN Deformable Facial Action Parts (DAP), and NIRExpNet were obtained from [37], and the result of LBP-TOP was obtained by implementing the algorithm using MatLab software (MathWorks, Natick, MA, USA). SETFNet and SETFNet + global were implemented by using Caffe. It is seen that LBP-TOP and 3D CNN DAP can achieve recognition rates of 69.32% and 72.12%, respectively, which are higher than that of DTAGN. NIRExpNet used the fusion information of local and global features, and therefore can achieve an even higher recognition rate than LBP-TOP and 3D CNN DAP. SETFNet uses only local information of three regions, but it can achieve a higher recognition rate (even higher than NIRExpNet, which uses local and global features). When a global face stream is added to SETFNet, it further improves the recognition rate to 81.67%. This indicates that the automatic allocation of the weight-of-features channel helps improve the recognition performance, which could be a promising method for NIR facial expression.

Table 5. Comparison of total recognition rates of different methods.

Method	Accuracy (%)
LBP-TOP [18]	69.32
DTAGN [43]	66.67
3D CNN DAP [27]	72.12
NIRExpNet [37]	78.42
SETFNet	80.34
SETFNet + global	81.67

4.4. Confusion Matrixes

To analyze the experimental results further, the confusion matrixes of SETFNet and SETFNet + global are shown in Tables 6 and 7, respectively. The labels on the left-hand side represent actual classes and those at the bottom represent the predicted classes; each percentage value in the matrix was calculated by dividing the number of a predicted class to the number of the corresponding actual class. After adding the global stream, the recognition rate of each expression is increased by 1–2%. It can be seen from Tables 6 and 7 that whether or not the global face stream is added, both happiness and surprise have high recognition rates, while fear and disgust have relatively low rates. The latter low recognition rates may be due to the slight movement of AUs for fear and disgust, which makes it more difficult to distinguish them from other expressions. Moreover, disgust is confused with anger, fear, and sadness, and fear is confused with anger, disgust, happiness, and surprise, perhaps because their appearance and movements are similar to each other.

Table 6. Confusion matrix of SETFNet. Labels on left-hand side represent actual classes; those on bottom represent predicted classes.

An	**77.64%**	**12.27%**	**1.25%**	**0**	**8.84%**	**0**
Di	15.06%	**72.91%**	9.53%	0	2.50%	0
Fe	7.45%	6.31%	**68.53%**	1.25%	0	16.46%
Ha	0	0	6.64%	**93.36%**	0	0
Sa	12.25%	3.52%	0	2.89%	**81.34%**	0
Su	0	0	8.46%	3.25%	0	**88.29%**
	An	**Di**	**Fe**	**Ha**	**Sa**	**Su**

Table 7. Confusion matrix of SETFNet + global. Labels on left-hand side represent actual classes; those on bottom represent predicted classes.

An	**78.43%**	**11.86%**	0	0	9.71% ↑	0
Di	13.38%	74.67%	7.87%	0	**4.08%** ↑	0
Fe	9.54% ↑	5.58%	71.08%	0	0	13.83%
Ha	0	0	5.74%	94.26%	0	0
Sa	9.46%	**8.25%** ↑	0	0	82.29%	0
Su	0	0	3.38%	7.31% ↑	0	89.31%
	An	**Di**	**Fe**	**Ha**	**Sa**	**Su**

SETFNet + global takes the entire face as input. The more input features there are, in general, should increase the true prediction values (values on the diagonal of the confusion matrix) and decrease the false prediction values (the zero value will be unchanged). It is seen from Table 6 that SETFNet + global does increase all true prediction values. However, more input does not always decrease the false prediction values. We can see from Table 7 that increased false prediction values do exist, which are indicated by up-pointing arrows. As the database is small in size, the prediction values could vary due to noise. To ensure that the located false prediction values are increased only as a result of more input features, we located their paired false prediction values as well. Each false prediction value pair appears in the same color in Table 7; for example, 9.54% (fear predicted as anger) and 0% (anger predicted as fear) in green. Only when both paired values are increased can the two expressions be considered as confused with each other more in SETFNet + global.

Under this criterion, we can see that sadness tends to be more recognized as disgust (8.25% versus 3.52%), or disgust tends to be more recognized as sadness (4.08% versus 2.50%), if SETFNet + global is used. The reason for this might be that, in sadness and disgust expression situations, lower cheek areas have an up-and-down movement pattern due to the movement of AU15 or AU10 [44]. When SETFNet + global takes these similar movement patterns as input, sadness will be recognized as disgust more.

Tables 8–11 show the confusion matrix of the comparison algorithms, with the labels on the left-hand side representing actual classes and those at the bottom representing the predicted classes. The confusion matrix of NIRExpNet (Table 8) was adopted from [37] directly. The other matrixes were obtained by implementing the algorithms with MatLab code on the database (tenfold cross-validation). Happiness and surprise again have higher recognition rates than the others in all algorithms. Fear has the lowest average recognition rate, and disgust has a similar average recognition rate to that of anger and sadness. This trend is in accord with what SETFNet reveals.

Table 8. Confusion matrixes of NIRExpNet.

An	**71.01%**	**14.43%**	**0**	**0**	**14.56%**	**0**
Di	20.56%	79.44%	0	0	0	0
Fe	0	8.00%	62.44%	0	0	29.56%
Ha	0	0	0	96.01%	0	3.99%
Sa	10.44%	0	14.44%	0	75.12%	0
Su	0	0	9.41%	4.04%	0	86.55%
	An	**Di**	**Fe**	**Ha**	**Sa**	**Su**

Table 9. Confusion matrixes of 3D CNN DAP.

An	**69.82%**	**16.23%**	**8.68%**	**0**	**5.27%**	**0**
Di	14.54%	73.41%	8.47%	0	3.58%	0
Fe	7.34%	7.46%	60.21%	8.32%	0	16.67%
Ha	0	0	8.58%	83.23%	0	8.19%
Sa	13.45%	9.93%	12.32%	0	64.30%	0
Su	**4.51%**	0	11.49%	2.45%	0	81.55%
	An	**Di**	**Fe**	**Ha**	**Sa**	**Su**

Table 10. Confusion matrixes of DTAGN.

An	**69.25%**	**15.28%**	**2.35%**	**3.30%**	**9.82%**	**0**
Di	18.72%	70.32%	10.96%	0	0	0
Fe	5.42%	3.13%	59.32%	5.62%	3.05%	23.46%
Ha	0	7.66%	12.57%	71.13%	0	8.64%
Sa	15.62%	0	14.52%	0	60.21%	9.65%
Su	0	0	13.46%	15.42%	0	71.12%
	An	**Di**	**Fe**	**Ha**	**Sa**	**Su**

Table 11. Confusion matrixes of LBP-TOP.

An	**63.45%**	**16.52%**	**7.66%**	**0**	**12.37%**	**0**
Di	15.33%	58.36%	10.67%	3.26%	12.36%	0
Fe	7.46%	6.89%	64.31%	0	3.89%	17.45%
Ha	0	11.68%	7.89%	75.86%	0	4.57%
Sa	10.62%	8.77%	10.43%	0	70.18%	0
Su	0	0	9.39%	6.85%	0	83.76%
	An	**Di**	**Fe**	**Ha**	**Sa**	**Su**

To further analyze the discrimination ability of different methods, we counted the number of zero false prediction values in each matrix. This number indicates that two corresponding expressions are perfectly recognized by the method. It is observed that NIRExpNet has 20 zero false prediction values, much more than other methods. 3D CNN DAP, DTAGN, and LBP-TOP have a similar number of zero false prediction values (approximately 12). These results indicate that NIRExpNet has the best performance in distinguishing one expression from others. This could be because NIRExpNet is designed specifically for the dataset. The features extracted by NIRExpNet are balanced so the possibility of confusing one expression with others is small.

Some zero false prediction values do not have zero paired values, e.g., the values in red in Table 9. 4.51% of the surprise expression was recognized as anger, but 0% anger was recognized as surprise using 3D CNN DAP. This could be due to the noise of the small dataset.

The F1 score and Matthews correlation coefficient (MCC) are calculated using the confusion matrixes, which are indexes considering accuracy and recall of the classification results and are fairer methods for assessing a classifier. The F1 score and MCC are summarized in Table 12. It is observed that SETFNet and SETFNet + global have the highest F1 and MCC, NIRExpNet has the second-highest values, and 3D CNN DAP the third highest. LBP-TOP and DTAGN have the lowest F1 and MCC. This indicates that SETFNet outperforms other methods in even more rigorous assessment. The order of the F1 and MCC performance of the methods is in accord with accuracy performance. This also indicates that the number of each sub-category is well balanced.

Table 12. Comparison of F1 score and MCC of different methods.

Method	F1 Score	MCC
LBP-TOP [18]	0.6712	0.6343
DTAGN [43]	0.6949	0.6077
3D CNN DAP [27]	0.7235	0.6702
NIRExpNet [37]	0.7828	0.7416
SETFNet	0.8034	0.7648
SETFNet + global	0.8164	0.7806

4.5. Potential Application and Improvement

SETFNet, which used three regions of the face as the input, can achieve higher recognition rates than NIRExpNet, which used the entire face as input, because an SE block can automatically allocate the weights to different streams. These results suggest that the automatic allocation of weights to

different features will help improve the recognition rate. This idea of automatic allocation may have potential use in other recognition tasks. The SE block can always be added after a feature fusion step to allocate weights to different features to further improve the recognition rate.

SETFNet + global has a slightly higher recognition rate than SETFNet, but consumes much more calculation time. This indicates that a small part of the face could carry most of the expression information. For any other type of facial expression recognition task, we may only analyze the parts of face carrying expression information, which can save much calculation time and make recognition a real-time application.

The highest recognition rate on the Oulu-CASIA NIR facial expression database (dark condition) is 98.6%, achieved by Rivera et al. [45]. A number transitional graph method (DNG) was proposed in [45]. The confusion matrixes achieved by DNG method were summarized in Tables 13 and 14 (adopted from [45] directly), with the labels on the left-hand side representing actual classes and those at the bottom representing the predicted classes. Table 13 is the confusion matrix of DNG using 3D Sobel (DNG_S), and Table 14 is the confusion matrix of DNG using nine-plane mask (DNG_P). It is seen that the recognition rate of each expression class is more than 97% and similar to each other. This may indicate that the DNG has obtained good enough features to discriminate one expression from others. In terms of zero false prediction values, DNG_S has 21 zero false prediction values, and DNG_P has 23 zero false prediction values, which are less than all other methods. This indicates that the DNG method can achieve the most un-confused matrix. The F1 and MCC of DNG are higher than other methods, as well (DNG_S: F1 0.9859, MCC 0.9830; DNG_P: F1 0.9879, MCC 0.9856). This indicates that DNG outperforms other methods in more rigorous assessment.

Table 13. Confusion matrixes of DNG_S.

An	**98.75%**	**1.25%**	**0**	**0**	**0**	**0**
Di	2.53%	97.47%	0	0	0	0
Fe	0	0	97.81%	0.63%	1.25%	0.31%
Ha	0	0.63%	0	98.73%	0.63%	0
Sa	0	0	0	0.63	99.38%	0
Su	0	0	0.63%	0	0	99.38%
	An	**Di**	**Fe**	**Ha**	**Sa**	**Su**

Table 14. Confusion matrixes of DNG_p.

An	**100%**	**0**	**0**	**0**	**0**	**0**
Di	1.9%	96.2%	0	0	1.9%	0
Fe	0	0	99.38%	0	0.63%	0
Ha	0	0	0	98.73%	0.63%	0
Sa	0.63	0	0.63	0	98.75%	0
Su	0	0	0	0	0.63	99.38%
	An	**Di**	**Fe**	**Ha**	**Sa**	**Su**

DNG consists of designed feature-extraction and feature-fusion methods, which make the extracted features robust in uneven illumination conditions. This could be the reason why DNG can achieve the best performance. According to the design of the DNG, two aspects could be considered in the future design of the SETFNet. Firstly, the uneven illumination conditions in the database could be taken into account when designing the network, such as using the features extracted from DNG as a stream to the network. Secondly, a more sophisticated fusion method could be used in future design, e.g., the concatenation operation used in this paper could be replaced by the fusion method in DNG.

However, a different form of DNG using hand-crafted features, SETFNet, proposed in this paper extracts features automatically. This design does not need the background knowledge of the data. Specifically, The feature extraction in this paper was finished by using a 3D CNN. Since the dataset used for training the CNN is small in size, the proposed network is not deep enough and may not

extract high-level features. To further improve the recognition rate, transfer learning could be used, i.e., training a deeper CNN on a larger dataset and then fine-tuning the network on the NIR database.

5. Conclusions

In this paper, we proposed a three-stream 3D CNN architecture with an SE block called SETFNet that can automatically learn spatio-temporal features simultaneously. We only used three local regions of the face as input to the network. The advantages of using local information as input to the network were the removal of some information unrelated to recognition and a reduction of the amount of computation. To enable the network to adaptively learn the weight of each feature channel, an SE block was added to the network after the fusion of three single sub-networks. Experimental results show that SETFNet can achieve an average recognition rate of 80.34%; when a global face stream was added to SETFNet, the recognition rate was further increased to 81.67%, which is higher than some state-of-the-art methods.

Author Contributions: Data curation, L.Z., J.C., and Y.Y.; Formal analysis, Y.C.; Methodology, Z.Z.; Supervision, T.C.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 61301297, 61502398), and the Southwest University Undergraduate Science and Technology Innovation Fund (No.20600901).

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

References

1. Anderson, K.; McOwan, P.W. A real-time automated system for the recognition of human facial expressions. *IEEE Trans. Syst. Man Cybern. Part BCyben.* **2006**, *36*, 96–105. [CrossRef]
2. Ip, H.H.; Wong, S.W.; Chan, D.F.; Byrne, J.; Li, C.; Yuan, V.S.; Wong, J.Y. Enhance emotional and social adaptation skills for children with autism spectrum disorder: A virtual reality enabled approach. *Comput. Educ.* **2018**, *117*, 1–15. [CrossRef]
3. Tulyakov, S.; Slowe, T.; Zhang, Z. Facial expression biometrics using tracker displacement features. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 18–23 June 2007; pp. 1–5.
4. Quintero, L.A.M.; Muñoz-Delgado, J.; Sánchez-Ferrer, J.C.; Fresán, A.; Brüne, M.; Arango de Montis, I. Facial emotion recognition and empathy in employees at a juvenile detention center. *Int. J. Offender Ther. Comp. Criminol.* **2018**, *62*, 2430–2446. [CrossRef] [PubMed]
5. Zeng, Z.; Pantic, M.; Roisman, G.I.; Huang, T.S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 39–58. [CrossRef]
6. Bartlett, M.S.; Littlewort, G.; Fasel, I.; Movellan, J.R. Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. In Proceedings of the 2003 IEEE Conference on Computer Vision and Pattern Recognition Workshop, Madison, WI, USA, 16–22 June 2003; Volume 5, p. 53.
7. Zhang, Z.; Wang, Y.; Zhang, Z. Face synthesis from low-resolution near-infrared to high-resolution visual light spectrum based on tensor analysis. *Neurocomputing* **2014**, *140*, 146–154. [CrossRef]
8. Wang, S.; Liu, Z.; Lv, S.; Lv, Y.; Wu, G.; Peng, P.; Wang, X. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Trans. Multimedia.* **2010**, *12*, 682–691. [CrossRef]
9. Li, S.Z.; Chu, R.; Liao, S.; Zhang, L. Illumination invariant face recognition using near-infrared images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 627–639. [CrossRef]
10. Qiao, Y.; Lu, Y.; Feng, Y.S.; Li, F.; Ling, Y. A new method of NIR face recognition using kernel projection DCV and neural networks. In Proceedings of the 2013 Fifth International Symposium on Photoelectronic Detection and Imaging, Beijing, China, 25 June 2013; pp. 89071M1–89071M6.
11. Ekman, P.; Friesen, W.V. *Manual for the Facial Action Coding System*; Consulting Psychologists Press: Palo Alto, CA, USA, 1978.
12. Zeng, N.; Zhang, H.; Song, B.; Liu, W.; Li, Y.; Dobaie, A.M. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing* **2018**, *273*, 643–649. [CrossRef]

13. Tsai, H.H.; Chang, Y.C. Facial expression recognition using a combination of multiple facial features and support vector machine. *Soft Comput.* **2018**, *22*, 4389–4405. [CrossRef]
14. Gu, W.; Xiang, C.; Venkatesh, Y.V.; Huang, D.; Lin, H. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recognit.* **2012**, *45*, 80–91. [CrossRef]
15. Majumder, A.; Behera, L.; Subramanian, V.K. Automatic facial expression recognition system using deep network-based data fusion. *IEEE transactions on cybernetics.* **2018**, *48*, 103–114. [CrossRef] [PubMed]
16. Otberdout, N.; Kacem, A.; Daoudi, M.; Ballihi, L.; Berretti, S. Deep Covariance Descriptors for Facial Expression Recognition. *arXiv*, 2018; arXiv:1805.03869.
17. Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach.Intell.* **2007**, *29*, 915–928. [CrossRef]
18. Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; PietikäInen, M. Facial expression recognition from near-infrared videos. *Image Vis. Comput.* **2011**, *29*, 607–619. [CrossRef]
19. Ghimire, D.; Jeong, S.; Lee, J.; Park, S.H. Facial expression recognition based on local region specific features and support vector machines. *Multimed. Tools Appl.* **2017**, *76*, 7803–7821. [CrossRef]
20. Yan, W.J.; Wang, S.J.; Chen, Y.H.; Zhao, G.; Fu, X. Quantifying micro-expressions with constraint local model and local binary pattern. In Proceedings of the European Conference on Computer Vision workshop, Zurich, Switzerland, 6–12 September 2014; pp. 296–305.
21. Ringeval, F.; Schuller, B.; Valstar, M.; Jaiswal, S.; Marchi, E.; Lalanne, D.; Pantic, M. Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. ACM, Brisbane, Australia, 26 October 2015; pp. 3–8.
22. Corneanu, C.A.; Simón, M.O.; Cohn, J.F.; Guerrero, S.E. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Trans. Pattern Anal. Mach.Intell.* **2016**, *38*, 1548–1568. [CrossRef] [PubMed]
23. Shan, C.; Gong, S.; McOwan, P.W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vis. Comput.* **2009**, *27*, 803–816. [CrossRef]
24. Khan, S.A.; Hussain, A.; Usman, M. Reliable facial expression recognition for multi-scale images using weber local binary image based cosine transform features. *Multimed. Tools Appl.* **2018**, *77*, 1133–1165. [CrossRef]
25. Liu, M.; Shan, S.; Wang, R.; Chen, X. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1749–1756.
26. Liu, P.; Zhou, J.T.; Tsang, I.W.H.; Meng, Z.; Han, S.; Tong, Y. Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 151–166.
27. Liu, M.; Li, S.; Shan, S.; Wang, R.; Chen, X. Deeply learning deformable facial action parts model for dynamic expression analysis. In Proceedings of the 12th Asian Conference on Computer Vision, Singapore, 1–5 November 2014; pp. 143–157.
28. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [CrossRef] [PubMed]
29. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
30. Chen, J.; Chen, Z.; Chi, Z.; Fu, H. Facial expression recognition in video with multiple feature fusion. *IEEE Trans. Affect. Comput.* **2018**, *9*, 38–50. [CrossRef]
31. Fonnegra, R.D.; Díaz, G.M. Deep Learning Based Video Spatio-Temporal Modeling for Emotion Recognition. In Proceedings of the International Conference on Human-Computer Interaction, Las Vegas, NV, USA, 15–20 July 2018; pp. 397–408.
32. Yan, H. Collaborative discriminative multi-metric learning for facial expression recognition in video. *Pattern Recognit.* **2018**, *75*, 33–40. [CrossRef]
33. Zia, M.S.; Hussain, M.; Jaffar, M.A. A novel spontaneous facial expression recognition using dynamically weighted majority voting based ensemble classifier. *Multimed. Tools Appl.* **2018**, 1–31. [CrossRef]
34. Farokhi, S.; Sheikh, U.U.; Flusser, J. Near infrared face recognition using Zernike moments and Hermite kernels. *Inf. Sci.* **2015**, *316*, 234–245. [CrossRef]

35. Taini, M.; Zhao, G.; Li, S.Z. Facial expression recognition from near-infrared video sequences. In Proceedings of the 2008 IEEE International Conference on Pattern Recognition, Tampa, FL, USA, 18–21 December 2008; pp. 1–4.
36. Jeni, L.A.; Hideki, H.; Takashi, K. Robust Facial Expression Recognition Using Near Infrared Cameras. *JACIII* **2012**, *16*, 341–348. [CrossRef]
37. Wu, Z.; Chen, T.; Chen, Y.; Zhang, Z.; Liu, G. NIRExpNet: Three-Stream 3D Convolutional Neural Network for Near Infrared Facial Expression Recognition. *Appl. Sci.* **2017**, *7*, 1184. [CrossRef]
38. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
39. Peng, M.; Wang, C.; Chen, T.; Liu, G. Nirfacenet: A convolutional neural network for near-infrared face identification. *Information* **2016**, *7*, 61. [CrossRef]
40. Peng, M.; Wang, C.; Chen, T.; Liu, G.; Fu, X. Dual temporal scale convolutional neural network for micro-expression recognition. *Front. Psychol.* **2017**, *8*, 1745. [CrossRef] [PubMed]
41. Asthana, A.; Zafeiriou, S.; Cheng, S.; Pantic, M. Robust discriminative response map fitting with constrained local models. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3444–3451.
42. Smolic, A.; Muller, K.; Dix, K.; Merkle, P.; Kauff, P.; Wiegand, T. Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems. In Proceedings of the 15th IEEE International Conference on Image Processing, San Diego, CA, USA, 12–15 October 2008; pp. 2448–2451. [CrossRef]
43. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint fine-tuning in deep neural networks for facial expression recognition. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2983–2991.
44. Ekman, P.; Friesen, W.; Hager, J. Facial Action Coding System The Manual. Available online: https://www.paulekman.com/product/facs-manual/ (accessed on 10 March 2019).
45. Rivera, A.R.; Chae, O. Spatiotemporal directional number transitional graph for dynamic texture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *10*, 2146–2152. [CrossRef] [PubMed]