Hindawi
Scientific Programming
Volume 2022, Article ID 9471334, 10 pages
https://doi.org/10.1155/2022/9471334

Research Article

Improved Generative Adversarial Networks for Student Classroom Facial Expression Recognition

Chaoyi Wang

Aviation Institute, Jilin Communications Polytechnic, Changchun 130000, Jilin, China

Correspondence should be addressed to Chaoyi Wang; wchyi@jljy.edu.cn

Received 5 May 2022; Revised 5 June 2022; Accepted 28 June 2022; Published 16 July 2022

Academic Editor: Lianhui Li

To assess students' learning efficiency under different teaching modes, we used students' facial expressions in the classroom as a study point. An enhanced generative adversarial network is presented. We designed a generator as an automatic coding-decoding combination in a cascade structure with a discriminator configuration. It can retain different expression intensity features to the maximum extent. We also added a new auxiliary classifier, which can classify different intensity features and improve the model's recognition of detailed features of similar expressions, thus improving the comprehensive facial expression recognition accuracy. Our approach has a great advantage over the other facial expression recognition approaches on public datasets. Finally, we conduct experimental validation on the self-made student facial expression dataset in all cases. The experimental findings showed that our approach's recognition accuracy is superior to that of other methods, demonstrating the method's efficacy.

1. Introduction

In classroom teaching, what the teacher explains and what the students understand is not visually represented in the current assistive teaching systems. It is also a topic of debate which teaching style students would prefer between the traditional classroom teaching style and the modern smart classroom teaching style. The literature [1] then mentions that smart teaching and intelligent learning environments can give full play to students' cognitive abilities, greatly increase their interactivity, and provide better mastery of new knowledge. In terms of the current investigation, there is no intuitive system to measure students' acceptance of different teaching methods. For this reason, we will concentrate on this problem, we set out to identify facial expressions, and by obtaining the emotional expressions of the teacher and the facial expressions between students and then performing facial expression analysis, we can determine the students' acceptance and satisfaction with the teaching method. Our research, to some extent, provides some reference value for the quality of teaching and can respond to the effectiveness of teaching at the biotechnical level.

In human communication, facial expression is an important communication tool. It often adds different emotional factors to nonverbal communication, and it is crucial in the process of comprehending one another's emotional expression. With the advancement of biotechnology and computer science, facial expressions are used in various industries. The most common application area is privacy and security, which is most directly demonstrated by the face unlocking feature on cell phones and computers. Second, in the field of transportation, driver fatigue and drunken driving detection are also predicted by capturing facial expressions. Also, facial expression recognition technology is also frequently integrated into the fields of virtual reality, medical care, and service robotics [2–4]. Of course, the facial expression recognition technology is not so simple, and there are several technical difficulties to be broken. Different countries have different language and cultural backgrounds, and their meanings conveyed by facial expressions are more or less different. In addition, the results of facial expression recognition are not sufficient due to the objective influence of nonstructural conditions, such as occlusion, illumination, and focus problems. Recently, many researches have arisen in the field of facial recognition to address these technical challenges, but the technological breakthroughs are all relatively limited [5].

The process of recording real-life student emotions is known as facial expression recognition, and the inner feelings can be mapped side by side from the fluctuations of emotions. The process is mainly based on video dynamic frames and still image sequences as the main recognition subject, and based on face recognition, it rises to a level to synthesize the linkage reaction among five senses, thus predicting facial expressions. The literature [3] starts the study from the simplest basic facial expressions, mainly the expressions of joy and sadness series. The authors, in order to obtain facial expressions accurately, first remove the noise from the images by preprocessing operations, followed by face detection to delineate the range of facial expression features. Then feature fusion is performed jointly with the linkage between the eyes, eyebrows, mouth, and cheeks, and finally, and facial expressions are predicted by matching with the training feature library.

To address the difficulties in facial expression recognition research, related researchers have made unremitting efforts. Some researchers have focused their research on manual features. For example, literature [6] proposed the use of Gabor filters to optimize manual features, and literature [7] proposed local binary patterns to break the limitations of manual features. The literature [8] proposed a gradient histogram method to extract features, which further enriched the artificial feature set. Some researchers put their research focus on deep neural networks. For example, the literature [9] innovatively improved the network structure in the approach using neural networks, and the authors picked to fine-tune the two-stage training algorithm to adapt the feature linkage between the five senses and enhance the expression recognition. The literature [10] both adopted generative adversarial networks, which further explored the intrinsic features of the face and eliminated the interference of nonsubjective factors. The literature [11], on the other hand, performed adaptive optimization on the constraint function and proposed island loss to determine the attribution problem between features by learning the connection between different expressions. The literature [12] places the research focus on the attention mechanism and proposes an adaptive regional attention network and validates the high efficiency of the network on the available dataset, and results proved that integrating the learnt model can increase the model's robustness.

However, facial emotion detection is not a simple work, so the previously mentioned studies ignore the direct connection between facial attributes and emotions, and the main reason for the poor recognition results is the inability to positively map the way of distortion among the five facial nodes, and the changes between specific locations cannot be responded. Some researchers have proposed setting up standard lines on the face for facial node calibration, and the literature [13] also mentions that using this approach can decrease the data variance and improve the stability of the model. The literature [14] also proposed model-aware flags for the automatic perception of facial position, and experiments demonstrated that this method not only reduces the workload but also preserves the robustness of the model. In the literature [15], it was unexpectedly found in the experiments that additional flagging of facial positions by predetermined trajectories could increase the recognition speed of the model without affecting the accuracy. All the above methods take an end-to-end form, and such methods also have certain limitations. Its recognition effect is limited by the quality of facial markers, and when facial expression features are captured, they can easily be incorporated into shallow features in a nonmaximal suppression operation.

To counteract the drawbacks of deep learning approaches, the literature [16] used a multitask learning strategy in neural network construction to enhance the primary task by shifting the learning number of different tasks. In addition, the literature [17, 18] added facial detection flags in the feature design of the facial action structure unit, which can aid in improving facial emotion recognition accuracy. In terms of multitask parameters, most of the previous studies launched optimization based on hard parameter sharing, but this approach limited the recognition efficiency of facial expressions to some extent. Nowadays, more soft parameters have started to be developed for sharing, such as the multitask convolutional partial sharing strategy in the literature [19] and the cross-stitch network proposed in the literature [20], which successfully break the efficiency limitation.

In our study, we consider various models comprehensively. We finally choose a generative adversarial network as the base method. To obtain the intensity features of different expressions hierarchically, we added a new auxiliary classifier and optimized the network structure. Finally, the effectiveness of our approach is demonstrated on both public and self-made datasets.

The rest of the study is arranged as follows. Section 2 presents the work related to different facial expression recognition methods. Section 3 introduces our adaptive improvement strategy and implementation process for generative adversarial networks. Section 4 presents the comparison of experimental databases and experimental methods. Finally, Section 5 presents research prospects and improvement directions.

2. Related Work

Traditional facial expression recognition research mainly relies on extracting geometric features, texture features, and hybrid features of the face as the basis [21]. The active shape model is the mostly used in facial expression recognition work and is the geometric feature method, which mainly uses facial feature points as a reference to construct geometric features and then localizes them. In practical application, the method is affected by lighting and occlusion and does not achieve better recognition results. The facial action unit is also a typical example of the geometric feature method. This method first divides the face into units and then compares them with the facial reference points by calculating the relative distance between units. However, this method requires intensive training in advance and has a very high computational complexity at the time level [22]. Texture feature-based facial expression recognition methods are more common and usually have faster computational speed, but they are not effective for motion scenes such as Gabor filter and local orientation pattern methods. In the face of occlusion, the most effective method is the scale-invariant feature variation, which can automatically find the spatial extrema and extract their position, scale, and rotation invariants and can circumvent the effect of occlusion by local mapping, but this method is not effective for the target smoothed by edges.

In facial expression recognition work, the input video frames or image information are subjected to preprocessing operations and then input to convolutional layers of different scales for feature extraction, and then the facial features are transformed into independent vectors, and finally, the classification is completed by fully connected layers [23]. Different application scenarios have different structural requirements for convolutional neural networks [24], and to address the influence of nonstructural environmental factors, facial expression recognition work often requires specific preprocessing operations, such as the HOG feature method [25], the LBP method [26, 27], and the ROI method [28–30]. Different features have different extraction stages, resulting in multiple features in different dimensions, which cannot be unified at the time level and affect the convergence efficiency of neural networks. Besides, convolutional neural networks are often used by researchers as a basic network. According to different requirements for different tasks, convolutional neural networks are optimized and upgraded accordingly to the increase in the adaptability and performance of deep networks. Some researchers have designed cascade networks to enhance the efficiency of the localization of facial nodes [31]. Some researchers tried to add auxiliary modules to improve the robustness of the model [32]. Some researchers divided the network into parallel or tandem networks of small modules to achieve the inclusion of features at the decision level [33, 34]. All of the above research methods aim to improve the depth and parameter tuning of the network, which invariably increases the number of parameters. Considering the computational cost, some researchers have proposed recurrent neural networks [15], capsule networks [35], deep belief networks [36, 37], and so on.

For deep learning methods, the recognition accuracy is proportional to the volume of training data, and the richer the dataset, the higher the recognition accuracy. For facial emotion detection, building a database of facial expressions is undoubtedly a difficult and long-lasting task. The features of facial expressions are deeply related to different background cultures, and the process of data annotation usually requires the annotators to have a certain understanding of national culture and background. In addition, the optimization process of neural networks is often not transparent enough, and most researchers rely on constant repetition of experiments and experience to verify the optimal parameter sizes [38]. Therefore, the period and computational cost factors of the project need to be considered before adopting a deep learning approach. To circumvent complex parameter tuning strategies, the literature [39] proposed the multi-granularity cascade forest method, an integrated neural network structure inspired by the cascade forest classification rule and the random forest rule. Compared with pure deep learning methods, this method has a smaller number of parameters and sets hidden layer hyperparameters to reduce the computational cost.

3. Method

3.1. Pipeline Overview. Researchers usually take an unsupervised approach to train the adversarial model, which belongs to the same deep neural network model and is divided into two parts in the phased design of the network. The generator part belongs to the front-end of the network and the discriminator belongs to the back-end. The generative adversarial network principle is simulated training at the neural network level, where different samples are iterated and generated in a random mode. The original samples are input at the input side, and the generator generates pseudosamples based on the original samples, and the usability of the generated samples is judged by comparing the difference between the original samples and the generated samples within a specified threshold of the pseudosamples. If the generated sample does not meet the standard value, by iterating this method, the pseudosamples can be approximated to the eigenvalues of the true samples in terms of eigenvalues. The structure of the generative adversarial network is shown in Figure 1.

In our study, face recognition systems can be made more robust by combining facial expression recognition with adversarial generating networks. Generative adversarial networks essentially play the facial expression details against each other by repeatedly updating iterations until the best facial expression features are obtained and then output to the terminal. Considering the facial expression details feature refinement, we define the classification of facial expressions to prevent the problem of increasing errors with different expression strengths.

3.2. Generator. The generator is in the front part of the adversarial network and its input is the real sample. After the real samples are input, the generator parses the real samples, divides the real samples into different feature nodes, and finally simulates the feature nodes to generate pseudosamples. The working process of the generator is shown in Figure 2.

We refer to the literature [40, 41] for an enhanced method to generative adversarial networks, where the generator is meant to work as an encoder and decoder in the tandem, which is a creative design. After several experimental verifications, we also apply the nested combination of encoding and decoding to the generator network. The encoder of the generator acquires different intensity facial expression features I^{low} by downsampling. Researchers in the literature [42] added a residual structure to the generator optimization to improve the efficiency of the generator encoding. We also verified the effectiveness of the method

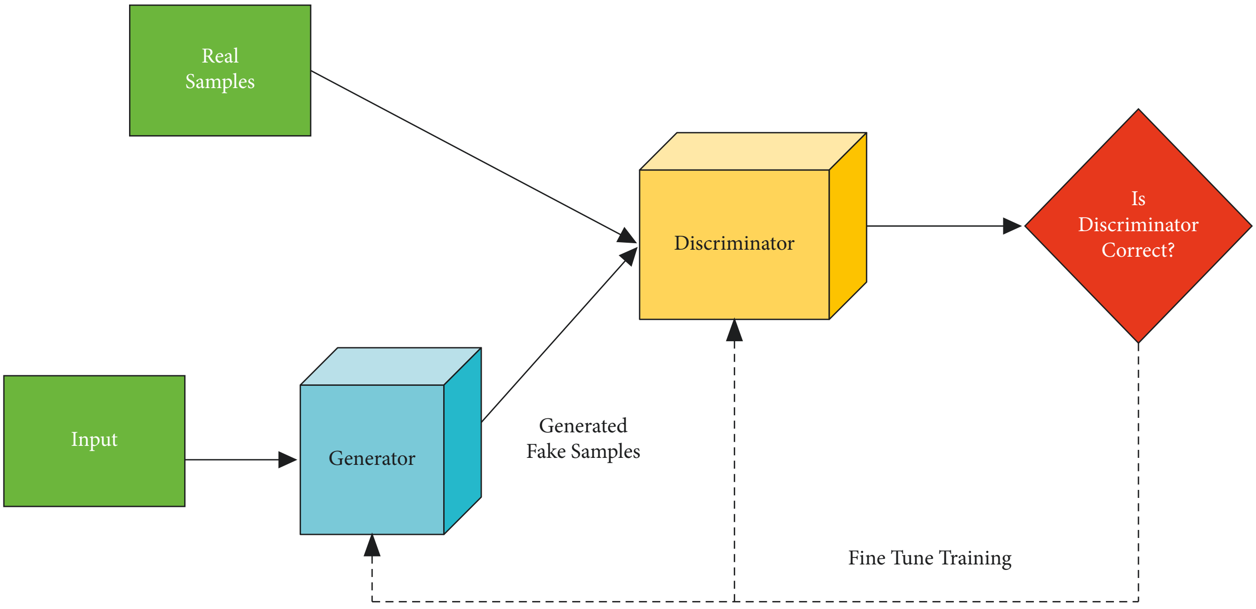

Figure 1: Generative adversarial network architecture.

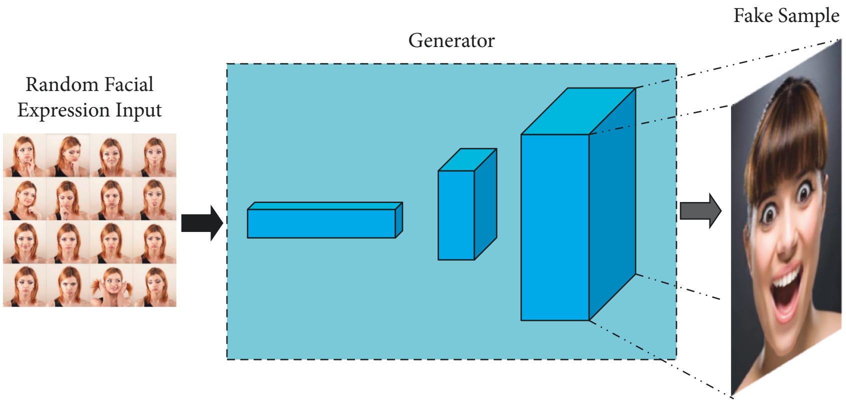

Figure 2: Facial expression generator process.

experimentally. In the decoder network layer, we use upsampling to transform the intensity features of facial expressions and then implement nonlinear activation by RELU. According to the decoder network optimization method in the literature [43], we implemented facial expression intensity figuration using the X-conv operator. Assuming the expression K input point $(p_1, p_2, \ldots, p_k)$, where K denotes the result of a multilayer perceptron of real samples, in a transformation matrix $\mathrm{X} = MLP(p_1, p_2, \ldots, p_K)$ of dimension $K \times K$ is computed, and the summation between feature elements can be simplified to the commonly used convolution operator. When X is performing the computation of the transformation matrix, different facial expression nodes have different effects, and we define the mathematical equation of the X-conv operator as follows:

$$
\begin{aligned}
F_p &= \mathrm{X\_}conv(K, p, P, F), \\
\mathrm{X\_}Conv(K, p, P, F) &= Conv\left(K, MLP(P - p) \times \left[MLP_\delta(P - p), F\right]\right),
\end{aligned}
\tag{1}
$$

where p represents the facial expression feature node, K represents the facial expression traversal function, $P = (p_1, p_2, \ldots, p_k)^T$ represents the nodes within the neighborhood expression feature node with K nodes, and $F = (f_1, f_2, \ldots, f_K)$ represents the expression feature nodes in different domains. In the nonlinear connection of the X-conv operator, facial expressions of different intensities will have different feature expressions in the generator, and the details of the X-conv operator at each level are shown in Figure 3.

3.3. Discriminator. The discriminator network consists of a combination of fully connected and deconvolutional layers. The discriminator is at the output port of the generator. In the discriminator, different threshold ranges are set and the pseudosamples are marked as invalid if they are below the threshold range. The feature information of the invalid sample will be fed back to the generator with the simulation side of the real sample. All the feedback methods will pass the correct feature values in this back propagation way, and

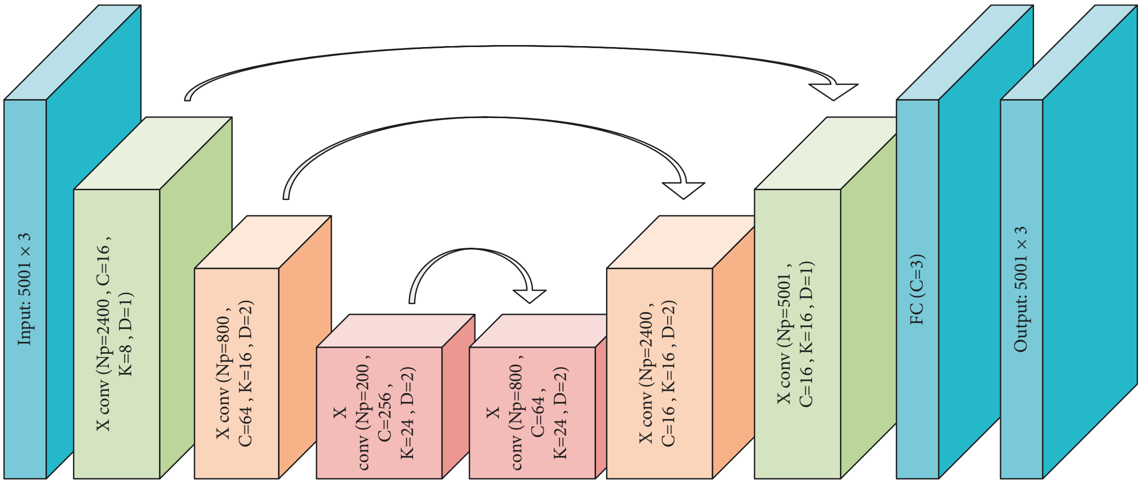
Input: 5001 × 3
X conv (Np=2400 , C=16 , K=8 , D=1)
X conv (Np=800 , C=64 , K=16 , D=2)
X conv (Np=200 , C=256 , K=24 , D=2)
X conv (Np=800 , C=64 , K=24 , D=2)
X conv (Np=2400 , C=16 , K=16 , D=2)
X conv (Np=5001 , C=16 , K=16 , D=1)
FC (C=3)
Output: 5001 × 3

Figure 3: Detailed hierarchy of generators.

the generator will automatically correct the newly generated expression features based on the feedback feature values. The discriminator principle is shown in Figure 4.

The intensity of facial expression features was not consistent according to the differences in facial expression types. Low-intensity expression features are less demanding on the generator and only need to filter the facial contour data density. For high-intensity expression features, it is necessary to first decompose the high-intensity expression features and then convert them into low-intensity feature combinations. Researchers in the literature [44] will have used an alternating training model to optimize the discriminator with threshold discretization detection of pseudosamples. We define min-max as follows:

$$\min_{Gen}\max_{Dis} = \mathrm{E}_{I}^{high}\log\left(Dis\left(I^{high}\right)\right) + \mathrm{E}_{I}^{low}\log\left(1 - Dis\left(Gen\left(I^{low}\right)\right)\right), \qquad (2)$$

where *Gen* denotes the twin sample of the generator and real sample and *Di s* denotes the threshold discrete detection of the discriminator and pseudosample. $\{I^{low}, I^{high}\}$ represents the feature intensity grading corresponding to facial expressions, and the generator *Gen* and discriminator *Di s* are distributed in a certain linear function, and the mathematical expression is as follows:

$$L_{G\_adv} = -\frac{1}{N}\sum_{n=1}^{N}\log\left(Dis\left(Gen\left(I_n^{low}\right)\right)\right),$$
$$L_{D\_adv} = -\frac{1}{N}\sum_{n=1}^{N}\left\{\log\left(Dis\left(I_n^{high}\right)\right) + \log\left(1 - Dis\left(Gen\left(I_n^{low}\right)\right)\right)\right\}, \qquad (3)$$

where N represents the expression feature intensity. During the intensity feature convergence process, the pseudosample features can be ranked with respect to the degree of threshold discretization under the detection of the discriminator. The generator fine-tunes the new features at a later stage based on the feature discretization values fed by the discriminator. The different levels of discriminator network layers we constructed are shown in Figure 5.

3.4. Auxiliary Classifier and Loss Function. The intensity of facial expression features can cause feature loss in the middle transition layer of the network layer. For this reason, we add auxiliary classifiers in the middle layer, which can retain the facial expression feature information under different intensities. In the actual course scenario, facial expressions will have different levels of facial muscle expressions. In order to maintain a stable mapping relationship between expression changes and feature intensities, the adversarial loss function is utilized to guide the feature decomposition of real expressions. Adaptive linear fitting function is added to the auxiliary classifier network layer, and all samples are configured with low intensity features combined with low intensity features by default during the production of classifier pseudosamples. It prevents the problem of feature intensity confusion in the process of expression feature perception. The mathematical equations of feature perception added in the auxiliary classifier are shown below:

$$L_{perceptual} = \frac{1}{N}\sum_{n=1}^{N}\left\|\phi\left(G\left(I_n^{low}\right)\right) - \phi\left(I_n^{high}\right)\right\|, \qquad (4)$$

where ϕ represents the expression feature intensity perceptron. In refining the pixel feature representation of 2-dimensional images of facial expressions, the high-intensity facial expression feature I^{high} and the linked expression feature $Gen\,(I^{low})$ generated by the generator take advantage of the point-by-point loss optimization to overcome the feature refinement and loss problems arising from the high-intensity feature decomposition. Researchers in the literature [45] performed experimental validation on the algorithm of point-by-point loss optimization, and the authors found that the L2 loss function is more stable. The mathematic functions are calculated as follows:

$$L_{pixel} = \frac{1}{N_{pixel}}\sum_{i=1}^{N_{pixel}}\left\|Gen\left(I^{low}\right)_i - I_i^{high}\right\|, \qquad (5)$$

where N_{pixel} denotes the intensity expression of the facial expression at the two-dimensional level. According to the

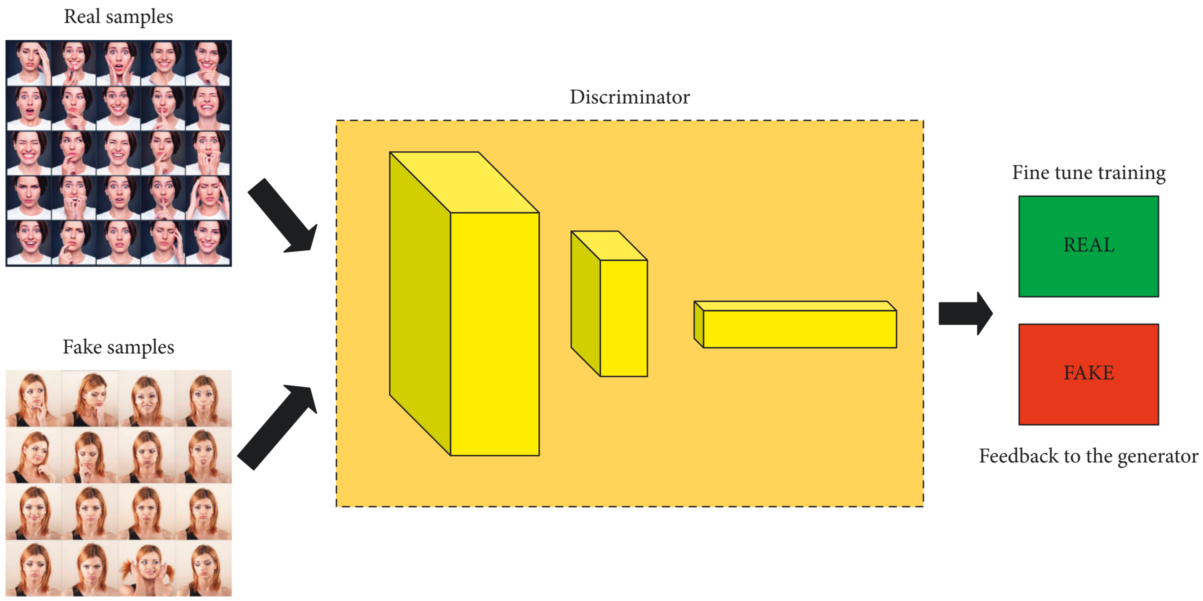

Figure 4: Facial expression discriminator process.

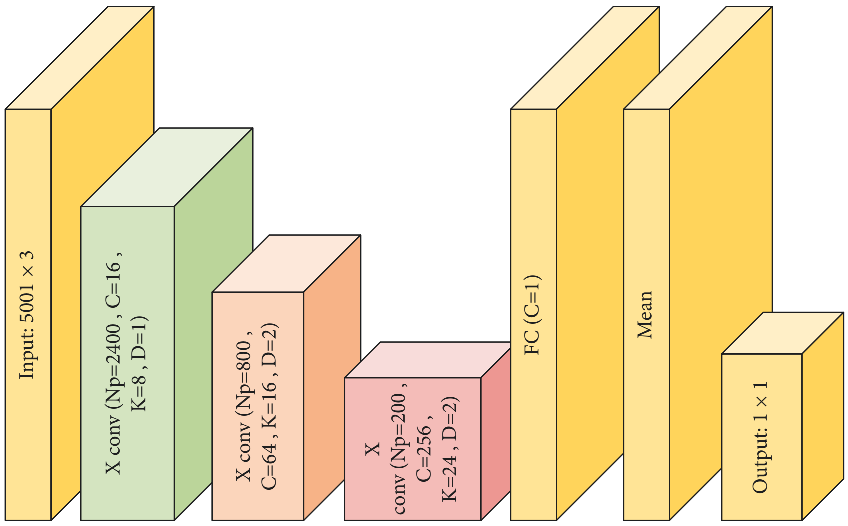
Input: 5001 × 3
X conv (Np=2400 , C=16 , K=8 , D=1)
X conv (Np=800 , C=64 , K=16 , D=2)
X conv (Np=200 , C=256 , K=24 , D=2)
FC (C=1)
Mean
Output: 1 × 1

Figure 5: Detailed hierarchy of discriminator.

constraint effect of the loss function, we designed the new loss function has the following mathematical expression:

$$L = \omega_1 L_{G\_adv} + \omega_2 L_{pixel} + \omega_3 L_{perceptual}, \qquad (6)$$

where ω_1, ω_2, and ω_3 denote the expression intensity feature weighting coefficients.

3.5. Improved Generative Adversarial Networks. In our study, to assess students' learning efficiency at the level of their facial expressions in the classroom, we present an enhanced generative adversarial network strategy for improving the accuracy of facial expression recognition models while also separating comparable expressions using feature intensity classification. The auxiliary classifier can provide feature generation guidelines and pseudosample feature discrimination to the generator and discriminator. At the pixel level, the auxiliary classifier middle layer neural network uses the X-conv operator to assist in synthesizing independent facial expression pseudofeatures, which are fed back in parallel with the generator in the joint output. The back propagation information from the discriminator will act as a filter in the auxiliary classifier to extract the feedback that aids in enhancing the effectiveness of the pseudofeatures into the real sample perception network. The facial expression detection network is shown in Figure 6.

4. Experiment

4.1. Datasets. We chose the well-known contemporary public facial expression datasets Oulu-CASIA (OC), Cohn–Kanade (CK+), and Facial multiview expression

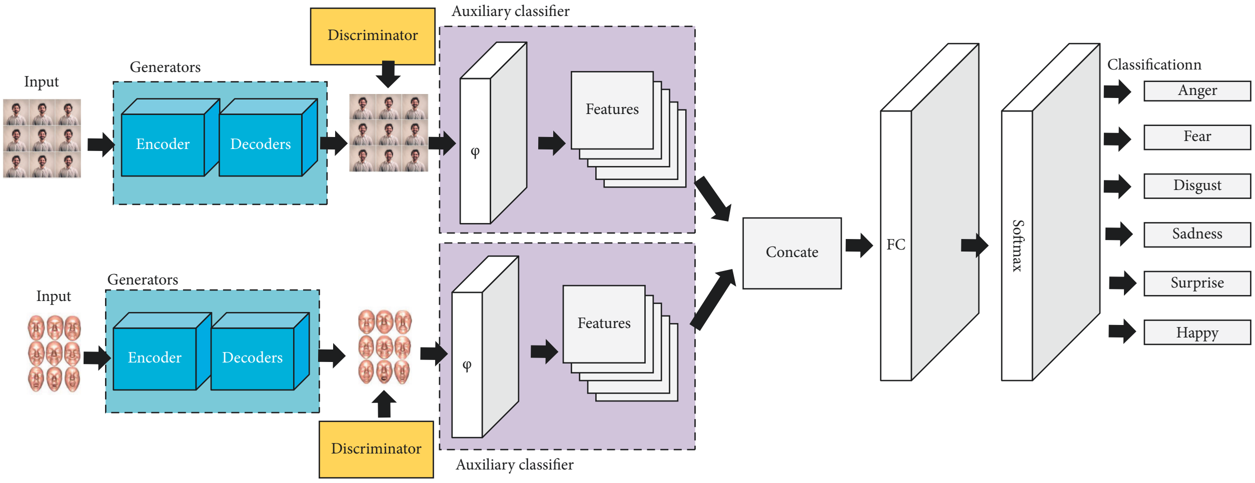

Figure 6: The structure of improved generative adversarial networks.

dataset with occlusion (FMEO) for the experimental test. Before performing expression classification operations on the above datasets, we collaborated with medical schools to manually standardize clear boundaries between expressions, and then we preprocessed all data to segment the images to specified sizes, with differences in the testing approach we took for different sizes of data.

The Oulu-CASIA dataset [46] contains a total of 2880 samples from the expression acquisition of 80 volunteers, which were captured using video recording and divided into visible light (VIS) series and near infrared (NIR) series according to the imaging system. Three different illumination methods were selected for the acquisition process to analyze the effect of detection methods on the structural environment. There are 480 videos of normal illumination samples, 60 videos of low illumination samples, and 15 videos of dark scenes. For the selection of the training set, we chose all the normal illumination video frame samples. The details of expression classification are shown in Table 1.

The Cohn–Kanade(CK+) dataset [47] contains a total of 593 video samples of facial expressions captured from 118 volunteers. Each piece of video is divided into 20–50 frames, and all video frame sequences are captured using a facial action coding system, which automatically classifies the expressions and labels them accordingly after the capture is completed. Its detailed facial expression classification information is shown in Table 2.

To evaluate the effectiveness of our strategy in complex situations such as occlusion, we chose FMEO to do the validation test. The dataset contains a total of 690 samples of data from 10 young volunteers, who were used in the experiment to collect facial expression samples by masking their faces with props, such as hats, glasses, and masks. The detailed classification of facial expressions in this dataset is shown in Table 3.

4.2. Experimental Settings. We trained the two-dimensional samples separately from the three-dimensional samples. The detailed parameter settings are shown in Table 4. In the validation process, we adopted the method mentioned in the literature [11]. For multitask learning training, to fairly compare random input expressions, we utilized a random search strategy with hyperparameter tuning.

4.3. Experimental Results. In the facial emotion detection work, we mainly analyze three metrics, such as accuracy (Acc), F1 score, and recall (R). To ensure that our method is effective, we conducted a test, and we choose traditional facial emotion detection approaches and a neural network series of facial emotion detection methods as control group experiments. We compared three methods, LBP_SVM, CNN, and LSTM. During the training and tuning phase, each network was trained independently without the recognition module to confirm the accuracy of each technique. The experimental results are shown in Table 5.

Table 5 proves the facial emotion detection effectiveness of our strategy. Considering the results of the experiments, CNN is the more commonly used method; however, it falls short of the LSTM approach in terms of facial expression recognition accuracy. This is mostly owing to the benefits provided by the LSTM's unique network topology, which can achieve local perception and maximize memory information fusion. Our method uses generative adversarial networks with a new CNN-based auxiliary classifier, which can recognize similar expressions hierarchically starting from the expression feature strength, further improving the accuracy of facial expression recognition while obtaining better robustness.

The experimental results show that the datasets OC and FEMO perform the best. Due to the computational cost, we mainly use the experimental results of datasets OC and FEMO as the main judging criteria. To test the efficiency of our approach for facial expression recognition in the classroom, we conducted experimental validation by self-made datasets. We collected classroom expression video data of 300 college students and manually labeled the homemade dataset according to the OC dataset labeling rules, and then tested it with the trained model. The results are shown in Table 6.

Table 1: Oulu–CASIA (OC) dataset facial expression classification.

	Anger	Fear	Disgust	Sadness	Surprise	Happy	Total
OC	799	790	765	794	768	784	4700
	Training set	3760		Test set	940		

Table 2: Cohn–Kanade (CK+) dataset facial expression classification.

	Anger	Fear	Disgust	Sadness	Surprise	Happy	Total
CK+	135	75	177	768	768	261	981
	Training set	785		Test set	196		

Table 3: FMEO dataset facial expression classification.

	Anger	Disgust	Happy	Sadness	Surprise	Total
FMEO	132	136	144	143	135	690
	Training set	552		Test set	138	

Table 4: Experimental parameter settings.

Parameter	Value
Initial learning rate	0.01
Decay rate	10
Weight decay	0.005
Epoch	80
Regularization	0.001
Margin loss discount	0.5
Dropout rate	0.1

Table 5: Results of text detection by different methods.

	OC			CK+			FMEO		
	Acc	R	F1	Acc	R	F1	Acc	R	F1
LBP_SVM	0.56	0.55	0.59	0.65	0.64	0.60	0.58	0.54	0.59
CNN	0.67	0.73	0.71	0.72	0.61	0.62	0.71	0.65	0.61
LSTM	0.75	0.82	0.80	0.76	0.72	0.73	0.81	0.77	076
Ours	0.89	0.94	0.83	0.86	0.81	0.83	0.89	0.95	0.93

Table 6: Results of text detection by different methods.

Method	Anger	Disgust	Happy	Sadness	Surprise
LBP_SVM	74.3	77.5	72.1	70.3	71.1
CNN	80.3	81.1	79.6	78.3	79.8
LSTM	85.3	86.4	82.1	84.3	81.8
Ours	95.3	96.3	93.1	92.7	93.6

In the students' facial expression recognition experiments, our improved generative adversarial network outperforms the others, and it further proves the effectiveness of our approach.

5. Conclusion

We offer a method for recognizing facial expressions based on an upgraded generative adversarial network. The method belongs to the deep training model, we divide the network into three stages. The front end of the network is the generator network layer, which relies on real sample features to generate pseudosamples. The middle of the network is the auxiliary classifier, which assists the generator in generating pseudosamples that are closer to the real samples. The end of the network is the discriminator network layer, which determines whether the pseudosamples satisfy the output conditions according to the degree of threshold discretization, and the pseudosamples that do not satisfy the conditions are fed back to the front layer for reconstruction. During the experiment, we test the efficiency of the strategy on the open-source datasets. In addition, we also test on the homemade student datasets. The experimental results prove that the facial expression detection accuracy of our method stays above 92%. Comprehensive performance of the model outperforms other methods.

Facial expressions are a very complex task to capture, and there are thousands of facial expressions in different scenes. In this paper, we tentatively select facial expressions with more prominent features as the study points. However, for many obscure expressions, our method still does not perform well. In further research, we are going to use a dual RNN framework to perceive the 3D features of facial expressions, and enhance the model's tolerance of high-intensity feature expressions.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The author declares that there are no conflicts of Interest.

Acknowledgments

The authors thank China Institute of Communications Education, The Fusion and Innovation of the Talent Training Model in International Cruise Service Professional Based on 1+x Certificate System (no. JTYB20-353).

References

[1] Z. T. Zhu, M. H. Yu, and P. Riezebos, "A research framework of smart education[J]," *Smart learning environments*, vol. 3, no. 1, pp. 1–17, 2016.

[2] G. R. Alexandre, J. M. Soares, and G. A. Pereira Thé, "Systematic review of 3D facial expression recognition methods," *Pattern Recognition*, vol. 100, Article ID 107108, 2020.

[3] S. A. Khan, A. Hussain, and M. Usman, "Facial expression recognition on real world face images using intelligent techniques: a survey[J]," *Optik*, vol. 127, no. 15, pp. 6195–6203, 2016.

[4] S. Li and W. Deng, "Deep facial expression recognition: a survey," *IEEE transactions on affective computing*, vol. 99, p. 1, 2020.

[5] Y. Miao, "Improved deep neural network for cross-media visual communication," *Computational Intelligence and Neuroscience*, vol. 2022, p. 1556352, Article ID 1556352, 2022.

[6] C. Chengjun Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467–476, 2002.

[7] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: a comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection,"vol. 1, pp. 886–893, in *Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE, San Diego, CA, USA, June 2005.

[9] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: regularizing a deep face recognition net for expression recognition," in *Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 118–126, IEEE, May 2017.

[10] G. Ian, P. A. Jean, and M. Mehdi, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 22, pp. 26772–32680, 2014.

[11] J. Cai, Z. Meng, A. S. Khan et al., "Island loss for learning discriminative features in facial expression recognition," in *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 302–309, IEEE, Columbia, USA, 2018.

[12] K. Wang, X. Peng, J. Yang, and Y MengQiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[13] Y. Wu and Q. Ji, "Facial landmark detection: a literature survey," *International Journal of Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019.

[14] H. Jung, S. Lee, J. Yim et al., "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE international conference on computer vision*, pp. 2983–2991, IEEE, Santiago, Chile, December 2015.

[15] K. Zhang, Y. Huang, Y. Du, and L Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.

[16] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, p. 1, 2021.

[17] G. Pons and D. Masip, "Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition," 2018, http://arxiv.org/abs/1802.06664.

[18] E. Paul, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, Oxford University Press, 2020.

[19] J. Cao, Y. Li, and Z. Zhang, "Partially shared multi-task convolutional neural network with local constraint for face attribute learning," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pp. 4290–4299, IEEE, Salt Lake City, UT, USA, December 2018.

[20] I. Misra, A. Shrivastava, and A. Gupta, *Cross-stitch Networks for Multi-Task Learning*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3994–4003, April 2016.

[21] T. Zhang, "Facial expression recognition based on deep learning: a survey," *Advances in Intelligent Systems and Computing*, Springer, in *Proceedings of the International conference on intelligent and interactive systems and applications*, pp. 345–352, 2017.

[22] S. Rajan, P. Chenniappan, S. Devaraj, and N Madian, "Facial expression recognition techniques: a comprehensive survey," *IET Image Processing*, vol. 13, no. 7, pp. 1031–1040, 2019.

[23] D. Canedo and A. J. R. Neves, "Facial expression recognition using computer vision: a systematic review," *Applied Sciences*, vol. 9, no. 21, p. 4678, 2019.

[24] N. Samadiani, G. Huang, B. Cai et al., "A review on automatic facial expression recognition systems assisted by multimodal sensor data," *Sensors*, vol. 19, no. 8, p. 1863, 2019.

[25] B. Sun, L. Li, and G. Zhou, "Combining multimodal features within a fusion network for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 497–502, ACM, WA, USA, November 2015.

[26] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 503–510, ACM, WA, USA, November 2015.

[27] M. M. Ghazi and H. K. Ekenel, "Automatic emotion recognition in the wild using an ensemble of static and dynamic representations," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 514–521, ACM, Tokyo, Japan, November 2016.

[28] Z. Zhang, P. Luo, C. C. Loy, and X Tang, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, no. 5, pp. 550–569, 2018.

[29] W. Hua, F. Dai, L. Huang, and G XiongGui, "HERO: human emotions recognition for realizing intelligent internet of things," *IEEE Access*, vol. 7, pp. 24321–24332, 2019.

[30] B. F. Wu and C. H. Lin, "Adaptive feature mapping for customizing deep learning based facial expression recognition model," *IEEE Access*, vol. 6, pp. 12451–12461, 2018.

[31] M. Liu, S. Li, S. Shan, and X Chen, "AU-Inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, 2015.

[32] A. Yao, D. Cai, and P. Hu, "HoloNet: towards robust emotion recognition in the wild," in *Proceedings of the 18th ACM international conference on multimodal interaction*, pp. 472–478, ACM, October 2016.

[33] B. K. Kim, S. Y. Dong, and J. Roh, "Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 48–57, IEEE, Las Vegas, NV, USA, June 2016.

[34] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: state of the art," 2016, http://arxiv.org/abs/1612.02903.

[35] F. Zhang, T. Zhang, and Q. Mao, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3359–3368, IEEE, Salt Lake City, UT, USA, June 2018.

[36] P. Liu, S. Han, and Z. Meng, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1805–1812, IEEE, Columbus, OH, USA, June 2014.

[37] D. Nguyen, K. Nguyen, and S. Sridharan, "Deep spatio-temporal features for multimodal emotion recognition," in *Proceedings of the 2017 IEEE winter conference on applications of computer vision (WACV)*, pp. 1215–1223, IEEE, Santa Rosa, CA, USA, March 2017.

[38] S. Khan and T. Yairi, "A review on the application of deep learning in system health management," *Mechanical Systems and Signal Processing*, vol. 107, pp. 241–265, 2018.

[39] Z. H. Zhou and J. Feng, "Deep forest," 2017, http://arxiv.org/abs/1702.08835.

[40] Y. H. Lai and S. H. Lai, "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition," in *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 263–270, IEEE, Xi'an, China, May 2018.

[41] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2168–2177, IEEE, UT, USA, June 2018.

[42] K. He, X. Zhang, and S. Ren, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, IEEE, June 2016.

[43] Y. Li, R. Bu, and M. Sun, "Pointcnn: convolution on x-transformed points," *Advances in Neural Information Processing Systems*, p. 31, 2018.

[44] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[45] R. Huang, S. Zhang, T. Li et al., "Beyond face rotation: global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proceedings of the IEEE international conference on computer vision*, pp. 2439–2448, IEEE, October 2017.

[46] G. Zhao, X. Huang, M. Taini, and M LiPietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[47] P. Lucey, J. F. Cohn, and T. Kanade, "The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression," in *Proceedings of the 2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pp. 94–101, IEEE, San Francisco, CA, USA, June 2010.